B | Economic Studies
at BROOKINGS

Evidence Speaks Reports, Vol 2, #5
January 8, 2017

Big bang for just a few bucks: The impact of math textbooks in California

Cory Koedel and Morgan Polikoff

Executive Summary

Textbooks are one of the most widely used educational inputs, but remarkably little is known about their effects on student learning. This report uses data collected from elementary schools in California to estimate the impacts of mathematics textbook choices on student achievement. We study four of the most popular books in the state from 2008-2013 and find that one—Houghton Mifflin *California Math*—consistently outperforms the other three. The superior performance of *California Math* persists up to four years after adoption and shows up in grades 3, 4, and 5.

The textbook impacts we identify are educationally meaningful and come at an extremely low cost. With regard to cost, textbooks are relatively inexpensive and tend to be similarly priced. The implication is that the marginal cost of choosing a more effective textbook over a less effective alternative is essentially zero. In terms of achievement impacts, our findings suggest non-trivial gains in student achievement are attainable simply by choosing more effective curriculum materials. The effect sizes we document are on par with what one could expect from a hypothetical policy that substantially increases the quality of the teaching workforce. But whereas there is much uncertainty about whether commensurate increases in teacher quality are attainable, and how they might be attained—at least in the near term—choosing a more effective textbook is a seemingly straightforward policy option for raising student achievement.

A critical factor limiting the capacity of school administrators to choose more effective textbooks is that there is virtually no evidence on how different textbooks affect student achievement. The fundamental problem limiting the development of an evidence base is that very few states track school and district textbook adoptions. This point bears repeating: most states do not know which curriculum materials are being used in which schools and districts. Without these data, it is not possible to perform evaluations of textbook efficacy. Thus, in most states, decisionmakers who wish to incorporate into their adoption decisions evidence on how textbooks affect student achievement are simply out of luck.

Our study adds to a small body of research showing that textbooks differ considerably in their effectiveness. It is an example of the type of analysis that could be performed much more broadly in other states and over time, if only data on textbook adoptions were routinely collected. By combining evidence from multiple, similar studies, we could build an evidence base for numerous sets of materials in a variety of contexts, and perform analyses to determine what features of curriculum materials matter for promoting student achievement. We conclude with suggestions for district officials and state policymakers to collect and make textbook adoption data available. The recommended data collection efforts would come with a relatively low burden and lead to substantially better evidence to drive school and district decisions with regard to an important and commonly-used schooling input.

A bit over four years ago, Russ Whitehurst and Matt Chingos put out a well-articulated call for greater research on curriculum materials and their effects.[i] The argument for this line of work is very straightforward (note: this is our version of the argument, not theirs):

1. Virtually every school in the country still adopts textbooks in the core subjects (whether traditional or digital), and large majorities of teachers use them as part of day-to-day instruction.
2. Existing research shows textbooks differ in their effects on student achievement in meaningful ways, with well-designed studies finding effects in the .1 to .2 standard deviation range.[ii]
3. Textbooks are not an expensive intervention. They are also typically priced very similarly, so the marginal cost of choosing one textbook over another is small. Thus, the potential return on investment for adopting more effective textbooks is quite large relative to other kinds of interventions.[iii]
4. Curriculum materials are a much more politically palatable reform than many other potential policy reforms such as teacher policies and school choice policies.

Chingos and Whitehurst recommend that states work to build out a data infrastructure to track textbook adoptions by schools and districts, which can be used to tease out the impacts of various materials on student outcomes, among other things.

While their argument is persuasive, there has been little progress. At the time of their writing, only two states were known to track textbook adoptions, Florida and Indiana. Data from both states have been used to study the efficacy of elementary mathematics curricula.[iv] While a few more states have either started collecting the data[v] or the data have been "discovered" in plain sight[vi] since, it remains the case today that most states do not track which curriculum materials are used by which schools and districts. Publishers are also loathe to provide these data (though we can only assume they know which districts buy their materials) because they could be used to perform comparative efficacy studies of the sort that we report on here.

Responding to Whitehurst and Chingos' call for more research on textbook efficacy, we have recently undertaken a project[vii] using data from California to investigate the impact of textbooks on student achievement. We stumbled across the California textbook data by accident. Due to a 2004 court settlement and subsequent legislation, it is a requirement that the textbook of record in each of the core subjects be reported by individual schools. The data are kept in School Accountability Report Cards (SARCs) as PDF files available online from the California Department of Education (CDE). We hope that this is the first of many studies using these newly-discovered data and other data that are in the process of being collected. The goal is to begin to assemble a larger set of evidence about textbook efficacy, drawing from multiple states and across time.

In this brief, we present evidence from the first of our textbook efficacy studies in California. Specifically, we present evidence on the effects of textbook adoption decisions made during the 2008 adoption cycle in elementary mathematics. This brief is based on a full working paper, which can be downloaded here.[viii] To preview our findings, we find that one of the four most commonly adopted books, Houghton Mifflin *California Math*, is more effective than the other three popular books used in the state. The effects persist across four years post-adoption and range from approximately .05 to .10 standard deviations of student achievement. A variety of falsification exercises convince us that these effects are causal.

In what follows, we describe the California textbook adoption data, present our analytical methods, and then summarize our results. We conclude with implications for policy and future research on textbook effects.

Data

Textbook data

The data we use on textbook adoption are collected as a result of a 2004 court case, *Eliezer Williams, et al., vs. State of California, et al.* In this court case, student plaintiffs from the San Francisco Unified School District sued the state, arguing that the state had failed to provide equal access to instructional materials, quality teachers, and safe and decent school facilities. The result of the case was 2007 legislation requiring that, among other things, schools annually report on the adequacy of their instructional materials in core subjects. In the lowest achieving 30 percent of schools, there are also inspections of textbook adequacy, but in other schools the data are merely for public accountability. Individual SARCs are the source of the textbook data for our analysis.

While the SARC textbook data are available for virtually all schools in the state,[ix] there are a number of problems with the data that made pulling together a

textbook panel quite challenging. The first is that the SARCs are only available in PDF form for more than 80 percent of the schools in the state (while there is a standard template, this is only used in approximately 20 percent of schools; thus, schools in different districts have different SARC formats with textbook information on different pages). The result is that the data must be gathered manually. Thus, over the past two years, we have paid undergraduate and master's students at USC to pull together the SARC data from all ~7600 schools in the state that serve grades K-8. This is obviously very time and labor intensive.

The second challenge is that there is no uniform requirement for how textbook titles are reported. Sometimes just the publisher is reported, sometimes the title, and sometimes the title and version. Occasionally titles are reported that, to the best of our knowledge, do not exist. California is a state-adoption state, meaning the state puts out a list of "approved" textbooks that it encourages schools and districts to buy. If we assume that when a school lists a publisher on the state list and an adoption year corresponding to the most recent state adoption that this means they adopted the book from the state list, then we can code the vast majority of textbooks.[x] Even with this assumption, however, around 14 percent of schools have one or more listed textbooks that can't be identified, comprising about 9 percent of all listed textbooks.

Even among books that we are confident have been correctly identified, there are often dozens (sometimes hundreds) of ways that a particular book can be listed in the data. For example, for Pearson Scott Foresman Addison Wesley *enVisionMATH California* (the most popular book in the state in 2012-13), we counted 142 unique ways that that book was listed in the SARCs. Very few of these titles actually provide enough information to know for sure which book was used without substantial knowledge of California textbook adoption policies and the textbook market. An important takeaway from the data collection phase of our project is that even in California, where state law mandates curriculum materials reporting, curriculum-materials data are far from ready for analysis and there appears to be little oversight of the data. This calls into question whether the textbook reporting can really be a source of public accountability as intended.

After a substantial data-cleaning effort, when all is said and done we identified approximately 240 unique textbooks in use in California across the approximately 7600 schools serving grades K-8 as of 2012-13. These books were mostly adopted in the years immediately surrounding the official state adoption, with the large majority entering use in fall of 2008 or 2009.[xi]

We construct our analytic sample from this list. In order to be included in our analysis, several criteria must be met: we must have school characteristics and outcome data available from the school before and after the adoption, the textbook must be identifiable from the SARC, the adoption must be on-cycle (2008 or 2009), the school must be a uniform adopter in the elementary grades under study (1-3 for the main analysis, 1-5 for the analysis up through fifth grade), and the school must not be in a district that is so large that finding appropriate comparison schools is impossible.[xii]

After we limit our sample using these criteria, our final cut is based on the prevalence of the different textbook options in California schools—only four textbooks were adopted in a sufficient number of schools such that we would have adequate statistical power to evaluate them: *enVisionMATH 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. In the end, our analytic sample includes 1878 schools that used one of these four books.

Student achievement data

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 California Department of Education (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. We focus most of the evaluation on grade 3 achievement, but we also extend our analysis to examine curriculum effects on test scores in grades 4 and 5.

Methods

To analyze our data, we apply three related analytic approaches. Specifically, we estimate curriculum effects using kernel matching, common-support-restricted ordinary least squares (restricted OLS),

and remnant-based residualized matching. The three methods are described in detail in the working paper.[xiii] In all cases, we match at the school level, though we use both school and district variables in the matching process. We match at the school level both because this improves our statistical power and because some districts are not uniform adopters (that is, different schools in the district adopt different books, and matching at the school level allows us to use rather than discard these cases).

The idea of all three approaches is to match schools based on their pre-adoption variables (e.g., size, student characteristics, prior achievement) and follow their achievement trajectories several years past adoption. In the paper, we show the results of all the typical balancing tests, which provide convincing evidence that our matching approach works well.

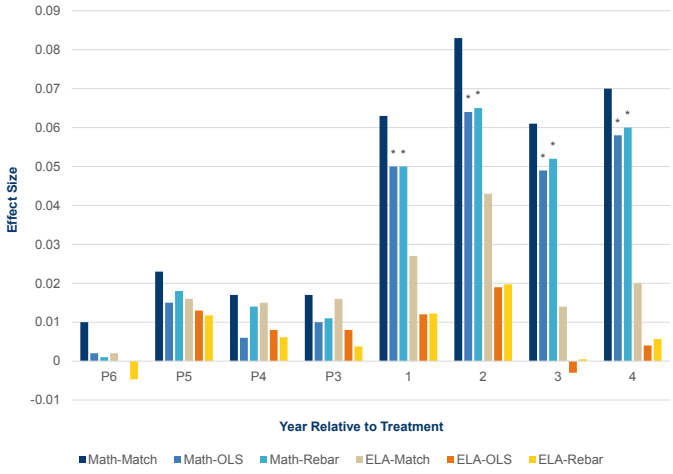
The key assumption for identification underlying all of our estimates is the conditional independence assumption (CIA), which assumes that there are no variables omitted from our matching process that affect both the choice of textbook and the outcome (student achievement). There are several conceptual reasons that the CIA is plausible in our application, which we discuss in detail in our paper. More formally, we also provide results from falsification tests that are designed to detect violations to the CIA. The falsification tests 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 the CIA is violated. In the falsification models, we estimate "curriculum effects" on test scores prior to the year of adoption of the focal curriculum materials, and on test scores in English Language Arts (ELA) instead of mathematics. If selection were driving our findings, we would expect to see time-inconsistent and/or off-subject "effects." We show these results below and in the paper; they provide no evidence to suggest that our findings are biased by unobserved selection.

Results

Our original analytic plan was to conduct pairwise comparisons among our four books of interest (i.e., six pairwise tests). However, we had to change course due to two problems. First, covariate balance is poor in some of the pairwise comparisons. Second, our statistical power is limited in the pairwise comparisons. That said, the pairwise comparisons (presented in the paper) point toward Houghton Mifflin *California Math* being more effective than the other three books, and that the other three books have similar levels of effectiveness. Based on the pattern of estimates from the pairwise comparisons, we restructured the analysis to present a comparison of Houghton Mifflin *California Math* with a composite of the other three books. This improves our statistical power and simplifies presentation (although again, we report findings from the pairwise comparisons in the paper for completeness).

The results of our main impact analysis on third grade math test scores are shown below in Figure 1 (the right four sets of bars labeled 1-4 on the horizontal axis, and the left three bars in each six-bar set). The results show that there is a positive impact of Houghton Mifflin *California Math* as compared to the other three books. That effect appears in the first year after adoption and persists through year four. The effect is in the range of .05 to .08 standard deviations, depending on the year and model. While the kernel matching estimates are similar in magnitude to the restricted OLS and remnant-based residualized matching estimates, the latter two are statistically significant because they gain statistical power from their additional assumptions (in the paper we discuss why we believe those assumptions are reasonable).

Figure 1. Effects of California math relative to the composite alternative on third grade test scores, over time and using different estimators



Notes: This figure corresponds to Figure 1 in our academic working paper, available at https://economics.missouri.edu/paper/wp-16-12. The vertical axis measures the effect of California Math relative to the composite alternative in student standard deviation units. Each set of bars along the horizontal axis is for a different year. The labels for years that pre-date the curriculum materials we study begin with a

*"P"; e.g., the year P3 denotes the year 3 years prior to the adoption of the materials we study (years P1 and P2 are omitted because we use data from these years, including achievement data, to match schools). The treatment years are years 1, 2, 3, and 4. The first three bars for years 1, 2, 3, and 4 are for the primary treatment effect estimates. All other bars are falsification estimates of various sorts as described in the text. An * indicates the effect estimate is statistically significant at the 5 percent level.*

These effects may seem modest in magnitude, but recall that they are for all students in a school. A policy targeted at the lowest 10 percent of students would have to have an effect ten times as large to generate the same total change in achievement. This effect size is equivalent to approximately a third to a half of a standard deviation in the distribution of teacher effectiveness; thus, it is equivalent to what would be a very large change in the average effectiveness of teachers in the workforce.

We might expect that the effects of the curriculum materials would get larger in later years as students have been exposed to the materials in more grades (i.e., higher dosage). However, we do not observe this pattern in the estimates. Of course, it is possible that there actually is a dosage effect, but we do not have the statistical power to observe it. It is also possible that is only (or primarily) the current grade's textbook that matters for student achievement on state tests. Thus, whether a student has been exposed to the same or different textbook in previous years may be of limited importance.

Our results from our analysis of math scores in the fourth and fifth grades, available in the paper, show generally similar patterns, with some differences across grades. In fourth grade, the effect sizes are somewhat smaller (.02 to .06 standard deviations) and are only statistically significant in certain years. In fifth grade, the effect sizes range from .03 to .05 standard deviations in year one up to .10 standard deviations in year four, again statistically significant in each year. The fifth grade effects do increase from year to year (though the increases are not always statistically significant) in a way that does suggest a dosage effect. In none of the models across any of the three grades or four years is there a negative coefficient on Houghton Mifflin *California Math* as compared to the composite of the other three textbooks.

Figure 1 also shows our falsification tests (which we replicate for grades four and five in the full paper as well). The impacts on ELA post-adoption are shown in the right four sets of bars and the right half of each set. Whether these estimates should be zero or not is unclear *ex ante* because there could be curriculum spillover effects, but at most they should be smaller than the math effects. Pre-adoption "effect" estimates are reported for both subjects in the first four sets of bars in the figure. The pre-adoption estimates should all be zero in the absence of selection bias because the treatments we study have yet to be implemented. All of the falsification estimates are substantively small and statistically insignificant, which is in line with expectations if the CIA is satisfied.

Discussion

Our work makes several important contributions. First, we have assembled a dataset of textbook adoptions in California, the largest U.S. state with the greatest number of schools. We have received funding to continue collecting these data moving forward. We will continue to analyze the data and go on to study other subjects and other grades. We also plan to make the data available to interested researchers so that others can pursue new lines of inquiry. There are many questions in this area of great import that do not have to do with impacts on student achievement—for instance, is there equitable access to current curriculum materials? How do charter and traditional public schools differ in their adoption patterns? We hope these newly available data can spawn a new wave of data-driven research on textbook adoptions and their effects. The current research literature is sorely lacking in quantitative analyses of textbooks in schools.

Second, our work again demonstrates a method (previously demonstrated by Bhatt, Koedel, and Lehmann[xiv]) that can be applied in other states, grades, and subjects. We believe at this point that the method is sufficiently well developed that it can be widely applied. By doing this—studying textbook effects across multiple settings—we can begin to develop a better understanding of what is working, where, and for whom. In addition to California, we have collected data on textbook adoptions in Texas, Illinois, New York, and Florida. Whether the data we have are sufficiently complete to allow this kind of investigation in each setting is unclear, but we will try.

One of the most common questions we get when we present this work is, "What is it about *California Math* that makes it effective?" As of now, there is simply no way to know the answer to this question,

and we therefore choose not to speculate. There are undoubtedly myriad ways in which *California Math* differs from the other textbooks under study, and it is impossible to know which of those dimensions matter, if any. It may be that *California Math* contains better pedagogical techniques for teaching core content, that it is better aligned to state standards or the state test, or that it is simply easier for teachers to implement with fidelity. If we had more studies of textbook effects, however, and if those studies were paired with analyses of the materials themselves, it is conceivable that we could begin to better understand which materials are working *and why.*

Third, our work extends the previous literature in several ways that we think are important. This analysis follows the achievement impacts for four years post-adoption, which is longer than any previous study. We also track achievement effects across three grades, again more than any previous study. We hope future studies can continue to look for these longer-term impacts.

Recommendations for district decisionmakers

School districts all around the country make decisions every few years about the textbooks they are going to adopt in the core academic subjects. These decisions are challenging with such little efficacy information at hand. When efficacy information does become available, of course we would recommend that district decisionmakers use it. This need not mean simply adopting the most effective book as judged by our analysis, but it would mean taking impact on student achievement into account in a serious way in the decisionmaking process. Our interviews with district leaders in California suggest that they often take multiple factors into account, and there is no reason to believe that they would not be interested in impact data if available.

We also believe that district decisionmakers could play a more active role in making curriculum materials information available for research. One way to facilitate this would be to put pressure on the state to devote more resources to data collection and analysis. Perhaps it would be possible for professional organizations to lead the way in this area (either district administrator organizations or subject-matter organizations like the National Council of Teachers of Mathematics) by collecting data from multiple districts; or, when circumstances are favorable, even helping to organize experiments to help us learn about curricular efficacy. There are many possible paths forward. At the end of the day, it comes down to the question of who is responsible for ensuring that students in our schools use effective curriculum materials. Whoever these individuals are, they should be alarmed by how little we know about the efficacy of various sets of materials, and all the more so given the handful of studies—including the one we describe here—showing great potential for more effective curriculum materials to improve student achievement.

Recommendations for states

While the findings of this analysis are, we believe, interesting and important, Houghton Mifflin *California Math* is no longer sold in California (or anywhere else, to our knowledge). Of the four textbooks under study, we believe only *enVisionMATH* is still sold, albeit in a "Common Core-aligned" version.[xv] Thus, the findings about this particular book are not as useful as they would be if the four books were still in wide use.

Future work must be timelier in order to influence policy. While we will always be limited in our ability to conduct more timely analyses because research of this nature is inherently backward looking (i.e., we need time to observe outcomes after treatment), there are several policy actions that can improve turnaround time. For one, SARC release dates could be pushed up—currently they come out with a one-year delay (e.g., the 2015-16 SARCs are posted from spring into summer of 2017). Another issue is that the work to gather and clean the California data takes multiple months per subject and grade span, even working at full capacity. It is our hope that California (and other states) make this work easier by creating and making available data systems that contain accurate textbook information in a timelier fashion. If they did, this kind of analysis could be done much faster, which would make it more useful in informing policy decisions on the ground.

The California data are of poor quality, and even the Texas data (which we do not discuss here but are much better) leave something to be desired. We believe it would be relatively straightforward to include textbook data collection as part of states' regular data collection activities from schools and districts. Specifically, with an ISBN number and an adoption year for each textbook used in each subject area and grade level, it would be straightforward to put together a panel of textbook adoptions statewide.

Such a dataset would have many uses for research, but it could also be used in policy and practice to pair schools or districts adopting the same curricula for purposes of professional development or sharing resources. This type of data collection would not be cost-free, but it is quite unobtrusive and given the significance of textbooks in the day-to-day operations of schools, we think almost certainly worth the effort and cost.

---

[i] Chingos, M.M., & Whitehurst, G.J. (2012). *Choosing blindly: Instructional materials, teacher effectiveness, and the Common Core*. Washington, DC: Brookings.

[ii] Agodini, R., Harris, B., Atkins-Burnett, S., Heaviside, S., & Novak, T. (2010). *Achievement effects of four early elementary school math curricula: Findings for first and second graders*. Washington, DC: National Center for Education Evaluation and Regional Assistance, U.S. Department of Education, Institute of Education Sciences.
Bhatt, R., & Koedel, C. (2012). Large-scale evaluations of curricular effectiveness: The case of elementary mathematics in Indiana. *Educational Evaluation and Policy Analysis*, 34(4), 391-412.
Bhatt, R., Koedel, C., & Lehmann, D. (2013). Is curriculum quality uniform? Evidence from Florida. *Economics of Education Review*, 34(1), 107-121.

[iii] Boser, U., Chingos, M., & Straus, C. (2015). *The hidden value of curriculum reform: Do states and districts receive the most bang for their curriculum buck?* Washington, DC: Center for American Progress.

[iv] See Bhatt & Koedel, 2012; Bhatt, Koedel, & Lehmann, 2013.

[v] To my knowledge, New Mexico and Louisiana have begun collecting the data.

[vi] California and Texas have been collecting the data for some time, as it turns out. The California data are described in our paper. In Texas the state collects all textbook purchase data at the state level, and the data can be obtained from the state website or via Freedom of Information request (for historical data).

[vii] This material 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 material are those of the author(s) and do not necessarily reflect the views of the funders.

[viii] https://economics.missouri.edu/paper/wp-16-12

[ix] There is a small number of schools that don't seem to provide SARCs at all, and another small number that provides SARCs but no textbook information. Together these comprise a little less than 7 percent of schools in 2014-15.

[x] This assumption is appropriate because the vast majority of California schools, especially pre-Common Core, adopted off the state approved list (until recently, categorical textbook funds could only be used for books on the state list).

[xi] We limit all analyses to schools that adopted books on-cycle (meaning these two years).

[xii] This eliminates Los Angeles and Long Beach Unified School Districts, which are much larger than other districts in the analytic sample.

[xiii] https://economics.missouri.edu/paper/wp-16-12

[xiv] See Bhatt & Koedel, 2012; Bhatt, Koedel, & Lehmann, 2013.

[xv] Other work suggests that the Common Core-aligned version of *enVisionMATH* is not actually particularly well aligned to the Common Core. Polikoff, M. S. (2015). How well aligned are textbooks to the Common Core Standards in mathematics? *American Educational Research Journal*, 52(6), 1185-1211.