10-2-2009

A Non-Experimental Evaluation of Curricular Effectiveness in Math

Rachana Bhatt
Georgia State University, rbhatt@gsu.edu

Cory Koedel
University of Missouri

W.J. Usery Workplace Research Group Paper Series

Working Paper 2009-10-2
October 2009

A Non-Experimental Evaluation of Curriculum Effectiveness in Math

Rachana Bhatt
Georgia State University

Cory Koedel
University of Missouri

Georgia State University | Andrew Young School of Policy Studies

A Non-Experimental Evaluation of Curricular Effectiveness in Math

Rachana Bhatt
Georgia State University

Cory Koedel*
University of Missouri

October 2009

This paper uses non-experimental data to evaluate curricular effectiveness. We show that non-experimental methods can be used to obtain causal estimates of curricular effects at just a fraction of what it would cost to produce analogous experimental estimates. Furthermore, external validity concerns that are particularly cogent in the context of curricular evaluations suggest that a non-experimental approach may be preferred. Our results provide important insights for educational administrators and policymakers. In the short term, we find large differences in effectiveness across some math curricula. However, like many educational inputs, the effects of math curricula do not persist over time, a result that would be quite costly to attain using experimental data. Across curricula adoption cycles, publishers that produce less effective curricula in one cycle do not lose market share in the next cycle. One explanation for this result is the dearth of information available to administrators about curricular effectiveness.

* We thank Emek Basker, Julie Cullen, Barry Hirsch, Josh Kinsler, David Mandy, Peter Mueser, Rusty Tchernis and seminar participants at Georgia State University, the University of Missouri, and the University of Rochester for useful comments and suggestions. We also thank Karen Lane and Molly Chamberlin at the Indiana Department of Education for help with data. This work was not funded or influenced by any outside entity.

I. Introduction

Curricular effectiveness has received much attention in the education literature, and justifiably so (see, for example, Slavin and Lake, 2008; National Research Board, 2004). The majority of instructional time and homework assignments are textbook oriented, and a substantial amount of school expenditures are devoted to curricula purchases. According to a 2002 survey sponsored by the National Education Association and the American Association of Publishers, 80% of teachers use textbooks in the classroom, and over half of students' in-class instructional time involves textbook use (Finn, 2004).[1] In 2006 alone, expenditures on K-12 instructional materials totaled close to $8.1 billion dollars.[2] Different curricula are developed using different theories for how students learn - this results in different content, organization and structure across curricula for the same subject and grade group. Given the central role that curricula play for students and schools, it is of interest to determine the extent to which different curricula differentially affect student achievement.

Although there have been many hundreds of studies evaluating the curricular alternatives facing school administrators, there are concerns about the reliability of the findings in the existing literature. For example, the What Works Clearinghouse (WWC), which was established in 2002 by the Institute for Education Sciences (IES) to serve as a filter for education research, evaluated over 200 studies of curricular effectiveness in elementary mathematics in 2007 and found that over 96 percent of these studies did not meet reasonable quality standards (WWC, 2007).[3] Likely in response to the dearth of reliable evidence in the literature, recent research has

[1] Textbooks are just one component of the curricula purchased by schools from publishers. Other aspects include teacher instructional support services and supplementary materials such as student workbooks, flashcards, and solution manuals.

[2] See http://www.aapschool.org/vp_funding.html

[3] The WWC reviews the literature on a variety of topics in education, including the effects of curricula adoptions, and classifies studies as either (1) meets evidence standards, (2) meets evidence standards with reservations or (3) does not meet evidence standards. Generally speaking, studies in category (1) use randomized controlled trials (RCTs) or quasi-experiments (e.g., regression discontinuity designs). Studies in category (2) may employ non-experimental techniques, but must be deemed by the principal investigator at WWC to have employed appropriate statistical tools such that causal inference is reasonable. Of the 237 studies on elementary math curricula reviewed

turned to randomized controlled trials (RCTs) to evaluate curricular effectiveness (see, for example, Agodini et al., 2009; Borman et al., 2008; Resendez and Azin, 2007). RCTs randomly assign curricula across schools (and/or classrooms) and will produce causal estimates of curricular effects that are internally valid – that is, valid within the context of the experiment. However, a general drawback of RCTs that is particularly cogent in the case of curricular evaluation is that the estimates may not extrapolate well outside of the experimental setting.

We highlight two concerns with RCTs in the context of curricular evaluation that will potentially limit the external validity of the results.[4] First, RCTs require voluntary participation by *schools and curricula publishers.* If the schools that select into the experiment differ from the general population of schools, then Manski's (1996) "experimentation on a subpopulation" concern is relevant, and the experimental results will not necessarily reveal anything about curricular effectiveness at schools not represented in the study. Equally importantly, there is also a selection problem with respect to publishers. With voluntary publisher participation, only publishers that expect their curricula to be successful in the setting of the RCT will agree to participate. Overall, the requirements of voluntary school and publisher participation limit the extent to which experimental designs can be used to evaluate the full curricular landscape.

A second threat to the external validity of RCTs is publisher responsiveness to evaluation, commonly referred to as Hawthorne effects. In the general evaluation literature, Hawthorne effects refer to the subjects of the experiment. In the case of curricular evaluation, the active role of publishers suggests that in addition to schools and students, they *are* subjects. Because curricula are generally packaged with teacher development and implementation services, publisher responses to evaluation along these margins can contaminate findings. Furthermore, given that experimental evaluations of curricula are high-stakes competitions for publishers, there is no reason to expect them to take a "business-as-usual" approach. Publisher

---

by the WWC as of July, 2007, just nine were deemed to be of sufficient quality by WWC to be included in categories (1) and (2) (WWC, 2007).

[4] See Heckman and Smith (1995) and Manski (1996) for general discussions about the strengths and weaknesses of experimental research designs.

Hawthorne effects raise questions about how well the results from RCTs will extrapolate to lower-stakes environments for publishers.

In addition to these threats to external validity, the costs associated with RCTs limit the amount of information that they can provide. For example, because RCTs are expensive to operate, they generally focus on just one or two curricula evaluated at small numbers of schools and districts.[5] The expenses associated with RCTs also limit their usefulness in evaluating long-term impacts because it is costly to maintain the validity of the experiment over time.

As an alternative to experimental analyses, we contribute to the literature by using non-experimental data from the entire state of Indiana to evaluate curricular effectiveness. We rely on data from Indiana because Indiana provides the most detailed information about curricula adoptions over time of any of the 50 states, and also provides thorough school- and district-level data about achievement, student demographics and school finances. With the exception of the information about curricula adoptions, the data used for our study are available in other states, suggesting that it would be straightforward to replicate our analysis elsewhere.[6]

We use school-level matching estimators in our evaluation. Drawing on the extensive methodological literature on matching, we show that the data conditions in Indiana are generally favorable to such an approach. Furthermore, because our dataset is particularly long and detailed, we are able to perform a series of falsification tests to evaluate the potential for bias in our non-experimental estimates. Overwhelmingly, our falsification tests confirm that our estimated curricula effects are *not biased*. In addition to producing causal estimates of curricula effects that are likely to extrapolate much more broadly than experimental estimates, our non-experimental analysis is performed at just a fraction of what an analogous experimental study would cost.

[5] In what is a relatively large-scale RCT, Agodini et al. (2009) evaluate four different curricula (more than the usual one or two curricula in other studies), but still only evaluate four school districts and 39 schools (in the first wave of their study). More typical RCTs are even more narrowly focused. Borman et al. (2008) and Resendez and Azin (2007) each evaluate just a single curriculum, looking across only five and four schools, respectively.

[6] It would not be expensive for states to track curricula adoptions, particularly when compared to the costs of tracking some of the other information that is commonly collected.

We highlight three primary findings from our analysis: (1) differences across some math curricula have large short-term effects on student achievement, (2) as has been found with other educational inputs (e.g., Jacob et al., 2008; Garces et al., 2002), math-curricula effects do not persist over time, and (3) curricula publishers that are relatively less effective in one adoption cycle do not lose market share in future adoption cycles. This latter result shares a common theme with prior research suggesting that educational administrators do not make optimal choices (Ballou, 1996). In this case, one explanation is the limited availability of reliable evidence on curricular effectiveness.

II. The Curricula Selection Process

We evaluate math curricular effectiveness in the state of Indiana. Curricula are adopted in Indiana for one subject in each year across the entire state, and rotate in six-year cycles by subject. For example, Indiana's districts adopted new math curricula in 1998 and 2004, with an upcoming adoption in 2010. Similarly, recent reading adoptions occurred in 1994, 2000 and 2006. We focus our attention primarily on the math adoption cycle from 1998 to 2004.

The curricula selection process in Indiana has centralized and decentralized components. The process begins with the state of Indiana's Department of Education (DOE) approving a list of selected curricula for use in the state. Upon receiving this list from the DOE, districts have three choices. First, and most commonly, they can adopt one or more of the state-approved curricula. At the elementary level in particular, the overwhelming majority of districts choose only a single curriculum for each grade, although there are no restrictions requiring them to do so.[7] Second, districts may choose to apply for alternate curricula that are not on the state-approved list, but this option is almost never used (e.g., no more than one out of the roughly 300 districts chooses this option during the adoption cycle that we study). Third, districts can apply for "continued use" where they continue to use the curricula that were adopted in the prior

[7] Further, most districts choose a single curriculum for *all elementary grades*, although again, there is nothing to preclude a district for choosing one curriculum for, say, grades one and two, and another for grades three, four and five.

adoption cycle in that subject. Overall, over 98 percent of the districts adopted new math curricula from the approved list during the 1998 adoption cycle in each grade.

We treat the DOE's approval process as exogenous to the districts, and focus our analysis on identifying differential curricular effects among the curricula that are included on the DOE's approved list. The centralized approval process adds a constraint to the environment whereby we cannot (feasibly) evaluate curricula that are not approved by the state. However, it is not clear that the DOE's constraint is binding for districts in any meaningful way. For example, although districts can apply to use curricula outside the state-approved list, this rarely happens in practice, suggesting that most districts are content to choose among the available options. Perhaps more telling, the majority of the curricula market share belongs to just a handful of publishers. Specifically, 86 percent of all curricula adoptions in the grades that we study involve just three of the ten state-approved curricula during the adoption cycle of interest.[8]

III. Data

We use a 17-year data panel of schools in Indiana to evaluate the effects of math curricula adoptions in grades one, two and three on grade-3 test scores in math (grade-3 is the first time that students are tested in Indiana). Among the 50 states, Indiana is the only state where curricula-adoption information is available at the district level for multiple statewide adoption cycles.[9] Upon request, Indiana also provides detailed school-level information on test scores (from the Indiana state test, the ISTEP), attendance rates and enrollment demographics (including language minorities and students on free and reduced price lunch – some of these data are readily available online). Indiana also collects district-level data for these same variables,

[8] Indiana is one of 22 states that have a state-level component to the adoption process. Tulley (1989) finds that in states where there is not a centralized component to the adoption process, the curriculum review processes and lengths of use are similar despite the lack of a formal process dictating textbook choice. In conjunction with the limited practical importance of the centralized constraint, this suggests that the centralized component to Indiana's curricula adoptions should not affect the generalizability of the results.

[9] In fact, in many states, the state department of education does not even have a readily available centralized database indicating which curricula are adopted by districts within the state during the *current* adoption cycle, let alone historical information.

along with financial information. Details on the district- and school-level information used in our analysis are provided in Table 1.

We estimate curricula effects for the three curricula that dominated the market during the adoption cycle of interest (1998-2004). These curricula were published by Saxon, Silver-Burdett Ginn and Scott-Foresman, and they accounted for 48, 23 and 15 percent of observed curricula adoptions, respectively, or 86 percent of adoptions overall. In the analysis below, we denote the Saxon curriculum as curriculum *A*, the Silver-Burdett Ginn curriculum as curriculum *B*, and the Scott-Foresman curriculum as curriculum *C*.

Because we first observe student outcomes in grade three, our estimates of curricular effects characterize the impacts of *sequences of treatments*. That is, grade-three test scores are presumably a function of the curricula to which students are exposed in grades one, two and three. To allow for cleanly identified curricula effects, we focus our analysis on districts that we refer to as "uniform curriculum adopters". These districts choose the same curriculum publisher in grades one, two and three in the relevant adoption cycle. To illustrate the assignment problem for non-uniform adopters, consider a district that adopted curriculum *A* in grade one and curriculum *B* in grades two and three. In identifying the effect of curriculum *A* relative to curriculum *B*, the schools in this district are not well-defined as either treatments or controls.

Similarly, we also exclude districts that adopted multiple curricula in any given grade because the data do not indicate which schools within each district used which curricula. Only in cases where a district used a single curriculum at all schools can we be sure that our treatment and comparison schools are properly identified.

Imposing the uniform-curriculum-adoption restriction ultimately reduces our district sample size by approximately eight percent and the analogous school sample size by seven percent. That is, most districts are "uniform adopters". Overall, our analysis includes data from 213 districts and 716 schools. Contrasted with the experimental literature, where studies often focus on just a handful of schools and districts, our non-experimental analysis allows for a much

broader evaluation of curricular effectiveness. Details on how we arrived at our final data sample are provided in Appendix Table A.1.

Table 1 reports differences in means across the schools and districts that adopted the different curricula, using pre-adoption information from 1997. There are only mild differences in test score performance and attendance outcomes across different curricula adopters, suggesting that selection bias may be mild. However, there are noticeable differences across adopters of the different curricula in terms of school demographics, district size and revenue, and to some extent, median household income (measured at the district-level from the US Census). Among other things, Table 1 indicates that Saxon adopters are disproportionately rural districts, as evidenced by their much smaller district sizes (and corresponding revenues) and their larger shares of white students.

It is unclear how selection into the different curricula might bias our estimates, which are conditioned on all the information detailed in Table 1. Clearly, any unobserved differences across observationally similar schools would generate bias, although we do not know what types of unobserved differences to expect given that our estimates depend on comparisons across large groups of students (i.e., schools).[10] One possibility is that differences in administrator quality may bias our results. For example, some administrators could choose better curricula and make other decisions that positively affect achievement. However, the complex curriculum adoption process, which involves many actors, likely limits any such bias. Although we cannot rule out the potential for bias in our estimates *a priori*, the falsification tests in Section X confirm that our results are not biased by observed or unobserved differences across curricula adopters.

IV. Curricula Descriptions

Are differences across math curricula important? Anecdotal evidence suggests that they are. For example, a 2002 story on charter schools in the Chicago Tribune reported that three elementary charter schools in Chicago were significantly outperforming local traditional schools,

[10] In results omitted for brevity we verify that students do not move across districts in response to curricula decisions.

and that school officials "suggest(ed) it may have to do with the Saxon Math program used at all of (these charter schools)." Parents also have strong opinions about curricula. On the Illinois Loop website (the Illinois Loop is an advocacy group of parents and teachers), parents can post comments about curricula. In reference to one math curriculum a parent wrote:

> We dealt with (this curriculum's) nightly visual assault of colors, graphics, fonts, and wildly irrelevant detail…(this curriculum) played a significant role in a terrible second grade experience.

Another parent was clearly upset about the lack of difficulty in the same curriculum:

> On page 2, as an intro to their fourth grade, students are asked to solve problems like 5+7 and 13-8, and are told, "You may use the [conveniently printed] number line to help." This is the fourth grade, yet kids are being told it's OK to count on a number line to solve simple problems!

In 1998, Mathematically Correct (MC), a national organization of mathematicians, scientists and engineers, qualitatively evaluated eight grade-2 math curricula, including the three curricula that we evaluate here. The MC evaluations were sponsored by the Texas Public Policy Foundation, a non-profit, non-partisan research institute. We briefly highlight the key differences between the Saxon, Silver-Burdett Ginn and Scott-Foresman curricula as indicated by MC. We also report the MC rating of each curriculum, which was based on a 5-point scale (all three curricula received a similar rating from MC).

Curriculum A: Saxon Math (overall rating: 3.6)

The program design is "easily implemented by teachers", and instructions to teachers are "clear and direct". In fact, the teacher's manual even includes scripted statements and questions for the teacher to ask to the class. The worksheets that students use are not necessarily related to the daily lesson, and contain a mixture of topics from prior lessons. One side of the worksheet is completed in class and checked, and the other side is assigned for homework. Oral assessments are given to individual students every 10 lessons, and are conducted while other students work on written work. Written assessments occur after every five lessons.

Saxon Math is very thorough in the topics that are covered, but more advanced topics are generally not covered. That is, this program supports learning effectively to a certain level but beyond that, achievement will be "very limited". As one example, of the three curricula of interest here, Saxon math is the only curriculum that does not cover addition and subtraction with three-digit numbers in the second grade. Overall, the MC evaluation suggests that Saxon Math may be the most effective curriculum for low-achieving students given its thorough coverage of the topics it covers, but will be less effective for high-achieving students.

Curriculum B: Silver-Burdett Ginn Math (overall rating: 3.4)

The teacher's manual provides guidance to teachers, although the guidance is not as direct as in Saxon Math. The teacher is given some discretion over how to present the material. In the example from the MC review, the teacher has two presentation choices for the lesson that are described as "visual/spatial learning" and one presentation choice that is described as "kinesthetic learning". In some cases, there is also a technology-based alternative. Student worksheets are tied to the daily lesson. No information is given about the regularity of assessments or homework assignments.

The MC review highlights that this curriculum relies heavily on "models" as teaching tools. Models provide an alternative way of teaching mathematics, using graphics to aid in the calculations. In teaching addition and subtraction, this curriculum relies heavily on models first, then models give way to pictures, and finally models become optional in later chapters. MC identifies the reliance of this curriculum on models as a weakness.

The level of this curriculum appears to be higher than that of Saxon Math – MC reports that students using this program have a "reasonable chance of moderate achievement levels" but also that the program is "not seen as supporting high achievement levels".

Curriculum C: Scott-Foresman Addison Wesley Math (overall rating: 3.8)

The teacher's edition receives mixed reviews within the MC evaluation. At one point, the evaluation indicates that it provides slightly more support to teachers than some of the other programs. At another, it indicates that the teacher's manual is "a bit thin in terms of aiding a teacher in actually teaching the lesson at hand". Like the Silver-Burdett Ginn curriculum, the lessons also involve some discretion for teachers in terms of the activities that they use to teach each lesson (although there appear to be fewer teacher choices). Vocabulary development is an important part of this curriculum – new vocabulary words are introduced at the beginning of each lesson, and a verbal skills assessment occurs after each lesson. A one page homework sheet is also attached to each lesson.

The level of this program appears to be somewhere in between the levels in the prior two curricula. On the one hand, the MC review indicates that "the level is low in a few topics" and "at the top level of students…some topics should be augmented". On the other hand, the review also notes that "some areas are very well taught and at an excellent level".

It is important to note that while the MC reviews provide useful insights, they are not based on empirical evidence. We present the descriptions simply to highlight the differences that can exist in organization, content, and presentation across math curricula. These differences have received considerable attention from parents, educators and other interested parties. Although qualitative curricular evaluation is outside of our area of expertise, we expect our quantitative analysis to provide useful insights to individuals who are interested in identifying the key components of effective math curricula.[11]

V. Methodology

We use school-level matching estimators to identify curricula effects. Matching is an increasingly common technique employed in empirical work, and the conditions under which

[11] Despite their limitations, the MC evaluations are the only independent reviews of the curricula of interest that we were able to obtain.

matching will identify causal estimates of treatment effects have been well-documented (see, for example, Rosenbaum and Rubin, 1983; Heckman, Ichimura and Todd, 1997). The key benefits of matching relative to simple regression analysis are (1) matching imposes weaker functional form restrictions and (2) matching resolves any "extrapolation" problems that may arise in regression analysis by limiting the influence of non-comparable treatment and control units in the data (Black and Smith, 2004).

Briefly, the key assumption under which matching will return causal estimates of treatment effects is the conditional independence assumption (CIA). The CIA requires that outcomes are independent of the curriculum uptake decision conditional on observable information. Defining Y as an outcome, T as the curriculum treatment and X as a vector of observable school- and district-level information, the CIA in our multi-treatment context can be written as:

$$Y \perp T \mid X \tag{1}$$

The CIA is actually a stronger assumption than is required to identify causal treatment effects, although it is difficult to imagine an environment where only the weaker but necessary condition of conditional *mean* independence is satisfied (Heckman et al., 1997, Imbens, 2003).

The most common source of failure of conditional independence in non-experimental settings is the existence of unobserved information that influences both treatment and outcomes. For example, if districts have access to information that is unobserved to the econometrician, Z, such that $\mathrm{P}(T = k \mid X, Z) \neq \mathrm{P}(T = k \mid X)$, and the additional information in Z also determines outcomes, matching will produce biased estimates of curricular effects.

The CIA is plausible here because curricula adoptions are determined on behalf of large groups of students (by district) for which we have detailed achievement and demographic information. Further, there is no evidence that students move across districts in response to curricula adoptions, suggesting that unobserved characteristics of students that might affect achievement are not likely to be correlated with curriculum exposure (these results are omitted

for brevity but available upon request). A common limitation of matching analyses is that even when the CIA is expected to hold, the data are generally insufficient to verify this to be the case. However, beyond simply asserting conditional independence, we use our data panel to provide *evidence* that conditional independence is satisfied in the form of a series of falsification estimates. That is, we estimate curricula effects for multiple cohorts of students who should not be affected by the curricula that we study. Our falsification tests overwhelmingly show that unobservables are not biasing our results.

Given conditional independence, we use matching to estimate causal curricula effects. To illustrate, consider a world with two possible curricula – *j* and *m*. In the data we observe only a simple difference in means, which estimates:

$$E(Y_j \mid T = j, X) - E(Y_m \mid T = m, X) \tag{2}$$

Under the assumption of conditional independence we can write:

$$E(Y_m \mid T = m, X) = E(Y_m \mid T = j, X) \tag{3}$$

Substituting into (2) with (3), we can use the data to estimate:

$$E(Y_j \mid T = j, X) - E(Y_m \mid T = j, X) \tag{4}$$

Equation (4) can be written as the average effect of treatment on the treated (ATTE), where treatment is defined as curriculum *j*:

$$ATTE_{j,m} : E(Y_j - Y_m \mid T = j, X) \tag{5}$$

Finally, treatment can be re-defined as curriculum *m*, and the same approach can be used to generate the corresponding $ATTE_{m,j}$, which need not equal $ATTE_{j,m}$:

$$ATTE_{m,j} : E(Y_m - Y_j \mid T = m, X) \tag{6}$$

We estimate treatment effects in our multi-treatment setting following this basic pairwise-comparison approach, suggested by Lechner (2002), where schools are matched using an estimated propensity score (Rosenbaum and Rubin, 1983). Defining P_j as the probability of choosing option *j*, we match schools in the comparison of curricula *j* and *m* by $(\frac{P_j}{P_j + P_m})$, where

P_j and P_m are estimated from a multinomial probit that simultaneously models all k heterogeneous treatment options (Lechner, 2002).

We use kernel and local linear regression (LLR) matching estimators.[12] These estimators construct the match for each treated school using a weighted average over multiple control schools. They take the general form (for $ATTE_{j,m}$):

$$\hat{\theta}_{j,m} = \frac{1}{N_1} \sum_{j \in I_1} [Y_j - \sum_{m \in I_{0j}} W(j,m) Y_m] \qquad (7)$$

In (7), N_1 is the number of treatments on the common support, I_1 indicates the set of these treated observations, I_{0j} the set of control observations in the neighborhood of observation j (determined by a bandwidth parameter – see Appendix B), Y_j and Y_m are outcomes for treated and control schools, respectively, and $W(j,m)$ weights each control school outcome depending on the distance between P_j and P_m. We omit a more detailed discussion of these matching estimators for brevity. For more information, see Heckman, Ichimura and Todd (1997, 1998), and Fan (1993).

We estimate relative *ATTE's* for the three math curricula in our data. As noted above, $ATTE_{m,j}$ need not equal $ATTE_{j,m}$. Nonetheless, in practice, we uncover little additional insight by estimating both. Therefore, we present treatment-effect estimates only in one direction, defining the most-adopted curriculum in each pairwise-comparison as the control curriculum (we briefly discuss our estimates from alternately defining treatment and control schools in Section IX). Letting N_x denote the number of schools adopting curriculum x, in the data $N_A > N_B > N_C$ (see Table 1). Therefore, we report estimates for $ATTE_{B,A}$, $ATTE_{C,A}$ and $ATTE_{C,B}$.

VI. Timing and Treatment Definition

Timing is an important issue in our analysis. Our data panel spans 17 years, starting with the 1991-1992 school year and ending with the 2007-2008 school year. The curricula of interest were adopted in the fall of 1998, and replaced with new curricula in the fall of 2004. We observe seven cohorts of grade-3 students who were never exposed to the curricula of interest

[12] Our results are robust to alternative estimators, see Section IX.

during the pre-period (1991-1992 through 1997-1998), one cohort that was exposed to the curricula in grade three only (1998-1999), one cohort that was exposed in grades two and three only (1999-2000), four cohorts that used the curricula in grades one, two and three and were thus "fully exposed" (2000-2001 through 2003-2004), one cohort that was exposed in grades one and two only (2004-2005), one cohort that was exposed in grade one only (2005-2006), and two additional cohorts that were never exposed to the curricula in the post-period (2006-2007 and 2007-2008).

Recall that the estimated curricula effects are based on grade-3 test scores, and as such represent the effects of sequences of treatments (T_{g1}, T_{g2}, T_{g3}). For the fully-exposed cohorts, the sequences for treatment and control schools are fully observed and as such these cohorts provide our cleanest estimates of curricular effectiveness. For the partially-exposed cohorts (the cohorts that were exposed to the curricula for at least one year, but less than three years), we can still estimate treatment effects because part of the curricula sequence is observed. For example, for the 1999-2000 cohort, we know with what curricula each school was treated in grade three. However, the full sequences of treatments are not observed for this cohort and before grade three, the treatment and control schools likely used heterogeneous curricula from the previous adoption cycle. A similar concern regarding out-of-cycle curricula adoptions is relevant for our falsification tests (using cohorts prior to 1998-1999, and after 2005-2006). This issue will be addressed in more detail in Sections IX and X when we present our results.

An additional concern related to timing in our study is that the exposure levels of the different cohorts overlap with "curricula experience" at schools. For example, the 1999-2000 cohort was exposed to the curricula for just one year, which was the year when the curricula were first introduced at districts, and perhaps most importantly, to teachers. Intertwined with the different levels of curricula exposure by cohort, therefore, are any effects of teachers' curricular familiarity. Only across the four fully exposed cohorts will any familiarity effects be separately distinguishable (e.g., if familiarity effects exacerbate differences in curricula, the latter two fully-exposed cohorts should experience larger differential curricula effects than the former two).

Finally, a third timing issue involves district restructuring over the course of the 17 years of our data panel. Specifically, there is a pattern of school consolidations in the data such that the number of individual elementary schools decreases over time. As will be discussed in the following section, we match schools based on their static characteristics from the 1996-1997 and 1997-1998 school years. School consolidations suggest that the populations of students served by the schools that remain in operation will change over time. This will reduce the quality of our matches, and potentially introduce bias into our estimates.

In order for the school consolidations to bias our estimates they must be correlated with curricula adoptions. However, this does not appear to be the case. Using a χ^2 test for independence, we fail to reject the null hypothesis that curricula adoptions are independent of whether a district experiences a school closing (p-value ≈ 0.40). As additional evidence that our results are unlikely to be biased by school consolidations, in the next section we evaluate the balance of the covariates across matched treatment and control schools over the entire course of the data panel. If the schools that drop out of our sample over time systematically adopted specific curricula, we should find that our treatment and control samples become less balanced as we move away from the matching years (1996-97 and 1997-98). However, this is not the case, which further supports our contention that school closings are not correlated with curricula adoptions (see Table 2).

Although we do not expect the school consolidations to bias our results, they will reduce the quality of our matches as we move away from the 1996-1997 and 1997-1998 school years in the data panel. This will add noise to our estimates. Ultimately, we simply report this issue as a caveat, and caution the reader to interpret results that are estimated far away from the matching years more liberally. In an omitted analysis, we also considered a more direct solution to this problem – at any point where a school closing was observed in a district, we dropped all school-level observations from that district for the remainder of the data panel.[13] This alternative

[13] We also performed an analogous procedure for schools that existed in 1996-1997, but came into existence between 1991-1992 and 1996-1997. If school closings re-shuffle student populations within districts, such an

approach produces estimates that are qualitatively similar to what we report below, although the efficiency costs associated with discarding data from entire districts may be higher than those from allowing the less accurate matches to occur.[14]

VII. Estimating the Propensity Score

We use a multinomial probit (MNP) specification to estimate the pairwise propensity scores. The covariates that we include in the MNP are documented in Table 1, and contain both school and district level information. At the school level, the propensity-score model includes controls for enrollment, demographics (race, free lunch, reduced lunch, language status) and outcomes (i.e., grade-three test scores in math and language, and attendance) from the 1996-1997 school year, and controls for enrollment and demographics from the 1997-1998 school year (for brevity, differences in means are not reported in the table for the 1998 information). At the district level, the model includes enrollment, outcome and finance controls from 1996-1997, and enrollment and finance controls from 1997-1998. We also use district-level zip codes to assign Census measures of local-area socioeconomic status to each school. Namely, we include controls in the model for median household income and the share of the adult population who do not have a high-school diploma, both obtained from the year-2000 census. We treat these census variables as fixed area characteristics.

The covariates in our MNP specification were selected based on the process by which the curricula were adopted, with the objective of replicating the relevant information set available to schools and districts at the time of the curricula-adoption decision (note that the curricula-adoption process in Indiana lasts approximately 18 months, and for the 1998 adoption this

---

approach will reduce the number of bad matches in the data. There is enough natural variation in the enrollment data that we cannot always identify which specific schools are affected by a school closing, particularly when the closing school is small. As such, the most straightforward solution is to drop all schools in the district where the school closing is observed.

[14] An additional problem with this alternative strategy is that when a school closes we cannot be certain that the only other schools that are affected are within the same district. For example, if the closing school is on the border of another district, its students may change districts, in which case the district-dropping procedure would be doubly harmful – it would retain the schools in the new district into which the students from the closed school were infused, and drop the schools from the district where the school closed, where the student populations at these schools were not actually affected by the closure.

process culminated with a final decision in the summer of 1998).[15] However, our findings are not qualitatively sensitive to reasonable adjustments to the MNP specification.[16]

In each comparison we match treatment and control schools based on the estimated pairwise propensity scores, and test for balance in the covariates among the treated and control samples used for estimation.[17] Balancing tests are motivated by Rosenbaum and Rubin (1983). The tests determine whether $X \perp T \mid P(T = k \mid X)$, a necessary condition if the propensity score is to be used to reduce the dimensionality of the matching problem to one. The results from our balancing tests are reported in Table 2 by comparison and year.

Our MNP specification uses 32 school and district-level covariates. As a rough summary statistic, we report average p-values from difference-in-means tests across the covariates for the schools that report test scores in each year. Additionally, we report the number of covariates for which we reject balance across treated and control units at the 10 percent level or better. As discussed in the previous section, we observe attrition from our sample of schools. If the attrition that we observe in the data is correlated with curricula adoptions, the balancing tests will highlight such a problem. The balancing properties of the covariates are roughly time invariant, suggesting this is not a concern.

For the most part, our treatment and control samples are balanced in our three comparisons. Only in our comparison between *C* and *A* do the balancing tests raise some concerns. In this comparison as many as three out of the thirty-two covariates are not balanced in any given year. When sufficiently large samples of treatment and control observations are available, researchers typically resolve imbalance by re-defining the propensity-score

[15] For example, we omit information about spring-1998 test scores and annual attendance rates because they were unlikely to be available to decision makers prior to the adoption decision for the fall of 1998. The timeline for the current math-curriculum adoption cycle is available at http://www.doe.in.gov/olr/docs/CHRONOLOGYFORTHE2009MATHEMATICSADOPTIONApr09.pdf.

[16] We also considered models that include 1997-1998 test scores and attendance rates for districts and schools, models that omit controls from the 1997-1998 school year altogether, and models that include a longer history of test scores. In all cases, the estimated curricula effects are very similar to what we report in the text.

[17] For brevity we do not report the results from the propensity-score specification, although we note that curricula adoption decisions are non-negligibly correlated with school and district-level observables. The estimates from the MNP specification are available upon request.

specification. Specifically, higher-order and interaction terms are added to the model and the balancing tests are repeated, and this process continues until balance is achieved. In our analysis, with relatively few treatment and control units (compared to the general matching literature), such an approach comes at the expense of valuable degrees of freedom.

Because of this limitation we present our primary results using the simple MNP specification, ignoring the mild imbalance in our comparison between *C* and *A*. In an omitted analysis, we constructed a separate propensity score model that balances the treatment and control schools in this comparison by adding interaction and higher-order terms. Ultimately, our estimates of $ATTE_{C,A}$ using this alternative propensity-score specification are noisier but qualitatively similar to the estimates reported in the text using the simple MNP.

VIII. Matching Performance

Taking the satisfaction of conditional independence as a starting point, a substantial body of recent research has evaluated the performance of matching estimators. This research considers the extent to which various data environments will be conducive to matching, and the efficiency properties of different estimators, both in general and given different data conditions.[18] We take two insights from the matching-performance literature that, based on evidence from simulation studies, should improve inference from our analysis. First, intuitively, research suggests that the distance between the densities of the propensity-score distributions for treated and control units will affect the precision of the estimates obtained from matching. Density-distance has been discussed in numerous studies, including Frölich (2004), who measures density distance using the Kullback-Leibler (KL) information criterion.[19] We follow his approach here to measure density-distance across each of our curricular comparisons.

We start by estimating kernel-density plots based on the Epanechnikov Kernel for the distributions of the propensity scores among treated and control units for each curricula

[18] See for instance: Caliendo and Kopeinig, 2005; Frölich, 2004; Imbens and Wooldridge, 2009; Lechner, 2002; Zhao, 2004.

[19] The KL information criterion is not technically a distance measure because it is not symmetric. More precisely, it is a density *divergence* measure.

comparison. We then measure the distance between the treatment and control densities. For example, using $\rho_{21} = (\frac{P_B}{P_B + P_A})$ to denote the probability of choosing *B* over *A*, we estimate:

$$KL = \int \ln(\frac{f_{p|T=B}(\rho_{21})}{f_{p|T=A}(\rho_{21})}) f_{p|T=B}(\rho_{21}) d\rho_{21} \quad (8)$$

In (8), $f_{p|T=B}(\rho_{21})$ is the probability density function of ρ_{21} among schools treated with *B*, and $f_{p|T=A}(\rho_{21})$ is the analogous probability density function for schools that used curriculum *A*. A KL-information-criterion measure of zero suggests that the densities are identical, and the measure increases with density distance.

Figure 1 plots the estimated kernel-density functions for treatment and control schools for each pairwise comparison, and Table 3 reports the corresponding KL information criteria.[20] The density comparisons provide interesting insights. First, note the atypical density function for schools that choose *B* over *A* – although there is a heavier weight in the upper tail of the distribution of propensity scores for these schools relative to the "untreated" Saxon schools, there is not an upper-tail peak in the distribution as is commonly observed. The KL-information-criterion value of 0.25 for this comparison is similar to the most favorable density design considered by Frölich (2004), suggesting that matching will perform relatively well in this comparison. The estimated KL-information-criterion for the comparison between *B* and *C* is also reasonable, and corresponds closely to Frölich's (2004) middle density design. However, the density comparison is less-favorable for the analysis of *C* relative to *A*, which is consistent with the balancing problems that we encounter in Table 2.[21] Frölich's (2004) work suggests that matching will perform relatively poorly in this comparison, which may limit inference.

The second insight we take from the matching literature is that improved bandwidth selection can improve estimation precision when using kernel and local linear regression matching algorithms. In the analysis below, we initially use conventional cross-validation to

[20] Our calculations here are somewhat coarse because the distributions are only estimated at 50 points.
[21] Coincidentally, the three density scenarios here are very similar to the three density designs constructed by Frölich (2004), which is useful for inference.

obtain fixed bandwidths for our matching estimators (Frolich, 2004; Li and Racine, 2007). However, because the cross-validation approach selects the bandwidth using only the distribution of control units, a variable bandwidth that varies in response to the location of the treatments in the propensity-score density function may improve estimation (Galdo, Black and Smith, 2007; Ham et al., 2006). Galdo, Black and Smith (2007) suggest multiplying the fixed-bandwidth obtained via cross-validation by the ratio $(\frac{\rho_i}{1-\rho_i})^{1/5}$ to obtain locally-varying bandwidths, where ρ_i is the estimated (pairwise) propensity score for treatment *i*. This approach uses wider bandwidths for treatments with higher propensity-score values, which generally correspond to points in the density function where there are fewer control observations, and is shown to improve efficiency by Galdo, Black and Smith (2007). Of course, a tradeoff is that the local varying bandwidths will introduce bias if the wider bandwidths for high-ρ observations pull in comparison units that are of limited comparability.

The efficiency gains from employing the locally-varying bandwidths should be largest when there is the least overlap in the propensity-score distributions between the treatment and control samples. Judging from Figure 1, our estimates from the comparisons between *C* and *A*, and *C* and *B*, should benefit most by moving from the fixed bandwidths to the locally-varying bandwidths, and in fact this is what we find (see Tables 4 and 5 below). Overall, our results do not differ qualitatively regardless of whether we use fixed or locally-varying bandwidths in any of our comparisons.

IX. Estimates of Curricular Effectiveness in Math

Rather than overwhelm the reader with estimates using the numerous matching algorithms available in the literature, we instead present estimates using kernel and local-linear regression (LLR) matching only (for details on these and other matching estimators, see, for example, Heckman Ichimura and Todd, 1997, 1998; Mueser, Troske and Gorislavsky, 2007). Frölich's (2004) analysis indicates that kernel matching in particular should perform well in our

context.[22] As for LLR matching, the evidence in the literature is mixed.[23] Although our estimates using LLR matching are less precise than the kernel-matching estimates, they are generally very similar. We present results using the Epanechnikov kernel for both types of matching estimators. In omitted analyses available upon request we show that our results are robust to alternative estimators, including kernel and LLR matching estimators that use the Gaussian kernel, other matching estimators like simple pair matching or radius matching using various radii, and regression-adjusted matching estimators.

Table 4 presents results for all grade-three cohorts who were ever exposed to the curricula of interest using fixed-bandwidth matching estimators where the bandwidths are obtained via conventional cross-validation (see Appendix B).[24] Table 5 reports analogous results using the locally-varying bandwidths suggested by Galdo, Black and Smith (2007). All of our matching estimators impose the common support condition. We also report OLS estimates where we regress test score outcomes on the covariates used in the propensity score model and indicator variables for curricula adoptions, retaining our pairwise comparisons (that is, when we compare *B* to *A*, we drop all schools at districts that adopted *C*). The standard errors for our matching and OLS estimates are clustered at the district level and our matching-estimator standard errors are bootstrapped using 250 repetitions. We obtain the optimal numbers of bootstrap repetitions to use for our standard error calculations following Ham et al. (2006), who use a special case of Andrews and Buchinsky (2001).[25]

[22] Frölich's (2004) study also suggests that ridge matching should perform well, but the ridge parameter will lead to bias in the case of multiple covariates (Frölich focuses on a single-covariate settting). See Heckman, Ichimura and Todd (1998) for details.

[23]For instance, Fan (1993) indicates that local linear regression is a more efficient estimator than the standard kernel estimator, and Caliendo and Koepinig (2005) suggest LLR is particularly useful when controls are distributed asymmetrically around treated observations. Frölich (2004) concludes that kernel regression is more robust to bandwidth misspecification than LLR in finite samples, but Ham et al. (2006) suggest this issue with LLR can be greatly improved by using a variable bandwidth.

[24] In some cases the cross-validation estimates of the loss function are fairly flat. In these cases, we combine "visual inspection" with cross-validation to choose the optimal bandwidth. See Appendix B for details.

[25] For our estimators, the optimal number of bootstrap repetitions is consistently near 200. We use 250 repetitions to insure that we meet or exceed the optimal repetition count in each year.

Each cohort is labeled in the tables according to the year of its spring test score (i.e., the 1998-1999 cohort is labeled "1999"). Recall that the 1999, 2000, 2005, and 2006 cohorts were only partially exposed to the curricula, while the cohorts from 2001 through 2004 were exposed for all three years. All of the curricula effects in the table are standardized using the distribution of school-level test scores. For example, the estimate in Table 4 for $ATTE_{B,A}$ in 2002 indicates that schools that selected *B* over *A* moved up 0.369 standard deviations in the distribution of school-level math test scores. More typically, researchers report effects that are standardized based on the distribution of *individual-level* scores rather than school-level scores, but we do not have access to the distributions of individual-level scores over the entire course of the data panel (specifically, we do not have these distributions for the years prior to 1999-2000). In Appendix Table A.2, for each year where we have access to the individual-level distribution of test scores (such that we could compute the standard deviation), we provide the scaling factors that convert the estimates in Tables 4 and 5 into the more common metric. Roughly speaking, dividing the estimates by three returns estimates approximately in the metric of standard deviations of the individual-level distribution of scores.

Focusing first on our largest comparison between *B* and *A*, and the estimates for the fully-exposed cohorts (2001 – 2004), *B* consistently outperforms *A*. The exception is 2004, where we find no statistically distinguishable differences in curricular effectiveness in that year. One of the most likely explanations for this finding is that the testing instrument in 2004 differed from the instruments in prior years in some unobservable way. For the other comparisons, *C* also generally outperforms *A* for the fully-exposed cohorts, and there is limited statistical evidence suggesting that *B* outperforms *C*.

The magnitudes of the curricula effects for the fully-exposed cohorts are economically meaningful, particularly when weighed against the marginal costs associated with choosing one curriculum over another. Fryer and Levitt (2006) find that between grades one and three, the black-white achievement gap grows at a rate of approximately 0.10 standard deviations per

year.[26] Our estimates suggest that the effect on student learning of using curriculum *B* instead of *A* is on the order of 0.11 standard deviations of the test over three years (averaging the point estimates across the four fully-exposed cohorts and using the scaling factors in Appendix Table A.2). That is, the effect is on the order of one year's worth of expansion of the black-white achievement gap. Given that the curricula are very similarly priced (the texts from *A*, *B* and *C* were, averaged over grades 1-3, $23.08, $24.80 and $25.34 each, respectively), our estimates imply that selecting a better curriculum is a cost-effective way to improve student achievement.

Recall that our ATTE estimates define the most-used curriculum as the control curriculum in each pairwise comparison. In omitted results where we reverse the treatment and control definitions within each pair, our findings are very similar to those reported in Tables 4 and 5. Specifically, curricula *B* and *C* continue to outperform curriculum *A*, with the qualification that the differential effect between curriculum *C* and curriculum *A* is smaller when we estimate $ATTE_{A,C}$ instead of $ATTE_{C,A}$. This suggests that although curriculum *A* underperforms relative to curriculum *C* in all schools, the degree of underperformance is smaller at schools that actually chose curriculum *A* relative to what would have happened at schools that actually chose *C*, had they instead chosen *A*.

We do not find any evidence of curricular-familiarity effects for the fully-exposed cohorts. If curricula familiarity is important for teachers, we might expect the 2001 and 2002 cohorts, for example, to be less affected by curricula differences than the cohorts in 2003 and 2004 (under the assumption that when familiarity is low, curricula implementation by teachers reverts toward a common mean). However, there is no distinguishable evidence of such a trend in curricular effectiveness across cohorts.

We also do not find any statistically significant curricula effects for the cohorts of students who were not fully exposed. This result may be partly explained by heterogeneity in out-of-cycle curricula adoptions within the treatment and control samples, which will attenuate

[26] Fryer and Levitt (2006) analyze a different testing instrument; however, similar estimates of the black-white achievement gap spread are available elsewhere (see, for example, Chubb and Loveless, 2002).

our estimates for the partially-exposed cohorts. That is, regardless of whether curricula quality is uncorrelated, (imperfectly) positively correlated, or negatively correlated for schools across adoption cycles, the estimates for the partially-exposed cohorts will be pushed toward zero.

We explore the correlations in curricula adoptions in Indiana across the 1998 and 2004 adoption cycles in Table 6 (recall that we do not have data for the prior cycle from 1992). The table takes our initial sample of districts that uniformly adopted one of the three primary curricula during the 1998 adoption cycle and reports average curricula adoptions in 2004. For brevity, the table only shows adoption shares in 2004 for the four most popular curricula from that adoption cycle (published by Saxon, Harcourt, Houghton-Mifflin and Scott-Foresman).

For students in the 2005 and 2006 cohorts, Table 6 provides direct information about the heterogeneity in the sequences of treatments at treatment and control schools. For students in the 1999 and 2000 cohorts, it is merely suggestive about the extent to which curricula adoptions are correlated across cycles more generally. The table shows that while Saxon adopters in 1998 were much more likely to be Saxon adopters in 2004, adopters of the other two curricula are quite dispersed across alternative options during the 2004 adoption cycle. Without knowing the respective qualities of the different curricula adopted outside of the 1998 adoption cycle, including those from the same publishers (there is no evidence that we are aware of on the persistence of publisher quality), it is difficult to form expectations based on the correlations in Table 6.[27] Ultimately, given the sizes of our point estimates and standard errors for the partially-exposed cohorts, we cannot rule out the possibility that there are partial-exposure curricula effects despite our inability to statistically distinguish such effects.

Table 6 is also informative about the changing market shares of curricula publishers over time. The publisher of curriculum *A*, despite its relative underperformance, slightly *increased* its

[27] Evidence on the persistence of publisher quality would be difficult to obtain without the availability of consistent comparisons over time. For example, because Silver-Burdett Ginn did not offer a curriculum in Indiana during the 2004 adoption, our most reliable comparisons (per Section VIII) cannot be replicated in the later adoption cycle. Even more, we cannot reliably compare Saxon and Scott-Foresman in 2004 because of the large decline in Scott-Foresman's market share across adoption cycles. Even in cases where curricula publishers are consistently represented across adoption cycles, the cycle durations imply that long data panels will be required to identify the persistence of publisher quality.

near-50-percent market share from the 1998 adoption cycle to the 2004 adoption cycle. Although curriculum *B* was the most effective curriculum during the 1998 adoption cycle, it did not appear in 2004. The publisher of curriculum *B* was bought out by Pearson Publishing during the 1998 cycle and Pearson phased out curriculum *B* in favor of curriculum *C*, which it also publishes. Curriculum *C's* market share fell from fifteen to nine percent despite the evidence here that it outperformed curriculum *A* during the 1998 adoption cycle.

Overall, our most reliable estimates of curricula effectiveness come from the four cohorts of "fully-exposed" students who used the curricula of interest in grades one, two and three. Our estimates based on these cohorts indicate that curriculum *A* underperformed relative to curricula *B* and *C*, although this did not impact its market share in the next adoption cycle. The magnitudes of the effects estimated in Tables 4 and 5 are non-negligible, suggesting that curricula are important determinants of student achievement.[28]

X. Falsification Tests

Perhaps the biggest weakness in many matching analyses is that it is difficult to empirically verify that the conditional independence assumption is satisfied. Our particularly long data panel from Indiana, which includes test scores in math and reading for students in multiple grade levels, allows us to overcome this limitation by providing a series of falsification tests for our primary results. For brevity, we only report falsification estimates using kernel matching with the Epanechnikov kernel and fixed bandwidths.[29]

We perform two types of falsification tests. First, we estimate curricula "effects" for cohorts of students who were never exposed to the curricula of interest. Examples include cohorts of grade-three, grade-six and grade-eight students from the mid 1990's (and later for the later grades). We expect to estimate curricula "effects" for these cohorts that are statistically

[28] We also note that our inability to follow individual students over time implies some downward bias in our estimates to the extent that students switch curricula between grades one and three. That is, across-district movers who are tested in grade-three may only be partially exposed to the "treatment" curricula. Because we have no way of identifying these students, we cannot exclude them from the analysis and they will bias our estimates toward zero.

[29] In unreported results we verify that our findings are robust to using the Gaussian kernel instead of the Epanechnikov kernel and to alternative matching estimators such as radius or nearest-neighbor matching.

indistinguishable from zero. We also estimate curricula effects for students who were exposed to the math curricula of interest, but we estimate the effects on reading test scores. For these latter estimates timing does not rule out the possibility of causal effects; however, at most, we would anticipate only small cross-subject spillover effects.

Potentially confounding both types of falsification estimates are correlations in curricula adoptions across grades, subjects, and adoption cycles. Recall from Table 6 that there are non-zero correlations in math-curricula adoptions across adoption cycles. Not surprisingly, in unreported results (omitted for brevity and available upon request) we also find that math curricula adoptions are correlated across grades within adoption cycles, and to a lesser extent, with curricula adoptions in other subjects (where the adoptions overlap imperfectly with the math adoptions – see Section II). The correlations between the curricula of interest and the other curricula to which the falsification cohorts were exposed could potentially confound the falsification tests. For example, if curricula quality is correlated across adoption cycles for districts, the falsification estimates will capture more than just bias, making them difficult to interpret. However, in practice, none of the correlations in curricular quality across adoption cycles appear to be strong enough to limit inference from our falsification exercise - almost all of our estimates are statistically indistinguishable from zero.

We present 117 falsification estimates in Tables 7 through 10, or 39 estimates for each comparison (although these tests are not independent). In summary, the pattern of the falsification estimates offers little suggestion of bias in our primary results. For our comparisons between *B* and *A*, and *C* and *B*, none of the falsification estimates are statistically significant at the 10 percent level. For our comparison between *C* and *A*, three estimates are statistically significant at the 10 percent level or better.[30]

We first estimate curricula "effects" on grade-three math test scores for cohorts of students from 1992 through 1996, and 2007 and 2008 (recall that we use data from 1997 and

[30] If the falsification tests were independent we would expect roughly four "false positives" in each comparison by chance. However, because treatment and control schools are uniformly defined over time the tests are not independent, making it unclear how many false positives to expect.

1998 to match schools). None of these cohorts were ever directly exposed to the curricula of interest. Our results are reported in Table 7, and as expected, the estimated curricula "effects" are statistically indistinguishable from zero (with the exception of the "effect" estimated in 1992 for the comparison between *C* and *A*).

Next we produce estimates using cohorts of *grade-six* students who were never exposed to the curricula of interest (cohorts from 1993-2001). Note that because many districts teach grades three and six in different buildings (i.e., elementary and middle schools), and multiple elementary schools generally feed into a single middle school, these falsification tests use samples of schools that only partly overlap with the grade-three samples and are much smaller.[31] We use the same basic matching approach to predict the same treatments (the uniform adoption of curriculum *A, B or C* in grades one, two and three), only we match schools that have a sixth grade class and estimate the effects of curricula on sixth grade achievement. Our results are reported in Table 8, and like Table 7, the estimates are again statistically indistinguishable from zero with the exception of two significant "effects" (at the 10-percent level) in our most tenuous comparison.

In Table 9 we replicate the analysis from Table 8 for cohorts of *grade-eight* students who were never exposed to the curricula of interest. All of the falsification estimates in Table 9 are statistically indistinguishable from zero.[32]

In Table 10 we estimate curricula effects for grade-3 cohorts in all years of the data panel, but change the outcome from math test scores to reading test scores. Students in the grade-three cohorts in 1992 through 1996, and 2007 and 2008, were never exposed to the curricula of interest. The other cohorts of students were exposed, and it is unclear *a priori* whether we should expect any cross-subject spillover effects. We suggest three possible mechanisms that may generate non-zero spillover effects. First, math curricula may directly

[31] As a consequence of observing fewer schools with grade-6 (and grade-8) classrooms, these estimates are noisier than the estimates using the cohorts of grade-3 students. In unreported results we also pooled the grade-6 and grade-8 falsification estimates across years to improve power – as is the case for the results reported below, the pooled falsification estimates are indistinguishable from zero.

[32] Note that data were unavailable for grade-8 test scores in 1996.

affect reading performance. As an example, math curricula may differentially use word problems, which could lead to differential effects on reading comprehension. Second, a better math curriculum may afford teachers more time to spend on reading instruction. Third, a better math curriculum may increase the return to math instruction and encourage teachers to substitute out of reading instruction and into math instruction. These latter two possibilities are analogous to income and substitution effects from basic microeconomic theory. The direction of the cross-subject spillover will depend on which effect dominates.

Although we do not have a strong prior about whether math curricula affect reading outcomes, one straightforward expectation is that the effects of math curricula on math test scores should be larger in magnitude than their analogous effects on reading test scores. Thus, at its most basic level, this final test should confirm this result. Table 10 presents estimates for the effects of math curricula on reading test scores throughout our data panel, and indeed, there are no statistically significant reading effects.

While all of the estimates in Table 10 are statistically indistinguishable from zero, note that the point estimates follow a pattern similar to what we find in our primary results (Tables 4 and 5) - they nominally peak for the fully-exposed cohorts. If we were to take these statistically insignificant point estimates at face value, one explanation for this pattern is that the more effective math curricula also improve reading scores. Alternatively, one could argue that the reading results are evidence of bias in our estimates. However, if estimation bias explains the pattern of estimates in Table 10, the source of bias would have to be unique in timing such that it is aligned with our fully exposed cohorts.

We briefly consider how a pure-bias interpretation of the reading estimates impacts our results. Under the assumption that math curricula have no causal relationship with reading achievement, we estimate math-curricula effects on de-trended math test scores. These de-trended scores were obtained by separately standardizing each school's math and reading test score, and subtracting the reading score from the math score. We omit the estimates for brevity, but note that they are in line with what would be expected by subtracting the stand-alone reading

estimates from the stand-alone math estimates. Specifically, for the fully-exposed cohorts in 2001, 2002 and 2003, the estimates decline in magnitude (by roughly half in most cases), but remain statistically significant in our comparisons between *B* and *A*, and *C* and *A* (at the 10 percent level or better). For our comparison between *C* and *B*, the de-trended results to not indicate any statistically significant curricula differences. We treat these de-trended results as lower-bound estimates because they assume that cross-subject spillover effects are zero.

Overall, the falsification exercise provides overwhelming evidence that our primary results are not biased, particularly for our comparisons between *B* and *A*, and *C* and *B*. Furthermore, even in our comparison between *C* and *A*, where the conditions are less favorable to the matching approach, the falsification estimates suggest that estimation bias is unlikely to significantly affect our findings.

XI. Persistence

Finally, we use the extended data panel in Indiana to evaluate the persistence of curricula effects over time. Specifically, we ask whether the cohorts of students who were exposed to the more-effective curricula in grades one, two and three perform better by grade six and grade eight. To estimate persistence effects we look at test score outcomes for cohorts of grade-six students between 2002 and 2008, and cohorts of grade-eight students between 2004 and 2008. Note that these cohorts correspond to the cohorts of grade-three students who were exposed to the curricula of interest in our primary analysis – for example, the 2005 cohort of grade-six students is also the 2002 cohort of grade-three students. The fully exposed cohorts were in grade six between 2004 and 2007, and in grade eight between 2006 and 2009 (recall that our data panel ends in 2008).

Two issues merit attention in our persistence analysis. First, if there are test-score ceilings in higher grades on the Indiana test, it will be difficult to detect persistence effects because the tests in later grades may not adequately differentiate student learning. We test for ceiling effects following Koedel and Betts (forthcoming) and find that the testing instruments should be sufficient to detect any persistence effects should such effects exist. A second concern

is that we cannot track individual students over time in the data, and as a consequence our assignment to curriculum treatment during grades 1-3 may not be accurate for all students in any given cohort. That is, because every school that contains a grade-6 or grade-8 classroom is attached to a district, we can identify the curriculum to which students would have been exposed in grades 1-3 if they attended a school in that same district. However, some students may have moved districts between grades 1-6 or 1-8. This churning implies that at least some of the students who contribute to a school's grade-6 or grade-8 test score were not actually treated with the district's curriculum in the early grades. In practice, this will add noise to our treatment classifications, attenuating any estimated persistence effects.[33]

Table 11 presents our persistence analysis. Leaving aside the concern that our results will be biased toward zero per the previous discussion, the estimates in the table provide little indication that curricular effects persist over time. For the estimates in the table to be driven by downward bias from student movement across districts, the amount of student movement would need to be inordinately large. Put differently, under the assumption that most students do not transfer districts during elementary school, the results in Table 11 indicate that math curricula effects do not persist over time. This result is consistent with a large body of evidence pointing to a general lack of persistence in the effects of educational inputs (see, for example, Jacob et al., 2008; Garces et al., 2002), and raises doubts about the extent to which administrators can improve student performance in the long run by choosing more effective curricula.[34]

XII. Conclusion

Our non-experimental analysis of curricular effectiveness offers both methodological and policy-oriented insights. Methodologically, we show that non-experimental methods can be used to identify causal curricular effects at just a fraction of the cost of experimental designs. Furthermore, particularly in the case of curricula, concerns about the external validity of

[33] Student churning across districts is also a problem in our primary analysis, although less so. For example, if a student changes districts in grade-2, she may change curricula. As such, all of our estimates will be biased toward zero to some extent.

[34] Also, as with the falsification tests involving grade-6 and grade-8 schools, in unreported results we pool across years to improve power. Similarly to our reported results, the pooled estimates provide no evidence of persistence.

experimental estimates suggest that non-experimental results may be preferred when the data conditions are favorable. We also provide useful insights for policymakers facing curriculum uptake decisions. While our results indicate that some curricula are more effective than others in the short run, they also show that curricula effects do not persist over time.

That a non-experimental approach to curricular evaluation can produce verifiably causal estimates is an important contribution to the literature. Non-experimental methods bypass some of the limitations inherent to experimental designs, including the experimentation on a subpopulation problem (Manski, 1996), and the possibility of publisher Hawthorne effects. Furthermore, the non-experimental analysis outlined here is not only feasible to replicate in other environments methodologically, but also fiscally. In contrast to the ongoing project by Agodini et al. (2009), a particularly well-designed RCT that is funded by the Institute for Education Sciences for roughly 21 million dollars over five years, our study was performed using publicly available data at only a small fraction of this cost.

Although our study is preferable to experimental designs along some dimensions, it also has weaknesses. First, we do not have enough data, or the right kind of data (i.e., student level), to evaluate the extent to which curricula differentially affect different types of students (e.g., high and low-achieving, English-proficient and ESL, etc.). This deficiency in our analysis is likely to be less problematic in the future because districts and states are continuing to develop longitudinal databases that track individual students. We also depend critically on the standardized test administered by the state of Indiana as our outcome measure (the ISTEP). While we expect our results to extrapolate well to other states or districts that use similar tests, they may not carry over into states or districts where the testing instrument differs greatly in content.[35]

Our analysis shows that students in Indiana who were exposed to curricula *B* or *C* outperformed students who were exposed to curriculum *A*. In our most compelling comparison,

[35] As we do not have any expertise in evaluating testing content, we do not make any judgments as to the validity of the Indiana test, or as to which other prominent tests in the United States are similar or different.

between curricula *B* and *A*, the effect of exposure to the better curriculum for three consecutive years is roughly 0.11 standard deviations of the grade-3 ISTEP test.[36] This effect is equivalent in magnitude to one year's growth of the black-white achievement gap over these grades (Fryer and Levitt, 2006). Interestingly, despite the consistent underperformance of curriculum *A* in our analysis, the publisher of curriculum *A* slightly *increased* its market share during the next adoption cycle in Indiana. There are many possible explanations for this finding, ranging from a lack of reliable information available to administrators about curricular quality (WWC, 2007), to poor decision making by educational administrators (also see Ballou, 1996).

Overall, our results are encouraging because choosing a better curriculum can non-negligibly improve student performance. Further, the near-zero marginal cost of choosing one curriculum over another suggests that implementing a better curriculum will be quite cost-effective. However, the lack of persistence of curricula effects (although not unique to curricula in education) dampens enthusiasm about the potential benefits of improved curricula. By grades six and eight, the benefits of the better curricula are no longer distinguishable.

References

Agodini, Robert and Barbara Harris and Sally Atkins-Burnett and Sheila Heaviside, and Timothy Novak. 2009. *Achievement Effects of Four Early Elementary School Math Curricula*. National Center for Education Evaluation and Regional Assistance, U.S. Department of Education, Institute of Education Sciences. NCEE 2009-4052.

Ballou, Dale. 1996. "Do Public Schools Hire the Best Applicants?" *Quarterly Journal of Economics* 111(1), pp. 97-133.

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.

Borman, Geoffrey D. and N. Maritza Dowling and Carrie Schneck. 2008. "A Multisite Cluster Randomized Field Trial of Open Court Reading," *Education Evaluation and Policy Analysis* 30 (4), 389-407.

Caliendo, Marco and Sabine Kopeinig. 2005. "Some Practical Guidance for the Implementation of Propensity Score Matching," IZA Discussion Paper No. 1588.

[36] Estimate obtained by averaging across the point estimates for the four fully-exposed cohorts.

Chubb, John and Tom Loveless. 2002. *Bridging the Achievement Gap,* Brookings Institution Press, Washington, D.C.

Fan, J. 1993. "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, 21, 196-216.

Finn, Chester. 2004. "The Mad, Mad World of Textbook Adoption" The Thomas B Fordham Foundation and Institute Report. Washington, D.C.

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

Fryer, Roland and Steven Levitt. 2006. "The Black-White Test Score Gap Through Third Grade," *American Law and Economics Review* 8 (2), 249-281.

Galdo, Jose and Jeffrey Smith and Dan Black. 2007. "Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data," IZA Discussion Paper 3095.

Garces, Eliana and Duncan Thomas and Janet Currie. 2002. "Longer-Term Effects of Head Start," *American Economic Review* 4, 999-1012

Ham, John and Xianghong Li and Patricia Reagan. 2006. "Propensity Score Matching, a Distance-Based Measure of Migration, and the Wage Growth of Young Men," Federal Reserve Bank, Staff Report No. 212.

Heckman, J.J. and Smith, J.A. 1995. "Assessing the case for social experiments," *Journal of Economic Perspectives* 9 (2), 85-110.

Heckman, James and Hidehiko Ichimura and Petra Todd. 1997. "Matching As An Econometric Evaluation Estimator," *Review of Economic Studies* 65 (2), 261-294.

Heckman, J. and Ichimura, H. and Todd 1998. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job-Training Programme," *Review of Economic Studies* 65, 261-294.

Imbens, Guido and Jeffrey Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature* 47 (1), 5-86.

Imbens, Guido. 2003. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," NBER Working Paper No. 294.

Jacob, Brian and Lars Lefgren and David Sims. 2008. "The Persistence of Teacher-Induced Learning Gains," NBER Working Paper No. 14065.

Koedel, Cory and Julian R. Betts (forthcoming). "Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation," *Education Finance and Policy*.

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.

Li, Qi and Jeff Racine. 2007. *Nonparametric Econometrics: Theory and Practices*, Princeton University Press, Princeton N.J.

Ludwig, Jens and Douglas Miller. 2007. "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design*," Quarterly Journal of Economics* 122 (1) 159-208.

Manski, Charles. 1996. "Learning About Treatment Effects from Experiments with Random Assignment of Treatments," *The Journal of Human Resources* 31 (4), 709-733.

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.

National Research Board. 2004. *On Evaluating Curricular Effectiveness: Judging the quality of K-12 Mathematics Evaluations*, The National Academies Press, Washington DC.

Resendez, Miriam and Mariam Azin. 2007. "The Relationship Between Using Saxon Elementary and Middle-School Math and Student Performance on California Statewide Assessments," Planning Research Evaluation Services.

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

Slavin, Robert E. and Cynthia Lake. 2008. "Effective Programs in Elementary Mathematics: A Best-Evidence Synthesis," *Review of Educational Research,* 78 (3), 427-515.

Smith, Jeffrey and Petra Todd. 2005. "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics* 125 (2), 305-353.

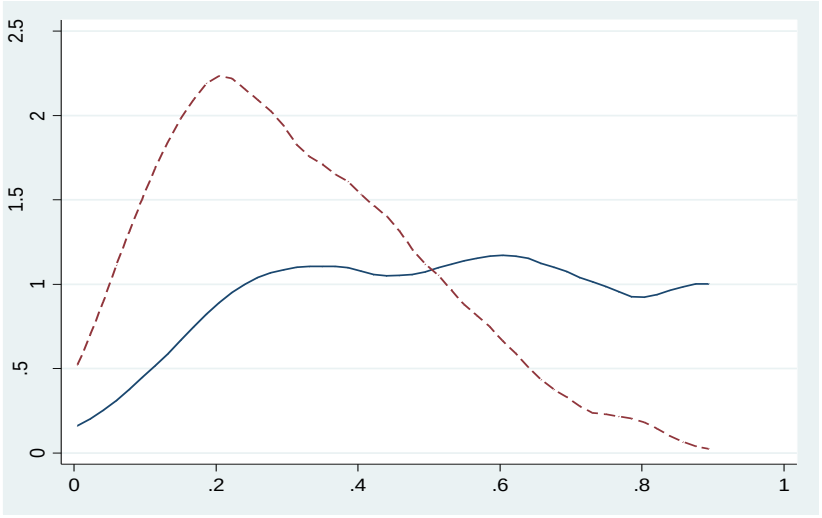
Tulley, Michael. 1989. "The Pros and Cons of State-Level Textbook Adoption," *Publishing Research Quarterly*, 5 (2)

What Works Clearinghouse. 2007. Topic Report: Elementary School Math. Available at: http://ies.ed.gov/ncee/wwc/reports/elementary_math/topic

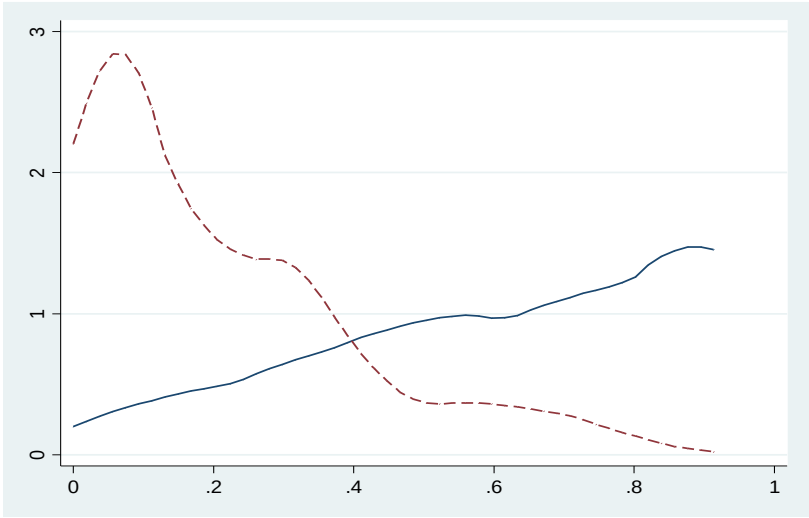
Zhao, Zhong. 2004. "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence" *Review of Economics and Statistics* 86 (1) 91-107.

Figure 1. Probability Density Functions for Estimated Propensity Scores for Treatment and Control Units in Each Comparison Using 2001 Data, Reported Where the Control Densities are Non-Zero (Solid Lines are Treatment Densities, Dashed Lines are Control Densities).

Treatment: Silver-Burdett Ginn (B) Control: Saxon (A)



Treatment: Scott-Foresman (C) Control: Saxon (A)



Treatment: Scott-Foresman (C) Control: Silver-Burdett Ginn (B)

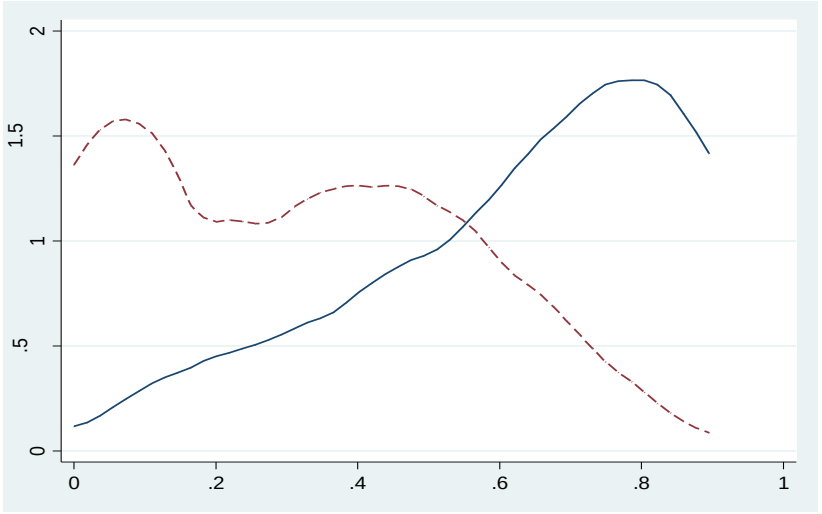


Table 1. Average Characteristics of Schools and Districts, by Adopted Curriculum (1997 values)

	All	Saxon (A)	Silver (B)	Scott (C)
School-Level Outcomes				
Attendance Rate	96.2	96.3[a]	96.1[a]	96.3
Grade-3 Math Test Score	496.6	496.5	494.2[c]	499.7[c]
Grade-3 Language Test Score	496.7	496.1	495.8	498.7
School-Level Demographics				
Percent Free Lunch	27.4	24.7[a,b]	28.5[a]	30.5[b]
Percent Reduced Lunch	6.7	7.1[a]	6.3[a]	6.6
Percent Not Fluent in English	1.2	0.7[a]	1.7[a]	1.2
Percent Language Minority	2.6	1.8[a]	3.9[a]	2.6
Percent White	91.3	95.4[a,b]	88.0[a]	88.4[b]
Percent Black	5.6	2.3[a,b]	7.2[a,c]	9.2[b,c]
Percent Asian	0.7	0.4[a,b]	0.9[a]	1.1[b]
Percent Hispanic	2.2	1.8[a,b]	3.7[a,c]	1.1[b,c]
Percent American Indian	0.2	0.1	0.2	0.2
Enrollment (logs)	5.95	5.92	5.97	5.96
N (Schools)	716	311	221	184
District-Level Outcomes				
Attendance Rate	95.8	95.72[b]	95.82	96.12[b]
Grade-3 Math Test Score	498.1	495.79[b]	498.12[a,c]	506.9[b]
Grade-3 Language Test Score	498.9	496.47[a,b]	500.60[a]	505.6[b]
District-Level Demographics				
Enrollment (logs)	7.72	7.56[a,b]	7.83[a,c]	8.17[b,c]
Total Revenue (logs)	16.55	16.37[a,b]	16.67[a,c]	17.03[b,c]
Local Revenue (logs)	14.96	14.73[a,b]	15.07[a,c]	15.64[b,c]
Census Information (District Level)				
Median Household Income (logs)	10.81	10.78[a,b]	10.82[a,c]	10.90[b,c]
Share of Population with Low Education	18.2	18.8[b]	19.2[c]	14.25[b,c]
N (Districts)	213	124	56	33

[a] Indicates statistically significant difference at the 10% level between Saxon and Silver-Burdett Ginn adopters.
[b] Indicates statistically significant difference at the 10% level between Saxon and Scott-Foresman adopters.
[c] Indicates statistically significant difference at the 10% level between Silver-Burdett Ginn and Scott-Foresman adopters.
Note: The propensity-score specification also uses italicized information from 1998 – differences in means for these years are not reported for brevity.

Table 2. Balancing details for the 32 covariates included in the multinomial probit specification.

	1992	1993	1994	1995	*1996*	*1999*	2000	2001	2002	2003	2004	2005	2006	2007	2008
Silver (B) to Saxon (A)															
# of unbalanced covariates (ten percent level or better)	0	0	0	0	*0*	*0*	0	0	0	0	0	0	0	0	0
Average p-value from balancing test, all covariates	0.65	0.62	0.59	0.56	*0.57*	*0.56*	0.60	0.55	0.54	0.54	0.57	0.54	0.54	0.55	0.56
Scott (C) to Saxon (A)															
# of unbalanced covariates (ten percent level or better)	1	3	3	3	*3*	*3*	1	3	3	3	3	3	1	1	0
Average p-value from balancing test, all covariates	0.55	0.57	0.52	0.53	*0.53*	*0.48*	0.52	0.54	0.49	0.42	0.49	0.41	0.41	0.40	0.46
Scott (C) to Silver (B)															
# of unbalanced covariates (ten percent level or better)	0	1	0	1	*1*	*1*	1	1	1	1	0	0	0	0	0
Average p-value from balancing test, all covariates	0.49	0.49	0.46	0.46	*0.49*	*0.47*	0.49	0.49	0.51	0.48	0.51	0.51	0.49	0.50	0.56

Note: Columns in italics are for years that are contiguous to the years from which the matching criteria are drawn.

Table 3. Kullback-Leibler (KL) Information Criteria by Curricula Comparison.

Comparison	KL Information Criterion
B to A	0.25
C to A	0.77
C to B	0.50

Note: Based on 2001 sample of schools.

Table 4. Estimates of Math Curricular Effectiveness on Math Test Scores for Partially and Fully-Exposed Cohorts (Fixed-Bandwidth Matching Estimators). All Comparisons.

	1999	2000	**2001**	**2002**	**2003**	**2004**	2005	2006
Treatment: B Control: A								
OLS	0.123	0.162	**0.355**	**0.357**	**0.374**	**0.269**	0.292	0.248
	(0.105)	(0.100)	**(0.095)****	**(0.087)****	**(0.099)****	**(0.131)***	(0.104)**	(0.110)*
Kernel Matching	0.056	0.055	**0.347**	**0.369**	**0.456**	**0.141**	0.251	0.145
	(.164)	(.176)	**(.145)***	**(.117)****	**(.175)****	**(.147)**	(.163)	(.154)
LLR Matching	0.070	-0.008	**0.343**	**0.317**	**0.430**	**0.075**	0.240	0.168
	(.196)	(.207)	**(.144)***	**(.160)***	**(.186)***	**(.164)**	(.182)	(.160)
Treatment: C Control: A								
OLS	0.132	-0.011	**0.189**	**0.263**	**0.208**	**0.015**	0.108	0.181
	(0.120)	(0.134)	**(0.103)†**	**(0.096)***	**(0.109)†**	**(0.118)**	(0.104)	(0.119)
Kernel Matching	0.171	-0.032	**0.268**	**0.419**	**0.453**	**-0.002**	0.190	0.174
	(.263)	(.221)	**(.179)**	**(.161)****	**(.171)****	**(.158)**	(.176)	(.176)
LLR Matching	0.171	0.260	**0.042**	**0.277**	**0.495**	**0.030**	0.111	0.126
	(.276)	(.249)	**(.216)**	**(.167)†**	**(.190)****	**(.146)**	(.175)	(.185)
Treatment: C Control: B								
OLS	0.008	-0.160	**-0.100**	**-0.186**	**-0.284**	**-0.271**	-0.181	-0.083
	(0.100)	(0.122)	**(0.117)**	**(0.129)**	**(0.166)†**	**(0.161)†**	(0.129)	(0.138)
Kernel Matching	-0.118	-0.236	**-0.172**	**-0.091**	**-0.398**	**-0.189**	-0.117	-0.066
	(.282)	(.304)	**(.318)**	**(.235)**	**(.212)†**	**(.215)**	(.242)	(.269)
LLR Matching	-0.084	-0.205	**-0.124**	**-0.035**	**-0.365**	**-0.152**	-0.055	-0.054
	(.291)	(.335)	**(.290)**	**(.278)**	**(.315)**	**(.204)**	(.252)	(.280)
N(A)	309	307	**307**	**305**	**300**	**294**	286	287
N(B)	220	219	**219**	**213**	**213**	**212**	210	207
N(C)	184	182	**182**	**181**	**176**	**174**	169	163

Note: Bolded years are for fully-exposed cohorts. Matching estimators impose the common support restriction. Standard errors are clustered at the district level and bootstrapped using 250 repetitions.

** Denotes statistical significance at the 1 percent level or better

* Denotes statistical significance at the 5 percent level or better

† Denotes statistical significance at the 10 percent level or better

Table 5. Estimates of Math Curricular Effectiveness on Math Test Scores for Partially and Fully-Exposed Cohorts. Locally-Varying Bandwidth Estimates. All Comparisons.

	1999	2000	**2001**	**2002**	**2003**	**2004**	2005	2006
Treatment: B Control: A								
Kernel Matching	0.045	0.079	**0.350**	**0.361**	**0.430**	**0.163**	0.259	0.157
	(0.123)	(0.138)	**(0.114)****	**(0.090)****	**(0.120)****	**(0.118)**	(0.123)*	(0.138)
LLR Matching	0.067	0.065	**0.345**	**0.357**	**0.476**	**0.101**	0.254	0.176
	(.118)	(.12 6)	**(.116)****	**(.094)****	**(.143)****	**(.121)**	(.122)*	(.133)
Treatment: C Control: A								
Kernel Matching	0.165	-0.008	**0.294**	**0.420**	**0.446**	**0.016**	0.208	0.207
	(0.141)	(0.172)	**(0.166)†**	**(0.158)****	**(0.157)****	**(0.120)**	(0.147)	(0.174)
LLR Matching	0.157	0.051	**0.148**	**0.386**	**0.537**	**0.081**	0.236	0.207
	(.137)	(.178)	**(.158)**	**(.147)****	**(.168)****	**(.126)**	(.157)	(.182)
Treatment: C Control: B								
Kernel Matching	-0.134	-0.246	**-0.216**	**-0.131**	**-0.427**	**-0.247**	-0.165	-0.114
	(0.148)	(0.171)	**(0.195)**	**(0.151)**	**(0.157)****	**(0.147)†**	(0.183)	(0.171)
LLR Matching	-0.085	-0.198	**-0.120**	**-0.036**	**-0.365**	**-0.154**	-0.059	-0.024
	(.136)	(.178)	**(.198)**	**(.161)**	**(.154)****	**(.139)**	(.191)	(.173)
N(A)	309	307	**307**	**305**	**300**	**294**	286	287
N(B)	220	219	**219**	**213**	**213**	**212**	210	207
N(C)	184	182	**182**	**181**	**176**	**174**	169	163

Note: Bolded years are for fully-exposed cohorts. Matching estimators impose the common support restriction. Standard errors are clustered at the district level and bootstrapped using 250 repetitions.

** Denotes statistical significance at the 1 percent level or better

* Denotes statistical significance at the 5 percent level or better

† Denotes statistical significance at the 10 percent level or better

Table 6. Average 2004 Curricula Adoptions for Uniform-Curricula Adopters (grades 1, 2 and 3) in Math in 1998 for the Four Most Common Curricula from the 2004 Adoption Cycle, by Grade.

		1998 Uniform Math Adoptions – Grades 1 Through 3		
	Across-Sample Average	Saxon (A)	Silver-Burdett Ginn (B)	Scott-Foresman (C)
2004 Math Adoptions				
Grade 1				
Saxon	0.52	0.74	0.24	0.13
Harcourt	0.18	0.08	0.31	0.33
Houghton Mifflin	0.10	0.07	0.11	0.23
Scott-Foresman	0.08	0.07	0.07	0.10
Grade 2				
Saxon	0.52	0.75	0.24	0.10
Harcourt	0.18	0.09	0.31	0.33
Houghton Mifflin	0.10	0.07	0.11	0.23
Scott-Foresman	0.07	0.06	0.07	0.13
Grade 3				
Saxon	0.51	0.74	0.22	0.10
Harcourt	0.18	0.09	0.31	0.33
Houghton Mifflin	0.11	0.07	0.14	0.23
Scott-Foresman	0.08	0.07	0.06	0.17
Grade 4				
Saxon	0.49	0.71	0.20	0.13
Harcourt	0.18	0.10	0.29	0.33
Houghton Mifflin	0.12	0.09	0.13	0.20
Scott-Foresman	0.09	0.07	0.11	0.17
Grade 5				
Saxon	0.51	0.73	0.20	0.17
Harcourt	0.17	0.09	0.30	0.30
Houghton Mifflin	0.10	0.07	0.11	0.20
Scott-Foresman	0.10	0.07	0.11	0.17
Grade 6				
Saxon	0.31	0.45	0.14	0.07
Glencoe	0.24	0.20	0.24	0.40
McDougal	0.15	0.11	0.18	0.23
Prentice Hall	0.10	0.11	0.04	0.17
N*	207	122	55	30

*N indicates the number districts in our primary sample for which we have data on adoptions for 2004.

Table 7. Falsification Estimates of Math Curricular Effectiveness, Estimated Using Math Test Scores for Grade-3 Cohorts Who Were Never Exposed to the Curricula of Interest (Fixed-Bandwidth Matching Estimators). All Comparisons.

	1992	1993	1994	1995	1996	2007	2008
Treatment: B Control: A							
Kernel Matching	-0.207	-0.02	-0.014	0.046	0.073	0.026	0.163
	(.144)	(.180)	(171)	(.163)	(.150)	(.159)	(.134)
Treatment: C Control: A							
Kernel Matching	-0.432	-0.135	-0.009	0.050	-0.072	0.065	0.025
	(.165)**	(.187)	(.163)	(.158)	(.204)	(.162)	(.208)
Treatment: C Control: B							
Kernel Matching	-0.105	0.122	0.145	0.018	-0.064	-0.126	-0.236
	(.294)	(.325)	(.325)	(.310)	(.313)	(.244)	(.264)
N(A)	301	304	304	306	308	284	280
N(B)	209	210	213	216	220	205	201
N(C)	179	179	182	182	182	163	162

Notes: Matching estimators impose the common support restriction. Standard errors are clustered at the district level and bootstrapped using 250 repetitions.

** Denotes statistical significance at the 1 percent level or better

* Denotes statistical significance at the 5 percent level or better

† Denotes statistical significance at the 10 percent level or better

Table 8. Falsification Estimates of Math Curricular Effectiveness, Estimated Using Math Test Scores for Grade-6 Cohorts Who Were Never Exposed to the Curricula of Interest (Fixed-Bandwidth Matching Estimators). All Comparisons.

	1992	1993	1994	1995	1996	1999	2000	2001
Treatment: B Control: A								
Kernel Matching	-0.211	-0.291	-0.058	-0.110	0.040	-0.021	-0.259	-0.143
	(0.192)	(0.220)	(0.193)	(0.180)	(0.187)	(0.177)	(0.185)	(0.165)
Treatment: C Control: A								
Kernel Matching	0.195	0.106	0.243	0.397	0.361	0.019	-0.382	-0.288
	(0.216)	(0.283)	(0.207)	(0.209)†	(0.251)	(0.246)	(0.216)†	(0.207)
Treatment: C Control: B								
Kernel Matching	-0.081	-0.040	-0.143	-0.150	-0.235	-0.336	-0.271	-0.247
	(.251)	(.307)	(.320)	(.289)	(.271)	(.318)	(.307)	(.309)
N(A)	205	208	213	213	218	212	205	204
N(B)	117	118	122	125	127	122	120	120
N(C)	90	89	93	95	101	79	78	76

Notes: Matching estimators impose the common support restriction. Standard errors are clustered at the district level and bootstrapped using 250 repetitions.

** Denotes statistical significance at the 1 percent level or better

* Denotes statistical significance at the 5 percent level or better

† Denotes statistical significance at the 10 percent level or better

Table 9. Falsification Estimates of Math Curricular Effectiveness, Estimated Using Math Test Scores for Grade-8 Cohorts Who Were Never Exposed to the Curricula of Interest (Fixed-Bandwidth Matching Estimators). All Comparisons.

	1992	1993	1994	1995	1996	1999	2000	2001	2002	2003
Treatment: B Control: A										
Kernel Matching	0.159	0.005	-0.064	-0.046		0.135	0.050	0.022	-0.022	-0.026
	(.181)	(.174)	(.187)	(.212)		(.170)	(.172)	(.184)	(.157)	(.161)
Treatment: C Control: A										
Kernel Matching	0.100	0.004	0.071	0.217		0.065	0.046	0.268	0.148	0.251
	(.263)	(.221)	(.243)	(.227)		(.246)	(.238)	(.244)	(.267)	(.298)
Treatment: C Control: B										
Kernel Matching	0.098	0.027	0.097	0.008		0.103	0.029	-0.083	-0.140	-0.005
	(.271)	(.288)	(.233)	(.287)		(.352)	(.365)	(.308)	(.286)	(.342)
N(A)	142	145	146	145		146	144	141	139	138
N(B)	80	79	81	82		79	79	79	79	78
N(C)	64	65	66	65		67	66	67	67	67

Notes: Matching estimators impose the common support restriction. Standard errors are clustered at the district level and bootstrapped using 250 repetitions. Information on grade-8 test scores in 1996 were not available.

** Denotes statistical significance at the 1 percent level or better

* Denotes statistical significance at the 5 percent level or better

† Denotes statistical significance at the 10 percent level or better

Table 10. Estimates of Math Curricular Effectiveness, Estimated Using Reading Test Scores for all Grade-3 Cohorts (Fixed-Bandwidth Matching Estimators). All Comparisons.

	1992	1993	1994	1995	1996	1999	2000	**2001**	**2002**	**2003**	**2004**	2005	2006	2007	2008
Treatment: B Control: A															
Kernel Matching	-0.238	-0.139	-0.087	0.026	0.161	0.107	0.104	**0.170**	**0.205**	**0.219**	**0.007**	0.032	0.036	-0.149	0.100
	(0.145)	(0.179)	(0.175)	(0.166)	(0.157)	(0.165)	(0.191)	**(0.175)**	**(0.136)**	**(0.177)**	**(0.188)**	(0.190)	(0.157)	(0.185)	(0.160)
Treatment: C Control: A															
Kernel Matching	-0.212	-0.174	-0.178	-0.037	-0.12	-0.060	-0.015	**0.097**	**0.078**	**0.196**	**0.065**	-0.003	0.263	0.211	0.200
	(0.172)	(0.190)	(0.181)	(0.182)	(0.217)	(0.248)	(0.260)	**(0.225)**	**(0.216)**	**(0.213)**	**(0.211)**	(0.220)	(0.204)	(0.219)	(0.229)
Treatment: C Control: B															
Kernel Matching	0.016	0.167	0.262	-0.008	-0.122	-0.227	-0.131	**-0.143**	**-0.181**	**-0.152**	**-0.123**	0.027	0.134	0.063	-0.051
	(0.303)	(0.333)	(0.354)	(0.316)	(0.351)	(0.299)	(0.318)	**(0.338)**	**(0.287)**	**(0.295)**	**(0.290)**	(0.276)	(0.297)	(0.299)	(0.318)
N(A)	301	304	304	306	308	309	307	**307**	**305**	**300**	**294**	286	287	284	280
N(B)	209	210	213	216	220	220	219	**219**	**213**	**213**	**212**	210	207	205	201
N(C)	179	179	182	182	182	184	182	**182**	**181**	**176**	**174**	169	163	163	162

Notes: Matching estimators impose the common support restriction. Standard errors are clustered at the district level and bootstrapped using 250 repetitions.

** Denotes statistical significance at the 1 percent level or better

* Denotes statistical significance at the 5 percent level or better

† Denotes statistical significance at the 10 percent level or better

Table 11. Persistence Effects. Estimated Curricular Effects for Cohorts of Grade-6 and Grade-8 Students that were Partially or Fully Exposed.

	Grade-6 Cohorts							Grade-8 Cohorts				
	2002	2003	**2004**	**2005**	**2006**	**2007**	2008	2004	2005	**2006**	**2007**	**2008**
Treatment: B Control: A												
Kernel Matching	-0.114	0.092	**0.071**	**0.051**	**-0.020**	**-0.005**	-0.002	-0.083	-0.042	**-0.126**	**0.048**	**-0.009**
	(0.171)	(0.154)	**(0.212)**	**(0.181)**	**(0.191)**	**(0.208)**	(0.194)	(0.203)	(0.203)	**(0.181)**	**(0.224)**	**(0.170)**
Treatment: C Control: A												
Kernel Matching	-0.236	0.030	**0.130**	**-0.024**	**-0.014**	**0.103**	-0.126	0.180	0.323	**0.372**	**0.186**	**0.409**
	(0.197)	(0.250)	**(0.251)**	**(0.280)**	**(0.229)**	**(0.218)**	(0.275)	(0.289)	(0.252)	**(0.304)**	**(0.220)**	**(0.268)**
Treatment: C Control: B												
Kernel Matching	-0.504	-0.281	**-0.114**	**-0.127**	**-0.153**	**-0.041**	-0.279	0.282	0.033	**0.085**	**0.080**	**0.077**
	(0.294)†	(0.243)	**(0.285)**	**(0.243)**	**(0.239)**	**(0.244)**	(0.224)	(0.289)	(0.282)	**(0.272)**	**(0.285)**	**(0.294)**
N(A)	200	189	**174**	**165**	**163**	**160**	156	135	135	**132**	**131**	**131**
N(B)	118	115	**105**	**101**	**97**	**94**	93	75	75	**75**	**73**	**71**
N(C)	75	73	**72**	**72**	**72**	**72**	67	67	66	**42**	**66**	**66**

Notes: Matching estimators impose the common support restriction. Standard errors are clustered at the district level and bootstrapped using 250 repetitions.

** Denotes statistical significance at the 1 percent level or better

* Denotes statistical significance at the 5 percent level or better

† Denotes statistical significance at the 10 percent level or better

Appendix A
Supplementary Tables

Appendix Table A.1. Data Sample Details.

	Schools	% of Universe	Districts	% of Universe
Universe*	1115		294	
Missing Information:				
District-reported curriculum adoption	3	0.3	3	1.0
District outcome variables	2	0.2	2	0.7
School outcome variables	23	2.2	1	0.3
District finance or enrollment data	2	0.2	1	0.3
School enrollment or demographic data	82	7.3	12	4.0
Did not use one of the primary curricula in grade one, two or three	211	18.9	38	12.9
Used only primary curricula, but did not uniformly adopt	76	6.8	24	8.2
Remaining Sample	*716*	*64.2*	*213*	*72.4*

* The universe consist of those schools and districts for which any information was reported in 1997, and at least one grade-3 math test score was reported for an exposed cohort (1999-2006).

Appendix Table A.2 Scaling Factors Used to Convert Estimation Metric from School-Level Distribution to Individual-Level Distribution for Grade-3 Math Scores.

Year	Standard Deviation of Distribution of School Scores	Standard Deviation of Distribution of Individual Scores	Approximate Scaling Factor
1992	2.8	N/A	N/A
1993	2.9	N/A	N/A
1994	2.8	N/A	N/A
1995	2.8	N/A	N/A
1996	1.9	N/A	N/A
1999	21.3	N/A	N/A
2000	20.5	61.0	0.34
2001	21.0	61.4	0.34
2002	19.9	59.7	0.33
2003	20.7	60.9	0.34
2004	22.5	63.1	0.36
2005	21.0	62.2	0.34
2006	20.0	64.3	0.31
2007	21.3	65.4	0.33
2008	22.5	63.7	0.35

Appendix B
Bandwidth Selection

We use standard leave-one-out cross validation (C-V) to obtain fixed bandwidths for the kernel and LLR matching estimators (the locally-varying bandwidth selection is also based on the fixed bandwidths). The grid search for kernel and LLR matching is over the range (0.005, 2.0). Using Frölich's (2004) notation, the C-V approach selects the optimal bandwidth, h_{CV}, by solving the following minimization problem:

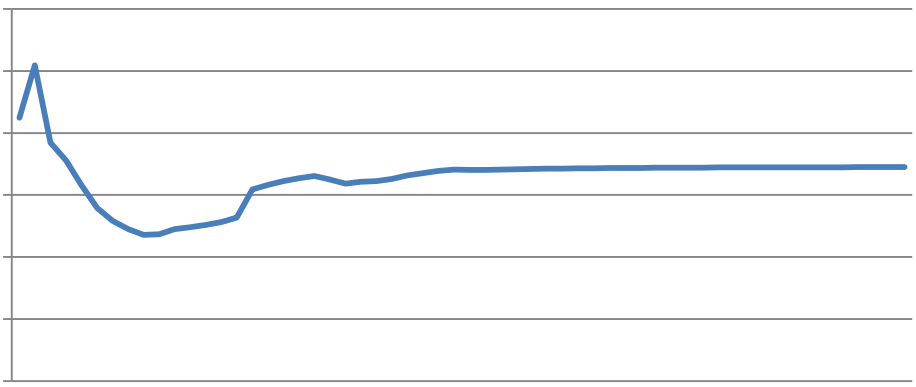
$$h_{CV} = \arg\min(h) \sum_{q=1}^{Q} (Y_q - \hat{m}_{-q}(p_q))^2$$

where q indexes the sample of control units, Y is the outcome (test score) and $\hat{m}_{-q}(p_q)$ is the estimate of the mean outcome among the control observations, excluding observation q, conditional on the estimated propensity score for unit q.

As has been reported in other contexts (see, for example, Ludwig and Miller, 2007), the loss function used to select the fixed bandwidth is fairly flat in most of our comparisons. As such, we use a combination of conventional C-V and "visual inspection" to identify the appropriate fixed bandwidth for each of our matching estimators.

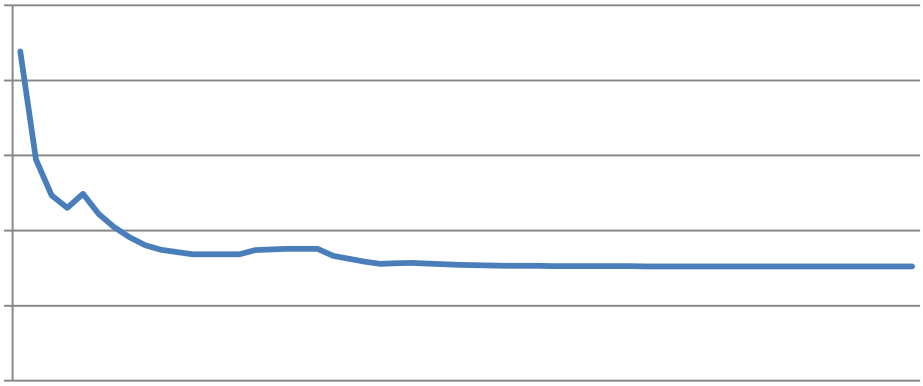
First, Figure B.1 illustrates a case where cross-validation produces a clear bandwidth choice at the global minimum of the loss function, for our comparison between *B* and *A* in 2000 using the kernel matching estimator. In this case we use bandwidth at the global minimum, 0.048.

Figure B.1
Loss Function: Silver-Burdett Ginn versus Saxon, 2000



Next, Figure B.2 illustrates a case where cross-validation suggests an optimal bandwidth at the edge of our grid search, for our comparison between *B* and *A* in 2005 using the kernel matching estimator. For this comparison we use a bandwidth of 0.062, which occurs just prior to the narrowly upward sloping portion of the curve.

Figure B.2
Loss Function: Silver-Burdett Ginn versus Saxon, 2005



We describe our bandwidth selection procedure for the comparison in Figure B.2 as a combination of cross-validation and visual inspection. Because the flat region of the curve has a mild negative slope, the mechanical application of the C-V procedure would produce a bandwidth of at the edge of our grid search, 2.0. However, by visual inspection, we can see that there is very little difference in the loss function between the bandwidth determined mechanically by the C-V procedure and a much narrower bandwidth selected after the initial drop in the loss function. We ultimately use the narrower bandwidth in this and similar cases because the efficiency gains associated with the wider bandwidth will be minimal, and the narrower bandwidth should reduce bias in the estimates.

Across our grade-3 comparisons spanning the entire data panel, our approach of combining C-V with visual inspection yields a bandwidth at the global minimum of the loss function 40 percent of the time. In the remaining cases where the global minimum occurs at the edge of our grid search, the average increase in the loss function that we observe by choosing an interior bandwidth is 1.3 percent, with a maximum increase of 2.9 percent in one instance. Details about our bandwidth selection process for each estimator in the paper are available upon request.

Finally, that cross validation produces large flat regions in the loss function in most of our comparisons provides some indirect evidence that curriculum adoptions are not meaningfully correlated with other, unobservable determinants of school performance. The flat regions suggest that as increasingly non-comparable units (as measured by the propensity score) are used as comparisons for one another, there is minimal change in their measured outcomes. Such conditions will certainly be favorable to a non-experimental analysis.