Mathematics Curriculum Effects on Student Achievement in California

Cory Koedel
Diyi Li
Morgan S. Polikoff
Tenice Hardaway
Stephani L. Wrabel

September 2016

We estimate relative achievement effects of the four most commonly adopted elementary-mathematics textbooks in the fall of 2008 and fall of 2009 in California. Our findings indicate that one book, Houghton Mifflin's *California Math,* is more effective than the other three, raising student achievement by 0.05-0.08 student-level standard deviations of the grade-3 state standardized math test. We also estimate positive effects of *California Math* relative to the other textbooks in higher elementary grades. The differential effect of *California Math* is educationally meaningful, particularly given that it is a schoolwide effect and can be had at what is effectively zero marginal cost.

Acknowledgement
This study is based upon work supported by the National Science Foundation under Grant No. 1445654 and the Smith Richardson Foundation. Any opinions, findings, and conclusions or recommendations expressed in this study are those of the author(s) and do not necessarily reflect the views of the funders.

1. Introduction

Several recent experimental and quasi-experimental studies point toward differences in curriculum materials having educationally meaningful effects on student achievement (Agodini et al, 2010; Bhatt and Koedel, 2012; Bhatt, Koedel and Lehmann, 2013). Chingos and Whitehurst (2012) argue that relative to other potential educational interventions – and in particular human resource interventions – making better-informed decisions about curriculum materials represents an easy, inexpensive and quick way to raise student achievement. However, the extent to which educational administrators can improve achievement by selecting better curriculum materials is hampered by a general lack of information about the content, quality, and efficacy of various materials. Given the wide variety of curriculum materials from which decision makers can choose, and the wide variety of implementation contexts (e.g., high/low poverty schools, states with different curricular goals and assessments, etc.), the handful of available efficacy studies is far from sufficient to inform those charged with selecting curriculum materials on behalf of students.

We contribute to the sparse literature on curricular efficacy by leveraging unique school-level data on textbook adoptions to estimate the relative effects on student achievement of four commonly-used elementary mathematics textbooks in California (we refer to curriculum materials as "curricula" and "textbooks" interchangeably throughout our study). Textbook adoptions are reported by individual schools as a requirement of the 2004 *Eliezer Williams et al. vs. State of California et al.* court ruling and resulting legislation. The plaintiff in the case argued that low-income students do not have access to the same high-quality resources available to their higher-income peers. As a result of the *Williams* ruling, each school in the state is required to report on the presence of various educational resources, including textbooks. These data are kept in School Accountability Report Cards (SARCs) as PDF files available online from the California Department of Education (CDE).

We manually collect textbook data from schools' SARCs and merge textbook adoptions with a longitudinal data file containing information about school achievement and characteristics. We use the merged file to perform a quasi-experimental evaluation of curriculum effects on grade-3 state standardized assessments. Our results indicate that one elementary mathematics textbook – *California Math* published by Houghton Mifflin – outperformed the other three popular textbooks during the period we study. Specifically, we estimate that *California Math* increased student test scores by 0.05 to 0.08 student-level standard deviations on the grade-3 test relative to the alternatives. We extend our analysis into grades 4 and 5 and find that *California Math* increased math achievement in these grades as well, particularly in grade-5.

The differential curriculum effects that we document in California are on the lower end of the range of estimates reported in similar recent studies, which have been between 0.08 and 0.17 student-level standard deviations (Agodini et al., 2010; Bhatt and Koedel, 2012; Bhatt, Koedel, and Lehmann, 2013). That said, the effect of *California Math* is still educationally meaningful, particularly given the scope of the intervention and low cost of implementation. With regard to scope, curriculum effects apply on average across entire cohorts of students in schools. With regard to cost, as noted by Bhatt and Koedel (2012) and Chingos and Whitehurst (2012), the marginal cost of choosing one textbook over another is so small that it is effectively zero.[1]

The fact that estimated differences in curriculum effects in California are smaller than in the handful of locales where other, similar evaluations have been conducted is interesting and worthy of further exploration. This could be due to differences in the curricula studied, evaluation contexts (including the assessments used to gauge impact), or simply sampling variability. Ideally, our efficacy

[1] We do not know the list price of the textbooks used in this study, but available research indicates that most textbooks adopted by a state are approximately the same unit cost. The elementary mathematics books in Boser et al. (2015) cost an average of $34 per pupil, or approximately 0.32% of per-pupil spending (the true per-pupil expenditure is even lower because textbooks are used for multiple years).

estimates could be compared to a much larger set of similar estimates for the same and different curricula, and in similar and different evaluation contexts, to gain further insight into this finding. However, given that so few states collect textbook adoption data, and correspondingly there are so few studies of curricular efficacy, we can do little more than speculate as to the source of the differential results. Our inability to contextualize our findings within a larger literature – which essentially does not exist – highlights the frustrating lack of information nationally about the effectiveness of different sets of curriculum materials.

2. Background & Data

California has what is best described as a partially centralized curriculum adoption process. The important centralized feature is that the state initiates the process for a particular subject in a particular year by assembling a list of "state approved" curriculum materials. This list then goes out to districts, but it is advisory only. Districts can choose any curriculum materials they would like – on list or off – or they can choose not to adopt curriculum materials at all. Like other states with partially centralized adoption processes, districts in California adopt new curriculum materials in each subject on roughly the same schedule – in math, California districts have recently adopted new textbooks on a six-year cycle (2008-09 to 2014-15), though again districts can choose when and whether to adopt. This cycle length is typical of other states. Districts are all prompted to move together by the state's initiation of the adoption process, so the large majority of districts make adoption decisions in the years immediately following the state adoption.

We focus our analysis on elementary mathematics textbooks adopted in California schools in either fall-2008 or fall-2009. Our curriculum materials data, which we collected manually from schools' 2013 SARCs, include information on textbooks from this adoption cycle that were still in use in 2013 in most schools (only a small fraction of schools adopted a new textbook after fall-2009 and before the publication of the 2013 SARCs, which we drop – see details below). The textbook adoption we study was intended for fall-2008 and the state-approved list was released in November of 2007, but based on data collected from individual schools' SARCs it appears that many schools and districts delayed the adoption one year. Thus, we refer to the adoption as occurring in 2009/2010 (for presentational convenience we refer to school years by the spring year throughout our study – e.g., "2009" for "2008-09").

We merge information on schools' curriculum adoptions from their SARCs with a longitudinal database containing school and district characteristics and achievement outcomes covering the school years 2003 to 2013, constructed based on publicly-available data from the CDE. We supplement the CDE data with data from the United States Census on the median household income and education level in the local-area for each school, which we link at the zip-code level. Achievement effects are estimated using school-average test scores on state standardized math assessments.[2] We focus most of the evaluation on grade-3 achievement, which aligns our study with previous related work focusing on early primary grades (Agodini et al., 2010; Bhatt and Koedel, 2012; Bhatt, Koedel and Lehmann, 2013), but also extend our analysis to examine curriculum effects on test scores in grades 4 and 5.

Appendix Table A.1 provides details about the construction of our analytic sample starting with the universe of elementary schools in California. There are several notable attrition points from the sample. First, although California provides a SARC template for schools, which some follow, the quality of information about curriculum materials reported on the SARCs varies greatly. Curriculum materials information was either not reported (perhaps because no book was used in some cases), or reported in such a way that the actual textbook used is indeterminate, for 20.8 percent of elementary

[2] Access to student-level test scores would offer little additional value for our evaluation because the curriculum-adoption data are at the school level, making schools the smallest feasible units of analysis. It is also unlikely that student-level data on test scores and curriculum exposure (we are not aware of the latter existing anywhere in the United States), even if available, would meaningfully improve inference from our evaluation because very few schools report using more than one set of curriculum materials in the same grade, which implies limited treatment variability within schools that could be exploited with student-level data (these schools are a small subsample of "non-uniform" adopters reported in Appendix Table A.1)

schools in the state. As an example of an indeterminate report, a district might only list a publisher's name for a publisher that produced more than one state-approved textbook (e.g., list "Houghton Mifflin", which published both Houghton Mifflin *Harcourt California HSP Math* and Houghton Mifflin *California Math*). In such a case, if no other information is provided, the actual textbook cannot be determined. We drop all schools from the sample that report no textbook information or indeterminate information.

A second notable reason schools were removed from the analytic sample is that they report a curriculum adoption year other than 2008 or 2009 on the 2013 SARC. Appendix Table A.1 shows that this applies to approximately 15.9 percent of schools. Schools may have delayed adoptions beyond 2009 for a variety of reasons, including budgetary issues or a lack of need. As an example of the latter, a school may have adopted off-cycle in a recent year prior to 2009/2010, and thus may not have needed to adopt new materials on the standard timeline.

A third significant source of attrition from our dataset, conditional on schools adopting textbooks in 2008 or 2009 and reporting identifiable materials, is that we drop approximately 8 percent of schools that either (a) explicitly indicate using more than one textbook in grades 1-3, or (b) indicate using more than one textbook in the school, and where the SARC was ambiguous about which curriculum materials were used in which grades. The reason for this restriction is that we focus primarily on estimating achievement effects on grade-3 mathematics tests. Schools that use more than one textbook in grades 1-3 have mixed treatments. While in principle these schools could be used to examine mixed-treatment effects, in practice there are too few observations for an effective analysis along these lines, so we simply drop them from the analytic sample.[3]

[3] When we extend our analysis to grades 4 and 5 later on, we also extend the restriction of constant materials usage to grades 4-5. Most schools that used constant materials in grades 1-3 also used the same materials in grades 4-5, but there is a small amount of sample attrition owing to this issue when we examine the later grades.

After imposing these data restrictions, plus a few other minor restrictions detailed in Appendix Table A.1, we are left with a sample of just over half of the schools in California. These schools clearly report which curriculum materials they use, and use the same materials in grades 1-3. Among them, 78 percent adopted one of these four textbooks: *enVision Math California* published by Pearson Scott Foresman, *California Math* published by Houghton Mifflin, *California Mathematics: Concepts, Skills, and Problem Solving* published by McGraw-Hill, and *California HSP Math* published by Houghton Mifflin Harcourt. We focus our evaluation on these textbooks and the schools that adopted them. In total, this group initially included 2,281 California schools spread across 311 districts; however, after our analysis began we also dropped data from the Los Angeles Unified School District (LAUSD) and Long Beach Unified School District (LBUSD). Both districts are much larger than all other districts in the state, which created comparability problems in our evaluation. After dropping LAUSD and LBUSD schools, our final analytic dataset includes 1,878 California schools in 309 districts.

Table 1 provides descriptive characteristics and sample sizes for all California schools in our initial universe and schools that were retained in our final analytic sample. We also report separate statistics for schools that adopted each of the four focal curricula. The initial universe of schools in column 1 includes all schools in the CDE data for which at least one grade-3 test score is available during the years 2009-2013, school characteristics are available for either 2007 or 2008, and the highest grade is 8 or lower.[4] The table shows that schools in our analytic sample are negatively selected relative to all schools in the state, but not substantially. Within our analytic sample, adopters of *California Math* are similar to, although somewhat more advantaged than, adopters of the other curricula. However, there is substantial distributional overlap in pre-adoption achievement and other school characteristics

[4] Prior to merging in the curriculum data and performing our analysis centered on when actual adoptions occurred, these conditions are the minimal conditions for inclusion into our analytic sample. For example, no school without a test score from 2009 or later can be included in our study because test scores from 2009 and later are the outcomes by which we evaluate curricular efficacy.

between *California Math* adopters and the comparison schools, which facilitates our analysis as outlined below. This overlap is illustrated in Appendix Figure B.1.

3. Focal Textbooks

As noted above, the textbooks we study were adopted in either fall-2008 or fall-2009 (the state refers to these textbooks as being a part of the 2007 adoption cycle – see California Department of Education, 2009). The adoption was to select books aligned with the state's 1997 mathematics content standards and 2005 mathematics framework. The multi-step adoption process, which is described in detail in the adoption report (California Department of Education, 2009), included 14 content experts (university professors) and 141 instructional materials experts (k-12 educators) divided into 26 panels. The chosen books were required to meet criteria in five categories: Mathematics content/alignment, program organization, assessment, universal access, and instructional planning and support. The final selections passed through a public comment period and were approved by the State Board of Education in winter-2007.

There were ten textbooks for grades K-3 that were approved. We study four of these books, which we chose because they were the most popular. In addition to their popularity making these books the most policy-relevant ones to study, it also affords sufficient sample sizes to support our empirical evaluation. In this section, we briefly describe the four books, drawing on available data from the What Works Clearinghouse, the state adoption report, and available web materials. All of the textbooks we study are the California editions of their respective book series. Because some of the information available online describes the national or Common Core versions of these series, we cannot always be confident that it applies to the California versions we study. We are hampered in our descriptions by the fact that there is little or no publicly available information about the differences between state-specific and national versions of textbooks.

Pearson Scott Foresman's *enVision Math California* is an early edition of the *enVision* series that is still marketed and sold by Pearson as Common Core and Texas editions. According to the WWC, *enVision* aims to help students develop an understanding of mathematics concepts through problem-based instruction, small-group interaction, and visual learning, with a focus on reasoning and modeling. Each lesson is intended to include small-group problem solving. The book's lead author, Randall Charles, was a coauthor of the National Council of Teachers of Mathematics' Focal Points, widely considered a reform-oriented mathematics document. Despite its seemingly reform-oriented description, analyses of other editions of *enVision* (the Common Core and Florida grade-4 versions) found them to be typical in terms of their cognitive demand coverage and far below the level of cognitive demand emphasized in the standards (Polikoff, 2015). The California state adoption report indicates that this curriculum met all five evaluative criteria.

We have far less information about the other three textbooks. The California state adoption report indicates that all three meet the five evaluative criteria (California Department of Education, 2009). Houghton Mifflin's *California Math* and Harcourt's *California HSP Math* are both updated versions of textbooks previously adopted by the state in the 2001 adoption, while McGraw-Hill's *California Mathematics* was not adopted previously by the state. Other than this, we were unable to find information about the Houghton Mifflin and Harcourt books. McGraw-Hill's *California Mathematics* has an evaluation report (Papa & Brown, 2007) that describes the book as including both conceptual understanding and guided practice and argues that it aligns with what is known about effective mathematics instruction. McGraw-Hill does not appear to have published any books in this series since 2009. In the conclusion we return to the challenge of characterizing these textbooks, and correspondingly, in interpreting our results based on student achievement in terms of their content and form.

4. Empirical Strategy

4.1 Methodological Overview

We estimate the achievement effects of *California Math* relative to a composite alternative of the three other focal curricula using three related empirical strategies: (a) kernel matching, (b) common-support-restricted ordinary least squares (restricted OLS), and (3) "remnant"-based residualized matching. The unit of analysis in our study is the school, but we cluster our standard errors at the district level to reflect data dependence within districts across schools, including along the dimension of curriculum adoptions. We describe our methods within the context of our evaluation of grade-3 test scores. The methods carry over directly when we extend our analysis to study test scores in grades 4 and 5, as we discuss briefly when we present those results below.

4.1.1 Matching

Our matching estimators follow Bhatt and Koedel (2012) and draw on the larger matching literature to identify the approach best-suited to our data (Caliendo and Kopeinig, 2008; Frölich, 2004). The key to identification is the conditional independence assumption (CIA). The CIA requires potential outcomes to be independent of curriculum choice conditional on observable information. Denoting potential outcomes by $\{Y_0, Y_1\}$, curriculum treatments by $D \in \{0,1\}$, and X as a vector of (pre-treatment) observable school, district and local-area characteristics, the CIA can be written as:

$$Y_0, Y_1 \perp D \mid X \tag{1}$$

Conditional independence will not be satisfied if there is unobserved information that influences both treatments and outcomes. For example, if districts have access to information that is unobserved to the researcher, Z, such that $\mathrm{P}(D=1 \mid X, Z) \neq \mathrm{P}(D=1 \mid X)$, and the additional information in Z influences outcomes, matching estimates will be biased. We discuss the plausibility of the CIA in our context and provide evidence consistent with it being satisfied below.

We match schools using propensity scores (Rosenbaum and Rubin, 1983; Lechner, 2002). The propensity score model predicts whether each school adopted *California Math* as a function of a variety of school, district and local-area characteristics. Specified as a probit, our propensity score model is as follows:

$$T_{sd} = \mathbf{X_s}\boldsymbol{\beta}_1 + \mathbf{X_d}\boldsymbol{\beta}_2 + \varepsilon_{sd} \qquad (2)$$

In Equation (2), T_{sd} is an indicator variable equal to one if school *s* in district *d* adopted *California Math* and zero if it adopted one of the other focal curricula. $\mathbf{X_s}$ and $\mathbf{X_d}$ are vectors of school and district covariates, respectively, that include the variables listed in Table 1. For schools, $\mathbf{X_s}$ includes pre-adoption student achievement in math and reading, the share of students by race, gender, language fluency and economic disadvantage, school enrollment (cubic), and whether the school adopted new materials in 2008 or 2009. The vector $\mathbf{X_s}$ also includes the log of median household income and the share of individuals over age-25 without a high school degree in the local area – these data are taken from the 2013 American Community Survey 5-year average (from the U.S. Census) and merged to schools at the zip-code level.[5] The vector $\mathbf{X_d}$ includes district level pre-adoption achievement in math and reading, and enrollment (cubic).[6]

With the estimated propensity scores from equation (2) in hand, we estimate the average treatment effect (ATE) of adopting *California Math*. Defining *California Math* as curriculum *j* and the composite alternative as curriculum *m*, where Y_j and Y_m are standardized test-score outcomes for

[5] We also include a binary variable to indicate CDE data quality for individual schools and an indicator for missing Census data. The CDE data-quality indicator is equal to one if the enrollment counts by subgroup (e.g., by race, gender, etc.) do not exactly match total reported enrollment for schools. For most schools the subgroup enrollments sum to total enrollment and this variable is of no practical consequence in our analysis (i.e., if we omit the variable entirely our results are unchanged).

[6] The non-test-score school and district covariates are averaged over the two years immediately prior to the adoption of the new materials, and the test-score covariates are from two years before the adoption. We follow Bhatt and Koedel (2012) in not using test score information from the year immediately before the new books were adopted because this information would not have been available to decision makers at the time of the decision per the above discussion. That said, none of our findings are substantively affected if we include lagged test score information from the year just before adoption into the selection models as well.

adopters of *j* and *m*, respectively, we estimate $ATE_{j,m} \equiv E(Y_j - Y_m \mid D \in \{j,m\})$. We use kernel matching estimators (with the Epanechnikov kernel), which construct the match for each "treated" school using a weighted average of "control" schools, and vice versa. The formula for our estimate of $ATE_{j,m}$ is:

$$\hat{\theta}_{j,m} = \frac{1}{N^S}[\sum_{j \in N_j \cap S_p} \{Y_j - \sum_{m \in I_{0j} \cap S_p} W(j,m)Y_m\} - \sum_{m \in N_m \cap S_p} \{Y_m - \sum_{j \in I_{0m} \cap S_p} W(m,j)Y_j\}] \tag{3}$$

In (3), N^S is the number of schools using *j* or *m* on the common support, S_p. I_{0j} indicates the schools that chose *m* in the neighborhood of observation *j,* and I_{0m} indicates the schools that chose *j* in the neighborhood of observation *m*. Neighborhoods are defined by a fixed bandwidth parameter obtained via conventional cross-validation (as in Bhatt and Koedel, 2012). *W(j,m)* and *W(m,j)* weight each comparison school outcome depending on its distance, in terms of estimated propensity scores, from the observation of interest. We compute separate *ATE* estimates by year based on the distance from the adoption year using the formula in Equation (3). All of our standard errors are estimated via bootstrapping using 250 replications and clustered at the district level (i.e., with district re-sampling). We omit a more detailed discussion of the matching estimators for brevity but more information can be found in Caliendo and Kopeinig (2008), Heckman et al. (1997), and Mueser et al. (2007).

4.1.2 Restricted OLS

We also use restricted OLS models to estimate curriculum effects for schools on the common support of propensity scores. We use the same school and district characteristics taken from pre-adoption data in the OLS models as we use to match schools, allowing the coefficients to change over time as follows:

$$\boldsymbol{Y_{sdt}} = \mathbf{X_s}\boldsymbol{\pi_{1t}} + \mathbf{X_d}\boldsymbol{\pi_{2t}} + T_{sd}\theta + u_{sdt} \tag{4}$$

In Equation (4), $\boldsymbol{Y_{sdt}}$ is a grade-3 math test score for school *s* in district *d* in year t, $\mathbf{X_s}$ and $\mathbf{X_d}$ are the vectors of pre-adoption school/district characteristics that we use for matching as described above (these variables do not change over time), T_{sd} is an indicator equal to one if the school adopted *California Math*, and u_{sdt} is the error term. The coefficient vectors $\boldsymbol{\pi}_{\mathbf{1t}}$ and $\boldsymbol{\pi}_{\mathbf{2t}}$ allow the pre-adoption school and district characteristics to differentially predict achievement over time.

The OLS estimates are very similar to the matching estimates. They rely on the same assumption of conditional independence for identification. The benefit of the OLS models is that they improve statistical precision by imposing a parametric form – linearity – on the outcome model. The cost is that if the linearity assumption is not justified it could introduce bias (Black and Smith, 2004). In our application, where California schools and districts are diverse and we have small samples (at least by the standards of matching analyses, and particularly when one accounts for district clustering), the efficiency benefit of imposing the linear functional form is substantial. This will become clear when we present our findings below. With regard to the potential for the linearity assumption to introduce bias into our estimates, we show results from falsification tests that provide no indication that our OLS estimates are biased.

4.1.3 Remnant-Based Residualized Matching

Remnant-based residualization is another way to improve statistical power. It blends aspects of the restricted-OLS and matching strategies. The fundamental idea, taken from Sales et al. (2014), is to pull in data from outside of the evaluation – i.e., "remnant data" – to improve statistical inference. Sales et al. (2014) suggest several potential uses of remnant-based residualization – in our application the appeal is that the procedure can remove noise from the outcome data prior to matching, thereby improving the precision of our estimates. Our evaluation is particularly well-suited for remnant-based residualization because we have access to substantial data from outside of the evaluation; e.g., from schools in California that use a textbook outside of the four focal curricula.

The "remnant" sample we use includes data from all schools in California that adopted a new curriculum in fall-2008 or fall-2009 uniformly, but chose a curriculum other than one of the four primary textbooks (there are 632 such schools per Appendix Table A.1). Thus, these schools are outside of our evaluation sample. Following Sales et al. (2014), we start by estimating the following linear regression model using the remnant data:

$$Y_{sdt} = \mathbf{X_s}\boldsymbol{\alpha}_{\mathbf{1t}} + \mathbf{X_d}\boldsymbol{\alpha}_{\mathbf{2t}} + \eta_{sdt} \tag{5}$$

In Equation (5), Y_{sdt} is a grade-3 math test score for school *s* in district *d* in year *t* and $\mathbf{X_s}$ and $\mathbf{X_d}$ are defined as above.[7] After estimating equation (5), we store the coefficient estimates $\hat{\boldsymbol{\alpha}}_{\mathbf{1t}}$ and $\hat{\boldsymbol{\alpha}}_{\mathbf{2t}}$ and construct the following residualized test score outcome for each school *in our analytic sample* in each year:

$$Q_{sdt} = Y_{sdt} - (\mathbf{X_s}\hat{\boldsymbol{\alpha}}_{\mathbf{1t}} + \mathbf{X_d}\hat{\boldsymbol{\alpha}}_{\mathbf{2t}}) \tag{6}$$

In Equation (6), Y_{sdt} is the grade-3 test score for school *s* in district *d* in year *t* for a school that adopted one of the four primary curricula. $\mathbf{X_s}$ and $\mathbf{X_d}$ continue to be defined as above. $\hat{\boldsymbol{\alpha}}_{\mathbf{1t}}$ and $\hat{\boldsymbol{\alpha}}_{\mathbf{2t}}$ are out-of-sample parameter estimates based on the remnant data that link the pre-adoption school and district characteristics to test-score outcomes by year.

Intuitively, Equation (6) can be described as specifying a set of general relationships between school/district characteristics and test-score outcomes in California as defined by $\hat{\boldsymbol{\alpha}}_{\mathbf{1t}}$ and $\hat{\boldsymbol{\alpha}}_{\mathbf{2t}}$. Implementing the matching procedure on the residualized outcomes, Q_{sdt}, is very similar to the restricted OLS approach, with the added benefit that the adjustment parameters $\hat{\boldsymbol{\alpha}}_{\mathbf{1t}}$ and $\hat{\boldsymbol{\alpha}}_{\mathbf{2t}}$ are

[7] The use of covariates from before the 2009/2010 adoptions is not particularly important given that none of these schools used any of the curricula of interest, but we follow the same timing convention as in other parts of our analysis for consistency. We obtain similar results if we estimate equation (5) using data from different years.

estimated entirely out of sample. Using an out-of-sample "training set" for the outcome model has several conceptual benefits over using in-sample data (as was the case with OLS) as described by Sales et al. (2014). In our application it addresses the concern that bias could be introduced by the OLS models if the covariate coefficients are disproportionately influenced by schools in the control condition, which dominate our sample. This in turn would result in asymmetric overfitting of the outcome model, potentially causing bias.[8]

A concern with remnant-based residualization is that the relationships between school/district characteristics and test scores may be different in the analytic sample and the remnant sample. Although in such a scenario the adjustment parameters $\hat{\boldsymbol{\alpha}}_{\mathbf{1t}}$ and $\hat{\boldsymbol{\alpha}}_{\mathbf{2t}}$ will be less useful, Sales et al. (2014) show that the procedure still improves inference, albeit by less. In practice, if $\hat{\boldsymbol{\alpha}}_{\mathbf{1t}}$ and $\hat{\boldsymbol{\alpha}}_{\mathbf{2t}}$ measure a relatively constant set of relationships between characteristics and outcomes in California schools within years, remnant-based residualization and restricted OLS should return similar results. Below, we show that this is the case in our application. Moreover, results from both approaches are as expected in two different falsification exercises, which lends credence to their agreement in the evaluation.

4.2 Conditional Independence

All three approaches outlined above rely on the assumption of conditional independence, or selection-on-observables, for identification (the restricted-OLS and remnant-residualization methods further impose a functional form on the outcome model to improve statistical power, utilizing either in-sample or out-of-sample data). Is conditional independence plausible in our evaluation context? Can it be examined empirically?

---

[8] We are not aware of a specific example of this particular problem causing bias, but the possibility is implied in related findings by Hansen (2008), who shows that bias can be caused when observations in one condition (either treatment or control) disproportionately supply identifying variation for covariates.

We begin by making the intuitive case for conditional independence in the context of evaluating curriculum-material effects. One reason that the CIA is plausible is that curriculum materials are adopted on behalf of large groups of students and teachers rather than being the product of individual choice (in practice most adoptions are districtwide; some are made at the school level). In evaluations where individuals choose whether to seek treatment, individual characteristics that are difficult to observe such as motivation and innate ability may influence treatment and outcomes. This would violate the CIA. However, in the case of school and district-level choices and conditional on the rich covariates to which we have access – pre-adoption test scores, demographic and socioeconomic status measures, etc. (see Table 1) – it is more difficult to tell a compelling selection story. For example, consider two school districts with similar shares of students by race, economic-disadvantage status and language status, and located in zip codes with similar socioeconomic conditions. It is harder to argue that there are substantial differences in group-average unobservable characteristics like motivation or innate ability across these districts that are not already accounted for by the group-level observed measures, certainly relative to the case of a treatment influenced by individual choice.

One could also argue that school- and district-level differences in teacher quality, which are not directly accounted for in our study, might lead to a violation of the CIA if teacher quality helps to determine curriculum adoptions (research is quite clear that teacher quality affects student achievement – e.g., see Koedel, Mihaly and Rockoff, 2015). However, many of the same arguments from the preceding paragraph apply – specifically, it would need to be the case that there are systematic differences in teacher quality across schools and districts after conditioning on the rich set of educational and economic characteristics of schools and their local areas used in our study. Such a condition is even less plausible when one recognizes that most of the variation in teacher quality occurs within schools (Aaronson, Barrow and Sanders, 2007; Koedel and Betts, 2011); not across schools, let alone districts. The limited cross-school variation in teacher quality leaves little scope for systematic differences in the quality of teachers across schools and districts to lead to significant violations of the CIA.

As noted by Bhatt and Koedel (2012), perhaps the biggest conceptual threat to the CIA in curriculum evaluations is that some high-level educational decision makers simply make better choices than others. For example, effective leaders might choose a more effective textbook and also make other decisions that improve student outcomes, which would violate the CIA because "decision-maker quality" is not observed in our data. While acknowledging this concern, Bhatt and Koedel (2012) offer two reasons why this problem might be minor in practice. First, the curriculum adoption process in most schools and districts is a complex and multi-faceted process, and based on available documentation it does not appear that any single decision maker has undue influence (Zeringue et al., 2010). Second, even if a single decision maker did have undue influence, it is not clear that this person would have sufficient information to make an informed choice. As noted previously, given the large number of textbook options and the many potential educational contexts under which materials are adopted, the evidence base on curricular efficacy is insufficient. While there are new resources available to help districts make decisions (e.g., the ratings available at EdReports.org), such resources did not exist in 2008, and it is not clear whether and how those resources actually capture textbook efficacy. Thus, while we view the "strong decision maker" hypothesis as the biggest conceptual threat to the CIA, there are reasons to expect it may not be practically important either.[9]

While it is useful to discuss the plausibility of the CIA in conceptual terms, the discussion thus far has been speculative. To address this issue more formally in our analysis, we complement our

---

[9] In work complementary to this evaluation, we have also interviewed 16 district administrators randomly chosen from across California about the curriculum adoption processes in their districts. These interviews confirm the complexity of the adoption process and indicate that decisions are driven by committees made up mostly of teachers. In none of the districts was there evidence of a strong decision maker.

primary estimates of curriculum effects with two different types of falsification estimates designed to examine the plausibility of the CIA. The falsification estimates cannot be used to confirm the satisfaction of the CIA (it is not possible to confirm with certainty that the CIA is upheld); however, they can be used to look for evidence consistent with the CIA being violated.

The falsification tests are designed to look for curriculum effects in situations where (a) we should not expect any effects at all, or (b) we should expect small effects at most. If, for example, we estimate non-zero "curriculum effects" in situations where we know the effects should be zero, this would be a strong indication that selection bias is problematic in our study. In the first set of falsification tests we estimate curriculum effects from textbooks adopted in 2009 and 2010 on student test scores in *previous years*, prior to the use of the textbooks of interest in schools. True curriculum effects in previous years should be zero. In the second set of falsification tests we estimate the effects of math curriculum adoptions on English/Language-Arts (ELA) scores. Math curriculum effects on ELA scores should be zero or near-zero (there may be small spillover effects).

Estimates from all of our falsification tests are as expected and provide no indication that the primary results are biased by unobserved selection. We elaborate on our falsification tests and their interpretation when we show the results below.

5. Results

5.1 Six Pairwise Comparisons

We compare *California Math* to a composite alterative of the three other focal curricula. To arrive at this final research design, we began our evaluation by examining all six possible pairwise comparisons across the four focal curricula described above. After performing the six comparisons it became clear that it would be difficult to obtain meaningful insight from the individual pairwise comparisons. Two issues arose: (1) covariate-by-covariate balance is mediocre in some of the pairwise comparisons with little scope for improvement given our small sample sizes (at least relative to typical matching applications), and (2) statistical power is limited. The statistical power issue is more problematic than we had initially anticipated because our point estimates suggest curriculum effects in California that are smaller than in previous, similar evaluations. Moreover, because of the diversity of curriculum materials adopted in California relative to other states – which means that there are fewer districts adopting any single book – and the district-level clustering structure necessitated in the evaluation, our effective sample sizes in the pairwise comparisons are no larger in California than in previous studies in smaller states.

Despite their limitations, collectively the pairwise comparisons suggestively point toward *California Math* being more effective than the other three commonly-adopted textbooks. Moreover, they also point toward the other three curricula having similar effects. It is these preliminary results that motivate our comparison of *California Math* to the composite alternative of the other three curricula. Rebuilding the evaluation in this way is advantageous because it allows us to identify better matches for *California Math* schools and to perform a better-powered study of the effectiveness of *California Math* relative to the other three books. For interested readers, Appendix Table B.1 presents disaggregated matching estimates for the six pairwise comparisons that led us to restructure our study to focus on *California Math*.[10]

5.2 Comparison of California Math to the Composite Alternative

5.2.1 The Propensity Score

The propensity score model as shown in Equation (2) explains roughly 12 percent of the variance in curriculum adoptions between *California Math* and the composite alternative. The limited scope for observed selection into curriculum materials implied by this R-squared value is notable given

[10] Appendix Table B.1 presents a lot of information tersely. It will be easier to interpret after reading through Section 5.2 and the corresponding tables showing our primary results from the composite comparison.

that our covariates are strong predictors of achievement.[11] For interested readers, Appendix Table B.2 reports results from the estimation of Equation (2) for our evaluation of *California Math*. The only statistically significant predictors of a *California Math* adoption are the linear, squared and cubed district-enrollment variables. Thus, collectively, the covariates do not predict adoptions of *California Math* well. The lack of predictive power in the selection model is consistent with qualitative accounts of the complexity of the curriculum adoption process and the lack of clear objectives and information to make decisions (Jobrack, 2011; Zeringue et al., 2010).

5.2.2 Covariate Balance

Table 2 presents information on covariate balance in our comparison between schools that adopted *California Math* and schools that adopted one of the other three curricula of interest after matching. The table reports results for each year of the data panel separately, including both pre- and post-adoption years. Subsequent tables follow a similar reporting format.

The results for each school are centered around the year of the curriculum adoption. In the case of a fall-2008 adoption, year-1 indicates the 2008-09 school year (the first year the new book was used), year-2 indicates the 2009-10 school year, etc.; for a fall-2009 adoption, year-1 indicates the 2009-10 school year. We use data from the two years preceding the adoption to match schools, as described in Section 3, so we do not perform any direct analysis in these years. Thus, the first pre-adoption year for each school shown in Table 2 and subsequent tables is year-P3 – three years prior to adoption. For schools that first used the new books during the 2008-09 school year, year-P3 is the 2005-2006 school year, year-P4 is 2004-2005, etc.

---

[11] There are several ways to empirically verify this statement, but we must be careful to not contaminate the predictive power of our covariates with their predictive power over curriculum materials. As one straightforward data point, we can use the remnant sample and estimate a model of achievement during the first year of a new adoption using our matching covariates. The R-squared from this regression is 0.74.

Although we split out the data year-by-year in Table 2, the years are strongly dependent. The practical implication of the data dependence is that balancing evidence from a second year of the data panel provides very little new information relative to what can be inferred from one year of data. Put differently, because the sample of schools is largely unchanged from year to year (except for changes due to school openings and closings and, for small schools, data reporting issues) and the treatment designation does not change over time (i.e., adoptions are static), covariate balance should be expected to change very little from one year to the next.[12] Nonetheless, for completeness we show balancing results year-by-year in Table 2.

As suggested by Smith and Todd (2005), we present results from several balancing tests. The first row of the table reports the number of unbalanced covariates using simple covariate-by-covariate t-tests among the matched sample. A covariate where the difference between treatment and control values is significant at the 5 percent level is reported as unbalanced. We use 22 covariates in total to match schools and none are individually unbalanced at the 5 percent level within the matched sample based on the t-tests. This indicates that the unconditional differences in school characteristics shown in Table 1 disappear completely in the matched comparisons.[13]

In Row 2, we report the average absolute standardized difference across all covariates. Following Rosenbaum and Rubin (1985), the formula for the absolute standardized difference for covariate X_k is given by:

$$SDIFF(X_k) = \frac{\left| \frac{1}{N^S}\left[\sum_{j \in N_j \cap S_p} \{X_{kj} - \sum_{m \in I_{0j} \cap S_p} W(j,m) X_{km}\} - \sum_{m \in N_m \cap S_p} \{X_{km} - \sum_{j \in I_{0m} \cap S_p} W(m,j) X_{kj}\} \right] \right|}{\sqrt{\frac{Var(X_{kj}) + Var(X_{km})}{2}}} * 100 \qquad (7)$$

[12] In fact, the only reason that covariate balance changes from year-to-year is because of changes to the analytic sample owing to school openings and closings. The sample is centered around the matching years and year-1; thus going forward the sample shrinks as some schools close, and going backward it shrinks because some schools that were matched at the time of the new adoption had not yet opened.

[13] The covariates are as listed in Table 1. As noted above, we also use cubics in school and district enrollment and include a variable to indicate CDE data quality for individual schools.

The numerator in Equation (7) is analogous to the formula for our matching estimators in Equation (3) where we replace Y with X_k and take the absolute value (note the denominator is calculated using the full sample). The absolute average standardized difference is complementary to the covariate-by-covariate t-tests reported in the first row of the table. Beyond measuring purely statistical differences as with the t-tests, the absolute average standardized difference provides an indication of the magnitude of potential imbalance.

A weakness of reporting on standardized differences is that there is not a clear rule by which to judge the results. Rosenbaum and Rubin (1985) suggest that a value of 20 is large, although recent studies have applied more stringent criteria (e.g., Sianesi, 2004). The average absolute standardized differences that we report in Table 2 are quite small compared to similar estimates reported in other studies, on the order of just 3-4 percent across the pre- and post-adoption years of our data panel. This corroborates the result from the t-tests that the covariates are well-balanced between *California Math* adopters and other schools. In Appendix Table B.3 we report standardized differences on a covariate-by-covariate basis for interested readers.

Rows 3 and 4 of Table 2 show results from alternative, regression-based balancing tests proposed by Smith and Todd (2005). Like with the standardized difference measure, we perform the regression test for each covariate in each year and aggregate the results. Specifically, we estimate the following regression on a covariate-by-covariate basis:

$$\begin{aligned} X_{ik} &= \beta_0 + \beta_1 p_i + \beta_2 p_i^2 + \beta_3 p_i^3 + \beta_4 p_i^4 \\ &+ \beta_5 D_i + \beta_6 D_i p_i + \beta_7 D_i p_i^2 + \beta_8 D_i p_i^3 + \beta_9 D_i p_i^4 + \xi_{ik} \end{aligned} \tag{8}$$

In Equation (8), X_{ik} represents a covariate from the propensity-score specification for school *i*, p_i is the estimated propensity score, and D_i is an indicator variable equal to one if the school adopted *California Math* and zero otherwise. The test for balance is a test for whether the coefficients β_5-β_9

are jointly equal to zero – that is, whether treatment predicts the *X*'s conditional on a quartic of the propensity score.

We report the number of unbalanced covariates at the 5 percent level and the average p-value from the joint test of significance for β_5-β_9 across the 22 covariates in each year. Although we see marginally more unbalanced covariates than would be expected by chance using the Smith-Todd tests (2-3 per year), the implied level of imbalance is small. Moreover, the average p-values from the regression tests are consistently around 0.50 across the covariates in each year, which is as expected in a balanced comparison.[14]

Overall, based on the balancing information in Table 2, we conclude that our comparison of *California Math* to the composite alternative is well-balanced along the observable dimensions of our data.

5.2.3 Estimated Curriculum Effects for Grade 3

Table 3 shows results for our comparison between *California Math* and the composite alternative for cohorts of students exposed to 1-3 years of the curriculum materials in grade-3. The year-1 results compare students who used these textbooks for grade-3 only (and used previously-adopted materials in grades 1 and 2), the year-2 results show results for students who used the books in grades 2 and 3, and the year-3 and year-4 results are for students who used the books in all three grades leading up to the grade-3 test.

The point estimates from all three estimation strategies are similar and indicate effect sizes on the order of 0.05-0.08 student-level standard deviations of achievement.[15] However, the standard

[14] We report covariate-by-covariate balancing results for the primary comparison in Appendix Table B.3.
[15] The analysis is performed using school-level achievement measures. Effect sizes are converted into student-level standard deviation units, which are more commonly reported for other educational interventions in the literature, by multiplying them *ex post* by the ratio σ_s / σ_i, where σ_s is the standard deviation of the distribution of school-averaged math test scores and σ_i is the standard deviation of the distribution of student-level scores. We calculate σ_s using data from all reporting schools in California each year; σ_i is provided for all students by the CDE in annual

errors for the matching estimates are much larger than for the OLS or remnant-residualized estimates. The standard errors decrease using the latter two methods because the linear regression model removes variation in outcomes attributable to observed covariates. The cost of the improved precision is that the linear specification may be wrong, which is a reason that the parametrically less-restrictive matching estimators are preferred conceptually. That said, below we provide evidence in the form of falsification tests suggesting that our use of the linear functional form to improve precision does not result in biased estimates.

It is somewhat surprising that the treatment effect estimates do not become more pronounced over time in Table 3. That is, one might expect cohorts of students who are exposed to the curricula for all three years during grades 1-3 (students in year-3 or year-4 in the table) to show larger test-score differences than cohorts of students who are exposed for just 1-2 years (students in year-1 or year-2), but no such pattern emerges. There are a number of potential explanations. One possibility is that there is a dosage effect, but it is small enough that we lack the statistical power to detect it. Given the sizes of our standard errors, even in the OLS and remnant-residualized models, moderate dosage effects cannot be ruled out. Another possibility is that the most-recent textbook is the dominant treatment. Given that even the year-1 students used the new textbooks in grade-3, which is the most recent year of instruction leading up to the grade-3 test, it is possible that increased dosage in earlier grades is not important enough to show up in contemporary achievement results (if it matters at all). This explanation is plausible and consistent with numerous other studies showing fade-out in educational interventions (e.g., Chetty, Friedman and Rockoff, 2014; Currie and Thomas, 2000; Deming, 2009; Krueger and Whitmore, 2001). Another explanation is that curriculum-materials quality is not stable from grade-to-grade. We are not aware of any research that directly informs this

reports. Our conversion follows Bhatt and Koedel (2012). The ratio of σ_s / σ_i averaged across years in our data panel is 0.45 (the ratio varies very little from year to year).

hypothesis, but Bhatt, Koedel and Lehmann (2013) show that math curriculum effects can vary by subtopic, documenting at least one dimension of non-uniformity in effects. Our analyses of grades 4 and 5, shown below, are also suggestive of some grade-to-grade variability in the efficacy of *California Math* in the late elementary grades.

6. Falsification Tests

In this section we present results from falsification models designed to detect violations to our key identifying assumption, conditional independence. The idea behind the falsification models is to estimate "curriculum effects" in situations where there should not be any, or only very small effects. If we detect "curriculum effects" in situations where there should be none, or if we detect large curriculum effects in situations where small effects (at most) are expected, it would be suggestive of bias in our primary estimates.

We estimate two types of falsification models. The first type is a time-inconsistent model where we estimate curriculum effects on student achievement for cohorts of students who pre-date the adoption of the curriculum materials we study: specifically, students in the cohorts from year-p3 to year-p6. If our matching and regression-adjusted models are resulting in truly balanced comparisons (on observed and *unobserved* dimensions), we would not expect to see achievement differences between cohorts of students in matched schools prior to the curriculum adoptions of interest. The second type of falsification model estimates math curriculum effects on achievement in ELA. We estimate some time-inconsistent ELA models and some ELA models using achievement data that overlap with the timing of the math curriculum adoptions. In the latter case we cannot rule out non-zero curriculum effects because math curricula may have spillover effects onto ELA achievement. However, we would expect smaller cross-subject effects.

One issue with the falsification models is that we do not know which curriculum materials were used by schools prior to the curriculum adoption we study. No such longitudinal data on curriculum-materials adoptions exist, which points to an underlying problem with the state of curriculum data and research. In our models we rely on lagged school-level test scores to capture the impacts of previous curriculum materials on achievement (and all other educational inputs that we do not observe, for that matter), despite our inability to directly observe these materials. Whether this strategy is sufficient is ultimately an empirical question, which our falsification tests are designed to inform. If lagged test scores (and our other controls) are not sufficient to control for previous curriculum effects, and if curriculum adoptions are correlated across cycles within schools (which seems likely, but again data are limited), serial correlation in adoptions would be expected to manifest itself in the falsification tests in the form of non-zero pre-adoption "curriculum effects."

Table 4 shows the first set of falsification results from the time-inconsistent models of math achievement. Across all three estimation strategies and in all pre-adoption years, the false "effects" of *California Math* relative to the composite alternative are substantively small and far from statistical significance. This is as expected if our methods are sufficient to generate balanced comparisons. Table 5 shows the complementary falsification results using ELA achievement as the dependent variable. As is the case in Table 4, all of our estimates in Table 5 are small and statistically insignificant.[16]

Figure 1 visually illustrates our treatment effect and falsification estimates side-by-side over time. The bars with asterisks are for estimates that are statistically distinguishable from zero at the 5 percent level.

7. Results for Grades 4 and 5

Figures 2 and 3 show results from the full replication of our methods applied to grade-4 and grade-5 test scores, respectively. We follow analogous procedures as in the grade-3 analysis to produce

[16] For the remnant-residualized estimates in ELA, we re-estimate Equation (5) using ELA scores from the remnant sample to obtain appropriate adjustment parameters analogously to the procedure we follow for math scores described in Section 4.1.3.

the results for grades 4 and 5.[17] Like with the sample we constructed for our analysis of grade-3 scores, our grade-4 and grade-5 samples are well-balanced between *California Math* and composite-alternative schools. This is not surprising because the samples are essentially the same.[18] We do not report the balancing details for the grade-4 and grade-5 samples for brevity, but they are available upon request. The falsification results for the higher grades are shown in Figures 2 and 3; as in Figure 1, they provide no indication of selection bias in our primary estimates.

Taken together, our findings in Figures 2 and 3 are broadly consistent with the interpretation that *California Math* outperformed the composite of the other three focal curricula in California. Specifically, the estimates in grade-4 are all nominally positive and sometimes statistically significant, and the estimates in grade-5 are larger than in grade-3 and statistically significant in all post-treatment years, at least using restricted OLS and remnant-based residualization. There is no evidence to suggest a negative relative effect of *California Math* in any grade or year.

The effect sizes we estimate are largest in grade-5, but it is somewhat puzzling that they are smallest in grade-4. The up-and-down pattern of estimates holds even for cohorts who were exposed to *California Math* in all three grades, which indicates something other than a linearly-progressing dosage effect.[19] This is consistent with the results in Table 3, which also show no evidence of dosage effects for cohorts with more and less exposure to *California Math* in the early primary grades.

Unfortunately, because the literature on curricular efficacy is so thin, there is little prior evidence on which we can draw to gain inference about dosage effects. In similar previous studies

---

[17] With obvious appropriate adjustments; e.g., in the matching model for the grade-4 analysis we match schools on grade-4 test scores instead of grade-3 test scores.

[18] Specifically, we lose 2.5 percent of the initial grade-3 sample of schools in the grade-4 analysis, and another 1.5 percent when we move to the grade-5 analysis. The small data loss is attributable to schools that did not have pre-adoption test scores in grades 4 and/or 5 and schools that did not continue to uniformly adopt a focal curriculum past grade-3. The schools dropped from the sample as we move to higher grades are much smaller than the typical California school.

[19] There is some overlap in the samples between grades. For example, the year-2, grade-4 cohort is the same as the year-1, grade-3 cohort; the year-3, grade-4 cohort is the same as the year-2, grade-3 cohort; and the year-4, grade-4 cohort is the same as the year-3, grade-3 cohort. Similarly, there are two overlapping cohorts between the grade-3 and grade-5 results.

there is suggestive evidence of increased effect sizes for greater dosages in the early-primary grades, but no study finds a statistically significant effect of longer exposure to a curriculum that is more effective on average (Agodini et al., 2010; Bhatt and Koedel, 2012; Bhatt, Koedel and Lehmann, 2013). Our study, which provides the longest range of curricular-efficacy estimates in the literature to date (up to four consecutive years of use for the cohorts we follow the longest), can be characterized similarly. On the one hand, suggestive evidence of positive dosage effects across four different studies is more compelling than suggestive evidence from any single study, but on the other hand it is interesting that evidence of dosage effects is not stronger. The dip in our estimates in grade-4 for *California Math* suggests a potential mechanism worthy of additional exploration: the presence of grade-to-grade variability in the relative efficacy of curriculum materials.[20] More research is needed to understand why dosage effects are not stronger than they appear in the handful of available studies, which has implications for understanding the scope for the adoption of more-effective curriculum materials to raise student achievement.

8. Conclusion

We use unique school-level data on curriculum adoptions in California to estimate the achievement effects of *California Math* relative to a composite alternative consisting of *enVision Math California*, *California Mathematics: Concepts, Skills, and Problem Solving*, and *California HSP Math*. Our findings indicate that *California Math* outperformed the composite alternative curriculum during the time of our study. The differential effect in grade-3 is on the order of 0.05-0.08 student-level standard deviations of the state standardized assessment in mathematics, which would move the 50th percentile school to roughly the 54th to 56th percentile in the school-level distribution of average test scores.[21]

[20] However, we caution against over-interpreting this one result, which may be unique to the particular curricula we evaluate or could be the product of sampling variability. Note that in some years the grade-4 estimates are substantially smaller than the grade-5 and grade-3 estimates, but in other years they are quite close, especially given the sizes of our standard errors.

[21] Per the conversions used in this paper as described in the table notes, a 0.05/0.08 student-level standard deviation move corresponds to a 0.11/0.18 school-level standard deviation move.

The grade-5 estimates are suggestively larger, but while the grade-4 estimates are always nominally positive, they are smaller and mostly insignificant. A potential explanation for the grade-by-grade variability in our estimates that merits attention in future research is that the effects of curriculum materials vary across grades.

The relative achievement effects that we estimate for *California Math* in grade-3 are smaller than in the handful of related efficacy studies that focus on the early primary grades, including technically similar quasi-experimental studies by Bhatt and Koedel (2012) and Bhatt, Koedel and Lehmann (2013), and the experimental study by Agodini et al. (2010). Nonetheless, our estimates imply that *California Math* has an economically and educationally meaningful effect, particularly given that (a) it is a school-wide effect and thus applies, on average, to each student in a treated school, and (b) the marginal cost of choosing one curriculum over another is so small as to be effectively zero (Bhatt and Koedel, 2012; Chingos and Whitehurst, 2012). On the issue of the scope of the effects we estimate, note that an alternative intervention targeted at 10 percent of the student population would need to have an effect 10 times as large as the *California Math* effect to generate as large an increase in student achievement overall (ignoring spillovers).

We can only speculate as to why we find seemingly smaller differential curriculum effects in California than in previous studies.[22] Candidate explanations include that the curriculum materials in California are more similar to each other than the curriculum materials that have been evaluated previously, the context in California is such that curriculum effects are smaller (e.g., curricular objectives, assessments, etc.), or simply sampling variance. Our ability to gain inference into the mechanisms underlying the differential curriculum effects that we estimate here, and related estimates elsewhere, is hampered by the lack of a larger literature by which our findings can be contextualized.

[22] Note that in addition to our comparison centered on *California Math* yielding a smaller differential effect size, the suggestive results from our initial pairwise analysis (Appendix Table B.1) imply that the other three curricula are similarly effective.

Indeed, ours is one of only a small handful of rigorous studies that test the impacts of textbooks on student achievement in mathematics. Moreover, we are not aware of any similar studies in a subject outside of mathematics. By this point, we believe the methods we use are well established enough that it would be straightforward to apply them in other contexts if textbook data were available. By replicating this study across states or within states over time, we could begin to gather enough impact data to explore variation in curricular impact estimates as a function of features thought to matter (e.g., textbook content, alignment to standards, approach to teaching the subject, author, content or form of the state test). However, currently there is not enough efficacy information to support such investigations and in the meantime, studies like ours contribute efficacy evidence for specific sets of materials and can be used to inform contemporary curriculum adoption decisions, even if the features that make some curricula outperform others remain unidentified.

We conclude by re-iterating the calls made by Bhatt and Koedel (2012) and Chingos and Whitehurst (2012) for improved efforts to collect data on curriculum materials. Curriculum materials are a substantial input into educational production and data consistently point toward high curriculum materials usage by students and teachers in the Common Core era (Opfer et al., 2016; Perry et al., 2015). However, it remains the case that in nearly all states, which curriculum materials are being used by which schools is not tracked. Even in California where reporting on curriculum materials is the law, we found that information provided by a significant fraction of schools does not actually identify the curriculum materials being used (Appendix Table A.1), which suggests little oversight of the data. This much is for certain: with no data, we are committed to leaving educational decision makers to adopt curricula without efficacy evidence.

References

Aaronson, Daniel, Lisa Barrow and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics* 25(1), 95-135.

Agodini, Robert and Barbara Harris and Sally Atkins-Burnett and Sheila Heaviside, and Timothy Novak. 2010. *Achievement Effects of Four Early Elementary School Math Curricula: Findings for First and Second Graders*. National Center for Education Evaluation and Regional Assistance, U.S. Department of Education, Institute of Education Sciences. NCEE 2011-4001.

Bhatt, Rachana and Cory Koedel. 2012. "Large-Scale Evaluations of Curricular Effectiveness: The Case of Elementary Mathematics in Indiana," *Educational Evaluation and Policy Analysis* 34(4): 391-412.

Bhatt, Rachana, Cory Koedel and Douglas Lehmann. 2013. "Is Curriculum Quality Uniform? Evidence from Florida," *Economics of Education Review* 34(1): 107-121.

Black, Dan and Jeffrey Smith. 2004. "How Robust is the Evidence on the Effects of College Quality? Evidence From Matching," *Journal of Econometrics* 121 (2), 99-124.

Boser, Ulrich, Matt Chingos and Chelsea Straus. 2015. *The Hidden Value of Curriculum Reform: Do States and Districts Receive the Most Bang for Their Curricular Buck?* Washington, DC: Center for American Progress.

Caliendo, Marco and Sabine Kopeinig. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching," *Journal of Economic Surveys* 22(1), 31-72.

California Department of Education. 2009. *2007 Mathematics Primary Adoption Report.* Sacramento, CA: California Department of Education Press.

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

Chingos, Matthew M. and Grover (Russ) J. Whitehurst. 2012. *Choosing Blindly: Instructional Materials, Teacher Effectiveness and the Common Core*. Policy Report, The Brown Center on Education Policy.

Currie, Janet and Duncan Thomas. 2000. School Quality and the Longer-Term Effects of Head Start. *Journal of Human Resources* 35(4), 755-774.

Deming, David J. 2009. Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics* 1(3), 111-134.

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

Hansen, Ben B. 2008. The Prognostic Analogue of the Propensity Score. *Biometrika* 95(2), 481-488.

Heckman, James and Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job-Training Programme," *Review of Economic Studies* 64 (4), 261-294.

Jobrack, Beverlee. 2011. *Tyranny of the Textbook: An Insider Exposes how Educational Materials Undermine Reforms*. Lanham, MD: Rowman & Littlefield Publishers.

Koedel, Cory and Julian R. Betts. 2011. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique," *Education Finance and Policy* 6(1), 18-42.

Krueger, Alan B. and Diane M. Whitmore. 2001. The Effect of Attending a Small Class in the Early Grades on College Test Taking and Middle School Test Results: Evidence from Project STAR. *Economic Journal* 111(468), 1-28.

Lechner, Michael. 2002. "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies," *The Review of Economics and Statistics* 84 (2), 205-220.

Mueser, Peter R. and Kenneth R. Troske and Alexey Gorislavsky. 2007. "Using State Administrative Data to Measure Program Performance," *The Review of Economics and Statistics* 89 (4), 761-83.

Opfer, V. D., Julia H. Kaufman, and Lindsey E. Thompson. 2016. *Implementation of K-12 State Standards for Mathematics and English Language Arts and Literacy*. Santa Monica, CA: RAND.

Perry, Rebecca R., Neal D. Finkelstein, Nanette Seago, Alberto Heredia, Sandy Sobolew-Shubin, and Cathy Carroll. 2015. *Taking Stock of Common Core Math Implementation: Supporting Teachers to Shift Instruction*. San Francisco, CA: WestEd.

Polikoff, Morgan S. 2015. "How Well Aligned Are Textbooks to the Common Core Standards in Mathematics?" *American Educational Research Journal 52* (6), 1185-1211.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observtional Studies for Causal Effects," *Biometrika* 70 (1), 41-55.

Rosenbaum, Paul R. and Donald B. Rubin. 1985. "The Bias due to Incomplete Matching," *Biometrika* 41 (1), 103-116.

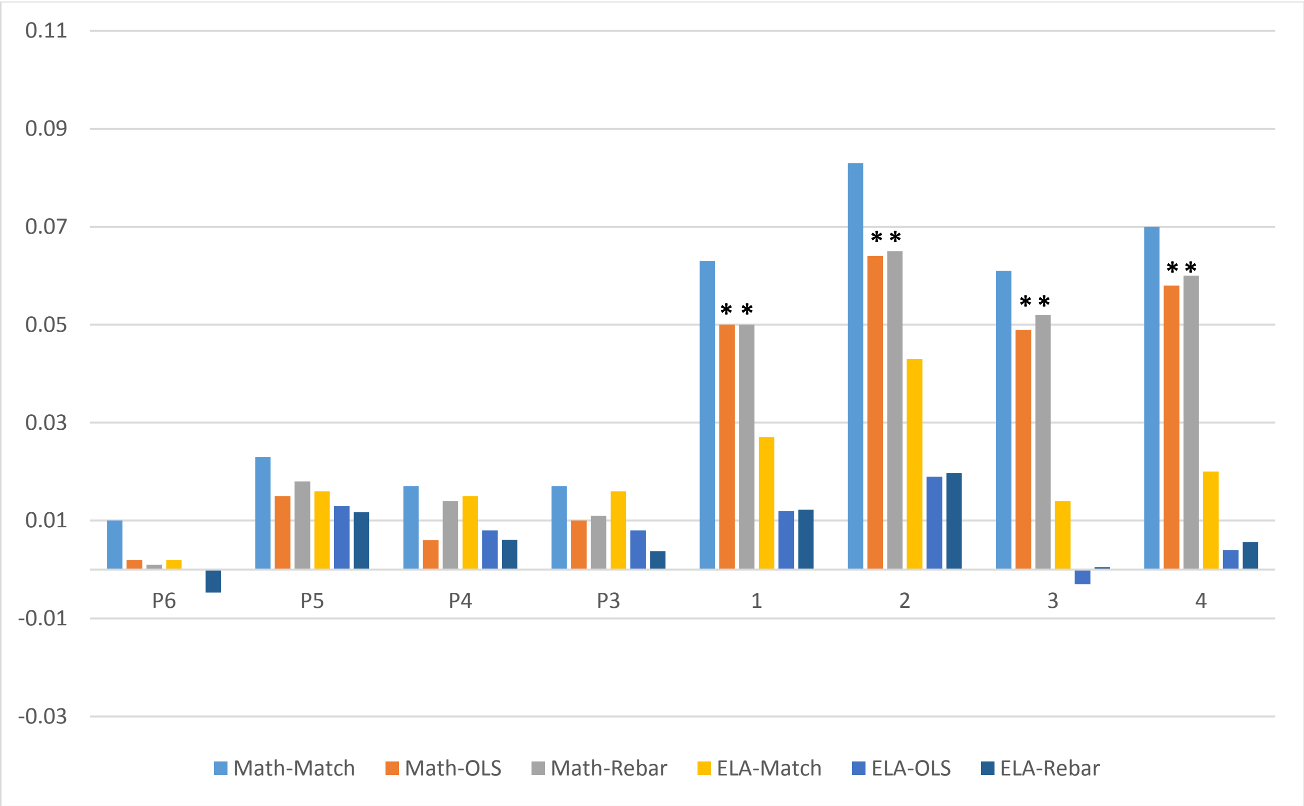
Sales, Adam, Ben B. Hansen and Brian Rowan. 2014. Rebar: Reinforcing a Matching Estimator with Predictions from High-Dimensional Covariates. Unpublished manuscript

Sianesi, Barbara. 2004. An evaluation of the Swedish system of active labor market programs in the 1990s. *Review of Economics and Statistics*, 86(1), 133-155.

Smith, Jeffrey and Petra Todd. 2005. "Rejoinder," *Journal of Econometrics* 125 (2), 365-375.

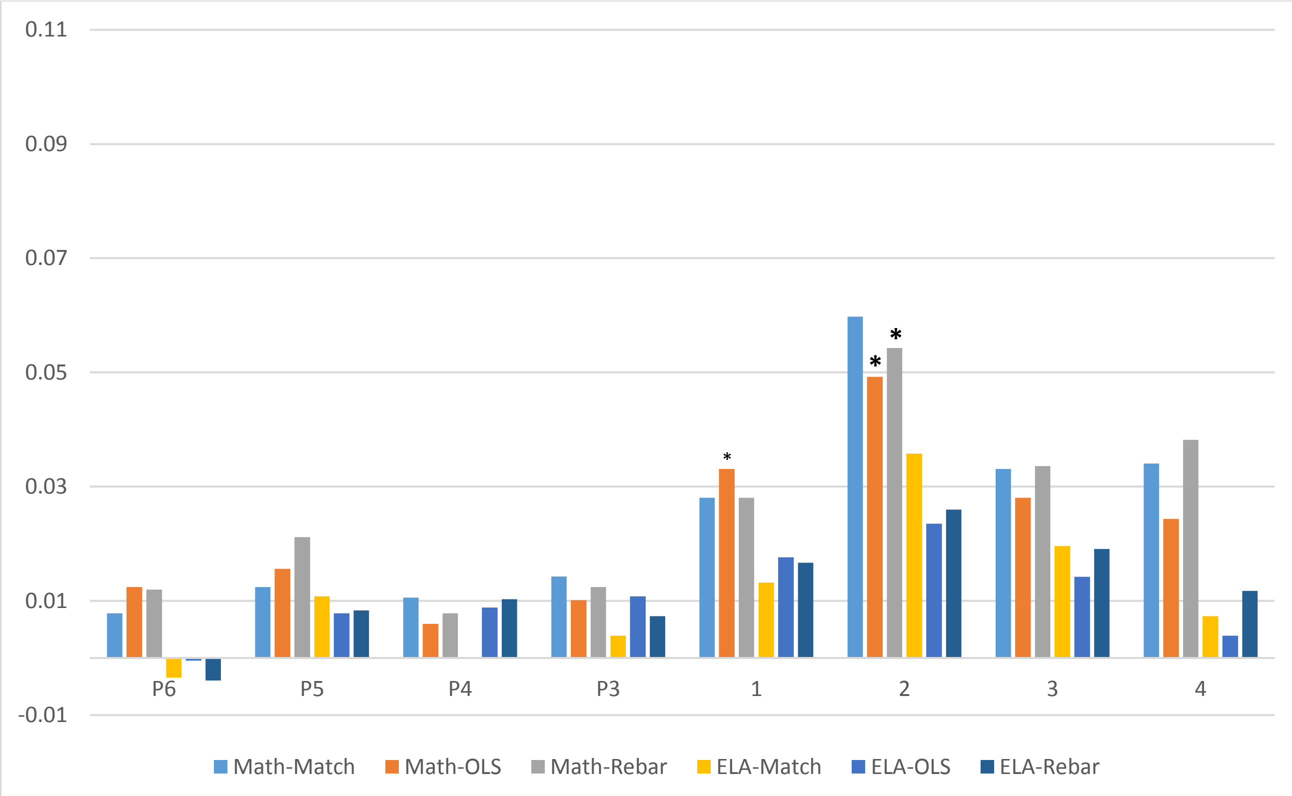
Zeringue, J. K., Spencer, D., Mark, J., & Schwinden, K. (2010). *Influences on Mathematics Textbook Selection: What Really Matters?* Paper presented at the Research Pre-session of the National Council of Teachers of Mathematics, San Diego, CA.
URL: http://mcc.edc.org/pdf/Final_Draft_Research_Presession_2010.pdf

Figure 1. Effects of California Math Relative to the Composite Alternative on Grade-3 Test Scores, Over Time and Using Different Estimators.



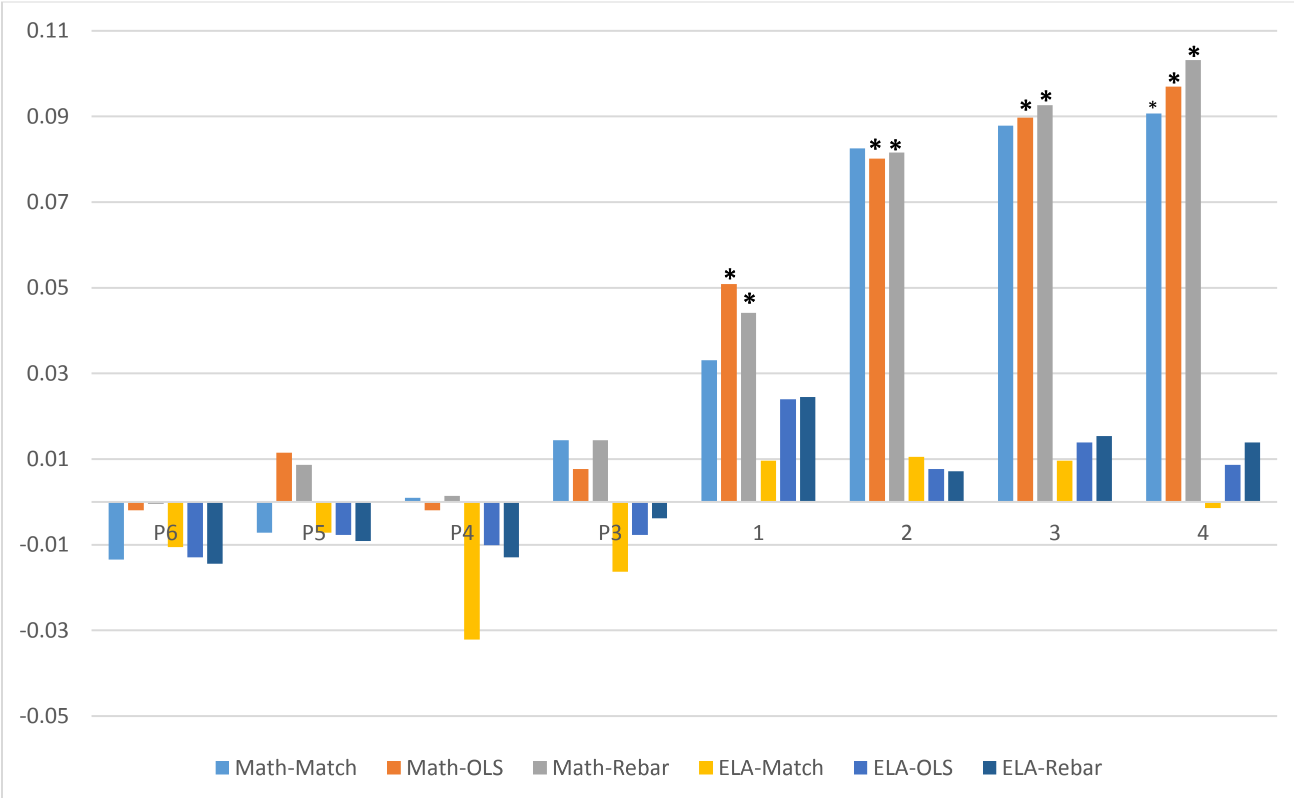
Notes: Each bar shows an estimate reported in the preceding tables. All estimates are converted to student-level standard deviation units. Bars with asterisks are for estimates that are statistically distinguishable from zero at the 5 percent level. Years P6-P3 are pre-treatment years; years 1-4 are post-treatment years.

Figure 2. Effects of California Math Relative to the Composite Alternative on Grade-4 Test Scores, Over Time and Using Different Estimators.



Notes: All estimates are converted to student-level standard deviation units. Bars with large, bolded asterisks are for estimates that are statistically distinguishable from zero at the 5 percent level; small, standard-font asterisks indicate statistical significance at the 10 percent level. Years P6-P3 are pre-treatment years; years 1-4 are post-treatment years. They year-2, grade-4 cohort in the post-treatment period corresponds to the year-1, grade-3 cohort; the year-3, grade-4 cohort corresponds to the year-2, grade-3 cohort; etc.

Figure 3. Effects of California Math relative to the Composite Alternative on Grade-5 Test Scores, Over Time and Using Different Estimators.



Notes: All estimates are converted to student-level standard deviation units. Bars with large, bolded asterisks are for estimates that are statistically distinguishable from zero at the 5 percent level; small, standard-font asterisks indicate statistical significance at the 10 percent level. Years P6-P3 are pre-treatment years; years 1-4 are post-treatment years. They year-3, grade-5 cohort in the post-treatment period corresponds to the year-1, grade-3 cohort; the year-4, grade-5 cohort corresponds to the year-2, grade-3 cohort.

Table 1. Descriptive Statistics for California Schools, Our Full Analytic Sample, and by Textbook Adoption.

	All Schools	All Schools without LAUSD or LBUSD	Analytic Sample	Within the Analytic Sample, by Textbook: Envision Math California	California Math	California Mathematics: Concepts, Skills, and Problem Solving	California HSP Math
School Outcomes							
Pre-Adoption Grade-3 Math Score	0.02	0.03	-0.03	-0.08	0.06	-0.09	-0.02
Pre-Adoption Grade-3 ELA Score	0.01	0.05	-0.03	-0.09	0.07	-0.10	-0.02
School Chars							
Percent Female	48.7	48.7	48.7	48.5	48.9	48.8	48.7
Percent Econ Disadvantaged	56.6	54.2	56.9	56.2	56.0	59.1	58.0
Percent English Learner	29.3	28.3	29.5	30.2	28.0	29.9	30.3
Percent White	31.4	33.6	29.5	30.1	29.9	29.2	26.4
Percent Black	7.8	7.1	7.3	8.0	6.3	8.6	4.8
Percent Asian	8.4	8.8	7.5	7.6	7.2	7.5	8.4
Percent Hispanic	47.9	45.8	50.5	48.6	51.7	49.7	56.3
Percent Other	4.6	4.7	5.2	5.7	4.9	5.0	4.1
Enrollment	385.7	378.1	410.5	399.9	429.5	405.7	399.0
2008 Adopter			50.2	49.0	53.7	53.5	36.2
School-Area Chars (Census)							
Median Household Income (log)	11.0	11.0	10.8	10.7	10.9	10.8	10.9
Share Low Education	17.8	17.2	19.5	17.6	19.3	22.6	23.9
Share Missing Census Data	3.1	3.3	1.8	2.7	1.2	0.8	0.0
District Outcomes							
Pre-Adoption Grade-3 Math Score	0.01	0.01	-0.02	0.03	0.00	-0.05	-0.05
Pre-Adoption Grade-3 ELA Score	0.02	0.02	-0.09	-0.03	-0.12	-0.09	-0.12
District Characteristics							
Enrollment	5138.0	4438.9	5690.4	6404.0	6075.5	5279.0	4339.9
N (Schools)	5,494	4,931	1,878	710	602	389	177
N (Districts)	825	823	309	107	92	69	48

Notes: The "all schools" sample is the universe of schools reported in Appendix Table A.1. It includes schools in the CDE data with characteristics from either 2007 or 2008, at least one grade-3 test score from 2009-2013, and where the highest graded is 8 or lower. The descriptive statistics for the analytic sample in column 3 are a weighted average of the textbook-by-textbook statistics reported in columns 3-6. Note that some districts have a uniformly-adopting school of more than one textbook, thus the sum of the district counts in the last four columns is greater than 309.

Table 2. Balancing Results for the Primary Comparison.

Treatment: *California Math* Control: *Composite Alternative*	Year-P6	Year-P5	Year-P4	Year-P3	Year-1	Year-2	Year-3	Year-4
No. Unbalanced Covariates, Matched T-tests (5 percent)	0	0	0	0	0	0	0	0
Mean Standardized Difference of Covariates (%)	2.8	3.0	3.0	3.6	3.5	3.4	3.1	3.6
No. Unbalanced Covariates, Smith-Todd Regression Tests (5 percent)	1	2	2	2	2	3	2	2
Average P-value, Smith-Todd Regression Tests	0.50	0.50	0.51	0.48	0.48	0.46	0.50	0.49
No. of Districts/Schools (California Math)	89/560	88/567	90/575	90/588	92/597	89/588	91/595	90/590
No. of Districts/Schools (Composite Alternative)	210/1,063	213/1,085	212/1,106	215/1,124	213/1,143	214/1,145	216/1,146	213/1,144

Notes: There are 22 covariates included in the balancing tests. The sample size fluctuates year-to-year due to school openings and closings, and data reporting issues for small schools. Note that there is a 2-year gap between Year-P3 and Year-1. We use data from the two gap years to match schools as described in the text.

Table 3. Effects of California Math on Grade-3 Mathematics Achievement for Exposed Cohorts Relative to the Composite Alternative, by Year After the Initial Adoption.

	Year-1	Year-2	Year-3	Year-4
Treatment: *California Math*				
Control: *Composite Alternative*				
Treatment Effect: Kernel Matching	0.063	0.083	0.061	0.070
	(0.054)	(0.051)	(0.059)	(0.059)
Treatment Effect: Restricted OLS	0.050	0.064	0.049	0.058
	(0.019)**	(0.023)**	(0.023)**	(0.023)**
Treatment Effect: Remnant-Residualized Matching	0.050	0.065	0.052	0.060
	(0.020)**	(0.024)**	(0.024)**	(0.026)**
No. of Districts/Schools (California Math)	92/597	89/588	91/595	90/590
No. of Districts/Schools (Composite Alt.)	213/1,143	214/1,145	216/1,146	213/1,144

Notes: Standard errors are estimated by bootstrapping using 250 repetitions and clustered at the district level. Year-1 denotes the first year the new curriculum was adopted (e.g., the 2008-2009 school year for textbooks adopted in fall-2008), year-2 denotes the second year, etc. All estimates are converted from school-level standard deviation units to student-level standard deviation units by multiplying them by a factor of 0.45, which is the ratio of standard deviations of the school-average test score distribution and the student-level test score distribution averaged across our data panel, as reported in the text. This transformation has no bearing on the results qualitatively or quantitatively – the rescaling is performed only to improve comparability of our findings to those in other studies that report effect sizes in student-level standard deviation units.
**/* Indicates statistical significance at the 5/10 percent level.

Table 4. Falsification Results: California Math "Effects" on Grade-3 Mathematics Achievement for Cohorts of Students in Years Prior to the 2008/2009 Adoption Cycle.

	Year-P3	Year-P4	Year-P5	Year-P6
Treatment: *California Math* Control: *Composite Alternative*				
Treatment Effect: Kernel Matching	0.010	0.023	0.017	0.017
	(0.053)	(0.051)	(0.054)	(0.056)
Treatment Effect: Restricted OLS	0.002	0.015	0.006	0.010
	(0.014)	(0.015)	(0.015)	(0.018)
Treatment Effect: Remnant-Residualized	0.001	0.018	0.014	0.011
Matching	(0.014)	(0.018)	(0.020)	(0.024)
No. of Districts/Schools (California Math)	89/560	88/567	90/575	90/588
No. of Districts/Schools (Composite Alt.)	210/1,063	213/1,085	212/1,106	215/1,124

Notes: Standard errors are estimated by bootstrapping using 250 repetitions and clustered at the district level. Year-P3 denotes the school year 3 years prior to the new curriculum being adopted (e.g., the 2005-2006 school year for textbooks adopted in fall-2008), year-P4 denotes the year 4 years prior, etc. Data from the two years preceding the adoption are used to match schools and thus not analyzed directly. All estimates are converted from school-level standard deviation units to student-level standard deviation units by multiplying them by a factor of 0.45, which is the ratio of standard deviations of the school-average test score distribution and the student-level test score distribution in math averaged across our data panel, as reported in the text. This transformation has no bearing on the results qualitatively or quantitatively – the rescaling is performed only to improve comparability of our findings to those in other studies that report effect sizes in student-level standard deviation units.
**/* Indicates statistical significance at the 5/10 percent level.

Table 5. Falsification Results: California Math "Effects" on Grade-3 English Language Arts Achievement for Exposed and Un-Exposed Cohorts.

Treatment: *California Math* Control: *Composite Alternative*	Year-P6	Year-P5	Year-P4	Year-P3	Year-1	Year-2	Year-3	Year-4
Treatment Effect: Kernel Matching	0.002	0.016	0.015	0.016	0.027	0.043	0.014	0.020
	(0.064)	(0.060)	(0.058)	(0.057)	(0.061)	(0.056)	(0.066)	(0.064)
Treatment Effect: Restricted OLS	-0.000	0.013	0.008	0.008	0.012	0.019	-0.003	0.004
	(0.016)	(0.016)	(0.014)	(0.013)	(0.017)	(0.021)	(0.020)	(0.022)
Treatment Effect: Remnant-Residualized	-0.005	0.012	0.006	0.004	0.012	0.020	0.001	0.006
Matching	(0.021)	(0.018)	(0.015)	(0.015)	(0.017)	(0.022)	(0.020)	(0.026)
No. of Districts/Schools (California Math)	89/560	88/567	90/575	90/588	92/597	89/588	91/595	90/590
No. of Districts/Schools (Composite Alt.)	210/1,063	213/1,085	212/1,106	215/1,124	213/1,143	214/1,145	216/1,146	213/1,143

Notes: Standard errors are estimated by bootstrapping using 250 repetitions and clustered at the district level. Year-P3 denotes the school year 3 years prior to the new curriculum being adopted (e.g., the 2005-2006 school year for textbooks adopted in fall-2008), year-P4 denotes the year 4 years prior, etc. Year-1 denotes the first year the new curriculum was adopted (e.g., the 2008-2009 school year for textbooks adopted in fall-2008), year-2 denotes the second year, etc. All estimates are converted from school-level standard deviation units to student-level standard deviation units by multiplying them by a factor of 0.47, which is the ratio of standard deviations of the school-average test score distribution and the student-level test score distribution in ELA averaged across our data panel. This transformation has no bearing on the results qualitatively or quantitatively – the rescaling is performed only to improve comparability of our findings to those in other studies that report effect sizes in student-level standard deviation units.
**/* Indicates statistical significance at the 5/10 percent level.

Appendix A
Data Appendix

Appendix Table A.1. Construction of the Analytic Sample.

	Schools	% of total	Districts	% of total
Initial Universe	5,494		825	
Reasons for data loss				
No record in textbook file	-339	6.2	-32	3.9
Indeterminate textbook information	-804	14.6	-134	16.2
Adoption year other than 2008 or 2009	-876	15.9	-119	14.4
Non-uniform adopter (or uncertain), grades 1-3	-481	8.8	-54	6.6
Gradespan conflict between CDE and SARC data	-33	0.6	-17	2.1
Missing school/district outcome data	-48	0.9	-19	2.3
Missing district/school covariate data	0	0	0	0
Did not use one of the four focal curricula	-632	11.5	-139	16.8
Initial Analytic Sample	2,281	41.5	311	37.7
Drop LAUSD and LBUSD	-403	7.3	-2	0.2
Final Analytic Sample	1,878	34.2	309	37.5

Notes: The initial universe includes all schools in the CDE data with characteristics from either 2007 or 2008, at least one grade-3 test score from 2009-2013, and where the highest graded is 8 or lower.

Appendix B
Supplementary Results

B.1 Pairwise Comparisons

Appendix Table B.1 summarizes initial results from the six pairwise comparisons. The first three comparisons involve what becomes the focal curriculum in our analysis: *California Math*. *California Math* is the treatment curriculum in the first comparison and the control curriculum in the other two (we use the convention of defining the most-adopted book as the "control" curriculum in each pairwise comparison). Notice that we obtain fairly large point estimates in all three comparisons involving *California Math*, and all three comparisons suggest that *California Math* is more effective. For the comparisons involving the other curricula, our point estimates are consistently small and do not suggest differential effects.

Like for the primary comparison in the text, we report balancing information in several ways for each pairwise comparison in Appendix Table B.1. As is clear from the table, a limitation of most of the pairwise comparisons is that the balancing results, while not indicative of egregious imbalance, are also not particularly compelling. Covariate balance using the matched t-tests generally looks good, but the mean standardized difference for several of the pairwise comparisons is large, and certainly much larger than in the comparison between *California Math* and the composite alternative. In all pairwise comparisons the Smith and Todd (2005) regression tests indicate imbalance in one form or another (i.e., either too many unbalanced covariates and/or average p-values that are too low).

As noted in the text, our small sample sizes in the pairwise comparisons (relative to sample sizes more typical of matching analyses in other contexts) limit our ability to improve covariate balance separately for each comparison. Thus, based on these initial results, and the suggestion that *California Math* is more effective than the other three textbooks (which all appear to be similarly effective), we focus our main evaluation on comparing *California Math* to a composite of the other three popular curricula. Reducing the dimensionality of the comparison in this way yields a more effective matching procedure, which can be seen by comparing the balance statistics shown in Appendix Table B.1 for the pairwise comparisons to the analogous numbers for the composite comparison in the main text (Table 2). The falsification tests shown in the main text offer additional evidence consistent with our final evaluation of *California Math* being balanced.

Appendix Table B.1. Balance and Estimation Results for the Six Initial Pairwise Comparisons During Treatment Years.

	Estimated Treatment Effects and Balancing Results by Year After Adoption			
	Year-1	Year-2	Year-3	Year-4
Comparison 1				
Treatment: California Math				
Control: Envision Math				
Treatment Effect (Kernel Matching)	0.048	0.059	0.041	0.054
	(0.063)	(0.066)	(0.058)	(0.061)
No. Unbalanced Covariates, Matched T-tests (5 percent)	2	3	2	2
Mean Standardized Difference of Covariates	6.0	5.8	6.1	6.1
No. Unbalanced Covariates, Smith-Todd (5 percent)	3	4	4	3
Average P-value, Smith-Todd	0.41	0.28	0.43	0.40
Comparison 2				
Treatment: California Mathematics: Concepts, Skills, and Problem Solving				
Control: California Math				
Treatment Effect (Kernel Matching)	-0.087	-0.152	-0.110	-0.091
	(0.072)	(0.077)**	(0.072)	(0.076)
No. Unbalanced Covariates, Matched T-tests (5 percent)	0	0	0	0
Mean Standardized Difference of Covariates	4.8	5.5	4.3	4.1
No. Unbalanced Covariates, Smith-Todd (5 percent)	11	5	3	3
Average P-value, Smith-Todd	0.21	0.30	0.31	0.31
Comparison 3				
Treatment: California HSP Math				
Control: California Math				
Treatment Effect (Kernel Matching)	-0.063	-0.065	-0.039	-0.059
	(0.063)	(0.057)	(0.072)	(0.076)
No. Unbalanced Covariates, Matched T-tests (5 percent)	0	0	0	0
Mean Standardized Difference of Covariates	6.2	6.4	5.9	5.6
No. Unbalanced Covariates, Smith-Todd (5 percent)	5	5	5	4
Average P-value, Smith-Todd	0.28	0.27	0.28	0.29
Comparison 4				
Treatment: California Mathematics: Concepts, Skills, and Problem Solving				
Control: Envision Math				
Treatment Effect (Kernel Matching)	0.010	-0.017	-0.003	0.005
	(0.066)	(0.066)	(0.058)	(0.065)
No. Unbalanced Covariates, Matched T-tests (5 percent)	0	0	0	0
Mean Standardized Difference of Covariates	5.6	5.2	5.4	5.2
No. Unbalanced Covariates, Smith-Todd (5 percent)	3	3	3	3
Average P-value, Smith-Todd	0.52	0.56	0.56	0.57
Comparison 5				
Treatment: California HSP Math				
Control: Envision Math				
Treatment Effect (Kernel Matching)	0.065	0.004	-0.009	0.013
	(0.090)	(0.078)	(0.106)	(0.104)
No. Unbalanced Covariates, Matched T-tests (5 percent)	0	0	0	0
Mean Standardized Difference of Covariates	6.4	6.7	6.5	6.3
No. Unbalanced Covariates, Smith-Todd (5 percent)	4	4	4	4
Average P-value, Smith-Todd	0.39	0.39	0.39	0.39

Comparison 6				
Treatment: California HSP Math				
Control: California Mathematics: Concepts, Skills, and Problem Solving				
Treatment Effect (Kernel Matching)	0.016	0.021	0.058	0.028
	(0.091)	(0.079)	(0.081)	(0.083)
No. Unbalanced Covariates, Matched T-tests (5 percent)	0	0	0	0
Mean Standardized Difference of Covariates	4.1	4.2	3.9	4.2
No. Unbalanced Covariates, Smith-Todd (5 percent)	6	6	6	6
Average P-value, Smith-Todd	0.31	0.34	0.33	0.34
No. of Districts/Schools (Envision Math California)	106/706	106/707	107/707	105/706
No. of Districts/Schools (California Math)	92/602	89/593	91/600	90/599
No. of Districts/Schools (California Mathematics: Concepts, Skills and Problem Solving)	67/387	69/389	69/389	69/389
No. of Districts/Schools (California HSP Math)	48/177	47/176	48/177	47/176

Notes: The balancing tests report results based on the same 22 matching covariates used in each pairwise comparison. Standard errors for matching estimators are estimated by bootstrapping using 250 repetitions and clustered at the district level. Year-1 denotes the first year the new curriculum was adopted (e.g., the 2008-2009 school year for textbooks adopted in fall-2008), year-2 denotes the second year, etc. All estimates are converted from school-level standard deviation units to student-level standard deviation units by multiplying them by a factor of 0.45, which is the ratio of standard deviations of the school-average test score distribution and the student-level test score distribution in math averaged across our data panel, as reported in the text. This transformation has no bearing on the results qualitatively or quantitatively – the rescaling is performed only to improve comparability of our findings to those in other studies that report effect sizes in student-level standard deviation units.
**/* Indicates statistical significance at the 5/10 percent level.

B.2 Matching Details for the Primary Comparison and Overlap of Propensity Scores

Appendix Tables B.2 and B.3 report details about the matching procedure for the primary comparison between *California Math* and the Composite Alternative. First, Table B.2 shows the output from the initial selection model from which the propensity scores are generated to give a sense of which covariates predict the adoption of *California Math*. The only statistically significant covariates are the three terms for district enrollment (linear, quadratic, cubic).

Second, Table B.3 shows covariate-by-covariate balancing results to complement the aggregate reporting in Table 2. For brevity, we show covariate-by-covariate balance using the Year-1 sample of schools and districts only (recall from the text that the balancing results fluctuate mildly from year-to-year because of sample changes due to building openings and closings and data reporting issues for small schools).

Figure B.1 shows the distributional overlap in propensity scores between *California Math* (treatments) and other focal-curricula adopters (controls). The propensity scores are summary measures of school and district characteristics, weighted by their predictive influence over the adoption of *California Math*. In any program evaluation where treatment is predicted at least to some degree by observable characteristics, treatment units will have higher propensity scores on average than controls, as in the case in Figure B.1. However, the figure shows considerable overlap in the distributions of propensity scores for treatment and control schools, which is conducive to our matching evaluation.

Appendix Table B.2. Probit Coefficients from the Propensity Score Model Predicting the Adoption of *California Math* Instead of the Composite Alternative.

Data Quality Indicator	-1.028
	(0.715)
Census Data Missing Indicator	-1.777
	(4.580)
Fall-2008 Adoption	0.305
	(0.222)
School Average Math Score (Standardized)	-0.049
	(0.059)
School Average ELA Score (Standardized)	0.189
	(0.126)
District Average Math Score (Standardized)	-0.078
	(0.434)
District Average ELA Score (Standardized)	0.338
	(0.410)
Share Female	0.438
	(1.148)
Share Economically Disadvantaged	0.633
	(0.673)
Share African American	-1.328
	(1.296)
Share Asian	-0.729
	(0.807)
Share White	-0.738
	(0.781)
Share Other	-1.539
	(1.721)
Share English Learner	-0.886
	(0.879)
School Enrollment (1000s)	4.091
	(2.640)
School Enrollment Squared (1000s)	-0.00615
	(0.00473)
School Enrollment Cubed (1000s)	0.00000302
	(0.00000251)
District Enrollment (1000s)	-0.167
	(0.089)*
District Enrollment Squared (1000s)	0.0000172
	(0.00000710)**
District Enrollment Cubed (1000s)	0.000000000382
	(0.000000000150)**
Share Low Education (U.S. Census)	-0.004
	(0.011)
Median Household Income (U.S. Census)	-0.153
	(0.398)
Constant	2.263
	(4.450)
Psuedo R-Squared	0.1221
N (total)	1,878

Notes: The data quality indicator is set to one if the sum of student subgroups does not equal total enrollment as reported by the CDE. This was not an issue for most schools, and even when it was, inequalities were small. *Ex post* this variable has no bearing on our findings, and all of our results are robust to excluding it. The omitted student categories are the share male, economically disadvantaged, Hispanic, and non-English Learner.
**/* Indicates statistical significance at the 5/10 percent level

Appendix Table B.3. Covariate-by-Covariate Balancing Details for the Comparison between *California Math* and the Composite Alternative, Year-1 Sample.

	Matched t-test, Significant Difference	Standardized Difference	Smith-Todd Test, Significant Difference	Smith-Todd p-value
Data Quality Indicator	No	12.2	No	0.85
Census Data Missing Indicator	No	2.6	No	0.27
Fall-2008 Adoption	No	2.7	No	0.69
School Average Math Score	No	1.8	No	0.50
School Average ELA Score	No	4.3	No	0.77
District Average Math Score	No	1.3	No	0.68
District Average ELA Score	No	5.5	No	0.92
Share Female	No	1.1	Yes	0.02
Share Economically Disadvantaged	No	-4.4	No	0.97
Share African American	No	-0.3	No	0.53
Share Asian	No	-2.1	No	0.29
Share White	No	7.4	No	0.91
Share Other	No	-1.9	No	0.45
Share English Learner	No	-5.9	No	0.73
School Enrollment	No	2.6	No	0.24
School Enrollment Squared	No	1.5	No	0.15
School Enrollment Cubed	No	0.7	No	0.13
District Enrollment	No	3.1	Yes	0.01
District Enrollment Squared	No	3.4	No	0.10
District Enrollment Cubed	No	4.3	No	0.14
Share Low Education (Census)	No	-5.5	No	0.79
Median Household Income (Census)	No	-2.3	No	0.36

Notes: This table provides full details for the balancing results shown in Table 2 for year-1. Detailed balancing results for other years are substantively similar. The average absolute standardized difference reported in Table 2 is the average of the absolute values of the standardized differences reported in this table.

Appendix Figure B.1. Kernel Densities of Estimated Propensity Scores for Treatment (*California Math*) and Control (Composite Alternative) Schools on the Common Support, Grade-3 Math.