Curriculum Reform in The Common Core Era: Evaluating Elementary Math Textbooks Across Six U.S. States

David Blazar
Blake Heller
Thomas J. Kane
Morgan Polikoff
Douglas O. Staiger
Scott Carrell
Dan Goldhaber
Douglas N. Harris
Rachel Hitch
Kristian L. Holden
Michal Kurlaender

Abstract

Can a school or district improve student achievement simply by switching to a higher-quality textbook or curriculum? We conducted the first multi-textbook, multi-state effort to estimate textbook efficacy following widespread adoption of the Common Core State Standards (CCSS) and associated changes in the textbook market. Pooling textbook adoption and student test score data across six geographically and demographically diverse U.S. states, we found little evidence of differences in average achievement gains for schools using different math textbooks. We found some evidence of greater variation in achievement gains among schools using pre-CCSS editions, which may have been more varied in their content than post-CCSS editions because they were written for a broader set of standards. We also found greater variation among schools that had more exposure to a given text. However, these differences were small. Despite considerable interest and attention to textbooks as a low-cost, "silver bullet" intervention for improving student outcomes, we conclude that the adoption of a new textbook or set of curriculum materials, on its own, is unlikely to achieve this goal. © *2020 by the Association for Public Policy Analysis and Management*

INTRODUCTION

The choice of textbook or curriculum is an enticing lever for improving student outcomes. Few central office decisions have such far-ranging implications for the work that students and teachers do together in classrooms every day. In our own survey, we found that teachers in 94 percent of elementary schools in six geographically and demographically diverse U.S. states reported using the official district-adopted textbook or curriculum in more than half of their lessons.[1] Given such widespread

[1] Throughout the paper, we use the terms "textbook" and "curriculum" interchangeably. We recognize, though, that the physical textbook may be just one of multiple materials that make up a given curricu-

Journal of Policy Analysis and Management, Vol. 39, No. 4, 966–1019 (2020)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.22257

usage, helping schools and districts switch from less to more effective materials offers a large potential return on investment (Kirst, 1982; Whitehurst, 2009). As Chingos and Whitehurst (2012) point out, "…whereas improving teacher quality…is challenging, expensive, and time consuming, making better choices among available instructional materials should be relatively easy, inexpensive, and quick" (p. 1).

Textbook choice has been especially salient and has gained national policy attention in recent years after many states adopted the Common Core State Standards (CCSS), which generally are considered to be more rigorous than prior state standards (Friedberg et al., 2018).[2] Curriculum reform is one of the primary mechanisms by which policymakers, practitioners, and researchers hypothesized that the introduction of the CCSS could improve student outcomes at scale (Carmichael et al., 2010; Porter et al., 2011). In the years since CCSS adoption, large publishing houses (e.g., Houghton Mifflin Harcourt, McGraw Hill, Pearson) have invested heavily in adapting existing textbooks and curriculum materials to new standards, and in writing new materials from scratch. New York State spent over $35 million dollars to develop a set of curriculum materials, *Engage NY*, which are now widely used across the country under this title and *Eureka* (Cavanaugh, 2015). Once new textbooks are written, the marginal cost to schools and districts of switching from one textbook to another is quite small. On average, elementary math textbooks cost roughly $35 per student, which represents less than 1 percent of per-pupil expenditures (Boser, Chingos, & Straus, 2015). As of 2017, over 80 percent of the schools in our sample had adopted a CCSS-edition textbook in elementary math.

Despite the potential value to districts and schools, the research literature on the efficacy of alternative textbooks or curricula is sparse. We are aware of one multi-textbook randomized trial (Agodini et al., 2010), two randomized trials assessing the effectiveness of a single textbook (Eddy et al., 2014; Jaciw et al., 2016), and a handful of non-experimental studies that rely on matching techniques to estimate textbook effects (Bhatt & Koedel, 2012; Bhatt, Koedel, & Lehmann, 2013; Koedel et al., 2017). However, most of the textbook editions or curricula materials in common use today have never been subjected to a rigorous test of efficacy. Further, no studies have examined the sensitivity of textbook effects across time or across states. Although some textbook editions are written for local markets (e.g., California, Texas), logic suggests that a high-quality curriculum or textbook should be effective across settings, especially when the materials are written to align with a common or similar set of standards. Yet, to our knowledge, no studies have assessed this claim empirically. Of the studies cited above, most are analyses of single districts or states. Two studies (Agodini et al., 2010; Eddy et al., 2014) recruited participants across states; but, schools and districts volunteered and so are not representative of those settings, let alone of U.S. states more broadly.

One reason for the weakness of the evidence base is the historic diversity in state standards and assessments. When each state had its own standards and assessments, single-state studies were relevant only for schools in a given state, and few states were sufficiently large to justify the cost of such an analysis. A second, more practical barrier has been the omission of textbook adoptions from state data collection

lum. Curricula can include student and teacher editions of the textbook, formative assessment materials, manipulative sets, etc. In our survey to schools and teachers, we referred to the "primary textbook or curriculum materials" used by teachers, which could consist of "a printed textbook from a publisher, an online text, or a collection of materials assembled by the school, district, or individual teachers [but] does not include supplemental resources that individual teachers may use from time to time to supplement the curriculum materials."

[2] Since 2010, many of the states that initially adopted the CCSS have since revised their standards. Yet, several of the states that revised standards from the CCSS have landed on a close facsimile (Friedberg et al., 2018).

efforts (Polikoff, 2018). As useful as textbook adoption data would be for estimating efficacy, states have concentrated their data collection efforts on fulfilling federal accountability requirements that focus on student test score performance rather than informing district purchasing decisions. States typically have stayed away from collecting data on curricula adoptions in deference to local authorities (Hutt & Polikoff, 2018). We are aware of only a handful of states that regularly collect information on the textbooks used by schools.[3] As a result, it has been difficult to bring to bear states' longitudinal data on student achievement to compare the achievement gains of similar schools using different curricula.

We designed our study as a field test of a replicable, low-cost approach to measuring curriculum efficacy. Although experiments may be the most convincing way to estimate the causal effect of textbooks, the estimated impacts might not generalize beyond the small subset of schools that are willing to have their textbooks randomly assigned. Instead, by combining publicly available administrative data on textbooks in two states (California and New Mexico) with a survey administered to a random sample of schools in four additional states (Louisiana, Maryland, New Jersey, and Washington state), we ensured that we had state-representative samples of schools and were studying the textbook editions that schools are using in the current CCSS era. By pairing textbook adoption data with student test-score data on state-administered CCSS-aligned assessments, we also eliminated the need to collect our own assessments. Further, coordination amongst a large set of researchers across states—with different teams estimating the same "value-added" model with student-level data and then sharing only aggregated data—reduced the need to share sensitive data across state lines. In short, our methodology could be used to update results as textbook editions come and go.

Although our value-added methodology was the only way to examine textbook efficacy at scale, it comes with a trade-off with regard to internal validity. We estimate the association between textbook adoption and aggregate student achievement gains, and so cannot account for all factors that led schools and districts to choose different textbooks (e.g., characteristics of school leadership). However, we were able to examine a model of textbook selection based on observable school and district characteristics. These models indicate that selection varies substantially across textbooks and states, and so is unlikely to lead to systematic bias. Intuition also suggests that it is highly unlikely that high- or low-growth schools are systematically choosing the most effective textbooks when they (and we) do not know which textbooks those are. Empirically, we found that results were robust to models with different sets of school- and district-level controls, as well as models that replaced observable characteristics with school fixed effects that account for fixed differences across schools.

Unlike the prior literature, we found little evidence of substantial differences in average math achievement gains in schools using different textbooks. Our null findings

[3] California schools are mandated under state law to report curriculum materials on school accountability report cards (Holden, 2016; Hutt & Polikoff, 2018). In Indiana and Florida, centralized adoption processes allow state agencies to capture information on districts' adoption of certain texts (Bhatt & Koedel, 2012; Bhatt, Koedel, & Lehmann, 2013). In the course of preparing for our study, we learned that New Mexico collects curriculum data based on purchasing records from a state-organized curriculum warehouse, and these records can be attached to individual schools. Recently, Louisiana started to collect data on textbook adoptions through district and school surveys. Texas tracks adoption data based on requisitions and disbursements, and posts this information on a public website. In our study, we partnered with California and New Mexico to leverage their existing administrative data on textbook adoptions. We also partnered with Louisiana, though we administered our own survey to ensure representativeness. The other three states (Indiana, Florida, and Texas) did not administer student assessments from one of the two CCSS-aligned testing consortium, which was a critical criterion for our work.

are similar for schools with varied student populations related to English Language Learner status, free or reduced-price lunch eligibility, and high- or low-baseline achievement. Neither did we find textbook effects in the subset of schools in which teachers reported above- or below-median levels of textbook usage, or where teachers received above- or below-median levels of support to implement the textbook. We found some evidence of greater variation in achievement gains among schools using pre-CCSS editions, which may have been more varied in their content than post-CCSS editions because they were written for a broader set of standards. We also found greater variation among schools that had more exposure to a given text. In both of these latter subgroups, though, effects were smaller than average effects identified in prior studies.

In our conclusion, we attempt to reconcile our findings with the prior literature, which did find meaningful differences in efficacy across textbooks. We discuss the role of possible biases in our value-added methodology, the limited generalizability of the randomized controlled trials relative to our multi-state analysis, the role of implementation, and the possible greater uniformity in textbook content following the CCSS. Ultimately, though, we conclude that the adoption of a new textbook or set of curricula materials, on its own, is unlikely to improve student outcomes at scale and in the way that some policymakers, practitioners, and researchers have suggested and hoped.

MOTIVATION

Textbooks as a Lever for Efficiency

Textbooks and curriculum materials comprise a roughly $10 billion-per-year industry (Boser, Chingos, & Straus, 2015; McFarland et al., 2017), driven largely by their widespread use across K-12 public schools. In a nationally representative survey of schools focused on teachers' instructional decisions in the CCSS era, Opfer, Kaufman, and Thompson (2016) found that 98 percent of elementary math teachers reported using instructional materials selected or developed by district leadership. Roughly 85 percent of teachers reported that their districts required (57 percent) or recommended (27 percent) that they use specific textbooks to teach mathematics. Given such widespread adoption across classrooms and the small marginal difference in cost between curricula (Boser, Chingos, & Straus, 2015),[4] textbook selection has been described as a uniquely powerful lever for reform (Chingos & Whitehurst, 2012; Kirst, 1982; Whitehurst, 2009).

Textbook selection matters for many reasons, but perhaps most important is the extent to which the choice between one text versus another is likely to influence student learning. A broad theoretical literature from multiple disciplinary perspectives theorize critical links between curricula and desired student outcomes (e.g., Altonji, Blom, & Meghir, 2012; Schmidt et al., 2001; Stein, Remillard, & Smith, 2007), highlighting in particular the role of content (*what* material is presented) and structure (*how* that material is presented). For example, in mathematics, which is the focus of the current study, the "math wars" of the last several decades resulted in new learning standards (i.e., National Council of Teachers of Mathematics, 1989, 1991, 2000)

[4] Boser, Chingos, and Straus (2015) examined price data for 114 elementary math materials from 17 states that appeared on an approved list from more than one state. These products had an average per-pupil cost of $34 (in 2014 dollars), excluding ancillary materials such as teachers' editions. Comparing costs for the same product across states, they found that the difference between the lowest and highest prices paid was less than 1 percent for 30 percent of products, and less than 10 percent for 85 percent of products.

focused on conceptual rather than procedural instruction (Schoenfeld, 2004), and new textbooks and curriculum materials that align with this approach (Reys, 2001). Reform-oriented materials that focus on higher-order thinking typically have competed with more conventional resources, focused on teaching standard algorithms in sequence and using these procedures to solve basic problems. Debates regarding the design of textbooks typically have occurred in the educator practice community, but align with a broader push for schools to help students build the sorts of non-routine problem-solving skills that increasingly are valued in the labor market (Autor, Levy, & Murnane, 2003).

Current state-adopted math content standards aligned to the CCSS and associated textbooks are the most recent iteration of longstanding efforts to increase the academic rigor of instruction in U.S. classrooms and, in turn, student outcomes. One desired consequence of the CCSS, as a national policy initiative, is that widespread adoption of the standards across U.S. states would drive the textbook market to be uniformly more rigorous in terms of the curricula available to schools, teachers, and students (Heck, Weiss, & Pasley, 2011; Porter et al., 2011). The evidence to date is mixed. A recent study (Polikoff, 2015) used a prominent alignment methodology, the Surveys of Enacted Curriculum (SEC), to examine the content and standards alignment of several pre- and post-CCSS mathematics textbooks, all of which show up in our sample. That study found moderate overall levels of alignment of textbooks to state standards or to the CCSS. While the textbooks were more similar to each other (including pre- and post-CCSS editions of the same series) than they were to standards, they varied in terms of their relative emphasis on different domains of math content, and on levels of rigor and cognitive demand. That study is the only recent research evidence on the alignment of textbooks to each other or to standards. SEC research provides detailed alignment evidence but is limited by the research burden of analyzing entire textbooks from cover to cover.

Because reviewing, analyzing, and comparing the entire content of each textbook in our sample was beyond the scope of this project, we instead relied on publicly available reviews conducted by the nonprofit organization EdReports (see https://www.edreports.org/). EdReports engages with educators to conduct analyses of textbook series that are meant to align to the CCSS regarding their focus and coherence, rigor and mathematical practices, and usability. As we illustrate in our results below, EdReports identifies variation in these metrics across the most commonly used CCSS-edition math textbooks in our six partner states. These differences suggest that there may also be differences in student outcomes across schools using different textbooks.

Prior Evidence on Textbook Efficacy

Despite a strong theoretical basis linking textbooks to student learning, limited evidence exists on this topic. Of the 38 textbooks we observe in our sample, only six have been evaluated in a manner meeting the highest evidence standards (i.e., experiments, regression discontinuity designs) of the What Works Clearinghouse (WWC), a federal repository for education research. Only four of these are among the top 15 most commonly used textbooks in our sample, and only two were textbook editions meant to align with the CCSS. Thus, in the vast majority of cases, districts must select curricula in the absence of evidence of efficacy, relying instead on the judgments of central office staff, selection committees, and the choices of neighboring districts (Polikoff et al., 2020).

We discuss the past research on math textbook efficacy in two broad categories: randomized trials and non-experimental studies.

Randomized Trials

To our knowledge, only one study has used a randomized design to compare the impact of multiple elementary math textbooks on student achievement.[5] Agodini et al. (2010) randomly assigned one of four curricula to 1st- and 2nd-grade classrooms in 111 schools in 12 districts across 10 U.S. states. The study began in the 2006/2007 school year, prior to the development of the CCSS, and so none of the texts were designed to align with these standards. However, all texts have since been updated and now are marketed as aligned to the CCSS. These texts include: *Investigations in Mathematics* and *Math Expressions*, which generally are considered to be reform-oriented with a focus on building conceptual understanding; and *Saxon Math* and *enVision Math*,[6] which are described as more conventional and focused on teacher-directed instruction of procedures (Remillard, Harris, & Agodini, 2014; Stein, Remillard, & Smith, 2007). Teachers received two to three days of training on the assigned textbook over the course of the study. Second-grade classrooms using *Math Expressions* or *Saxon Math* outperformed those using *enVision* by 0.12 standard deviations (SD) and 0.17 SD, respectively. These effect sizes are large relative to the majority of educational interventions (Fryer, 2017). They would be larger than the effect of having an experienced versus a novice teacher (roughly 0.08 SD) and roughly equivalent to a 1 SD increase in teacher efficacy (Rockoff, 2004).

Two other studies used randomized designs to study the effect of individual curricula, both meant to align to the CCSS. Jaciw et al. (2016) evaluated the effectiveness of the post-CCSS edition of *Math in Focus* (there also is a pre-CCSS edition) by randomly assigning 22 clusters of 3rd- through 5th-grade teachers in 12 schools in Clark County School District (Las Vegas) in Nevada during the 2011/2012 school year. This text is modeled on a Singaporean math curriculum, emphasizing student-directed problem solving. Teachers attended a short training session (1.5 to 3 hours) during the summer and four half- or full-day sessions throughout the school year. The authors found that, after the first year of usage, students in grade-level teams randomly assigned to adopt *Math in Focus* outperformed students in the control group, who used the math curricula already in place in their school, by 0.11 to 0.15 SD on the Stanford Achievement Test. However, the study team found no impact of *Math in Focus* on the criterion-referenced test required by the state of Nevada.

[5] Several additional studies that attempted to use randomized designs to evaluate specific curricula were rejected from the WWC for failing to meet inclusion standards, generally due to imbalance between groups at baseline. Two doctoral dissertations cited by WWC used experimental designs to evaluate textbook effectiveness but never were published and rely on extremely small samples ($n < 100$ students). WWC reviewed two additional studies with randomized designs that meet their evidence standards; however, these studies are not available online (Beck Evaluation & Testing Associates Inc., 2005; Gatti & Giordano, 2010). In a currently unpublished review, Pellegrini et al. (2018)—updated from an earlier published review of the same topic (Slavin & Lake, 2008)—also cite a recent randomized evaluation of a post-CCSS edition of *enVision* that is not available online (Strobel, Resendez, & DuBose, 2017). At the time of writing, we were unable to obtain access to review these studies. Additionally, Pellegrini et al. (2018) cite results from randomized trials evaluating *Everyday Mathematics* and *JUMP Math* textbooks that were gleaned from conference presentations, where a full description of each study's experimental design and associated balance tests are not available online.

There also are several randomized evaluations of math materials, which we see as different from the textbooks we examine in our studies. Math software products that sometimes are described as curriculum, including *Cognitive Tutor Bridge to Algebra*, *Compass Learning's Odyseey Math*, *PLATO Achieve Now*, and *Larson Pre-Algebra*, have been subjected to randomized evaluations. In our study, we define these materials as supplemental and not the primary curriculum to teach mathematics. Jackson and Makarin (2018) experimentally evaluated the effectiveness of "off-the-shelf" curriculum materials for middle school math teachers, which we distinguish from complete textbooks.

[6] Prior to being updated to the CCSS, *enVision* was referred to as *Scott Foresman-Addison Wesley Elementary Math*. This is the title used in several evaluations cited in our review (i.e., Agodini et al., 2010; Bhatt & Koedel, 2012; Bhatt et al., 2013).

Eddy et al. (2014) randomly assigned *Go Math* to 1st- through 3rd-grade classrooms in nine schools across seven states during the 2012/2013 school year. *Go Math* was written for the CCSS and aims to balance conceptual and procedural understanding. After one year, the authors did not find statistically significant differences in average student achievement in classrooms using *Go Math*, as compared to control classrooms using the mathematics program already in place in their schools. However, the study was underpowered to detect effects smaller than 0.2 SD.

Non-Experimental Studies

A handful of non-experimental studies have identified moderate effects of textbooks on student achievement gains (upwards of 0.14 SD). Koedel and co-authors used matching methods and school-level aggregate achievement to measure textbook efficacy in three states: California, Florida, and Indiana (Bhatt & Koedel, 2012; Bhatt, Koedel & Lehmann, 2013; Koedel et al., 2017). Given the large number of texts observed in California and Florida, the analysts used a two-step process in these states. They first identified a differentially effective text based on an initial exploratory analysis, and then compared that text against a composite comparison group. Although this process helps to narrow the focus of inquiry, the danger is that the initial exploration may identify the "winning" text due to chance differences in achievement. The authors were careful to conduct a number of validity tests in the second step—e.g., verifying that the timing of any achievement increase aligned with the textbook adoption and that achievement did not grow in English Language Arts (ELA). Yet, such tests would not necessarily reveal a within-sample anomaly.[7]

Two textbooks appear in both the set of non-experimental studies and the randomized trials, yet with different conclusions about their relative efficacy. While *Saxon Math* outperformed *enVision* in the multi-textbook randomized trial (Agodini et al., 2010), the pattern was switched in one non-experimental matching study (Bhatt & Koedel, 2012).[8] These differences in ranking might be due to the methodology used to estimate textbook efficacy, or to differences in generalizability. As Bhatt and Koedel (2012) discuss, *Saxon Math* is a highly scripted curriculum and was designed for implementation in schools where the teachers have weak math backgrounds, the very type of schools that were willing to have their textbook randomized.

DATA COLLECTION

Ours is the first multi-state, multi-textbook study in the CCSS era, and aims to provide generalizable knowledge on the efficacy of math textbook editions in common use today. To achieve this goal, we assembled data from over 6,000 schools across six states: California, Louisiana, Maryland, New Jersey, New Mexico, and Washington. Our collaboration with these six states was strategic. At the time of our study, all six had adopted the CCSS; they also had ready access to administrative data on textbook adoptions or agreed to take part in our survey. Nonetheless, the states are geographically and demographically diverse, capturing variation across the U.S. As shown in Appendix Table A1,[9] three states (Maryland, New Jersey,

[7] WWC and Pellegrini et al. (2018) cite several non-experimental evaluations of single textbooks. We omit these evaluations from our literature review, preferencing the highest-quality research designs (randomized trials) or multi-textbook evaluations that are most similar to our own study.

[8] Another text, *Investigations,* showed up in both the multi-textbook randomized trial (Agodini et al., 2010) and one of the non-experimental studies (Bhatt et al., 2013). However, in the latter study, schools using *Investigations* were grouped together with schools using several other texts as part of a composite comparison group.

[9] See the Appendix at the end of this article.

and Washington) scored above the national average on the National Assessment of Educational Progress (NAEP), while the other three states (California, Louisiana, and New Mexico) scored below average. Two states (Maryland and New Jersey) also had above-average instructional expenditures and household income, while other states (Louisiana and New Mexico) fell below the national average on both measures. All six states were similar with regard to the percent of students eligible to receive special education services. California and New Mexico had an above-average share of Hispanic students and English Language Learners, while Louisiana and Maryland had an above-average share of African-American students.

We used the Common Core of Data (CCD) to construct state-specific sampling frames of public schools enrolling both 4th- and 5th-grade students between the 2014/2015 school year (the first year of testing using new CCSS-aligned assessments) and the 2016/2017 school year (the year we began data collection). Focusing on upper-elementary grades ensured that students across all six partner states had prior-year test scores. We also hypothesized that we would be more likely to see effects of textbooks in elementary grades rather than middle or high school, given that students, on average, tend to make larger academic gains in earlier rather than later grades (Hill et al., 2008). We included public charter schools but excluded private schools. We also limited the sampling frame to schools with test score data for more than 10 students, as many of our data use agreements made this a requirement for security and confidentiality purposes.[10]

Within this sampling frame, we assembled three types of data. First is textbook adoption data, which came both from administrative records covering close to full populations in two states (California and New Mexico) and a project-administered survey to a representative sample of schools in the other four states (Louisiana, Maryland, New Jersey, and Washington). Second is a survey administered to teachers in a representative sample of schools using one of the seven most frequently used CCSS-edition curricula, which asked teachers about their use of the textbook. In Table 1, we provide a summary of the state sampling frames and subsequent samples generated for each of these two sources of data. The third data source is state-collected information on students (excluded from Table 1, as the data were not sampled), including demographics and test scores on CCSS-aligned assessments. We describe each of the three sources of data in turn below.

Textbook Adoptions

In two of our partner states, textbook adoption data came from administrative records. In California, state law requires that schools report on textbook adoptions each year as part of school accountability report cards (Hutt & Polikoff, 2018). Our raw data came from reports hosted on the California Department of Education website, in which schools reported the title of textbooks they adopted by subject and grade (see Koedel et al., 2017, for additional information). Those data allowed us to identify the math textbook adopted for 87 percent of elementary schools in our sampling frame ($n = 5{,}107$ of 5,841 schools; see Table 1). Given the state mandate, very few schools were missing these reports. Instead, our inability to identify the math textbook for 13 percent of schools in our California sampling frame was due to the quality of information provided (e.g., reporting the name of a publisher rather than the textbook title). Given the lag in school-level reporting and state-level release of

[10] Restricting the sampling frame to schools with 10 or more students with test score data was relatively trivial, dropping fewer than 50 schools for whom we were able to capture textbook information (most from California).

Table 1. Sample of schools and teachers.

		Textbook Adoption Data		Teacher Survey Data		
States (by Source of Textbook Data)	# Schools in Sampling Frame	# Sampled Schools	# Schools with Textbook Data	# Originally Sampled Schools (Eligible Teachers)	# Sampled Schools (Eligible Teachers) after Replacement	# Schools (Teachers) with Survey Data
A. Administrative Data States						
California	5,841	N/A	5,107	95 (479)	150 (727)	100 (324)
New Mexico	439	N/A	297	24 (94)	32 (116)	24 (67)
B. Sampled States						
Louisiana	668	192	161	42 (150)	58 (218)	38 (79)
Maryland	853	139	121	23 (126)	34 (190)	15 (71)
New Jersey	1,146	427	322	116 (661)	134 (774)	107 (434)
Washington	1,054	316	247	68 (292)	78 (345)	61 (220)
Full Sample of Schools (Teachers)	10,001	1,074	6,255	368 (1,802)	486 (2,370)	345 (1,195)

Notes: The sampling frame includes public schools that enrolled 4th- and 5th-grade students between the 2014/2015 and 2016/2017 school years, and had achievement data for 10 or more students. Each school is counted once, even though most were observed in more than one school year; we leverage school-year observations in subsequent analyses. In California and New Mexico, textbook adoption data come from administrative records (school report cards in California, and purchasing data in New Mexico) and so cover close to full populations. "N/A" indicates that schools were not sampled in these states to collect textbook adoption data. In the remaining four states, textbook adoption data come from a survey administered to a representative sample of schools within the sampling frame. Schools were randomly selected using a two stage process in which we first sampled districts within states (all school districts with two or more schools in our sampling frame and half of the districts with just one school in our sampling frame). Then, we sampled schools within districts with the total number of schools proportional to district size. For the teacher survey, in all six states we randomly sampled a subset of schools for whom we had textbook adoption data and that were identified as using one of the top seven CCSS-edition textbooks by market share (see Table 3). The final sample includes 50 to 60 schools using each of these seven texts. The target number of schools using a given textbook per state was proportional to the distribution of that textbook across states, with the California sample downweighted by a factor of three. We randomly selected schools within strata based on textbook, state, and a median split of the proportion of students eligible for free or reduced-price lunch. We allowed for sampling with replacement, and so show the orignially sampled set of schools and the full sample after replacement. The final teacher survey sample records all valid reponses that successfully merged to state test-score records.

these data, we captured textbook adoptions in 2014/2015 and 2015/2016 but not in 2016/2017.

In New Mexico, the second administrative data state, we relied on purchasing data to capture textbook adoptions. Schools received a discount when they purchased textbooks from a centralized warehouse. Likely as a result, we observed textbook orders for a large share (68 percent) of public elementary schools in our New Mexico sampling frame ($n = 297$ of 439 schools; see Table 1). Presumably, schools not found in the purchasing records did not use a regular textbook or accessed textbooks or curricula from other sources, including open educational resources available online that would not be found in a physical warehouse. The purchasing data, available for 2010 through 2017, included ISBNs, textbook titles, quantity purchased,

and school addresses.[11] To distinguish between one-off purchases and official textbook adoptions, we limited our analysis to instances where the number of texts purchased was at least 50 percent of 4th- and 5th-grade enrollment in a given school and year. If we observed this sort of qualifying textbook purchase between 2010 and 2017, we imputed that data to future years, until/unless we observed a subsequent qualifying purchase. Of the 67 teachers from New Mexico whose schools also were randomly selected to participate in our teacher survey (see below), 91 percent indicated that the textbook identified for their school from the purchasing data was correct.

In the four remaining states (Louisiana, Maryland, New Jersey, and Washington), we surveyed schools in the winter of the 2016/2017 school year in order to identify the textbooks they adopted that year and the two years prior. We took a multi-step process to select schools in order to account for the fact that schools within most districts share a single text (see Koedel et al., 2017, and our own data). If we took a one-step process to sample schools proportional to their enrollments, we would have ended up with a large number of urban schools using a small number of texts. Instead, we first selected a random sample of districts within each state, stratified by the number of elementary schools (i.e., districts with just one school that fit our sampling frame criteria, two to three schools, four to seven schools, eight or more schools) and the percentage of students receiving free or reduced-price lunches (above and below the state median). By state, we randomly selected half of the one-school districts as part of our sample; we selected all districts with two or more schools. We then sampled one school from our sampling frame from districts with two to three schools, and two schools from districts with four to seven schools. In districts with eight or more elementary schools, each school had a 20 percent chance of being selected.[12] Under this schema, we arrived at a total sample of 1,074 schools in our sample across the four survey states (see Table 1).

To recruit these randomly selected schools to participate in our study, we sent an electronic version of the textbook adoption survey to the principal or curriculum coordinator in each school, also providing a description of the project and a promise of a $50 gift card for successful completion of the survey. We followed up with an e-mail and called the school every 10 days until we received a response. We also contacted district personnel to help with outreach. After multiple rounds of follow-up over five months, we achieved response rates of 79 percent overall and between 75 and 87 percent for each state (derived from numbers in Table 1). We attribute these high response rates to our multi-pronged approach, and to support for our study from state and district leaders who helped ensure that school personnel saw and responded to our e-mails or phone calls.

In Table 2, we compare observable school and district characteristics for schools with and without textbook data. In California and New Mexico, where we relied on administrative records, we see some differences between schools with and without

[11] Purchasing data that link to individual schools are rare. In this study, we also examined the quality of purchasing data captured by the agency GovSpend, which uses Freedom of Information Act (FOIA) requests to capture purchases that are part of the public record. However, the data proved to be limited, as purchases generally came from district-level offices, had no shipping information to connect to schools, and often did not provide sufficient information to identify the number of students covered by a given purchase. Agreement rates between these data and other adoption data at the school level generally were lower than 50 percent, and sometimes substantially lower than that.

[12] One exception was in Maryland, where initial conversations with state and district leaders indicated that very large districts in the state (several with over 100 schools that met our inclusion criteria) adopted a single textbook across the entire district. Therefore, we limited the random sample of schools to just 10, rather than 20 percent.

Table 2. Comparing schools with and without textbook adoption data.

	Administrative Data States				Sampled States	
	California		New Mexico		Louisiana, Maryland, New Jersey, and Washington	
School or District Characteristics	Schools with Known Textbook	Difference for Schools with Unknown Textbook	Schools with Known Textbook	Difference for Schools with Unknown Textbook	Respondents	Difference for Non-Respondents
Free or Reduced-Price Lunch (%)	62.0	−3.9~	84.3	−10.6**	50.5	3.5
Special Education (%)	12.6	5.6***	17.6	−0.9	14.5	0.0
English Language Learner (%)	24.2	−5.9***	16.1	−4.0*	605	1.5
Male (%)	51.4	2.5***	50.7	0.2	51.3	−0.2
African American (%)	5.8	0.5	1.6	0.0	22.7	1.6
Asian (%)	10.5	−3.4***	0.9	0.1	6.4	−2.1**
Hispanic (%)	52.4	−9.4***	63.0	−5.7	17.4	5.1
Native American (%)	0.9	0.4*	11.8	−1.9	0.6	0.5
Mixed Race or Other (%)	3.5	0.4~	1.9	0.2	3.1	0.0
Prior Math Test Score (Standardized)	−0.026	−0.089*	−0.023	0.027	−0.032	−0.09
Prior Reading Test Score (Standardized)	−0.021	−0.049	−0.017	0.042	−0.024	−0.059
Per-pupil Instructional Expenditure ($)	5,859	1,179***	5,505	−108	7,956	40~
Parents Married (%)	65.3	0.4	55.4	2.2	62.2	−1.0
Speaks Language Other than English (%)	45.2	−4.6*	33.8	−3.9	20.0	1.2
Median Household Income ($)	62,380	1,197	46,444	4,833~	70,898	−142
Parent Attended College, no BA (%)	29.1	1.0	33.0	2.9*	30.2	1.1~
Parent Holds BA Degree + (%)	23.7	1.2	19.9	−1.3	30.1	−2.1~
P-Value from Test of Joint Significance		0.000		0.000		0.097
School Observations	5,107	734	297	142	851	223

Notes: For states where schools were randomly sampled, estimates are weighted by the inverse of a school's sampling probability. Expenditure data and family characteristics come from the 2010 to 2014 American Community Survey, captured at the district level. Other characteristics are captured at the school level. ~p < .10; *p < .05; **p < .01; ***p < .001.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

textbook data. In California, the schools without textbook data had higher percentages of special education students, higher per-pupil expenditures, and lower baseline math achievement. However, overall coverage of schools is high (87 percent). The large sample of schools in California also makes us more likely to detect statistically significant differences even when the magnitude of those differences is small. In New Mexico, the schools without textbook data had somewhat lower percentages of students eligible for free or reduced-price lunch, but other differences were small or not statistically significant. In the states where we conducted our own survey, the non-respondents differed from respondents on only one out of 17 characteristics. Though not shown in Table 2, our randomly selected sample of schools in these four states looks similar to the larger sampling frame, as we expected given the randomized sampling process. On observable measures, the schools in our analyses generally are representative of their states.

We briefly describe the distribution of textbooks across schools in each state, which motivates our sampling strategy for a subsequent teacher survey. In Table 3, we show the percent of schools in each state using specific texts. Because some schools switched textbooks across years, we calculated means in a school-year dataset. However, we weighted estimates by the inverse of the probability that an individual school was selected for the sample.[13] In this and subsequent tables, we use "CC" to refer to textbooks that were meant to align to the CCSS. Some titles (e.g., *enVision*) are listed twice—with and without "CC"—because they originally were written prior to the rollout of the CCSS and then adapted in some way to align to the new standards.[14]

Despite the large number of options available (38 in our sample), we find that the elementary math textbook market remains dominated by a small number of texts. Over 70 percent of elementary schools in the six states used one of seven texts, and roughly 90 percent used one of 15 texts. Nevertheless, the market share for a particular curriculum varied by state. For instance, in New Mexico, the market was fairly split between three textbook series, all written for or adapted to the CCSS. Comparatively, in Louisiana, almost 60 percent of schools used just one open-source curriculum (*Engage NY*/*Eureka Math*) that was written specifically for the CCSS. In California, a sizeable share of school-year observations were attached to a non-CCSS edition textbook, driven largely by the Los Angeles Unified School District, where schools generally updated to a CCSS-edition text during the time frame of our study. In Maryland, over 30 percent of schools reported using materials developed by districts, rather than materials developed by outside publishers.[15]

Teacher Survey

To gain insight into teacher use of adopted textbooks, we conducted a teacher survey asking about their frequency of textbook use for different activities (e.g., preparation of lessons, classroom assignments), use of supplementary materials (e.g., those found online, developed by the teacher), reasons for supplementation, and access to professional development. We administered the teacher survey in the fall of 2017

[13] In California and New Mexico, all schools had a weight of one as there was no sampling conducted in these two states.

[14] To distinguish between pre- and post-CCSS editions of the same textbook series, we consulted with EdReports, did a high-level content comparison of the content of individual contexts, and examined the distribution of textbook editions by year. Based on this information, we categorized new editions as "CC" if they were published during or after 2011. The start of the CCSS was in 2010.

[15] We include all district-developed curricula in a single category, despite variation in materials within or across districts.

Table 3. Textbook market share by state.

Textbook (Sorted by Share in Pooled Sample)	Pooled 6 States	California	Louisiana	Maryland	New Jersey	New Mexico	Washington
enVision CC (% share)	14.8	11.5	15.1	14.7	27.7	28.4	7.9
Engage NY/Eureka CC	12.7	8.8	58.8	16.6	1.4	0.0	19.3
Go Math CC	12.5	17.3	10.2	1.5	19.1	2.6	0.6
My Math CC	12.0	17.1	8.3	2.1	6.0	29.4	3.3
enVision	7.8	13.7	0.9	4.9	1.2	0.0	1.6
Math Expressions CC	7.3	9.7	0.4	0.0	2.1	0.0	16.0
Everyday Mathematics CC	4.6	3.6	0.0	0.5	15.7	5.5	1.1
Everyday Mathematics	2.7	3.5	0.0	3.4	3.1	0.2	1.2
Houghton Mifflin Math	2.3	4.5	0.0	0.0	0.2	0.0	0.0
Math in Focus CC	2.2	0.0	0.0	0.0	10.8	1.4	4.0
Bridges in Mathematics CC	2.2	1.9	0.0	0.0	0.0	0.0	9.5
Stepping Stones CC	2.0	0.3	0.0	0.9	0.0	31.1	2.1
Ready Common Core CC	1.7	0.1	1.6	14.0	0.0	0.0	0.0
Math Connects	1.2	0.0	0.0	2.1	1.2	0.0	5.8
Math Expressions	1.0	0.2	0.0	0.0	0.2	0.0	7.1
Other Known Textbooks	6.4	7.9	2.0	0.9	3.6	1.2	12.4
District Developed	4.1	N/A	0.4	31.0	2.3	N/A	2.4
No Consistent Textbook	2.5	N/A	2.3	7.3	5.5	N/A	5.7
School*Year Observations (All Categories)	11,816	8,711	310	352	944	832	667
School*Year Observations (All Known Textbooks)	11,516	8,711	298	195	874	832	606
School*Year Observations (Top 15 Textbooks)	10,797	8,121	292	191	840	822	531

Notes: For states where schools were randomly sampled (Louisiana, Maryland, New Jersey, and Washington), estimates are weighted by the inverse of a school's sampling probability. When pooling data across states, schools in California and New Mexico, which were not sampled, are given a sampling probaility of 1. While our sampling approach selected individual schools (see Table 1), here we use a school-year dataset given that some schools switched textbooks across years. The school-year sample sizes listed at the bottom of the table are those used in later analyses. In California and Louisiana, we have at most two years of data on each school due to availability either of textbook adotion data (California) or state test-score data (Louisiana). In the other four states, we have at most three years of data on each school. "CC" indicates that a given textbook was written for or adapted to the CCSS standards. Some series titles are listed twice because they were written prior to the CCSS and then updated in some way after the rollout of the policy. The "other known textbooks" category includes 23 texts, each used by a small number of schools (in some instances, by just one school). We include all district-developed curricula in a single category, despite variation in materials across districts. "N/A" indicates that the nature of the administrative records in California and New Mexico could not identify the categories listed in these rows (i.e., district-developed materials, no consistent textbook).

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

but asked teachers about activities and practices in the prior academic year (i.e., the year in which we administered the school survey).[16]

Given our interest in the CCSS textbook market, we focused teacher survey data collection on a subset of schools that reported using one of the seven most commonly adopted CCSS-edition texts (see Table 3). Four of these—*enVision*, *Everyday Mathematics*, *Math Expressions*, and *Math in Focus*—were in press prior to the rollout of the CCSS and then were adapted to align to the new standards. The other three—*Engage NY/Eureka, Go Math*, and *My Math*—were written specifically for the CCSS. We excluded three other CCSS-edition textbooks that were part of the top 15 textbooks by market share because their use was dominated by schools in just one rather than multiple states. Our target was to receive surveys from 50 to 60 schools using each of these seven texts, for a total of 350 to 420 schools.[17]

To begin the sampling process, we identified a target number of schools per textbook such that the distribution of textbooks across states in the teacher survey sample was proportional to the distribution in our larger sample. For example, roughly 25 percent of schools in our full sample that reported using *enVision* (CCSS edition) came from New Jersey, and so roughly 25 percent of the schools using this textbook sampled for the teacher survey came from this state. In making these calculations, we downweighed the California sample by a factor of three, given that this state was larger than all others combined and would have led to a teacher survey sample that was dominated by this state. In analyses, we adjusted our estimates for this sampling design.

Next, we randomly selected that number of schools, stratifying on textbook, state, and an indicator of whether they were above or below the median percentage of students receiving free or reduced-price lunch within each state. We allowed for sampling with replacement in instances where districts or schools turned down our request to survey teachers, where we were unable to identify the roster of 4th- and 5th-grade teachers in a particular school, or where no teachers in the school responded within a three-month period. When a school needed to be replaced, we picked the next school on the randomly sorted list that reported using the same textbook, was in the same income bracket, and (where possible) was in the same state. If we selected a school for participation, we searched the school's website or called the school office for contact information on all 4th- and 5th-grade teachers from the 2016/2017 school year. We provided a survey link by e-mail and a $30 gift card for each respondent. We stopped replacing schools once we met our target of 50 to 60 schools per textbook that had at least one teacher complete our survey. We began with an original randomly selected sample of 368 schools (1,802 eligible teachers within these schools), and added 118 schools as part of our replacement strategy (see Table 1). Of the 486 total schools and 2,370 eligible teachers we contacted for this portion of the project, 1,195 teachers from 345 schools completed the survey and were successfully linked to administrative records (see Table 1). These numbers

[16] We originally planned on administering the teacher survey in the spring of 2017, to occur during the same school year as the school survey. However, we decided to delay in order to first ensure sufficiently high response rates on the school survey, which provided data for our main analyses. This delay likely is one reason for lower response rates on the teacher survey relative to the school survey. For example, our population of interest was all 4th- and 5th-grade math teachers working in randomly selected schools during the 2016/2017 school year; however, some of these teachers no longer were working in the same schools in the fall of 2017/2018 school year.

[17] For the teacher survey, we identified a target of 50 to 60 schools per textbook based on general benchmarks about sample sizes needed to approximate a population, as well as logistical constraints in recruiting, following up with, and compensating, teachers in these schools.

represent a school-level response rate of 71 percent and a teacher-level response rate of 50 percent.[18]

While the teacher-level response rate is comparable to other recent surveys with a similar focus (e.g., Opfer, Kaufman, & Thompson, 2016), it is lower than desired to draw generalizable conclusions about individual teachers. However, we remind readers that, given our sampling design, the primary unit of analysis is the school, where coverage was substantially higher. Low within-school response rates could lead to measurement error (and possible bias) about practices in those schools. As such, we view our analyses of the teacher survey data as exploratory. At the same time, patterns of textbook use were similar when we analyzed the teacher-level data versus school-level data that averaged teacher responses to the school level, indicating that variation in within-school response rates is unlikely to be a first-order concern. Further, in analyses where we used the teacher survey data to examine textbook effects across subgroups of schools (e.g., in schools with above- versus below-median levels of use), we found that results were similar when limiting the sample to schools with high teacher response rates.

Student Achievement and Demographic Data

All six partner states administered end-of-year mathematics assessments created by one of the two CCSS assessment consortia, the Partnership for Assessment of Readiness for College and Career (PARCC) or the Smarter Balanced Assessment Consortium (SBAC). At the time of our study, Louisiana, Maryland, New Jersey, and New Mexico were part of PARCC,[19] and California and Washington were part of SBAC. Although each consortium constructed its own test, both were designed to align to the same CCSS standards, and analyses of test items suggest that they cover similar content and with a similar level of rigor (Herman & Linn, 2014). Test scores were collected by state agencies, alongside a set of student demographic information including gender, race/ethnicity, eligibility for free or reduced-price lunch, eligibility to receive special education services, English Language Learner status, and grade level in school.

The primary research team worked with student achievement and demographic data for three states—Maryland, New Jersey, and New Mexico—with which we signed data use agreements. In the three remaining states—California, Louisiana, and Washington—the primary team coordinated with additional researchers who had access to the student-level data through their own agreements with state agencies. In these states, the partner researchers implemented similar statistical specifications to those estimated by the primary team and sent either data aggregated to the school level (in Louisiana and Washington) or parameter estimates (in California), both of which could be pooled with data/estimates from the remaining states.[20]

[18] Thirteen schools declined participation in the teacher survey as part of their response to our original school survey; 48 schools were from districts that declined participation; 20 school principals declined participation upon outreach; 21 schools did not provide sufficient information to identify teachers for recruitment; and 17 schools were no longer eligible to participate given that they switched textbooks from the prior year when we administered the school survey, the school closed from the prior year, or there no longer were eligible 4th- and 5th-grade teachers who also worked in the school the prior year. School-level participation rates in the teacher survey range from 44 to 80 percent across states (derived from numbers in Table 1). Participation rates by textbooks—which was the primary information we used to stratify the sample—are less variable, ranging from 59 to 86 percent (not shown in Table 1). As with school-level participation rates, teacher response rates vary more across states (36 to 64 percent; derived from numbers in Table 1) than across textbooks (42 to 63 percent; not shown in Table 1).

[19] Louisiana developed and administrated a hybrid PARCC/state test.

[20] Per data use agreements with state agencies, our partners in Louisiana and Washington were allowed to share annual school-level aggregates (derived from student-level value-added models described in the

In Louisiana, test score data were not yet available for the 2016/2017 school year by the time of our analyses.

EMPIRICAL METHODOLOGY

Value-Added Model

To examine differences in textbook efficacy, we used a two-step process where we first estimated school-level "value-added" and then used these scores as the outcome measure in a second-stage equation in which textbooks were the key predictors of interest. This two-step process allowed some partner states to share aggregated rather than student-level data.

First, we estimated school-level differences in achievement growth using the following model for student i in school j and grade g in year t, estimated separately for each state:

$$S_{ijgt} = \beta_0 + \beta_{1g} f\left(S_{it-1}{}^{Math}\right) + \beta_{2g} f\left(S_{it-1}{}^{ELA}\right) + \beta_3 X_{ijgt} + \theta_{gt} + \delta_{jt} + \varepsilon_{ijgt}. \quad (1)$$

We modeled students' current-year math achievement score (S) as a cubic function of prior achievement in math and English Language Arts (ELA).[21] We interacted these with grade fixed effects, allowing for different relationships between prior and current test scores across grades. In addition, we included grade-by-year fixed effects, θ_{gt}, to account for differences in scaling of tests at this level. We controlled for a vector of student characteristics, X_{ijgt}, including all demographic information provided by states and an indicator for students that repeated the current grade (at time t). Finally, we included school-by-year fixed effects, δ_{jt}, which were our parameters of interest from equation (1). The estimated school-by-year effects, $\widehat{\delta_{jt}}$, are commonly referred to as "value-added" estimates because they measure the degree to which students in a given school outperform or underperform other students with similar characteristics (Angrist et al., 2017). In equation (1), we conditioned on students' prior achievement scores to measure a student's achievement growth in the current year. Thus, even if the students used the same textbook in the prior year, we are measuring the effect of that textbook on a student's current year achievement growth. In a set of robustness tests, we also estimated textbook effects using school-by-grade-by-year value-added as the outcome of interest—estimated with one- versus two-year lagged achievement—to examine the possibility that textbooks may be a cross-grade intervention.

Empirical Methodology section). These data were pooled with school-level aggregates from Maryland, New Jersey, and New Mexico to estimate models of textbook effects using all five states. In California, the agreement between partner researchers and the state agency did not allow for sharing of any student- or school-level data with external teams. Therefore, the California team estimated textbook effects in those data alone and shared the estimates with the primary research team to pool with those from the other states.

[21] In California, the state did not record student-level test scores in the spring of 2014, as schools prepared for the new CCSS-aligned assessments to be administered in the spring of 2015. However, because California administers assessments to students in grades 2 through 8, we were able to use twice-lagged scores for both fourth and fifth graders in 2014/2015. Other work using the California data shows that the school-level value-added scores are not particularly sensitive to use of once- or twice-lagged achievement as a control (Carrell et al., 2018). For all states, if the prior score was missing in math (the primary outcome), we dropped that observation. If the prior score was missing in ELA, we created a flag for missing prior score, imputed the missing score to zero, and included the missing flag in the specification. We took the same approach for missingness of student demographic characteristics.

Next, we used the estimated school-by-year effects in a second stage to estimate a vector of textbook effects, μ_k, controlling for mean school characteristics, $\overline{X_{jt}}$, and characteristics of public school parents in the school district, Z_d, from the 2010 to 2014 American Community Survey (ACS):

$$\widehat{\delta_{jt}} = \gamma_0 + \sum_k \mu_k \, textbook\_k_{jt} + \gamma \overline{X_{jt}} + \lambda Z_d + \omega_{jt}. \quad (2)$$

We used *enVision* (CCSS edition) as the left-out category, since it was one of the few textbooks in high use in all states and years (see Table 3). Our parameter estimates, $\widehat{\mu_k}$, thus estimate differences in student achievement growth for a given textbook relative to schools using this left-out text. Coefficients for specific textbooks would differ if we changed the left-out category, but statistics from the test of whether textbook effects are jointly equal to zero—our primary estimand of interest—would not. When pooling data across years or states, we included state-by-year fixed effects.

Because textbooks typically vary at the district level, we needed to account for the fact that the school-by-year error term, ω_{jt}, is not independent for schools in the same district. We also needed to account for correlated errors for the same school over time. As a result, we separated the error term into three components: a district component, ϕ_d, a school component, χ_j, and an independent school-by-year error, ξ_{jt}:

$$\omega_{jt} = \phi_d + \chi_j + \xi_{jt}.$$

We estimated equation (2) using a hierarchical (random effects) model to account for the error terms at the district and school levels.[22]

When estimating the efficacy of individual textbooks using equation (2), we treated μ_k as a vector of fixed effects. While each should provide an unbiased estimate of the effect of an individual textbook, the variance across these parameter estimates overstates the underlying heterogeneity in textbook effects, since collectively they include sampling error at the state, district, and school levels. As a result, we also specified μ_k as a set of textbook random effects, with state-by-year fixed effects (in pooled analyses), and state, district, and school random effects nested within each textbook. Because the textbook random effect variance estimate is adjusted for the school-, district-, and state-level errors, we interpret it as an estimate of the "true" underlying variance in textbook efficacy. This specification allows for variation in value-added within textbook across states, districts, and schools, and estimates the component of variation in student achievement gains attributable to textbook effects that are common across states and over time.[23] While fixed effects specification allows the textbook effects to be correlated with covariates, the random effects specification assumes no correlation. Given use of observational data,

[22] An alternative approach would be to cluster standard errors at the district level. However, clustering can lead to overly optimistic standard errors when clusters have small numbers of observations (Cameron & Miller, 2015; Pustejovsky & Tipton, 2018). This is apparent in our data. In Figure A1, we plot the ratio of standard errors clustered at the district level to the random effects standard errors for each textbook, against the number of districts (clusters) using each textbook. For textbooks used in more than 15 districts, the standard errors were quite similar from the two methods. However, for the textbooks used in fewer than 15 districts, the standard error estimates differed dramatically depending upon the methods used, with standard errors for two of the textbooks falling by more than half when we clustered at the district level. We took this as evidence in favor of the random effects model.

[23] Because there are some districts using more than one text, we also explored models that crossed the state, district, and school random effects with the textbook random effects (e.g., non-nested) with similar results.

this may be too strong of an assumption. However, below we provide evidence to suggest that this assumption does not lead us to different answers across the two methods used to estimate textbook effects.

Identifying Assumptions

The key identifying assumption of our value-added approach is that textbook selection is uncorrelated with characteristics of districts, schools, and students, above and beyond those included in our model. However, since textbooks can vary in instructional approach, level of rigor, marketing materials, etc., we might also expect some differences in the baseline characteristics of schools adopting different texts (Bianchini & Kelly, 2003; Polikoff, 2018; Seeley, 2003; Tulley, 1985). In Table 4, we report the baseline characteristics of schools and districts adopting different textbooks, pooling across all six states. There are statistically significant differences in the observed characteristics of schools adopting different textbooks. (All tests of significance control for state fixed effects.) For example, the schools using *Stepping Stones* (CCSS edition), *My Math* (written for the CCSS), *enVision* (either pre- or post-CCSS edition), and *Engage NY/Eureka* (written for the CCSS) tended to be somewhat more disadvantaged. They had higher percentages of students receiving free or reduced-price lunches, had lower expenditures per student, and lower levels of parental education. In contrast, *Everyday Mathematics* (CCSS edition), *Bridges in Mathematics* (CCSS edition), and *Ready Common Core* (written for the CCSS) seemed to be used in somewhat more affluent schools, with lower percentages of students receiving federal subsidized lunches and higher levels of parental education. As a result, we included statistical controls for a variety of school-by-year demographic and test score measures and for district characteristics.[24]

At the same time, two additional sources of information suggest that nonrandom selection may not be as concerning when estimating textbook efficacy. First, patterns of selection do not appear to correlate with information published by EdReports. Recall that, with a lack of evidence on efficacy for the majority of CCSS-edition textbooks, publicly available reviews such as those provided by EdReports are likely to be a primary source of information for textbook adoption decisions (Polikoff et al., 2020). In Table 4, we sort textbooks by EdReports' ratings for alignment to the CCSS and find no clear pattern related to the observable characteristics of schools and districts using different texts. For example, of the four textbooks in our sample that received the highest rating by EdReports, two were used by schools that had well-below average prior math achievement; the two other texts from our sample in this high EdReports' rating category were used by schools that had well-above average prior math achievement.

Similarly, we observe that patterns of nonrandom selection are inconsistent across states. In Table A2, we disaggregate mean characteristics of schools using different textbooks across states, focusing on two characteristics that may be most likely to bias our textbook efficacy estimates: average prior math test scores (our outcome of interest), and percent of students eligible for free or reduced-price lunch (where we see some of the largest variation across textbooks; SD = 10 percentage points). In Table A3, we correlate these characteristics across states, finding weak negative

[24] The full set of school-by-year characteristics included all available student-level demographic characteristics averaged to the school-year level. The full set of district characteristics from the census included instructional expenditure per-pupil, median household income, percent of households that spoke a language other than English at home, percent of parents of school-aged children who were married, percent of parents who attended some college or hold an associate's degree (but no bachelor's degree), and percent of parents who held at least a bachelor's degree.

Table 4. Differences in school and district characteristics adopting different textbooks (top 15).

	School Characteristics						District Characteristics	
Textbooks (Sorted by EdReports Ratings for Alignment to CCSS)	Free or Reduced-Price Lunch (%)	Special Education (%)	English Language Learner (%)	African American (%)	Hispanic (%)	Prior Math Test Score (Standardized)	Per-pupil Instructional Expenditure ($)	Parent Holds BA Degree or Higher (%)
A. Minimum EdReports CCSS Alignment Rating: Meets Expectations								
Bridges in Mathematics CC	44.0	15.5	13.3	8.9	29.2	0.038	7,916	34.1
Engage NY/Eureka CC	58.7	13.2	20.4	10.3	45.5	−0.036	6,509	24.0
My Math CC	70.7	13.5	23.0	9.2	52.7	−0.136	6,187	19.6
Ready Common Core CC	43.4	13.2	14.7	3.1	27.0	0.102	6,072	34.4
B. Minimum EdReports CCSS Alignment Rating: Partially Meets Expectations								
Go Math CC	57.7	14.9	14.5	14.0	35.6	0.002	7,199	28.9
Math Expressions CC	54.7	13.6	23.5	7.2	38.3	0.022	5,991	29.3
C. Other CCSS Editions								
enVision CC	64.6	15.0	33.3	9.3	51.4	−0.047	6,093	23.7
Everyday Mathematics CC	47.3	13.7	19.5	7.9	20.3	0.006	5,982	36.9
Math in Focus CC	47.6	12.9	25.9	6.2	31.3	0.176	6,841	35.9
Stepping Stones CC	74.7	18.9	16.6	5.3	56.8	−0.042	5,424	25.8
D. Non-CCSS Editions								
enVision	65.3	12.8	13.9	27.5	26.8	−0.123	6,606	19.6
Everyday Mathematics	50.8	11.8	29.4	3.6	46.8	0.172	5,478	31.9
Houghton Mifflin Math	49.9	17.2	13.7	12.5	31.4	0.018	7,241	33.0
Math Connects	49.1	14.3	4.6	5.0	17.4	−0.024	6,802	25.7
Math Expressions	44.5	10.6	3.0	33.5	12.2	0.133	8,037	33.2
P-Value from Test of Joint Significance	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
School*Year Observations	10,797	10,797	10,797	10,797	10,797	10,797	10,797	10,797

Notes: Estimates are weighted by the inverse of a school's sampling probability. The sample is restricted to school-by-year observations known to have adopted one of the top 15 textbooks by market share. Joint tests of significance are estimated from models that include state-by-year fixed effects and cluster standard errors at the district level.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

to moderate positive relationships. For the three states with the broadest coverage of textbooks (i.e., California, New Jersey, Washington), cross-state correlations for school-average prior achievement associated with different textbooks range from −0.28 to 0.57. For average percentage of students eligible for free or reduced-price lunch in schools using different textbooks, correlations across these states are no higher than 0.49. Looking across all six states, several correlations are statistically significant and negative in magnitude. These findings are consistent with intuition suggesting that it is highly unlikely that certain types of schools are systematically choosing the most effective textbooks when school and district leaders (and we) do not know which textbooks those are.

Although we can control for observed school and district characteristics, it is possible that there are unmeasured determinants of student performance related to textbook adoption. Therefore, we also specified models that included school fixed effects. These models implicitly control for fixed, time-invariant differences between schools using different texts by focusing on changes in student achievement gains associated with changes in textbook adoptions. We also continued to control for time-varying, observable school characteristics. For our switcher models, our preferred sample comes from California, where we not only observed a large number of schools switching textbooks between the two years of textbook data available ($n = 1{,}060$) but also found switchers and non-switchers to look similar in terms of observed school and district characteristics ($p = 0.857$ on joint test of significance; see Table A4). Thus, estimates from the California switcher models are likely to generalize across elementary schools in this state. In the other five states, only 161 schools switched textbooks during the time frame of our study, and these schools differed from non-switchers ($p = 0.053$ on joint test of significance; see Table 4). Nonetheless, we show results across both samples with the caveat that school fixed effect estimates from our five-state sample excluding California may have limited generalizability.

RESULTS

Teacher-Reported Use of Textbooks

In Tables 5 and 6, we report estimates of teachers' use of textbooks in schools. These analyses include the sample of schools that adopted one of the top seven CCSS-edition textbooks by market share, which were the focus of the teacher survey sampling scheme. Given our sampling approach—which randomly selected schools (not teachers) for participation—we aggregated teacher-level responses to the school level, and then weighted these data by the inverse of the probability that a given school was included in the teacher survey (i.e., the product of the probability that a school was included in the textbook adoption data and the probability that the same school also was selected for the teacher survey). This approach aims to ensure that results generalize across our six partner states, at least for schools using one of these top seven CCSS-edition textbooks.[25]

In most schools, teachers reported using the adopted textbook frequently in their classes (see Table 5). Teachers in 78 percent of schools reported using their textbook for one of the listed purposes (e.g., creating tasks and activities, selecting examples,

[25] Use of a school-level dataset in Tables 5 and 6 masks variation in teacher-level response rates. However, patterns of textbook use are similar when drawing on the original teacher-level dataset (and also applying sampling weights), suggesting that lower-than-desired teacher response rates in some schools is unlikely to alter our primary conclusions.

Table 5. Teacher-reported use of textbooks (top 7 CCSS-editions) or other materials.

In what percentage of your lessons did you use textbook/materials for this purpose?	0% of lessons	1%–25%	26%–50%	51%–75%	More than 75%
Used textbook to:					
Choose objective (% of teachers)	2.3	5.7	7.5	20.4	64.1
Refresh content knowledge (%)	4.5	16.6	16.4	23.1	39.4
Create tasks and activities (%)	3.3	12.4	13.9	27.8	42.6
Select examples (%)	1.7	7.9	12.2	31.3	46.9
Assign independent classwork (%)	0.7	5.2	9.8	26.6	57.7
Choose homework problems (%)	4.4	9.6	8.9	23.2	54.0
Build assessments (%)	4.4	9.5	10.2	21.5	54.4
Any of the above (%)	0.3	2.8	3.0	16.1	77.8
Used other materials:					
State, district, or charter-produced materials (%)	17.4	25.3	10.6	17.9	28.9
Repositories on the web (%)	7.2	44.3	23.1	19.6	5.9
Materials created by teacher or with colleagues (%)	5.0	48.9	21.9	16.3	7.9
Materials created by other teachers in the school (%)	33.7	45.3	11.4	6.5	3.1
Materials from personal library (%)	19.6	46.6	17.8	10.8	5.2
Released test items (%)	13.5	48.5	15.8	14.5	7.7
Test-preparation books purchased by school/district (%)	53.4	26.9	8.4	6.2	5.1
Online content videos (%)	15.2	42.5	19.5	14.2	8.6
Online software (%)	35.2	27.7	14.9	12.5	9.7

Notes: Sample includes 1,194 teachers in 345 schools; one additional teacher with a valid survey skipped all questions reported in this table. Means were estimated by averaging teacher responses to the school level, and then applying a teacher survey sampling weight derived by dividing the school sampling weights by the probability that a given school was also selected to participate in the teacher survey. The percentages for "any of the above" reflect the maximum for the seven uses for each teacher. The sample is restricted to schools using one of the top seven CCSS-edition textbooks by market share, which were the focus of the teacher survey.

and building assessments) in at least 75 percent of lessons; teachers in 94 percent of schools used their textbook for one of these purposes in 50 percent or more of their lessons. Teachers also reported covering an average of 82 percent of chapters over the course of the school year (not shown in Table 5). At the same time, we found that teachers often supplemented or substituted into their lessons content from other sources. For example, teachers in 44 percent of schools used materials provided to them by the state, district, or school in more than half of their lessons. Teachers in about a quarter of schools reported using materials they found online or developed themselves in more than half of their lessons (see Table 5). Overall, teachers in only 8 percent of schools used their textbooks exclusively (see Table 6).

While Table 5 reports teachers' usage characteristics pooling across all seven texts that were a focus of the teacher survey, Table 6 examines differences by textbook. We found that teachers were likely to use supplemental materials with some textbooks more than others. In schools that adopted *Engage NY/Eureka,* teachers were more likely to use the textbook exclusively (18 percent of schools) than in schools that adopted other textbooks. For other texts such as *Go Math*, *Math Expressions,* and *Math in Focus*, teachers in fewer than 4 percent of schools used the textbook exclusively. Supplementing or substituting materials often was related to teachers' perception of the level of rigor of that text. For example, teachers in 32 percent of

Table 6. Teacher-reported supplementation or substitution, and amount of professional development by textbook (top 7 CCSS-editions).

	All 7 Textbooks	*Engage NY/Eureka* CC	*enVision* CC	*Everyday Mathematics* CC	*Go Math* CC	*Math Expressions* CC	*Math in Focus* CC	*My Math* CC
Reasons teachers used materials other than main textbook (not mutually exclusive):								
School or district requires use of other materials (%)	9.2	7.4	14.0	12.7	5.8	8.2	11.4	7.7
Textbook is too easy (%)	15.8	7.6*	20.3	8.3*	8.4*	8.5*	5.3***	31.5**
Textbook is too hard (%)	28.8	41.2	19.9	26.9	39.3	26.6	40.8	18.4~
Textbook does not cover all of the standards (%)	18.0	9.4*	23.8	9.0*	14.2	11.5~	35.8*	22.7
Textbook is not user friendly (%)	13.5	20.1	8.1	10.7	18.3	13.7	12.2	10.9
Examples in textbook are not sufficiently engaging for students (%)	51.7	40.3~	45.9	36.6*	51.4	48.5	49.9	72.4***
Access to materials used in the past (%)	41.7	32.0	38.3	48.0	52.5~	51.0	39.4	38.3
N/A – textbook used exclusively (%)	8.0	17.7~	8.9	5.2	3.6*	3.9~	3.4*	6.9
Access to professional development (PD):								
PD received this year (days)	5.4	5.4	5.6	5.0	4.7	5.3	4.9	6.1
PD specific to math received this year (days)	1.9	2.6*	1.5	1.9	2.0	1.6	2.3	1.8
PD specific to math textbook received this year (days)	1.1	1.8*	0.8~	1.2	1.0	0.8~	1.4	0.9
PD specific to math textbook received over entire career (days)	3.4	4.6*	2.7~	4.3	2.8	4.0	5.5*	2.7*
Provided with math coach this year (%)	37.1	51.0	31.7	39.3	33.6	48.9	41.2	27.3
Met with math coach (count)	1.3	1.9~	1.0	1.8	1.1	1.2	1.5	1.0
Observed teaching this year (count)	1.9	3.0***	1.8	2.1	1.8	1.3***	1.7	1.8
Received feedback about teaching this year (count)	14.6	16.6	18.2***	15.6	12.8	12.0*	13.8	12.0~

Notes: Sample includes 1,195 teachers in 345 schools. Means estimated by averaging teacher responses to the school level, and then applying a teacher survey sampling weight derived by dividing the school sampling weights by the probability that a given school was also selected to participate in the teacher survey. The sample is restricted to schools using one of the top seven CCSS-edition textbooks by market share, which were the focus of the teacher survey.
~p < .10; *p < .05; **p < .01; ***p < .001, comparing means for one textbook to all others combined.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

schools that adopted *My Math* indicated that they used other materials because they perceived the textbook to be "too easy," compared to teachers in 5 to 20 percent of schools using other textbooks. In contrast, teachers in roughly 40 percent of schools that adopted *Engage NY/Eureka, Go Math,* or *Math in Focus* reported using other materials because they perceived these textbooks to be "too hard" for their students. Teachers whose schools adopted *Math in Focus* also supplemented their textbook with other materials because they did not feel that it "covered all of the [CCSS] standards" (teachers in 36 percent of schools using this text, compared to teachers in 9 to 24 percent of schools using other textbooks).

Access to professional development—particularly math-specific programming and those aligned to textbook implementation—also differed by textbook. Teachers in schools that adopted *Engage NY/Eureka* reported receiving the most professional development tailored to the curriculum, but it was still modest: 1.8 days, on average, in the most recent school year, compared to 0.8 to 1.4 days, on average, for teachers in schools using other textbooks. Across all schools, teachers reported an average of just 3.4 days of textbook-aligned professional development over the course of their career. By design, teachers in schools using textbook series that were in print prior to the CCSS (i.e., *enVision*, *Everyday Mathematics, Math Expressions, Math in Focus*) tended to have more days of textbook-aligned professional developmemt, on average. Yet, for teachers in schools using *Engage NY/Eureka*—which is a newer text written after the rollout of the CCSS—the average amount of textbook-aligned professional development (4.6 days over the course of their career) was higher than the sample average and higher than the average in schools that adopted two other newer textbooks (i.e., *Go Math, My Math*).

Differences in use and support could influence a given textbook's efficacy. However, we did not attempt to "control for" such differences given that these factors almost certainly are endogenous. That is, implementation differences likely are a result of a given textbook's perceived strengths and flaws. Instead, we report below the average achievement gains of the schools using a given textbook *as adopted*. We do not report—because we cannot validly estimate—the magnitude of gain that a given school or district *could* have achieved with a given text *if* implemented in the ideal manner. The estimates presented below also are those most relevant to our policy question of interest: What is the effect of *adopting* a higher-quality textbook? The adoption *and* high-quality implementation of new textbook is a different, much more complicated, and much more expensive intervention to assess. We return to this topic in our conclusion.

Differences in Average Student Achievement Gains Across Textbooks

In Table 7, we report estimates from equation (2) of the average student achievement gains associated with different textbooks. In all models, estimates are reported in student-level SD of math achievement in a given state, grade, and year. All analytic samples are limited to school-years using one of the top 15 textbooks by market share (covering roughly 90 percent of school-year observations in our sample). In supplementary analyses, we also found similar results when expanding the sample to the full set of 38 known textbooks, as well as further restricting the sample to the top five or 10 textbooks that all had substantial shares of the market (see Table A5).

In some models in Table 7, we separate out the California estimates because the sample size is larger than all other states combined and data use agreements meant that the California team shared only these second-stage estimates. To pool second-stage estimates from California with those from others states, we used a precision-weighted average of the estimates from each (that is, weighting each set of estimates by the inverse of the variance-covariance matrix of the parameters). Estimating models by subsets of states and school years also allows us to examine the

Table 7. Differences in math achievement growth by textbook (top 15).

	Pooled 6 States: All Years (1)	Pooled 5 States (Excluding California): All Years (2)	California: All Years (3)	Pooled 6 States: 2014-15 (4)	Pooled 6 States: 2015-16 (5)	Pooled 4 States (Excluding California and Louisiana): 2016-17 (6)
A. Minimum EdReports CCSS Alignment Rating: Meets Expectations						
Bridges in Mathematics CC	0.010	0.069*	−0.020	0.016	0.051**	0.041
	(0.018)	(0.033)	(0.022)	(0.033)	(0.017)	(0.034)
Engage NY/Eureka CC	−0.003	0.021	−0.030*	−0.005	0.018	0.016
	(0.011)	(0.018)	(0.014)	(0.018)	(0.012)	(0.023)
My Math CC	0.019~	0.018	0.018	−0.018	0.003	−0.001
	(0.010)	(0.017)	(0.013)	(0.017)	(0.011)	(0.019)
Ready Common Core CC	−0.020	−0.015	−0.060	−0.039	0.015	−0.008
	(0.039)	(0.047)	(0.075)	(0.067)	(0.051)	(0.041)
B. Minimum EdReports CCSS Alignment Rating: Partially Meets Expectations						
Go Math CC	0.001	0.021	−0.006	0.009	0.013	−0.014
	(0.010)	(0.017)	(0.013)	(0.017)	(0.011)	(0.019)
Math Expressions CC	0.039**	0.022	0.046**	0.063**	0.034**	0.049~
	(0.013)	(0.025)	(0.016)	(0.020)	(0.013)	(0.027)
C. Other CCSS Editions						
Everyday Mathematics CC	−0.009	0.015	−0.018	−0.012	0.059***	0.021
	(0.012)	(0.019)	(0.017)	(0.021)	(0.015)	(0.021)
Math in Focus CC	0.002	0.000	0.003	0.039	0.013	−0.016
	(0.017)	(0.022)	(0.028)	(0.027)	(0.019)	(0.025)
Stepping Stones CC	0.010	−0.019	0.071	−0.024	0.037	−0.054*
	(0.031)	(0.036)	(0.061)	(0.060)	(0.034)	(0.026)

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table 7. Continued.

	Pooled 6 States: All Years (1)	Pooled 5 States (Excluding California): All Years (2)	California: All Years (3)	Pooled 6 States: 2014-15 (4)	Pooled 6 States: 2015-16 (5)	Pooled 4 States (Excluding California and Louisiana): 2016-17 (6)
D. Non-CCSS Editions						
enVision	−0.029**	−0.076*	−0.028*	0.024	0.017	0.077
	(0.010)	(0.034)	(0.013)	(0.017)	(0.017)	(0.052)
Everyday Mathematics	0.086***	−0.021	0.117***	0.028	0.014	0.030
	(0.015)	(0.028)	(0.019)	(0.024)	(0.023)	(0.048)
Houghton Mifflin Math	−0.020	0.046	−0.027	−0.060**	−0.001	
	(0.015)	(0.090)	(0.017)	(0.023)	(0.022)	
Math Connects	0.015	0.017		0.062	−0.003	0.046
	(0.028)	(0.029)		(0.048)	(0.035)	(0.036)
Math Expressions	−0.016	−0.014	−0.005	0.025	−0.003	0.000
	(0.034)	(0.039)	(0.082)	(0.052)	(0.043)	(0.040)
P-Value from Test of Joint Significance	0.000	0.323	0.000	0.000	0.000	0.223
SD of Textbook Fixed Effects	0.029	0.035	0.048	0.036	0.020	0.035
School*Year Observations	10,797	2,676	8,121	4,854	5,100	843

Notes: Estimates in each column come from separate linear regression models of school value-added. We report coefficients for a set of binary indicators for each textbook. The omitted textbook category is *enVision* CC, which is in the "other CCSS editions" category of EdReports ratings. All models are multi-level, mixed-effects models that include random effects for schools nested within districts. We control for school-by-year demographic characteristics, 2010 to 2014 district census characteristics, and state-by-year fixed effects (restricted to year fixed effects only or state fixed effects only in specifications limited to a single state or year). The sample is restricted to school-by-year observations with value-added data for the 2014/2015, 2015/2016, or 2016/2017 school years who are known to have used one of the top 15 textbooks by market share. The 2016/2017-specific estimates come from Maryland, New Jersey, New Mexico, and Washington, and exclude California (which did not have textbook data in that year) and Louisiana (which did not have value-added data in that year). Empty cells indicate that no school in a given set of states used that textbook.
~p < .10; *p < .05; **p < .01; ***p < .001.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

sensitivity of textbook effects across states and time. California and Louisiana are excluded for the 2016/2017 estimates due to data constraints (no textbook adoption data in California in this year, and no student achievement data in Louisiana in this year).

In the pooled sample across all six states in model (1), three texts have student achievement gains statistically significantly different from the left-out category, *enVision* (CCSS edition): two with larger gains (*Everyday Mathematics* pre-CCSS and *Math Expressions* CCSS edition), and one with smaller gains (*enVision* pre-CCSS). However, these results are driven by the sample of schools in California. When we exclude California in model (2), we see two individual coefficients that are statistically different from zero (*enVision* pre-CCSS and *Bridges in Mathematics* CCSS edition). When we test the stronger hypothesis that *all* differences between textbooks are equal to zero (the p-value reported at the bottom of the table), we cannot reject it ($p = 0.323$). In other words, when comparing such a large number of textbooks, it is possible that one or two are significantly different from each other by chance. However, when looking across the full set of textbooks, there were so few such instances that we could not reject the hypothesis of no difference between textbooks.

In model (3), we report estimates for the California sample on its own. Here, *Math Expressions* (CCSS edition) and *Everyday Mathematics* (pre-CCSS edition) are associated with better math test-score gains relative to *enVision* (CCSS edition), while schools using *Engage NY/Eureka* or the pre-CCSS edition of *enVision* have lower test-score gains. Compared to the five-state sample that excludes California, we could reject the null hypothesis that textbook effects were jointly equal to zero ($p < 0.001$). The same is true in the pooled six-state sample (model 1), driven by the large sample size in California. However, even for textbooks where we do see statistically significant differences, estimates generally are far smaller than those reported in prior evaluations.

Notably, no single text stands out as a consistent high or low performer in multiple states, nor in multiple years. In models (4) through (6), individual textbooks that appear to be effective in one year (e.g., *Everyday Mathematics* CCSS edition, *Stepping Stones*) often are not identified as effective in another year. In Table A6, we report the correlation between textbook efficacy estimates across states and years. The estimated correlation between the California estimates and those in the other five states, for example, is mildly negative (-0.16). The correlations between textbook efficacy estimates across years range from -0.02 to 0.56 and also point to a lack of stability.

We also see little to no relationship between textbook efficacy and alignment to standards. In California, for example, *Engage NY/Eureka* has a negative coefficient relative to *enVision* (CCSS edition), but was rated by EdReports as "meeting expectations" for alignment to the CCSS. The pre-CCSS edition of *Everyday Mathematics* has the largest observed relationship to student achievement gains, but, by design, was not rated by EdReports because that edition was not meant to align to the CCSS. (The updated CCSS edition of *Everyday Mathematics* did not meet nor partially meet expectations, as rated by EdReports.) Similar discrepancies between efficacy and alignment to standards are observed in other states and samples.

Next, we describe results from our textbook switcher models to examine whether our primary estimates are driven by unobserved characteristics at the school level. In models presented in Table 8, we expand the analytic sample to include schools using any known textbook in order to capture instances where a school switched from a non-CCSS textbook (relatively rare in the top 15 textbooks) to a CCSS edition. However, we continue to show estimates for the top 15 textbooks by market share. Textbook coefficients in Table 8 are identified off of those schools that switched texts. Therefore, we report estimates from models that include all schools (and school

Table 8. Marginal effects for textbook switchers (all known textbooks).

	Pooled 5 States: With School Fixed Effects (1)	Pooled 5 States: Switchers Only, No School Fixed Effects (2)	California: With School Fixed Effects (3)	California: Switchers Only, No School Fixed Effects (4)
A. Minimum EdReports CCSS Alignment Rating: Meets Expectations				
Bridges in Mathematics CC	0.546***	0.038	−0.025	−0.011
	(0.050)	(0.079)	(0.070)	(0.036)
Engage NY/Eureka CC	0.048	0.004	−0.009	−0.049*
	(0.047)	(0.037)	(0.062)	(0.025)
My Math CC	0.070	−0.022	0.038	0.069**
	(0.067)	(0.053)	(0.027)	(0.021)
Ready Common Core CC		0.169	0.077	−0.136
		(0.111)	(0.120)	(0.110)
B. Minimum EdReports CCSS Alignment Rating: Partially Meets Expectations				
Go Math CC	0.062	0.005	0.017	0.006
	(0.048)	(0.040)	(0.024)	(0.022)
Math Expressions CC	−0.068	−0.113	0.175	0.097*
	(0.105)	(0.072)	(0.121)	(0.040)
C. Other CCSS Editions				
Everyday Mathematics CC	−0.004	0.008	0.026	−0.009
	(0.049)	(0.043)	(0.039)	(0.025)
Math in Focus CC	0.026	−0.044	−0.038	0.010
	(0.040)	(0.062)	(0.088)	(0.059)
Stepping Stones CC	0.262***	0.007		
	(0.052)	(0.066)		
D. Non-CCSS Editions				
enVision	−0.149**	−0.242***	0.016	−0.010
	(0.057)	(0.058)	(0.024)	(0.023)
Everyday Mathematics	−0.009	−0.074~	0.134***	0.137***
	(0.033)	(0.045)	(0.040)	(0.030)
Houghton Mifflin Math	0.087	0.027	0.032	−0.014
	(0.060)	(0.094)	(0.038)	(0.028)
Math Connects	0.147***	−0.022		
	(0.044)	(0.057)		
Math Expressions			−0.037	−0.172
			(0.119)	(0.170)
School Fixed Effects	Yes	No	Yes	No
P-Value from Test of Joint Significance	0.000	0.002	0.000	0.000
SD of Textbook Fixed Effects	0.271	0.144	0.078	0.100
School*Year Observations	2,805	379	8,711	2,115

Notes: Estimates in each column come from separate linear regression models of school value-added. We report coefficients for a set of binary indicators for each textbook. The omitted textbook category is *enVision* CC, which is in the "other CCSS editions" category of EdReports ratings. Models (1) and (3) include school fixed effects, and the sample is restricted to school-by-year observations with value-added data for the 2014/2015, 2015/2016, or 2016/2017 school year who used a known textbook (excluding the distict-developed naterials or no known textbook categories from Table 3). We also control for school-by-year demographic characeristics. Models (2) and (4) limit the sample to schools that are known to have switched textbooks sometime between the 2014/2015 and 2016/2017 school years, but exclude school fixed effects. Empty cells indicate that no school switched to or from that textbook to or from another known textbook. Switcher models cluster standard errors at the district level.
~p < .10; *p < .05; **p < .01; ***p < .001.

fixed effects) as well as from models that limit the sample to switchers only (excluding school fixed effects).

In California, which is our preferred sample for the school fixed effects analysis given a large number of switchers that look similar to non-switchers (see Table A4) textbook efficacy estimates are similar with or without conditioning on school fixed effects. The correlation between the estimates in California from models that include or exclude school fixed effects is 0.58 (see Table A6). The correlation between the textbook efficacy estimates from the California switcher model and the pooled six-state sample that excludes school fixed effects is 0.63. The primary difference is that the positive estimate for *Everyday Mathematics* (pre-CCSS edition) is larger. Because many districts in California were switching out of the pre-CCSS edition of *Everyday Mathematics,* this may simply reflect the fact that achievement fell in those districts in the first year of a new text. Given the fact that we do not see similar results in the other states for this same textbook, we hesitate to take these as evidence of the unusual efficacy of the pre-CCSS edition of *Everyday Mathematics*. Rather, we take the findings from our school fixed effect model as consistent with what we see in the remainder of the table: that there are no substantial, consistent differences between the textbooks in our sample.

We also report estimates from our textbook switcher models in the other five states (pooled). However, we are cautious in interpreting these estimates given the small number of schools in these five states that switched textbooks during the time frame of our study and the fact that these schools are not representative of broader state populations (see Table A4). Here, we observe statistically significant variation in student achievement gains across textbooks when we include school fixed effects (see model 1), though this estimate is driven entirely by the unique set of schools that switched textbooks. When we estimate the model excluding school fixed effects and limiting the sample to switchers (see model 2), we observe greater variation across textbooks than we observed in the full set of schools in these states (see model 2 in Table 7).

The Underlying Variation in Textbook Efficacy

It is easy to lose track of the underlying story amidst the large number of parameter estimates presented in Tables 7 and 8. Individual estimates may appear large in a given state or year even if they are being driven by estimation error. As a result, rather than report estimates of the efficacy of individual textbooks, in Table 9 we report estimates of the SD of the textbook random effect, which we take as an estimate of the underlying heterogeneity in textbook efficacy. The estimates in Table 9 are essentially the square root of the underlying variance in textbook efficacy, after adjusting for state, school, and district sampling errors. With no individual textbook as a consistent positive or negative outlier (see Tables 7 through 9), the model-based estimate of the textbook-level variance is, in fact, the exact estimand of interest.[26]

We report random effect estimates for California separately from the remaining five states, as the data use agreement did not allow us to have access to the

[26] As described in our Empirical Methodology section, fixed and random effects specifications make different assumptions—namely, whether the textbook effects are correlated with covariates included in the model. Several tests indicate that comparison of fixed and random effects specifications is reasonable in our data. First, we find correlations between textbook effects specified as a set of fixed versus random effects of 0.86 (for pooled five-state sample excluding California) and 0.57 (for California). For this calculation, our "random effect" estimates are residuals averaged to the textbook level. We took this approach because hierarchical, random effects models employ a shrinkage factor, while fixed effect parameters do not. This shrinkage factor (a ratio less than one) would artificially attenuate the correlation between textbook fixed versus random effect estimates. In the teacher effectiveness literature, random effects and average residual models are described as comparable because they both calculate effects from the error

Table 9. The standard deviation in textbook efficacy (top 15).

Random Effects Parameters	Pooled 5 States (Excluding California) (1)	California (2)	Louisiana (3)	Maryland (4)	New Mexico (5)	New Jersey (6)	Washington (7)
Textbook	0.000	0.027***	0.000	0.000	0.016	0.017	0.028~
	–	(0.008)	–	(0.000)	(0.011)	(0.026)	(0.015)
State	0.013						
	(0.010)						
District	0.054***	0.083***	0.000	0.095**	0.049**	0.062**	0.052*
	(0.007)	(0.005)	–	(0.031)	(0.015)	(0.021)	(0.022)
School	0.071***	0.060***	0.073**	0.091***	0.052***	0.081***	0.065***
	(0.006)	(0.005)	(0.024)	(0.017)	(0.015)	(0.011)	(0.017)
Residual	0.139***	0.146***	0.169***	0.113***	0.118***	0.150***	0.133***
	(0.003)	(0.002)	(0.011)	(0.009)	(0.004)	(0.005)	(0.005)
School*Year Observations	2,676	8,121	292	191	840	822	531

Notes: Estimates in each column come from separate models. Random effects are estimated from a multilevel mixed-effects linear regression of school-level value added on school-by-year demographic characteristics, 2010 to 2014 district census characteristics, and state-by-year fixed effects (restricted to year fixed effects only in specifications limited to a single state). The model also includes nested random effects for textbook, state, district, and school, nested in that order, with textbook as the top level of the nesting structure (state random effects are excluded from regressions that are limited to a single state). The sample is restricted to school-by-year observations with value-added data for the 2014/2015, 2015/2016, or 2016/2017 school year who are known to have used one of the top 15 textbooks by market share. Robust standard errors are in parentheses. "–" indicates that the relevant parameter could not be estimated. Empty cells indicate that the random effect parameter was not included in the model. ~z >1.64, *z >1.96, **z >2.58, ***z >3.29, where z equals the ratio of a given random effects parameter estimate to its standard error. These z-scores do not correspond precisely to p-values as in a traditional linear regression framework, as the confidence interval for a random effect estimate is not symmetric around the estimate (random effect estimates have a lower bound of zero). These traditional markers of significance are included as an aid to reader, but they should be interpreted with caution.

school-level value-added data for that state. Unlike for fixed effects specifications, where we could use variance-covariance matrices to aggregate estimates across samples, we would have needed the unit-level data to estimate a pooled random effect across all six states. For our textbook random effect models, we also focus on our primary specification that excludes school fixed effects given similarity in patterns of results for textbook fixed effect estimates that include versus exclude school fixed effects.

In the pooled sample of five states (excluding California), we estimate that the variance (and SD) in efficacy across textbooks is zero SD.[27] This is consistent with our failure to reject the joint hypothesis of no textbook differences in many of the

rather than the fixed portion of the model (Guarino et al., 2015; Kane & Staiger, 2008). Second, we find a correlation of 0.99 (for both samples) on the coefficients for the set of school- and district-level covariates included in both the random and fixed effect models. Finally, when conducting a Hausman test, we cannot reject the null hypothesis that one specification produces more or less efficient estimates than the other, at least in the pooled five-state sample excluding California ($p = 0.952$). In California, despite superficially similar estimated textbook effects, we can reject this null hypothesis ($p < 0.001$), potentially driven by the very large number of school-year observations in that state.

[27] Random effects models have known challenges when estimates are close to zero (Harville, 1977; Raudenbush & Bryk, 2002). For example, when the estimated variance approaches zero, the standard error is

models presented in Table 7. Although we found some individual differences with respect to the reference textbook, the overall pattern of differences between texts was not inconsistent with zero difference between texts. In California, while we found evidence of differences in textbook efficacy with our textbook fixed effects models (see Table 7), the underlying SD in textbook efficacy from our random effects specification is much smaller at 0.027 SD. In other words, California schools using a textbook at the 95th percentile—roughly 2 SD above the mean—would expect to perform 0.054 SD above schools using an average textbook. Although positive, this estimate still is smaller than the differences estimated in the two experiments that found textbook effects (i.e., Agodini et al., 2010; Jaciw et al., 2016). Our estimate of the SD in textbook efficacy for the state of Washington is similar to that in California (0.028 SD). The estimated SD in textbook efficacy in the four remaining states all are quite close to zero.[28]

Even though we see little variation in textbook efficacy overall, we could be overlooking small differences between subsets of texts due to lack of statistical power. To investigate, we conducted a series of simulations in which we assigned one text an efficacy estimate ranging from 0.02 to 0.15 student-level SD above the reference textbook, and with a market share ranging from 1 to 25 percent (see Table A7).[29] We assigned all other texts an efficacy of zero. We then used equation (2) to estimate the SD in underlying textbook efficacy. Above an estimated effect of 0.1 SD and 5 percent market share, or above 0.05 SD effect and 20 percent market share, detection rates reach 99 percent or higher. It was when the single textbook was equal to or below 0.03 SD more effective that we would have failed to reject zero variance in textbook efficacy more than 80 percent of the time (no matter the market share). Yet, effects of this magnitude are of less substantive significance relative to critical benchmarks including differences in teacher quality and the human resources investment that likely is needed to switch textbooks.

undefined (see Table 9, model 1, for example). To confirm that our estimates are true zeros, we estimated results to 10 decimal places, finding similar results. We also fit models using both full and restricted maximum likelihood. Because full maximum likelihood tends to produce estimates that are biased downward (Harville, 1977), we prefer (and present) estimates fit using restricted maximum likelihood. Indeed, in this exercise, estimates generated using restricted maximum likelihood are the same or slightly larger than estimates generated from full maximum likelihood.

[28] One explanation for differences in random effect estimates between California and Washington versus the other four states may be that the former two states used the SBAC assessment, while the remaining states used the PARCC test. However, California and Washington do not identify the same textbooks as most or least effective. The correlation of textbook fixed effect estimates between and fixed effect estimates from Washington is 0.3.

[29] To estimate statistical power, we generated 100 simulations for each combination of market share (1 percent, 5 percent, 10 percent, 20 percent and 25 percent) and single-text effect size (0.02, 0.03, 0.05, 0.10, 0.15 SD). In each run, we stripped 2,676 school-by-year observations of their textbook data and randomly assigned one of 15 fake curricula within district-by-textbook clusters. That is, all schools within a given district that were observed using a given textbook were assigned a common fake textbook in each simulation for all years they appear in the data. To replicate real textbook adoption behavior, 6 percent of schools were chosen to switch to a new textbook in year two, and an additional 6 percent of schools were chosen to switch to a new textbook in year three. Schools that switched to a new text in year two kept that text in year three unless they were randomly chosen to switch in both years. Of the fake textbooks that schools were assigned, 14 of 15 were designed to have no effect on student outcomes in the simulation. As such, school-years assigned one of these 14 textbooks kept their original value-added. For school-years randomly assigned the single textbook identified as effective in the simulation, their value-added was increased by the amount in the effect size for that particular simulation. By design, this increase in value-added is attributable to that textbook, so the simulation assesses whether our random effects model is able to detect and correctly attribute systematic variation in value-added to a school's choice of textbook, for larger and smaller textbook effects distributed over larger and smaller shares of the sample of schools. The market-share percentage indicates what percentage of schools were assigned the effective textbook for a set of simulations. The random effect estimate for each simulated run was stored, and this process was repeated 100 times for each combination of effect size and market-share.

Heterogeneity in Textbook Efficacy

In addition to accounting appropriately for sampling error, our random effects specification allows us to easily disaggregate textbook effects by subgroups. It is possible, for example, that there may be larger differences in student achievement gains across textbooks for groups of students with specialized academic needs. In Table A8, we present estimates of the SD in underlying textbook efficacy by English Language Learner status, eligibility for special education services, eligibility for free or reduced-price lunch, and median split of baseline math achievement. Here, we estimated equations (1) and (2) in a single step with student-level data, limiting the sample to specific subgroups of students. Because we needed student-level data to do so, we specified these models for the three states for which the primary research team had access to student-level data: Maryland, New Jersey, and New Mexico; our partners in California did the same for that state on its own. The estimates of the SD in textbook efficacy were not statistically different by subgroup. In the pooled three-state sample, point estimates and standard errors are zero to three decimal places; we generally observe non-zero values at the fifth or sixth decimal place.

All estimates presented thus far focus on the average efficacy of textbooks *as implemented* in classrooms. Accordingly, they reflect the difference in average achievement gains for schools adopting each text averaged across varying levels of fidelity of implementation. However, if teachers in a subset of schools were substituting other materials—as we observe in Tables 5 and 6—or if a subset of schools just adopted their text and were not yet familiar with it, or if teachers received little professional development in the use of the text, we may be understating the differences. As a result, in Table 10, we also present estimates of the underlying variation in textbook efficacy after dividing schools with regard to four dimensions of implementation: teacher-reported usage, amount of textbook-aligned professional development, pre- versus post-CCSS textbook editions, and years of experience with a given text. Because of the endogeneity of implementation, we view these analyses as exploratory, highlighting possible mechanisms to explain our null findings presented thus far.

Variation in Textbook Efficacy by Levels of Use

First, we created an index of textbook usage for different purposes, including choosing lesson objective, creating tasks and activities, and selecting examples (see Table 5 for survey item text),[30] averaging across items within teachers and then across teachers within schools. Then, we split schools into two groups, depending on whether they were above or below the median in teacher-reported textbook use. On average, teachers in above-median usage schools used the textbook for one of a range of purposes in roughly 81 percent of lessons, while teachers in below-median usage schools used the textbook in roughly 53 percent of lessons. Information on usage comes from our teacher survey, and thus was available for a subset of schools using one of the top seven CCSS-edition textbooks by market share. To maximize the sample size, we categorized schools as above- or below-median usage based on their reported usage in 2016/2017, but then used all available years of achievement gains.[31]

[30] In our usage index, we excluded two questions that refer to teachers' use of a textbook outside of class rather than for instructional activities with students: "choose objective" and "refresh content knowledge."
[31] Also to maximize statistical power, we included all possible schools no matter within-school teacher response rate. Estimates for all subgroup analyses that split schools based on teacher survey responses were similar when we re-ran estimates restricting to schools with two or more teacher responses, three or more teacher responses, 50 percent or higher teacher response rates, and 70 percent or higher teacher response rates.

Table 10. Heterogeneity in textbook random effect estimates by measures of implementation.

Random Effects Parameters	Pooled 5 States (Excluding California)		California	
Panel A (Top 7 CCSS-Edition Textbooks)	Above-Median Teacher-Reported Use	Below-Median Teacher-Reported Use	Above-Median Teacher-Reported Use	Below-Median Teacher-Reported Use
Textbook	0.000	0.000	0.000	0.037
	–	–	–	(0.046)
School*Year Observations	309	315	90	89
Panel B (Top 7 CCSS-Edition Textbooks)	Above-Median Teacher-Reported PD Aligned to Textbook	Below-Median Teacher-Reported PD Aligned to Textbook	Above-Median Teacher-Reported PD Aligned to Textbook	Below-Median Teacher-Reported PD Aligned to Textbook
Textbook	0.000	0.000	0.000	0.000
	–	–	–	(0.000)
School*Year Observations	330	294	91	88
Panel C (Top 15 Textbooks)	CCSS Edition	Not CCSS Edition	CCSS Edition	Not CCSS Edition
Textbook	0.000	0.013	0.022*	0.042~
	–	(0.043)	(0.008)	(0.023)
School*Year Observations	2,476	200	6,221	1,900

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table 10. Continued.

Random Effects Parameters	Pooled 5 States (Excluding California)		California	
Panel D (Top 15 Textbooks)	2+ Years Since Adoption	Year 1 of Adoption	2+ Years Since Adoption	Year 1 of Adoption
Textbook	0.005	0.000	0.035**	0.000
	(0.023)	–	(0.011)	–
School*Year Observations	1,590	81	2,482	979

Notes: Estimates in each cell come from separate models. We estimate the standard deviation in textbook effects using a multilevel mixed-effects linear regression of school-level value added on school-by-year demographic characteristics, 2010 to 2014 district census characteristics, and state-by-year fixed effects (restricted to year fixed effects only in specifications limited to a single state). The model also includes nested random effects for textbook, state, district, and school, nested in that order, with textbook as the top level of the nesting structure (state random effects are excluded from regressions that are limited to a single state). The sample is restricted to school-by-year observations within the indicated subgroup with value-added data for the 2014/2015, 2015/2016, or 2016/2017 school year. In panels A and B, the sample includes schools that were sampled for the teacher survey and used one of the top seven CCSS-edition textbooks by market share. In panels C and D, the sample includes schools known to have used one of the top 15 textbooks by market share. To identify above- and below-median levels of textbook usage, we averaged teachers' response to five survey questions about frequency of textbook use (measured on a Likert scale from 0–4; see Table 5) across items and then across teachers within the school; then we split the sample of schools at the median value of this average (3.4, corresponding to 72.5 percent of lessons). Schools whose average response to the textbook usage questions were at or above the median are identified as "above-median usage" schools (median = 3.7, corresponding to 80 percent of lessons), while schools whose average response were below the median are identified as "below-median usage" schools (2.8, corresponding to 57.5 percent of lessons). To identify schools with above- or below-median levels of professional development (PD) aligned to the textbook, we averaged teachers' response regarding the number of total days across their career they participated in textbook-aligned PD, and we split the sample at the median value of this average (3 days). Schools whose average response to PD questions were at or above the median are identified as "above median" schools (median = 6 days), while schools whose average response were below the median are identified as "low PD" schools (median = 1.5 days). Textbooks identified as a "CCSS Edition" are those identified with "CC" in prior tables (i.e., published after 2011, and publisher indicated textbook was written for or adapted to the CCSS). Robust standard errors in parentheses; "–" indicates that standard errors could not be estimated. ~z > 1.64, *z > 1.96, **z > 2.58, ***z > 3.29, where z equals the ratio of a given random effects parameter estimate to its standard error. These z-scores do not correspond precisely to p-values as in a traditional linear regression framework, as the confidence interval for a random effect estimate is not symmetric around the estimate (random effect estimates have a lower bound of zero). These traditional markers of significance are included as an aid to reader, but they should be interpreted with caution.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

If teacher substitution of materials was blurring the differences between more and less effective texts, we might expect to see greater variation in textbook effects in the schools with above-median levels of usage. However, we find little evidence of differences in textbook efficacy in either above- or below-median usage schools. In panel A of Table 10, we estimated zero variance in underlying textbook efficacy for above-median usage schools in our pooled sample of five states (excluding California), as well as in California on its own. We also estimated zero variation in textbook efficacy for below-median usage schools in the five-state sample. In California, the estimate for below-median usage schools in California is 0.037 SD.

Variation in Textbook Efficacy by Days of Textbook-Aligned Professional Development

To get the most out of high-quality curriculum materials, districts or schools may have to provide more training and support for teachers (for a review of this large body of research, see O'Donnell, 2008). In panel B of Table 10, we disaggregated results for schools where teachers reported that they received above-median levels of professional development aligned to their textbook versus schools where teachers reported below-median levels of support. Across all schools in the teacher survey sample, the median amount of textbook-aligned professional development reported by teachers was three days over the course of their careers. In the above-median support schools, the median number of days of textbook-aligned professional development was six; in the below-median support schools, it was 1.5 days. Despite their differing levels of support, we found no difference in textbook efficacy within either of these two sets of schools, either in our pooled sample of five states (excluding California) or in California on its own.

Textbook Efficacy Among Pre- and Post-CCSS Texts

With the advent of the CCSS, publishers had a single set of standards around which to write their texts. Before the CCSS, when states each had their own standards, publishers had an incentive to broaden their coverage to contain nearly every states' standards in a given grade. Every textbook was "a mile wide and an inch deep" to accommodate multiple state standards (to paraphrase Schmidt et al., 2001; Schmidt, McKnight, & Raizen, 1997), but on any given standard some textbooks may have gone deeper than others. One of the goals of the CCSS was to allow textbook publishers to focus on a shorter list of standards in more depth. Our failure to find a difference in textbook efficacy may simply mean that the policy goal was achieved: that textbooks differed less in their content and coverage—and, in turn, in their efficacy—under the CCSS than they had previously.

Even after states adopted the CCSS, a nontrivial share of schools still were using pre-CCSS textbook editions: 37 percent of schools across our six states were using pre-CCSS editions in 2014/2015, and 16 percent still were using pre-CCSS editions in the latest year of data available. In panel C of Table 10, we estimated the underlying variance in textbook efficacy separately for the samples of schools using pre- and post-CCSS editions. Here, we could leverage our original school-survey sample, and not be limited to the schools with teacher surveys. In California as well as in the remaining five states, the point estimate of the underlying variance in efficacy is larger among the pre-CCSS texts than among the CCSS editions: 0.042 SD versus 0.022 SD in California, and 0.013 SD versus zero SD in the remaining five states. Although consistent with the hypothesis that the pre-CCSS texts differed more from each other than the post-CCSS texts, the differences are not large.

Variation in Textbook Efficacy by Years Since Adoption

Finally, it could be that the variation in textbook efficacy is muted during the first year following adoption, as teachers gain facility with the text and incorporate the new material into their lessons. Teachers and schools may have to use a given text for more than one year before the efficacy differences emerge. In panel D of Table 10, we split schools into two subgroups based on whether the school was in its first year of usage or not. Here, we leveraged our full school-survey sample, but limited to the top 15 textbooks by market share. In the pooled five-state sample (excluding California), there was no difference in the underlying variation in textbook efficacy, whether schools were in their first year of use or in their second year or higher. In California, however, we do see evidence that the SD in textbook efficacy is larger in the schools using a text for two or more years (0.035 SD, compared to zero SD for those schools using a textbook for one year). In other words, the non-zero estimate of textbook efficacy in California from prior analyses (see Table 9) is coming from the schools that had been using the text for more than one year.

In addition to returns for *schools* of using a text for more than one year, it also is possible that there is a benefit when *students* use the same textbook series across years and grades. The content would not be exactly the same, but the approach to the material would be similar. To examine this possibility, in Table A9 we disaggregate our efficacy estimates by grade (fourth versus fifth) and examine the sensitivity of these estimates to use of one- versus two-year lagged test scores. If textbooks are a cross-grade intervention, controlling for the prior-year test score may attenuate the textbook effect. Because most states administer end-of-year tests starting in third grade, we could only compare results for one- versus two-year test score lags for fifth grade. We present results for all school-year observations, as well as those schools that kept the same text for at least two consecutive years; this approach ensures that 5th-grade students were exposed to the same series in that grade and in the prior grade. We excluded Louisiana from these analyses, as the data use agreement did not allow for disaggregation and sharing of value-added estimates by school, grade, and year. We also had to exclude some estimates for California due to data constraints (i.e., two rather than three years of textbook data, and a one-year hiatus in test score reporting in the lead-up to the new CCSS-aligned assessment).

Consistent with results presented in Table 10, in Table A9 we find some evidence of greater variation amongst 5th-grade students, who had more exposure to a given text, relative to 4th-grade students. In the pooled four-state sample (excluding California and Louisiana) and in California on its own, point estimates for 4th-grade students are zero. For 5th-grade students in California, estimates range from 0.02 to 0.04 SD, with larger estimates when controlling for 3rd-grade rather than 4th-grade achievement and for 5th-grade students in non-switcher schools. In the other four states, we find small, non-zero point estimates for 5th graders in some models.

Additional Robustness Checks to Test Identifying Assumptions

Interpreting textbook estimates as causal effects requires an assumption of conditional exogeneity. That is, controlling for observable school and district characteristics accounts sufficiently for nonrandom selection of textbooks by schools and districts. In addition to specifying models that replace observable characteristics with school fixed effects (see Table 8), we found that our primary random effect estimates were robust to different subsets of the school- and district-level controls (see Table A10, panels A through C).

Finally, we conducted a placebo test to examine whether math textbooks had an "impact" on ELA achievement, which they should not. To do this, we needed to

condition our estimates on the ELA textbook used. Otherwise, we might simply find that the districts and schools that succeeded in choosing more effective math textbooks also were good at choosing ELA textbooks. It is only conditional on the ELA textbook that the math textbook should be irrelevant for gains in students' ELA test scores. In California, the only state where we had data on ELA textbooks, we found that math textbooks do appear to have a statistically significant relationship to ELA achievement (see Table A10, panel D). As expected, though, the relationship is small and roughly 50 percent of the magnitude of math textbooks on math achievement for this state (see Table 9).

RECONCILING WITH THE PREVIOUS LITERATURE

How do we reconcile our results with the previous literature, which suggested larger differences in student achievement gains between schools using different texts? There are several possibilities. First, the literature on the efficacy of alternative curricula is still in its infancy. While we did not find evidence of large differences in achievement gains for schools using different texts, it could be that the inclusion of more years and more states would point to larger differences than we have seen. Since we completed our data collection, new textbooks and curriculum materials have entered the market. Given the potential value of the curriculum lever, we hope future research will continue to try to resolve the differences between our findings and the earlier research.

A second possibility is that our value-added methodology is biased. As a result, we could be understating the differences in textbook efficacy. In contrast to the value-added methodology, the randomized trials would have ensured that schools using different texts were similar. Our school fixed effects models aim to account for fixed, time-invariant differences across schools, and we found that estimates from these models were similar to those from other models that excluded school fixed effects. However, school fixed effects cannot account for time-varying characteristics such as turnover of school leadership that may be correlated with textbook selection and textbook efficacy. At the same time, if we were understating the efficacy of textbooks based on bias due to unmeasured school characteristics, it would be an unusual form of bias in which low-growth schools were using better textbooks, and high-growth schools were using less effective ones. The bias due to unmeasured school characteristics would have had to be of equal and opposite magnitude to the textbook effects. Typically, we would have expected to see unmeasured traits exaggerating the efficacy of interventions, with more advantaged (or better-managed) schools compounding their advantage by purchasing more effective textbooks.

A third possibility is that the findings from the randomized trials are not generalizable. Although randomized trials accurately reflect the causal effect of textbooks for the population studied, the population of schools that were willing to have their textbook randomly assigned may have been unusual. In the Agodini et al. (2010) experiment, the small percentage of districts (2.5 percent) that were willing to participate in such a study may have been particularly dissatisfied with the texts they were using and may have benefited more from the change than other districts would have. Similar to Agodini et al. (2010), Eddy et al. (2014) had to contact over 6,000 school districts and principals to recruit nine participating schools (less than 0.15 percent participation rate). Just as other non-experimental studies have found, we see evidence of differences in achievement for some textbooks when focusing on one state. However, no single text proved to be more or less effective across multiple states. The fact that *Saxon Math* was among the most effective texts in the randomized trial by Agodini et al. (2010) but among the least effective texts in Bhatt and

Koedel (2012) may reflect the same lack of generalizability.[32] We recognize the usual trade-off between internal and external validity. In our study, we placed a stronger priority on the latter, while still assessing identifying assumptions and threats to internal validity.

A fourth, related possibility is that the answer is different in upper-elementary grades (where our study focused because of our need for prior achievement controls) than in lower-elementary grades (where much of the prior research was concentrated). In other words, the results of a given study may not generalize across grades. The randomized trial conducted by Agodini et al. (2010) focused on first and second graders, and the non-experimental studies by Koedel and his co-authors focused primarily on 3rd-grade achievement (Bhatt & Koedel, 2012; Bhatt, Koedel, & Lehmann, 2013; Koedel et al., 2017). The types of math skills tested in early elementary grades (e.g., single digit addition and subtraction) may be more sensitive to interventions of all types than the math skills measured on the 4th- and 5th-grade assessments (e.g., fractions, multi-digit multiplication and division). In one of the randomized studies, Jaciw et al. (2016) did find modest textbook effects in third through fifth grade; but their analysis was limited to a single textbook, *Math in Focus*.

A fifth possibility is that the answer has changed since the earlier studies were completed. In the pre-CCSS era, when textbooks were covering a broader range of topics, textbooks may have differed more in their alignment with any specific test being used to measure efficacy. For instance, if a test included several items measuring students' ability to add fractions with unlike denominators, then the textbooks that emphasized that standard may appear more effective than another text. Yet, on a different test, with fewer items measuring that standard, the textbook rankings may change. This may explain inconsistent results in the Jaciw et al. (2016) randomized trial, where schools using the CCSS edition of *Math in Focus* outperformed on the Stanford Achievement Test but not on the state test. In our study, on the CCSS-aligned tests, we also saw some evidence that this explanation may have been true. In California, where we had a large sample of schools using pre-CCSS editions of textbooks, we did find more variation in student achievement gains for schools using these texts, compared to less variation for schools using a post-CCSS edition textbook.

CONCLUSION

The adoption of the Common Core State Standards led schools in many states to switch curricula. Thus, the choice of textbook and curriculum has become much more salient, and often is perceived as a decision with high stakes (Pondiscio, 2017; Steiner, 2017). Contrary to prior research, though, we found little evidence of differences in average student achievement growth for elementary schools using different math textbooks in six states that adopted the CCSS standards and were using CCSS-aligned assessments. We did find some evidence of greater variation in achievement gains among those using pre-CCSS texts. Yet, the effect size was small.

[32] A related possibility is that the randomized trial by Agodini et al., while being unbiased, was conflating the efficacy of certain textbooks with schools' prior experience with the textbook. One-quarter of schools had been using *Saxon Math* in the year before the experiment. *Saxon Math* also was the textbook with the highest measured "efficacy" in that experiment. Although the authors included a control for whether or not an *individual teacher* had used the text before, the final results did not include controls for whether or not *the school* had been using the randomly assigned curriculum previously. An individual teacher new to *Saxon Math* working in a school that had been using *Saxon Math* previously may not be as disadvantaged as those in schools where no one had been using this text. This raises the question of construct validity: Were the authors measuring the efficacy of *Saxon Math* or were they witnessing the advantage of not having to transition to a new curriculum?

Some may interpret our findings as implying that curriculum choice does not matter. We believe that would be an overstatement. We remind readers that our estimates compare student achievement gains between schools using different textbooks, which say nothing about the difference between adopting a standard textbook versus adopting nothing. (Table 3 shows that very few schools adopted no textbook.) In exploratory analyses, we found some evidence of greater variation among schools and students who had spent more than one year using a specific text, suggesting that there may be returns to experience.

It may also be that we just did not see sufficiently intensive usage or training in our sample to detect the differences in student achievement between texts. Although teachers in the vast majority of sampled schools (94 percent) reported using the official textbook for *some* purpose in a majority of their classes, few teachers hewed closely to the text. Teachers in just 25 percent of schools reported using the textbook in nearly all of their lessons and for multiple purposes: "to create mathematical tasks and activities," "to select examples to present in class," "as a source of practice problems that students work on independently during class time," *and* "as a source of problems for students to complete outside of class." Similarly, teachers reported modest amounts of training in the use of their texts. The average teacher received just one day of training in the current year, and fewer than four days over their entire careers. Even in the schools with above-average levels of training, teachers reported receiving six days of training in their text over the course of their careers. Given districts' investments in curricula, these do not seem like large expenditures of time or funding.

In light of the patterns of implementation, some readers may continue to advocate for the importance of curriculum choice, despite our overall null results related to efficacy. However, those who want to hold on to the importance of curriculum as a primary lever for reform need to be able to identify the level of support and training required for such curriculum changes to actually bear fruit in the classroom. It is possible that closer adherence to a high-quality curriculum would produce benefits, but we still need to answer several questions: What levels of support are required to produce greater levels of adherence? Do the desired student achievement benefits appear afterwards? What are the costs associated with these supports, and are they justified given the observed effects?

Citing the earlier research, Chingos and Whitehurst (2012) posed a choice between "challenging, expensive, and time consuming" efforts to improve teaching quality and the "relatively easy, inexpensive, and quick" choice of a higher-quality curriculum. While our findings certainly cast doubt on the proposition that there are quick and easy payoffs to curriculum changes, the bigger error may be in thinking of curriculum choice and teaching reforms as alternatives. It could be that in order to gain the benefits of either, districts must do both.

DAVID BLAZAR is an Assistant Professor in the College of Education at the University of Maryland, College Park, 2205 Benjamin Building, 3942 Campus Drive, College Park, MD 20742 (e-mail: dblazar@umd.edu).

BLAKE HELLER is a PH.D. Candidate in the Kennedy School of Government at Harvard University (e-mail: bheller@g.harvard.edu).

THOMAS J. KANE is a Professor in the Graduate School of Education at Harvard University, Center for Education Policy Research, 50 Church Street, 4th Floor, Cambridge, MA 02138 (e-mail: tom_kane@gse.harvard.edu).

MORGAN POLIKOFF is an Associate Professor at the Rossier School of Education at the University of Southern California, Waite Phillips Hall #904A, 3470 Trousdale Parkway, Los Angeles, CA 90089 (e-mail: polikoff@usc.edu).

DOUGLAS O. STAIGER is a Professor in the Department of Economics at Dartmouth College, HB 61606, Rockefeller Hall, Dartmouth College, Hanover, NH 03755 (e-mail: douglas.o.staiger@dartmouth.edu).

SCOTT CARRELL is a Professor in the Department of Economics at the University of California at Davis, 1 Shields Avenue, Davis, CA 95616 (e-mail: secarrell@ucdavis.edu).

DAN GOLDHABER is Director of the Center for Analysis of Longitudinal Data in Education Research at the American Institutes for Research and the Center for Education Data and Research at the University of Washington at Seattle, 3876 Bridge Way N., Suite 201, Seattle, WA 98103 (e-mail: dgoldhab@uw.edu).

DOUGLAS N. HARRIS is Professor and Chair of the Department of Economics at Tulane University, 302 Tilton Memorial Hall, 6823 St. Charles Avenue, New Orleans, LA 70118 (e-mail: dharri5@tulane.edu).

RACHEL HITCH is Director of Program Management at Meteor Learning, 301 Edgewater Place, Suite 210, Wakefield, MA 01880.

KRISTIAN L. HOLDEN is a Researcher at the Center for Analysis of Longitudinal Data in Education Research at the American Institutes for Research, 1050 Thomas Jefferson Avenue, Washington, DC 20007 (e-mail: kholden@air.org).

MICHAL KURLAENDER is Professor and Department Chair in the School of Education at the University of California at Davis, 1 Shields Avenue, Davis, CA 95616 (e-mail: mkurlaender@ucdavis.edu).

ACKNOWLEDGMENTS

We gratefully acknowledge funding from the Bill & Melinda Gates Foundation, the Charles and Lynn Schusterman Foundation, the William and Flora Hewlett Foundation, and the Bloomberg Foundation. The research reported here also was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305B150010 to Harvard University and Grant R305E150006 to the Regents of the University of California (Michal Kurlaender, Principal Investigator, UC Davis School of Education) in partnership with the California Department of Education (Jonathan Isler, Co-Principal Investigator). The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences, the philanthropic funders, the departments of education in any of the participating states, the research institutions, nor the advisory board members.

Thomas Kelley-Kemple, Jake Kramer, and Virginia Lovison at Harvard University; Lihan Liu at Tulane University; Matthew Naven and Derek Rury at the University of California–Davis provided excellent research assistance. Rachel Urso and Sophie Houstoun at the Center for Education Policy Research at Harvard University led the recruitment of schools and teachers for our surveys, and Nate Brown at the University of Washington led additional outreach in Washington state. Eric Hirsch, Lauren Weisskirk, and Mark LaVenia at EdReports provided invaluable support and feedback in understanding the breadth of curriculum choices and textbook alignment to the Common Core. Our project depended upon the collaboration and support of our state partners, including John White, Jessica Baghian, Kim Nesmith, Rebecca Kockler, and Alicja Witkowski at the Louisiana Department of Education; Carol Williamson and Debra Ward from the Maryland Department of Education; Peter Shulman, James Riddlesperger, LaShona Burke, and Jessica Merville at the New Jersey Department of Education;

and Christopher Ruzkowski and Anthony Burns from the New Mexico Public Education Department. The study also benefited from the experience and feedback from an advisory board of experts on curriculum design and value-added methodology, including: Matthew Chingos (Urban Institute), Erin Grogan and Dan Weisberg (TNTP), Cory Koedel (University of Missouri), Darleen Opfer and Julia Kaufman (RAND Corporation), Grover J. "Russ" Whitehurst (Brookings), and Jason Zimba (Student Achievement Partners). Finally, we thank the thousands of district leaders, school principals, administrators, and classroom teachers who generously provided input about their curriculum choices to ensure that our project was a success.

REFERENCES

Agodini, R., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). Achievement effects of four early elementary school math curricula: Findings for first and second graders. National Center for Education Evaluation and Regional Assistance. Retrieved June 24, 2020, from https://files.eric.ed.gov/fulltext/ED512551.pdf.

Altonji, J. G., Blom, E., & Meghir, C. (2012). Heterogeneity in human capital investments: High school curriculum, college major, and careers. Annual Review of Economics, 4, 185–223.

Angrist, J. D., Hull, P. D., Pathak, P. A., & Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. The Quarterly Journal of Economics, 132, 871–919.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. The Quarterly Journal of Economics, 118, 1279–1333.

Beck Evaluation & Testing Associates, Inc. (2005). Progress in Mathematics 2006: Grade 1 pre-post field test evaluation study. New York, NY: William H. Sadlier, Inc.

Bhatt, R., & Koedel, C. (2012). Large-scale evaluations of curricular effectiveness: The case of elementary mathematics in Indiana. Educational Evaluation and Policy Analysis, 34, 391–412.

Bhatt, R., Koedel, C., & Lehmann, D. (2013). Is curriculum quality uniform? Evidence from Florida. Economics of Education Review, 34, 107–121.

Bianchini, J. A., & Kelly, G. J. (2003). Challenges of standards-based reform: The example of California's science content standards and textbook adoption process. Science Education, 87, 378–389.

Boser, U., Chingos, M., & Straus, C. (2015). The hidden value of curriculum reform: Do states and districts receive the most bang for their curriculum buck? Washington, DC: Center for American Progress. Retrieved June 24, 2020, from https://cdn.americanprogress.org/wp-content/uploads/2015/10/06111518/CurriculumMatters-report.pdf.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. Journal of Human Resources, 50, 317–372.

Carmichael, S. B., Martino, G., Porter-Magee, K., & Wilson, W. S., (2010). The state of state standards—and the Common Core—in 2010. Thomas B. Fordham Institute. Retrieved June 24, 2020, from https://files.eric.ed.gov/fulltext/ED516607.pdf.

Carrell, S., Kurlaender, M., Martorell, P., & Naven, M. (2018). The impacts of high school quality on postsecondary outcomes: Evidence from California. Working Paper. Davis, CA: California Education Lab, University of California–Davis.

Cavanaugh, S. (2015, June 9). N.Y. "open" education effort draws users nationwide. Education Week. Retrieved June 24, 2020, from https://www.edweek.org/ew/articles/2015/06/10/ny-open-education-effort-draws-users-nationwide.html.

Chingos, M. M., & Whitehurst, G. J. (2012). Choosing blindly: Instructional materials, teacher effectiveness, and the Common Core. Washington, DC: Brookings. Retrieved June 24, 2020, from https://www.brookings.edu/wp-content/uploads/2016/06/0410_curriculum_chingos_whitehurst.pdf.

Eddy, R. M., Hankel, N., Hunt, A., Goldman, A., & Murphy, K. (2014). Houghton Mifflin Harcourt GO Math! Efficacy study year one final report. La Verne, CA: Cobblestone Applied Research & Evaluation, Inc. Retrieved June 24, 2020, from https://www.hmhco.com/~/media/sites/home/educators/education-topics/hmh-efficacy/HMH_Go_Math_RCT_Yr1_2014.pdf.

Friedberg, S., Barone, D., Belding, J., Chen, A., Dixon, L., Fennell, F., … Shanahan, T. (2018). The state of state standards post-Common Core. Thomas B. Fordham Institute. Retrieved June 24, 2020, from https://files.eric.ed.gov/fulltext/ED592393.pdf.

Fryer Jr., R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In Handbook of Economic Field Experiments (Vol. 2, pp. 95–322). North-Holland.

Gatti, G., & Giordano, K. (2010). Pearson Investigations in Numbers, Data, & Space efficacy study: Final report. Pittsburgh, PA: Gatti Evaluation, Inc.

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P. N., & Wooldridge, J. M. (2015). An evaluation of empirical Bayes's estimation of value-added teacher performance measures. Journal of Educational and Behavioral Statistics, 40, 190–222.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American statistical Association, 72, 320–338.

Heck, D. J., Weiss, I. R., & Pasley, J. D. (2011). A priority research agenda for understanding the influence of the Common Core State Standards for Mathematics. Technical Report. Chapel Hill, NC: Horizon Research, Inc. Retrieved June 24, 2020, from https://files.eric.ed.gov/fulltext/ED544302.pdf.

Herman, J., & Linn, R. (2014). New assessments, new rigor. Educational Leadership, 71, 34–37.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. Child Development Perspectives, 2, 172–177.

Holden, K. L. (2016). Buy the book? Evidence on the effect of textbook funding on school-level achievement. American Economic Journal: Applied Economics, 8, 100–127.

Hutt, E., & Polikoff, M. (Spring 2018). Reasonable expectations: A reply to Elmendorf and Shanske 2018. University of Illinois Law Review Online, 194–208. Retrieved June 24, 2020, from https://illinoislawreview.org/wp-content/uploads/2018/05/HuttPolikoff.pdf.

Jaciw, A. P., Hegseth, W. M., Lin, L., Toby, M., Newman, D., Ma, B., & Zacamy, J. (2016). Assessing impacts of Math in Focus, a "Singapore Math" program. Journal of Research on Educational Effectiveness, 9, 473–502.

Jackson, K., & Makarin, A. (2018). Can online off-the-shelf lessons improve student outcomes? Evidence from a field experiment. American Economic Journal: Economic Policy, 10, 226–254.

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. NBER No. w14607. Cambridge, MA: National Bureau of Economic Research. Retrieved June 24, 2020, from https://www.nber.org/papers/w14607.pdf.

Kirst, M. W. (1982). How to improve schools without spending more money. The Phi Delta Kappan, 64, 6–8.

Koedel, C., Li, D., Polikoff, M. S., Hardaway, T., & Wrabel, S. L. (2017). Mathematics curriculum effects on student achievement in California. AERA Open, 3. Retrieved June 24, 2020, from https://journals.sagepub.com/doi/full/10.1177/2332858417690511.

McFarland, J., Hussar, B., de Brey, C., Snyder, T., Wang, X., Wilkinson-Flicker, S., … Hinz, S. (2017). The condition of education 2017. U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved June 24, 2020, from https://files.eric.ed.gov/fulltext/ED512551.pdf.

National Council of Teachers of Mathematics. (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

National Council of Teachers of Mathematics. (1991). Professional standards for teaching mathematics. Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. Reston, VA: Author.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. Review of Educational Research, 78, 33–84.

Opfer, V. D., Kaufman, J. H., & Thompson, L. E. (2016). Implementation of K–12 state standards for mathematics and English language arts and literacy. Santa Monica, CA: RAND. Retrieved June 24, 2020, from https://www.rand.org/content/dam/rand/pubs/research_reports/RR1500/RR1529-1/RAND_RR1529-1.pdf.

Pellegrini, M., Lake, C., Inns, A., & Slavin, R. E. (2018). Effective programs in elementary mathematics: A best-evidence synthesis. Baltimore, MD: Best Evidence Encyclopedia. Retrieved June 24, 2020, from http://www.bestevidence.org/word/elem_math_Oct_8_2018.pdf.

Polikoff, M. S. (2015). How well aligned are textbooks to the Common Core standards in mathematics? American Educational Research Journal, 52, 1185–1211.

Polikoff, M. S. (2018). The challenges of curriculum materials as a reform lever. Washington, DC: Brookings. Retrieved June 24, 2020, from: https://www.brookings.edu/research/the-challenges-of-curriculum-materials-as-a-reform-lever/.

Polikoff, M. S., Campbell, S., Rabovsky, S., Koedel, C., Le, Q. T., Hardaway, T., & Gasparian, H. (2020). The formalized processes districts use to evaluate mathematics textbooks. Journal of Curriculum Studies, 52, 451–477.

Pondiscio, R. (2017). Louisiana threads the needle on ed reform: Launching a coherent curriculum in a local-control state. Education Next, 17, 8–15.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core standards: The new U.S. intended curriculum. Educational Researcher, 40, 103–116.

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. Journal of Business & Economic Statistics, 36, 672–683.

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

Remillard, J. T., Harris, B., & Agodini, R. (2014). The influence of curriculum material design on opportunities for student learning. International Journal on Mathematics Education (ZDM), 46, 735–749.

Reys, R. E. (2001). Curricular controversy in the math wars: A battle without winners. Phi Delta Kappan, 83, 255–258.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. American Economic Review, 94, 247–252.

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H. C., Wiley, D. E., Cogan, L. S., & Wolfe, R. G. (2001). Why schools matter: A cross-national comparison of curriculum and learning. San Francisco, CA: Jossey-Bass.

Schmidt, W. H., McKnight, C. C., & Raizen, S. (1997). A splintered vision: An investigation of U.S. science and mathematics education. Dordrecht, Netherlands: Kluwer.

Schoenfeld, A. H. (2004). The math wars. Educational Policy, 18, 253–286.

Seeley, C. L. (2003). Mathematics textbook adoption in the United States. In G. M. A. Stanic & J. Kilpatrick (Eds.), A history of school mathematics (Vol. 2, pp. 957–988). Reston, VA: National Council of Teachers of Mathematics.

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. Review of Educational Research, 78, 427–515.

Stein, M. K., Remillard, J., & Smith, M. S. (2007). How curriculum influences student learning. In F. K. Lester (Ed.), Second handbook of research on mathematics teaching and learning (Vol. 1, pp. 319–370). Reston, VA: National Council of Teachers of Mathematics.

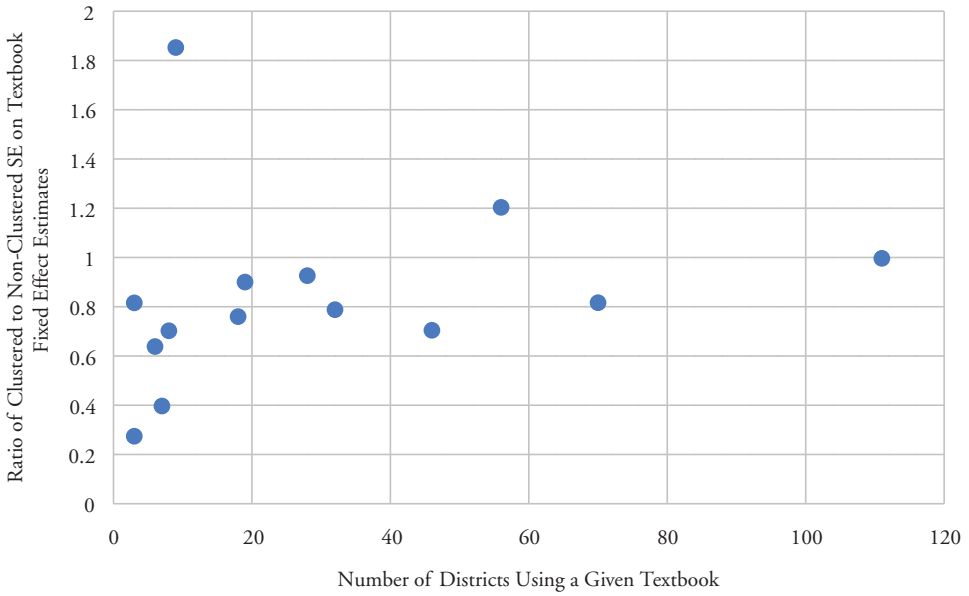
Steiner, D. (2017). Choosing a curriculum: A critical act. Education Next. Retrieved June 24, 2020, from https://www.educationnext.org/choosing-curriculum-critical-act/.

Strobel, A., Resendez, M., & DuBose, D. (2017). enVision Math 2.0 Year 2 RCT Study Final Report. Thayne, WY: Strobel Consulting, LLC.

Tulley, M. A. (1985). A descriptive study of the intents of state-level textbook adoption processes. Educational Evaluation and Policy Analysis, 7, 289–308.

Whitehurst, G. J. (2009). Don't forget curriculum. Washington, DC: Brookings. Retrieved June 24, 2020, from https://www.brookings.edu/research/dont-forget-curriculum/.

APPENDIX



Notes: Textbook fixed effects are estimated from a mixed-effects linear regression of school value-added on a set of binary indicators for whether a school used a given textbook, school-by-year demographic characteristics, 2010 to 2014 district census characteristics, and state-by-year fixed effects. The omitted textbook category is *enVision* CC. The graph plots the ratio of standard errors clustered at the district level to the random effects standard errors for each textbook, against the number of districts (clusters) using each textbook. The sample is restricted to school-by-year observations in Louisiana, Maryland, New Jersey, New Mexico, and Washington with value-added data for the 2014/2015, 2015/2016, or 2016/2017 school year that are known to have used one of the top 15 textbooks by market share.

Figure A1. Clustered versus Non-Clustered Standard Errors by Textbook. [Color figure can be viewed at wileyonlinelibrary.com]

Table A1. Characteristics of participating states.

State Characteristics	U.S.	California	Louisiana	Maryland	New Jersey	New Mexico	Washington
Public School Enrollment, 4th and 5th Grade	7,736,742	958,634	109,825	136,269	201,131	52,050	169,753
Free or Reduced-Price Lunch (%)	52.3	58.1	63.0	46.7	37.9	71.4	43.6
Special Education (%)	13.7	12.2	11.8	12.1	16.9	15.8	12.9
English Language Learner (%)	9.6	20.2	3.2	7.8	5.0	13.4	11.1
Male (%)	51.3	51.3	51.2	51.0	51.3	51.1	51.5
African American (%)	15.2	5.5	43.6	33.7	15.3	1.9	4.4
Asian (%)	5.2	11.6	1.6	6.6	10.0	1.1	7.7
Hispanic (%)	26.7	54.3	6.9	17.4	28.7	61.7	23.2
Native American (%)	1.0	0.5	0.7	0.3	0.1	10.0	1.2
Mixed Race or Other (%)	3.9	4.5	2.5	4.6	2.0	1.9	8.0
NAEP, 4th Grade Math (Standardized)	0.000	−0.258	−0.355	0.032	0.258	−0.323	0.065
NAEP, 4th Grade Reading (Standardized)	0.000	−0.184	−0.263	0.079	0.289	−0.368	0.026
Per-pupil Instructional Expenditure ($)	7,445	7,203	6,354	9,515	11,762	5,694	6,940
Parents Married (%)	62.7	65.5	51.9	61.3	67.0	57.1	66.9
Speaks Language Other than English (%)	22.8	47.0	6.2	17.3	29.4	32.7	22.9
Median Household Income ($)	55,600	64,200	46,200	76,700	74,500	46,400	63,600
Parent Attended College, no BA (%)	32.3	29.3	32.8	28.8	25.7	34.9	35.4
Parent Holds BA Degree + (%)	29.8	26.5	19.7	39.2	40.6	22.1	31.7

Notes: Enrollment, student demographic, and expenditure data come from the 2016/2017 school year. Test score data come from the 2017 National Assessment of Educational Progress (NAEP), which is standardized based on the national mean and standard deviation. Additional data on parents and families come from the 2010 to 2014 American Community Survey.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table A2. Observable characteristics of schools using different textbooks (top 15) by state.

Textbooks (Sorted by EdReports Ratings for Alignment to CCSS)	California	Louisiana	Maryland	New Jersey	New Mexico	Washington
Panel A	Prior Math Test Score (Standardized)					
A. Minimum EdReports CCSS Alignment Rating: Meets Expectations						
Bridges in Mathematics CC	0.159					0.062
Engage NY/Eureka CC	0.012	0.007	−0.624	−0.559		−0.147
My Math CC	−0.150	0.024	−0.017	−0.262	−0.063	−0.117
Ready Common Core CC	0.278	−0.258	0.160			
B. Minimum EdReports CCSS Alignment Rating: Partially Meets Expectations						
Go Math CC	−0.027	−0.018	0.084	−0.065	−0.178	0.07
Math Expressions CC	0.056	0.165		0.040		−0.061
C. Other CCSS Editions						
enVision CC	−0.014	−0.141	−0.092	0.037	0.05	0.183
Everyday Mathematics CC	0.023		−0.061	0.054	−0.135	0.366
Math in Focus CC	0.161			−0.21	−0.181	0.011
Stepping Stones CC	0.184		−0.077		−0.049	−0.101
D. Non-CCSS Editions						
enVision	−0.078	−0.518	0.116	0.328		0.327
Everyday Mathematics	0.159		0.55	−0.01	1.491	−0.067
Houghton Mifflin Math	0.173			0.326		
Math Connects			0.068	0.006		−0.059
Math Expressions	0.247			0.292		−0.024

Table A2. Continued.

Textbooks (Sorted by EdReports Ratings for Alignment to CCSS)	California	Louisiana	Maryland	New Jersey	New Mexico	Washington
Panel B	Free or Reduced-Price Lunch (%)					
A. Minimum EdReports CCSS Alignment Rating: Meets Expectations						
Bridges in Mathematics CC	43.4					43.4
Engage NY/Eureka CC	59.9	69.1	74.6	61.1		62.7
My Math CC	72.4	72.4	45.5	51.6	81.9	53.5
Ready Common Core CC	47.9	93.4	40.9			
B. Minimum EdReports CCSS Alignment Rating: Partially Meets Expectations						
Go Math CC	64.2	59.0	33.4	39.9	95.5	68.4
Math Expressions CC	59.2	40.9		25.7		48.8
C. Other CCSS Editions						
enVision CC	65.0	84.0	50.9	38.0	89.5	38.2
Everyday Mathematics CC	50.4		65.2	35.5	83.1	18.7
Math in Focus CC	50.3			49.0	67.5	47.9
Stepping Stones CC	53.5		76.4		78.2	64.2
D. Non-CCSS Editions						
enVision	67.7	83.0	57.4	11.2		30.2
Everyday Mathematics	54.7		20.2	41.8	12.6	50.3
Houghton Mifflin Math	51.1			12.8		
Math Connects			29.9	44.4		56.0
Math Expressions	51.3			15.9		48.1
School*Year Observations	8,121	292	191	840	822	531

Notes: Estimates are weighted by the inverse of a school's sampling probability. The sample is restricted to school-by-year observations known to have adopted one of the top 15 textbooks by market share.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table A3. Pairwise correlations between observable characteristics of schools choosing different textbooks (top 15) across states.

States	California	Louisiana	Maryland	New Jersey	New Mexico	Washington
Panel A			Prior Math Test Score			
California	1.000					
Louisiana	−0.082	1.000				
Maryland	0.296	−0.496	1.000			
New Jersey	0.324	−0.606	0.642~	1.000		
New Mexico	0.365	−0.714	0.935**	0.279	1.000	
Washington	−0.281	−0.865*	0.142	0.566~	−0.297	1.000
Panel B			Free or Reduced-Price Lunch (%)			
California	1.000					
Louisiana	−0.162	1.000				
Maryland	−0.086	0.031	1.000			
New Jersey	0.191	0.000	0.028	1.000		
New Mexico	0.367	−0.477	0.544	−0.119	1.000	
Washington	0.152	−0.503	−0.147	0.493	−0.001	1.000

Notes: Correlations are estimated from a textbook-level dataset where variables are the mean background characteristic (i.e., prior math test score, percent eligible for free or reduced-price lunch) in each of six states, and each observation is one of the top 15 textbooks by market share.
~p < .10; *p < .05; **p < .01.

Table A4. Comparing schools that switched textbooks to those that did not.

	Pooled 5 States (Excluding California)		California	
School or District Characteristics	Switchers	Difference for Non-Switchers	Switchers	Difference for Non-Switchers
Free or Reduced-Price Lunch (%)	47.4	11.7	64.4	−2.3
Special Education (%)	15	0.7	11.7	0.1
English Language Learner (%)	7.3	1.5	24.1	1.1
Male (%)	51.5	−0.3	50.9	0.1
African American (%)	15.3	−0.9	7.1	−1.7
Asian (%)	5.4	−0.5	11.8	−0.9
Hispanic (%)	17.9	13.1*	55.8	−2.6
Native American (%)	1.9	2	0.6	0.2
Mixed Race or Other (%)	3.2	−0.4	2.9	0.6
Prior Math Test Score (Standardized)	0.001	−0.009	−0.023	0.012
Prior Reading Test Score (Standardized)	0.016	−0.013	−0.037	0.029
Per-pupil Instructional Expenditure ($)	8,073	−783	5,818	−162
Parents Married (%)	65.6	−4.1	63.9	1.7
Speaks Language Other than English (%)	18.7	5.4*	51.6	−7.2
Median Household Income ($)	75,433	−12,047	60,658	1,972
Parent Attended College, no BA (%)	30.6	0.2	26.6	3.0~
Parent Holds BA Degree + (%)	31.7	−4.3	23.7	−0.2
P-Value from Test of Joint Significance		0.053		0.857
School Observations	161	987	1,060	4,047

Notes: Expenditure data and family characteristics comes from the 2010 to 2014 American Community Survey, captured at the district level. Other characteristics captured at the school level.
~p < .10; *p < .05.

Table A5. Sensitivity of fixed effect estimates to textbook sample (pooled 6 states, all years).

	All Known Textbooks (1)	Top 15 Textbooks (2)	Top 10 Textbooks (3)	Top 5 Textbooks (4)
A. Minimum EdReports CCSS Alignment Rating: Meets Expectations				
Bridges in Mathematics CC	0.011	0.010	0.01	
	(0.017)	(0.018)	(0.018)	
Engage NY/Eureka CC	−0.004	−0.003	−0.001	−0.008
	(0.011)	(0.011)	(0.011)	(0.011)
My Math CC	0.018~	0.019~	0.018*	0.013
	(0.010)	(0.010)	(0.01)	(0.01)
Ready Common Core CC	−0.019	−0.020		
	(0.039)	(0.039)		
B. Minimum EdReports CCSS Alignment Rating: Partially Meets Expectations				
Go Math CC	0.003	0.001	0.006	0.007
	(0.010)	(0.010)	(0.01)	(0.01)
Math Expressions CC	0.038**	0.039**	0.038***	
	(0.013)	(0.013)	(0.013)	
C. Other CCSS Editions				
Everyday Mathematics CC	−0.009	−0.009	−0.009	
	(0.012)	(0.012)	(0.012)	
Math in Focus CC	0.003	0.002	0.004	
	(0.017)	(0.017)	(0.017)	
Stepping Stones CC	0.009	0.010		
	(0.030)	(0.031)		
D. Non-CCSS Editions				
enVision	−0.028**	−0.029**	−0.028***	−0.013
	(0.010)	(0.010)	(0.011)	(0.011)
Everyday Mathematics	0.088***	0.086***	0.089***	
	(0.015)	(0.015)	(0.015)	
Houghton Mifflin Math	−0.018	−0.020		
	(0.015)	(0.015)		
Math Connects	0.012	0.015		
	(0.027)	(0.028)		
Math Expressions	−0.018	−0.016		
	(0.033)	(0.034)		
P-Value from Test of Joint Significance	0.000	0.000	0.000	0.026
SD of Textbook Fixed Effects	0.085	0.029	0.032	0.011
School*Year Observations	11,516	10,797	9,911	7,618

Notes: Estimates in each column come from separate linear regression models of school value-added. We report coefficients for a set of binary indicators for each textbook. The omitted textbook category is *enVision* CC, which is in the "other CCSS editions" category of EdReports ratings. All models are multilevel, mixed-effects models that include random effects for schools nested within districts. We control for school-by-year demographic characteristics, 2010 to 2014 district census characteristics, and state-by-year fixed effects. The sample is restricted to school-by-year observations with value-added data for the 2014/2015, 2015/2016, or 2016/2017 school year. Model (1) includes school-by-year observations using any of the 38 known textbooks in our sample, though we only show estimates for the top 15 by market share given that most of these textbooks where used by very few schools; coefficients on these texts are not particularly informative. Empty cells indicate that a given textbook was not in the relevant sample for a given analysis (i.e., Top 10 or Top 5 by market share).
~p < .10; *p < .05; **p < .01; ***p < .001.

Table A6. Correlations between textbook fixed effect estimates across states and years.

	Average Textbook Effects (Top 15 Textbooks)						Marginal Effects for Textbook Switchers (All Known Textbooks)			
	Pooled 6 States: All Years	Pooled 5 States (Excluding California): All Years	California: All Years	Pooled 6 States: 2014-15	Pooled 6 States: 2015-16	Pooled 4 States (Excluding California and Louisiana): 2016-17	Pooled 5 States: With School Fixed Effects	Pooled 5 States: Switchers Only, No School Fixed Effects	California: With School Fixed Effects	California: Switchers Only, No School Fixed Effects
Average Textbook Effects (Top 15 Textbooks)										
Pooled 6 States: All Years	1									
Pooled 5 States (Excluding California): All Years	0.092	1								
California: All Years	0.872***	−0.159	1							
Pooled 6 States: 2014–15	0.419	−0.122	0.367	1						
Pooled 6 States: 2015–16	0.07	0.201	0.058	−0.023	1					
Pooled 4 States (Excluding California and Louisiana): 2016–17	0.098	−0.062	−0.125	0.564*	0.058	1				
Marginal Effects for Textbook Switchers (All Known Textbooks)										
Pooled 5 States: With School Fixed Effects	0.006	0.606*	−0.059	−0.167	0.322	−0.251	1			
Pooled 5 States: Switchers Only, No School Fixed Effects	−0.182	0.503~	−0.331	−0.568*	0.055	−0.581*	0.657*	1		
California: With School Fixed Effects	0.631*	−0.104	0.575~	0.169	0.098	0.312	−0.414	−0.145	1	
California: Switchers Only, No School Fixed Effects	0.771**	0.111	0.760**	0.346	0.193	0.326	−0.248	−0.568~	0.571~	1.000

Notes: Each cell calculates the unweighted correlation of textbook fixed effect estimates from Table 7 (Average Textbook Effects) or Table 8 (Marginal Effects for Textbook Switchers).
~p < .10; *p < .05; **p < .01; ***p < .001.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table A7. Power analyses.

	Market Share of Single "Effective" Textbook				
Effect Size	1%	5%	10%	20%	25%
0.02 SD	Mean Effect = 0.002 SD = (0.003) 1% p<0.05	0.003 (0.004) 7%	0.004 (0.005) 14%	0.007 (0.005) 33%	0.007 (0.004) 39%
0.03 SD	0.002 (0.003) 2%	0.003 (0.004) 8%	0.006 (0.005) 34%	0.009 (0.004) 61%	0.010 (0.003) 73%
0.05 SD	0.002 (0.003) 1%	0.008 (0.006) 34%	0.013 (0.004) 82%	0.016 (0.002) 99%	0.017 (0.002) 100%
0.10 SD	0.003 (0.004) 3%	0.023 (0.005) 99%	0.026 (0.003) 100%	0.028 (0.002) 100%	0.028 (0.002) 100%
0.15 SD	0.006 (0.010) 10%	0.036 (0.004) 100%	0.038 (0.003) 100%	0.040 (0.002) 100%	0.040 (0.002) 100%

Notes: Each cell represents a summary of 100 simulations designed to test the sensitivity of our textbook random effect estimator in a simulated distribution of textbooks that vary in effectiveness. Rows vary the size of the simulated textbook effect, and columns vary the simulated market share of that text in the sample. The first value in each cell is the mean textbook random effect estimated in that set of 100 simulations (including zeros). The next value is the standard deviation of the simulated textbook random effects. The final value in each cell represents the proportion of simulated runs in that cell where the textbook random effect parameter was greater than 1.96 times larger than its standard error.

Table A8. Heterogeneity in textbook random effect estimates by student subgroups (top 15 textbooks).

Random Effects Parameters	Pooled 3 States (Maryland, New Jersey, New Mexico)		California	
Panel A	English Language Learner	Non-English Language Learner	English Language Learner	Non-English Language Learner
Textbook	0.000	0.000	0.026**	0.030***
	(0.000)	(0.000)	(0.009)	(0.009)
Student*Year Observations	22,727	232,586	307,952	711,559
Panel B	Special Education	Non-Special Education	Special Education	Non-Special Education
Textbook	0.000	0.000	0.018*	0.029**
	(0.001)	(0.000)	(0.008)	(0.009)
Student*Year Observations	40,416	214,897	108,476	911,030
Panel C	Free or Reduced-Price Lunch	Non-Free or Reduced-Price Lunch	Free or Reduced-Price Lunch	Non-Free or Reduced-Price Lunch
Textbook	0.000	0.000	0.029***	0.028**
	(0.000)	(0.000)	(0.008)	(0.009)
Student*Year Observations	136,646	118,667	650,167	369,344
Panel D	High Prior Math Achievement	Low Prior Math Achievement	High Prior Math Achievement	Low Prior Math Achievement
Textbook	0.000	0.000	0.022**	0.014**
	(0.000)	(0.000)	(0.007)	(0.005)
Student*Year Observations	129,667	125,646	505,981	513,530

Notes: Estimates in each cell come from separate models. We estimate the standard deviation in textbook effects with a multilevel mixed-effects linear regression of student-level standardized math test scores on student prior year math test scores, student demographic characteristics, school-by-year demographic characteristics, 2010 to 2014 district census characteristics, and state-by-year fixed effects (restricted to year fixed effects only in specifications limited to a single state). The model also includes nested random effects for textbook, state, district, and school, nested in that order, with textbook as the top level of the nesting structure (state random effects are excluded from regressions that are limited to a single state). Each subsample is restricted to student observations with value-added data for the 2014/2015, 2015/2016, or 2016/2017 school year who are known to have used one of the top 15 textbooks by market share. Robust standard errors are in parentheses. ~z > 1.64, *z > 1.96, **z > 2.58, ***z > 3.29, where z equals the ratio of a given random effects parameter estimate to its standard error. These z-scores do not correspond precisely to p-values as in a traditional linear regression framework, as the confidence interval for a random effect estimate is not symmetric around the estimate (random effect estimates have a lower bound of zero). These traditional markers of significance are included as an aid to reader, but they should be interpreted with caution.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

Table A9. Heterogeneity in textbook random effect estimates by grade level (top 15 textbooks).

Random Effects Parameters	Pooled 4 States (Excluding California and Louisiana)		California	
	5th Grade, Controlling for 4th Grade			
Panel A	All Schools	Non-Switchers	All Schools	Non-Switchers
Textbook	0.000	0.024	0.017**	0.032**
	(0.000)	(0.015)	(0.006)	(0.012)
School*Year Observations	2,322	1,431	3,507	2,109
	5th Grade, Controlling for 3rd Grade			
Panel B	All Schools	Non-Switchers	All Schools	Non-Switchers
Textbook	0.020	0.000	0.037***	NA
	(0.036)	(0.000)	(0.011)	
School*Year Observations	1,985	1,411	3,415	
	4th Grade, Controlling for 3rd Grade			
Panel C	All Schools	Non-Switchers	All Schools	Non-Switchers
Textbook	0.000	NA	0.000	NA
	–		(0.000)	
School*Year Observations	2,310		3,505	

Notes: Estimates in each cell come from separate models. We estimate the standard deviation in textbook effects using a multilevel mixed-effects linear regression of school-level value added on school-by-grade-by-year demographic characteristics, 2010 to 2014 district census characteristics, and state-by-year fixed effects (excluded from California estimates, which are limited to a single state and year). The model also includes nested random effects for textbook, state, district, and school, nested in that order, with textbook as the top level of the nesting structure (state random effects are excluded from regressions that are limited to a single state). The sample is restricted to school-by-year observations within the indicated subgroup with value-added data for the 2014/2015, 2015/2016, or 2016/2017 school year who are known to have used one of the top 15 textbooks by market share. "NA" indicates that estimates were not possible to estimate with available data. In all states, we did not collect textbook data in third grade and so cannot identify fourth grade non-switchers. In California, the state did not record student-level test scores in the spring of 2014, as schools prepared for the new CCSS-aligned assessments to be administered in the spring of 2015; therefore, we did not have multiple years of test score and textbook data for fifth graders who also had third grade prior achievement, necessary for the non-switcher model. Robust standard errors are in parentheses. ~z > 1.64, *z > 1.96, **z > 2.58, ***z > 3.29, where z equals the ratio of a given random effects parameter estimate to its standard error. These z-scores do not correspond precisely to p-values as in a traditional linear regression framework, as the confidence interval for a random effect estimate is not symmetric around the estimate (random effect estimates have a lower bound of zero). These traditional markers of significance are included as an aid to reader, but they should be interpreted with caution.

Table A10. Robustness of textbook random effect estimates to different sets of controls to limit selection bias (top 15 textbooks).

Random Effects Parameters	Pooled 5 States (Excluding California)	California
Panel A	School and District Covariates	
Textbook	0.000	0.027**
	–	(0.008)
School*Year Observations	2,676	8,121
Panel B	School Covariates	
Textbook	0.000	0.033***
	–	(0.009)
School*Year Observations	2,676	8,121
Panel C	No Covariates	
Textbook	0.000	0.037***
	–	(0.011)
School*Year Observations	2,676	8,121
Panel D	ELA Achievement as Outcome, Controlling for ELA Textbooks	
Textbook	NA	0.017**
		(0.006)
School*Year Observations		8,121

Notes: Estimates in each cell come from separate models. Random effects are estimated from a multilevel mixed-effects linear regression of school-level value added on state-by-year fixed effects (restricted to year fixed effects only in specifications limited to a single state), and school-by-year demographic characteristics and/or 2010 to 2014 district census characteristics where indicated. The model also includes nested random effects for textbook, state, district, and school, nested in that order, with curriculum as the top level of the nesting structure (state random effects are excluded from regressions that are limited to a single state). The sample is restricted to school-by-year observations with value-added data for the 2014/2015, 2015/2016, or 2016/2017 school year who are known to have used one of the top 15 textbooks by market share. Robust standard errors are in parentheses; "–" indicates that standard errors could not be estimated. ~z > 1.64, *z > 1.96, **z > 2.58, ***z > 3.29, where z equals the ratio of a given random effects parameter estimate to its standard error. These z-scores do not correspond precisely to p-values as in a traditional linear regression framework, as the confidence interval for a random effect estimate is not symmetric around the estimate (random effect estimates have a lower bound of zero). These traditional markers of significance are included as an aid to reader, but they should be interpreted with caution.