Is Curriculum Quality Uniform? Evidence from Florida

Rachana Bhatt
Georgia State University

Cory Koedel
University of Missouri

Douglas Lehmann
University of Michigan

We construct a large panel dataset of schools and districts in Florida to evaluate curricular effectiveness in elementary mathematics. A key innovation of our study is that we allow for curriculum quality to be non-uniform across various mathematics subtopics. We find evidence of variability in curricular effectiveness across different subtopics *within* the same curriculum. Our findings suggest that educational administrators should consider the topical performance of their various curricular alternatives when making adoption decisions.

I. Introduction

Federal and state accountability policies tied to standards-based testing have led to renewed interest in curriculum selection. A series of studies evaluating state-, district- and school-level responses to the 2001 No Child Left Behind Act show that improvement efforts have, in part, been focused on aligning curricula with state standards for learning and assessment (Padilla et al., 2005; Srikantaiah et al., 2009). Because education agencies face increasing accountability requirements, emphasis has been placed on selecting a curriculum that produces large gains in student achievement.[1]

Based on available data we know that most school districts adopt a single curriculum for all elementary schools (by subject).[2] One explanation for the prevalence of the "single-curriculum model" is that it better facilitates mobile students (Tyson, 1997). Single-curriculum adopters may also enjoy cost savings and/or lower administrative costs associated with bulk sales. Furthermore, because multiple elementary schools generally feed into fewer middle and high schools, using the same curriculum materials at all elementary schools likely simplifies the delivery of instruction in later grades. However, a potential problem with adopting a single curriculum in each subject in the early grades is that curriculum quality may not be uniform.[3] That is, a given curriculum may lead to higher achievement in some content areas relative to others.

State standards mandate that students master learning across multiple subtopics within major-subject areas. For instance, in Florida there are five subtopics for grade-3 mathematics: (1)

[1] Throughout the paper, we use the words "curriculum" and "curricula" to identify the set of instructional materials used by a school or district. Some district personnel may define a "curriculum" as a set of goals and objectives that is separate, at least conceptually, from the instructional materials themselves. Our analysis is strictly focused on the former definition.

[2] This is true in our dataset, which includes information on schools and districts in Florida, and in data from Indiana (see Bhatt and Koedel, forthcoming). In addition, one of the coauthors of this work has curriculum-adoption data from Missouri, which shows the same result. We are not aware of any other state-level curriculum-adoption data.

[3] Of course, this is also likely to be true in higher grades. There is currently little evidence on curriculum adoptions in higher grades. We conducted an informal review of adoptions in higher grades in Indiana (pre-algebra and algebra I) and found that there are fewer single-curriculum adopters, but there are still many. Improving our understanding of how curriculum adoptions occur in schools and districts for higher grades is an important area for future research.

number sense, (2) measurement, (3) algebraic thinking, (4) data analysis, and (5) geometry. A particular grade-3 curriculum may produce large gains in say, data analysis and geometry (higher order skills) relative to another curriculum, but produce smaller gains in numbers sense and measurement. This curriculum would be of particular value for high-achieving students who have already mastered the simpler mathematics subtopics, but may be less useful for students who are still learning basic concepts.

The purpose of our study is to empirically investigate the potential for curriculum quality to be non-uniform. This line of inquiry is important for administrators and other stakeholders involved in the curriculum-adoption process. Knowledge about how curricula affect student performance in particular subtopics within major subjects, beyond simply understanding performance differences overall, can help administrators and other stakeholders to make better curriculum-related choices. Diverse districts, for example, might be encouraged to move away from the single-curriculum model if curriculum effects are sufficiently non-uniform. Topical information on curriculum quality will also be useful to guide the use of supplemental instructional materials.

We construct a nine-year data panel of schools and districts in Florida for our evaluation. We focus on the elementary-mathematics curricula that were first used in Florida during the 2004-05 school year and remained in use through 2009-10. Our measure of achievement is the Florida Comprehensive Assessment Test (FCAT) in mathematics, for which we observe school-level achievement on the entire test as well as achievement on the five mathematics subtopics listed above. We estimate curriculum effects on test scores in each subtopic to evaluate the uniformity of curriculum quality.

Our analysis shows that curriculum effects can be non-uniform. The most-popular curriculum in Florida during the adoption cycle that we study excelled in two of the five subtopics on the mathematics portion of the FCAT, but was indistinguishable from other curricula in the

other three topics. To the best of our knowledge, our results provide the first empirical evidence showing that curriculum quality can be non-uniform. A general implication of our findings is that better information about curricular effectiveness *within* subject areas can aid education agencies in making better curriculum-related decisions.

II. Background

The Curriculum Selection Process in Florida

Curriculum materials in Florida are adopted for specific subjects and grades on a rotating schedule. Each adoption cycle lasts for six years. Curriculum adoptions in mathematics for grades K-12 occurred in 1998-99, 2004-05 and 2010-11. We focus our analysis on the curriculum adoption that occurred in 2004-05 (hereafter, the 2004 adoption cycle).[4] The adoption process in Florida is centrally organized by the state and occurs for all districts at the same time. Administrative tasks such as the purchase of curriculum materials are officially performed by the state on behalf of districts (National Association of State Textbook Administrators, 2012).[5] The adoption process for the 2004 cycle began in 2001 with the state Department of Education (DOE) issuing standards for student learning (referred to as the *Sunshine State Standards*) in each grade-cluster in mathematics (i.e. K-2, 3-5, etc.). The DOE then placed a call for bids from publishers. Along with curriculum materials, publishers were required to submit "correlation reports" detailing where in the proposed curriculum each of the standards for learning were met. Publisher bids for the 2004 cycle were submitted to the state by June 2003.

During the spring of 2003, the DOE formed a State Instructional Materials Committee (SIMC) comprised of math teachers, administrators, school board members and community

[4] We evaluate the 2004 adoption cycle because curriculum-adoption information from the prior cycle is unavailable, and, of course, we do not have outcome data from the cohorts of students who will use the 2010 curricula.

[5] As reported by Bhatt and Koedel (forthcoming), roughly half of states use a partially-centralized adoption model and the other half are completely decentralized. The adoption process can change over time within states; for example, Indiana has gone from a process similar to the process in Florida to a decentralized process in recent years (State of Indiana House Bill No. 1429).

members. Through the fall, the committee evaluated the curriculum submissions and held public meetings where audience members were given the opportunity to review and comment on the curricula. After a series of meetings, the SIMC voted on whether to approve each curriculum. The official state-approved list was available by the winter of 2003.

In January of 2004 the state-approved list and copies of the curriculum materials were circulated to school districts for review. At this point the review process was decentralized to the district level. The typical district review involved teachers, administrators, and community members. For instance, in the St. Johns School District there were elementary, middle, and high school committees that were formed to review the curricula and then report their recommendations to school principals. Principals from across the district met and agreed on which curriculum they would like to adopt for each grade and subject. The involvement of multiple stakeholders in the adoption process is reflective of the feature that district curriculum choices "generally take into consideration many local factors" (Levin, 2006).

Single Curriculum Adopters

For our empirical analysis we estimate the effects of the elementary-mathematics curricula on grade-3 FCAT scores (students first take the FCAT in grade-3). Grade-3 test scores are presumably a function of the curricula to which students are exposed in grades 1, 2 and 3. To identify the treatment effects of the curricula, therefore, we exclude districts that used more than one curriculum in these grades from our study. We refer to districts that purchased a single curriculum at all schools across grades 1, 2 and 3 as "single-curriculum adopters" per the preceding discussion. Appendix Table A.1 shows that restricting our attention to single-curriculum adopters reduces our district sample size by 21 percent and our school sample size by 24 percent. The finding that most districts are single-curriculum adopters in Florida is consistent with earlier findings from Indiana (Bhatt and Koedel, forthcoming). The Florida results are perhaps more interesting because

Florida districts are among the largest in the country – even so, most adopt a single curriculum for all schools in the early grades.[6]

Table 1 shows the numbers of schools and districts that adopted each curriculum in our final dataset. As can be seen in the table, the Harcourt curriculum was the most popular curriculum in the state during the 2004 adoption cycle – it was used by 28 of the 45 single-curriculum adopters. No other curriculum comes close in terms of usage, with the remaining 17 districts using one of six other curricula. Only one school district in the final sample chose to obtain a waiver to use a curriculum that was not on the state-approved adoption list.[7]

Although data on curriculum adoptions across states are limited (Chingos and Whitehurst, 2012), national market-share data indicate that a number of the curricula listed in Table 1 were commonly in use in the United States during the time of our study. For example, in 2004 Harcourt Math had a 40% market share in states where market-share information can be tracked (Reed Elsevier, 2004). Scott Foresman Investigations and Scott Foresman Addison Wesley also had sizeable market shares in the early grades (Agodini et al. 2010). Moreover, the curricula that we study share the same pedagogical, content and stylistic features as other popular elementary-level mathematics curricula that are used nationwide (see discussion below).

Because only the Harcourt curriculum is individually well-represented in the data we do not perform one-to-one curriculum comparisons. Instead, we compare the Harcourt curriculum to a "composite alternative" comprised of the other six curricula. Our estimates, therefore, should be

[6] There are four districts in Florida where at least 85 percent of purchase orders were for the same curriculum. In an unreported robustness analysis we include these districts in our models and treat them as single-adopters of the most-used curriculum. Our results are qualitatively insensitive to including these districts and their schools (65 schools in total). We also exclude schools that are designated as "non-regular" by the DOE, but we do not exclude charter schools because they are not always easily identified in the data. If charter schools make different curriculum choices than the regular public schools in their districts, this will result in treatment misclassification, and attenuate our estimates. In a robustness test we dropped any school that could be identified as a charter school based on its name (approximately 5 percent of our sample), and re-estimated our models. Our findings were unaffected qualitatively.

[7] This is not unique to Florida. Bhatt and Koedel (forthcoming) find that very few districts in Indiana requested a waiver during that state's mathematics adoption in 1998.

interpreted as showing the effects of using the Harcourt curriculum *relative to the weighted-average effects of the alternative curricula in Table 1*.[8] Despite our inability to perform one-to-one comparisons, the data are still well-suited for our analysis. In particular, we can compare the Harcourt curriculum – which was and remains widely popular – to a composite of alternatives, and examine the extent to which the quality of the Harcourt curriculum is non-uniform.

Curriculum Descriptions

In this section we briefly describe the salient features of the curricula listed in Table 1. We draw information from numerous sources including the Institute of Education Sciences, publisher websites, and independent curriculum reviews (as cited throughout). A key dimension along which curricula are typically compared is pedagogical – that is, whether the curricula are traditional- or reform-oriented. Traditional curricula emphasize teacher-led instruction, specific problem solving methods, and rely on "drill-based" learning. Reform-based curricula emphasize student inquiry, real-world application, and often incorporate technology. Four of the curricula in our study, including Harcourt, are best-described as traditional; the other three are reform-based.

The traditional/reform distinction has received much attention in policy circles and from the general public, but there are also other notable differences between curricula. One distinction is that some curricula use a "massed" approach and others use a "distributive" approach in their topical coverage. In the former topics are taught in self-contained units, whereas in the latter each topic is taught at multiple points over the course of the year. All but one of the curricula in Florida follows a massed approach to instruction (SRA/McGraw Hill).

Our comparative reviews shed light on the differences and similarities between the curricula included in our evaluation along these and other dimensions. It is important to recognize that our

[8] While this is a clear limitation of our project, we are not aware of any alternative datasets where our analysis could be performed. To the best of our knowledge, Florida is the only state where information on both curriculum adoptions and topic-specific test scores is available. Our study is not the first to use the "composite alternative" evaluation framework (e.g., see Resendez and Manley, 2005).

analysis is not able – nor intended – to pinpoint the particular aspects of different curricula that determine performance. For instance, below we show that Harcourt outperforms the composite alternative in geometry and data analysis. However, we cannot attribute the performance difference to any single aspect of the Harcourt curriculum. Rather, our focus is on examining how different "whole-package" curricula affect student learning. The curriculum effects will reflect any differences in pedagogy, content, organization of material, and time spent on instruction inherent to the curricula. Whole-package curriculum effects are of direct relevance because education agencies typically do not custom-build curricula. Documenting whole-package effects represents an important first step in developing a research basis (Chingos and Whitehurst, 2012). That said, we note that our findings can serve as a platform for more in-depth analyses that aim to pinpoint the specific aspects of the different curricula that lead to differences in student achievement.[9]

Harcourt Math is a traditional curriculum that uses a massed approach to instruction (Harcourt Publisher, 2012; Xin et al., 2011). Teachers provide direct instruction and then students practice using worksheets and/or computer programs. Teacher-led, whole-class instruction plays a prominent role in the Harcourt curriculum. Teachers provide a single method for problem solving and students replicate the teacher's approach. The Harcourt curriculum has been characterized as having a pace of instruction and opportunities for practice that work well for students who have below average math performance (Nelson et al., 2007).

Scott-Foresman Addison Wesley (SFAW) also uses a massed approach to instruction and is defined as a traditional curriculum (Slavin and Lake, 2007), but it includes some reform aspects as well. Teachers intersperse instruction with student practice. Problems in this curriculum are "real-world" oriented, and this curriculum uses a variety of tools such as transparencies and workbooks

[9] This is an example of how quantitative and qualitative work in this area can be greatly complementary.

for instruction.[10] This curriculum has been described as useful for students of heterogeneous ability levels (Agodini et al., 2009; Resendez and Manley, 2005; WWC, 2010).

Every Day Mathematics from SRA/McGraw Hill uses a distributive approach and fits in the reform mold of instruction (Slavin and Lake, 2007).[11] Students are taught using an "exploratory" approach where they first inquire about a topic and then apply their knowledge using games and interactive problem-solving. Students are exposed to multiple methods for solving problems, and teachers provide guidance as students work in small groups. This curriculum has been characterized as being challenging for students who are average or above average (Nelson et al., 2007).

Cambium is a reform curriculum that uses a massed approach to instruction. It is targeted towards at-risk or special education populations. Lessons are composed of two parts, "Anchors" and "Excursions." In the former, students are given structured, explicit instruction on mathematical concepts, with a strong emphasis on math vocabulary and pictures. "Excursions" are activities that include real-world connections. There is a "home learning" component to this curriculum which is designed to promote family discussions (Cambium Publishers, 2012).

Scott Foresman Investigations is a reform curriculum that uses a massed approach to instruction (Agodini et al., 2009; Slavin and Lake, 2007). It minimizes the role of algorithms, and teaches students that there are multiple solutions to a problem. Discussion is heavily emphasized; once teachers introduce a topic, they spend much of their time guiding students who are working in pairs

[10] Prior research indicates that using visual aids for teaching is an effective instructional practice (Mayer and Anderson, 1992; Mayer, 2001).

[11] The distributed approach has been argued by some to benefit students by increasing the amount of information they retain and understand (Bloom and Shuell, 1981; Rea and Modigliani, 1985).

or small groups (Agodini et al., 2009; WWC, 2009).[12]

MacMillan/McGraw Hill uses a massed approach to instruction and is considered to be a traditional curriculum (Slavin, 2005), but it too contains reform elements. Instruction is structured so that students develop intuition first, and then teachers introduce formal concepts. The curriculum focuses on "real-world" applications coupled with direct instruction on how to solve problems. Students are given multiple opportunities for practice, and technology is used in instruction (MacMillan/McGraw Hill Publishers, 2012; Idaho Department of Education, 2006).

Houghton Mifflin: is a traditional curriculum that uses a massed approach to instruction (Slavin and Lake, 2007; Houghton Mifflin Publishers, 2012). Each lesson is structured in three steps: First, teachers introduce the material and then guide students through a practice problem, and following that students practice individually using worksheets, or computer software. Math vocabulary lessons are a core component of the Houghton-Mifflin curriculum. It has been described as appropriate for students at all ability levels (WWC, 2007).

III. Data

Panel of Florida Schools and Districts

We construct a data panel of schools and districts in Florida that spans the academic years 2000-01 to 2008-09. The data panel includes information about which math curriculum each district adopted during the 2004 adoption cycle along with details about school- and district-level enrollment shares by race, language status, migrant status, disability status and eligibility for free and reduced-price lunch. We additionally collect district financial information, and supplement the school and district data with year-2000 Census data on local-area median household incomes and

[12] Slavin and Karweit (1985) find that students who were exposed to individualized or (ability) grouped instruction showed higher skill acquisition in math than those who received traditional whole-group instruction.

education levels, linked to schools by zip codes. The census variables control for fixed-area characteristics in our models.

Table 2 divides our data sample into Harcourt and non-Harcourt adopters, and compares the average characteristics for schools and districts across the two groups. The table uses data from the 2002-03 school year, which was two years prior to the adoption cycle that we study. Harcourt and non-Harcourt adopters are similar in some ways (share of limited English proficiency and white and black students) and differ in others (share of free/reduced price lunch students). Our matching procedure, which we describe in detail below, is designed to remove any bias stemming from these observable differences across adopters of the Harcourt and non-Harcourt curricula.

There are two notable aspects of the data panel. First, the large-district structure in Florida is a limitation in terms of our empirical work. Our final dataset includes over 1,200 schools, making it the largest curriculum evaluation of which we are aware in terms of the numbers of schools and students, but the data include just 45 districts and consequently there are only 45 unique sources of variation in curriculum adoptions. This affects the precision of our estimates; however, despite this limitation we retain enough statistical power to formally test for the presence of curriculum-quality uniformity.

The second notable data issue is that exposure to the curricula varies across student cohorts in the data. The curricula that we evaluate were first used in schools during the 2004-05 school year. Denoting cohorts of students by the spring of the year in which they were in grade-3 (i.e. students who were in grade-3 during the 2004-05 school year are denoted 2005), we have data from three cohorts – 2007, 2008, and 2009 – that were exposed to the curricula of interest in grades 1, 2, and 3. The 2005 cohort was exposed to the curricula of interest in grade-3 only, and the 2006 cohort was exposed in grades 2 and 3. Cohorts in earlier years were exposed to curricula from the prior adoption cycle. Information about mathematics adoptions prior to 2004 are not available in Florida.

Therefore, we do not know which curricula these cohorts used in earlier grades. Similarly, the cohorts from 2000 to 2004 were only exposed to the curricula from the prior adoption cycle.

We estimate treatment effects for the 2005-09 cohorts for our main analysis. We use the 2003 and 2004 cohorts to match schools (see Section V), and the 2001 and 2002 cohorts to perform falsification tests (see Section VI). The estimates for the 2007, 2008 and 2009 cohorts provide the cleanest estimates of curriculum effects because the students in those cohorts used the curricula of interest in all three grades. Our estimates for the 2005 and 2006 cohorts may be confounded by curriculum quality from the previous adoption cycle. A similar concern arises with the 2001 and 2002 cohorts that we use for the falsification exercise. We return to this issue in Section VI.

The Florida Comprehensive Assessment Test

We examine curricular effectiveness using the Florida Comprehensive Assessment Test (FCAT). The FCAT was administered annually to students in grades 3-10 in mathematics and reading throughout our data panel. We use data from the mathematics assessment for grade-3 from 2004-05 through 2008-09 for our main analysis. We use the mathematics assessment from the earlier years, and the reading assessment from all years, for our falsification exercise.

The FCAT is a norm-referenced test that measures student progress toward achieving the *Sunshine State Standards*. The mathematics assessment is structured to measure achievement in five areas, where the values in parentheses represent the emphasis placed on each of these subtopics in the grade-3 assessment: (1) Number Sense, Concepts, and Operations-students are tested on operations and using numbers (30%), (2) Measurement-students are tested on units of measurement and conversion (20%), (3) Geometry and Spatial Sense-students are tested on coordinate geometry and analyzing shapes (17%), (4) Algebraic Thinking-students are tested on using equations, inequalities, and graphs (15%), and (5) Data Analysis and Probability-students are tested on analyzing data, making predictions and conclusions (18%). For each school and year of our data

panel we know how students performed on the FCAT overall, as well as how they performed on each of the five subtopics.

IV. Research Design

Methodological Overview

Our empirical approach closely follows Bhatt and Koedel (forthcoming). We estimate the effects of the Harcourt curriculum relative to the composite alternative using school-level matching estimators. The key assumption under which matching will return causal estimates of curriculum effects is conditional independence. Conditional independence requires that potential outcomes are independent of the curriculum uptake decision conditional on observable information. To illustrate, define the Harcourt curriculum as curriculum *A* and the composite alternative as curriculum *B*. Denoting potential outcomes by $\{Y_A, Y_B\}$, curriculum treatment options by $T \in \{A,B\}$, and X as a vector of observable school- and district-level covariates, conditional independence can be written as:

$$Y_A, Y_B \perp T \mid X \tag{1}$$

Conditional independence will not be satisfied if there is some information unobserved by the econometrician that is used in the district's decision to adopt a particular curriculum, and this information is also correlated with student achievement. Examples of potential confounders include unobserved student ability, administrator quality, and differential emphasis placed on test-score performance across schools and districts. We discuss these possibilities in more detail below.

Following Rosenbaum and Rubin (1983) we match schools using an estimated propensity score. The propensity score for each school indicates the probability that it used the Harcourt curriculum as predicted by the set of conditioning variables (i.e., *X*). We use kernel matching to estimate average treatment effects. Kernel matching estimators construct the match for each "treated" school using a weighted average over multiple "control" schools, and vice versa. We use

the Epanechnikov kernel to assign the weights. Defining P_A and P_B as estimated propensity scores for schools that actually chose *A* and *B*, respectively, the Epanechnikov kernel function is a symmetric function that puts the highest weights on comparison observations where $|P_A\text{-}P_B|$ is close to zero. We estimate the average treatment effect of using *A* relative to using *B*, $ATE_{A,B}$, by the following formula:

$$\hat{\theta}_{A,B} = \frac{1}{N^S}[\sum_{a \in N^A \cap S_p} \{Y_a - \sum_{b \in I_{0a} \cap S_p} W(a,b)Y_b\} - \sum_{b \in N^B \cap S_p} \{Y_b - \sum_{a \in I_{0b} \cap S_p} W(b,a)Y_a\}] \qquad (2)$$

In (2), N^S indicates the total number of schools that chose *A* or *B* on the common support, S_p, and N^A and N^B indicate the numbers of adopters of *A* and *B*. I_{0a} indicates the set of schools that chose *B* in the neighborhood of observation *a,* and I_{0b} indicates the set of schools that chose *A* in the neighborhood of observation *b*. The neighborhoods are defined based on the estimated propensity scores and determined using a bandwidth parameter that we obtain via conventional leave-one-out cross validation (see Appendix B for details about our cross-validation procedure). Y_a and Y_b are outcomes for treated and control schools, respectively, and *W(a,b)* and *W(b,a)* are weighting functions that weight each comparison school. The weights on comparison observations that are outside of the bandwidth for each school are set to zero, and within the bandwidth the weights are assigned based on the kernel function.[13]

Our matching estimators condition on all of the observable information in Table 2. As we show below, after matching the differences in observables between the treatments and controls in our study are very small. However, any unobserved differences across schools and districts will introduce bias if the conditional independence assumption is violated. *A priori* there are reasons to expect unobserved selection to play only a limited role in our analysis. For example, while different curriculum adopters may have student populations that differ in unobservable ways, it seems

[13] Our results are robust to alternative matching estimators. See Section VI for more detail.

unlikely that there will be significant differences in underlying student ability or motivation across large groups of students (i.e., schools) once we condition on group-level prior test scores and other information. Another possibility is that curriculum selection is correlated with administrator quality. However, the curriculum selection process is complex and involves a number of actors with different interests, and thus it is plausible that the correlation between administrator quality and curriculum choice is small. That said, beyond simply asserting our expectation that there is a limited scope for bias from unobserved selection in our analysis, in Section VI we provide empirical evidence in the form of a series of falsification tests. These tests do not provide any indication that our main findings are biased by unobserved selection.

Why Match at the School Level?

Schools are the most disaggregated level at which sub-topical achievement data is available in Florida. This is an important practical reason to perform our analysis at the school level. However, in principle we could also perform a district-level analysis. We use school-level matching estimators rather than district-level matching estimators for two reasons. First, matching is a data intensive procedure, and constructing matches among a sample of 1,205 schools rather than 45 districts improves match-quality. Second, as described above, individual schools play an important role in the curriculum adoption process. Performing our analysis at the school level allows us to directly control for school-specific features in addition to features of districts. That is, we can match schools on their own characteristics in addition to the characteristics of the larger districts. It is not possible to do the reverse – for example, if we were to match districts it would not be straightforward to control for disaggregated characteristics of schools. Noting these benefits of the school-level approach, it is still important to acknowledge the role that districts play in the adoption process. The fact that schools within a district all move together creates a clustering structure to the data; accordingly, we cluster

our standard errors at the district level throughout the analysis.[14]

The issue of aggregation bias also merits brief attention. Because the school-level data are aggregated up from students, our findings could potentially be subject to aggregation bias. We indirectly test for aggregation bias when we perform our falsification exercise in Section VI. We find no evidence to suggest that aggregation bias is a concern in our study.[15]

V. Matching Procedure and Validation

We estimate the probability that each school uses the Harcourt curriculum (propensity score) using a probit model. The model conditions on the pre-adoption characteristics of schools and districts from the 2002-03 and 2003-04 school years. We match schools in each year of our data panel using this "static" propensity score.[16]

The covariates in the propensity-score model are listed in Table 2 along with their mean values (for brevity we do not report means for 2003-04). We chose these variables to reflect the information that was likely available to decision makers during the curriculum selection process. At the school level, the propensity-score model includes controls for enrollment (proxied by the number of test takers at each school), socioeconomic status (shares of students by race, free and reduced-price lunch status, language status, migrant status, disability status) and grade-3 test scores (in math and reading) from the 2002-03 school year, along with controls for enrollment and socioeconomic status from the 2003-04 school year. At the district level, the model includes enrollment, test-score and finance controls from 2002-03, and enrollment and finance controls from

[14] For a more detailed discussion about the appropriate level of analysis in curriculum evaluations, see Bhatt and Koedel (forthcoming).

[15] Like our main analysis, our falsification exercise uses school-level data. If aggregation bias were driving our results, we should see evidence of this in the falsification results, but we find no indication that this is the case.

[16] The static matches will only introduce bias in our estimates if schools exit our data panel over time in a way that is correlated with curriculum adoptions. Defining a school closing as occurring whenever a school fails to produce outcomes for three consecutive years, or through the end of our data panel, less than two percent of the schools in our data closed (some of these were probably not closings – for example, all schools without reported scores in 2009 are counted as having closed by our definition, but if our data panel were extended, some would surely resurface). There is no evidence that the school closings are correlated with curriculum adoptions.

2003-04. We do not include school- or district-level test scores from the 2003-04 school year in the propensity-score specification because these scores were not available to decision makers while they were making the curriculum-adoption decisions. As noted above, the propensity-score model also includes measures of local-area socioeconomic conditions for each school from the U.S. Census.

Our findings are not qualitatively sensitive to reasonable adjustments to the propensity-score specification, including the addition of the 2004 test scores and/or additional years of lagged test scores. A key reason that we favor the more-parsimonious specification is that it allows us to use as many years of data as possible for our falsification tests.[17]

We match schools based on the estimated propensity scores, and test for balance in the covariates across the treatment and control samples used for estimation.[18] Balancing tests are motivated by Rosenbaum and Rubin (1983). The tests determine whether $X \perp T \mid P(T = A \mid X)$, a necessary condition if the propensity score is to be used to reduce the dimensionality of the matching problem to one. Achieving covariate balance is important for any matching analysis that relies on a propensity score. We test for covariate balance using two different tests. The first test is proposed by Smith and Todd (2005) and is a regression-based test, where we regress each covariate in the propensity score model on a quartic of the estimated propensity score:

$$\begin{aligned} X_k = {} & \beta_0 + \beta_1 \hat{P}(X) + \beta_2 \hat{P}(X)^2 + \beta_3 \hat{P}(X)^3 + \beta_4 \hat{P}(X)^4 \\ & + \beta_5 T + \beta_6 * T * \hat{P}(X) + \beta_7 * T * \hat{P}(X)^2 + \beta_8 * T * \hat{P}(X)^3 + \beta_9 * T * \hat{P}(X)^4 + \varepsilon \end{aligned} \tag{3}$$

In (3), X_k represents a covariate from the propensity-score specification, $\hat{P}(X)$ is the estimated propensity score, and T indicates treatment. We test whether the coefficients $\beta_5 - \beta_9$ are jointly

[17] The matching literature shows that under-specifying the propensity-score model can lead to biased estimates of treatment effects, which is why we are careful to check that our findings are robust to more-detailed specifications (Millimet and Tchernis, 2009).

[18] For brevity we do not report the results from the propensity-score model, but they are available upon request. To provide a sense of the predictive power of the covariates in the model, we estimate a linear-regression model where the dependent variable indicates the adoption of Harcourt, and the independent variables are the covariates from the probit model. The covariates explain roughly 22 percent of the variation in curriculum adoptions.

equal to zero in each regression; conditional on the propensity score, there should be no remaining difference in covariates across treated and control schools.

The second test measures the absolute standardized difference in observables after matching (Rosenbaum and Rubin, 1985). The formula for the absolute standardized difference for covariate X_k is given by:

$$SDIFF(X_k) = \frac{\left| \frac{1}{N^S} \left[\sum_{a \in N^A \cap S_p} \{X_{ka} - \sum_{b \in I_{0a} \cap S_p} W(a,b) X_{kb}\} - \sum_{b \in N^B \cap S_p} \{X_{kb} - \sum_{a \in I_{0b} \cap S_p} W(b,a) X_{ka}\} \right] \right|}{\sqrt{\frac{Var(X_{ka}) + Var(X_{kb})}{2}}} * 100 \qquad (4)$$

We report estimates of the average and median standardized difference across the full set of X's from the propensity-score specification. These estimates can be interpreted as measuring the differences in observables that remain in our comparisons after matching. Rosenbaum and Rubin (1985) suggest that a value of 20 should be interpreted as large.

The results from the balancing tests are reported in Table 3 by comparison and year.[19] Overall, the tests suggest that the covariates are generally well balanced. In the regression tests the average p-values from the F-tests are close to 0.50 in each year, and the standardized-difference estimates are small (the standardized differences are similar in magnitude to estimates in other matching contexts; e.g., see Bhatt and Koedel, forthcoming; Sianesi, 2004). We also provide more-detailed balancing information in Appendix Table A.2, which shows covariate-by-covariate balance for a single year of the data panel. The appendix table shows that the treatment and control schools

[19] There are some fluctuations in our data sample from year-to-year, which is the only reason that the balancing properties in Table 3 change over time (recall that we use a static match). From 2003 onward, the fluctuations are related to a very small number of school closings, and some reporting issues. The primary reporting issue involves small schools that do not make the 10-student cutoff to report scores in all years. Also note that in relative terms, there is a large change in sample size going backward from 2003. The schools where scores went unreported in 2001 or 2002 but were available from 2003 forward do not have an obvious explanation, although there did appear to be more school openings than normal in Florida in the fall of 2002. Neither DOE officials nor the school districts themselves could offer any additional insight on this. Regardless of the cause, the change in sample size raises some concerns about our falsification tests, which use the 2001 and 2002 samples. To ensure that our falsification results are not tilted in our favor by the schools that disappear from the data going backward, we re-estimate our primary models after dropping these schools moving forward as well. Our results are qualitatively unaffected.

are well-matched on key characteristics such as prior achievement and student disadvantage.[20]

In addition to our investigation of covariate balance, we also calculate the divergence between the densities of the estimated propensity scores for treatment and control schools. Density divergence will affect the precision of the estimates obtained from matching: the larger the divergence, the less comparable are treatments and controls. Following Frölich (2004) we measure density divergence using the bidirectional Kullback-Leibler (KL) information criterion (Kullback and Leibler, 1951). We calculate:

$$KL = \int \ln\left(\frac{f_{p|T=A}(P)}{f_{p|T=B}(P)}\right) f_{p|T=A}(P)dP + \int \ln\left(\frac{f_{p|T=B}(P)}{f_{p|T=A}(P)}\right) f_{p|T=B}(P)dP \qquad (5)$$

where P is the probability of choosing A over B, $f_{p|T=A}(P)$ is the density function of P among schools that used curriculum A, and $f_{p|T=B}(P)$ is the analogous density function for schools that used B. A KL-information-criterion measure of zero indicates that the densities are identical, and the measure increases with density divergence.

Figure 1 plots the estimated density functions (we use kernel-density plots based on the Epanechnikov kernel) of the propensity scores for treatment and control schools over the common support, and reports the corresponding KL information criterion, which we estimate to be 0.99. This estimate indicates non-trivial divergence, meaning that there are important differences in observables between treatments and controls in the data. These differences need not introduce bias into the estimates because they are corrected for by the matching procedure; however, they will result in noisier estimates. The problem of limited overlap between treatment and control schools can be hidden by the linear functional form, so our use of matching estimators rather than simple

[20] The largest remaining difference between treatment and control schools after matching is in terms of enrollment. That is, the matching procedure does not fully resolve the baseline enrollment difference shown in Table 2. However, our falsification estimates do not suggest that this is an important area of concern. It is also noteworthy that if enough covariates are considered, some surely will not balance. In aggregate, the balance in the data is in line with what is reported in other similar studies.

linear regression is important (for further discussion, see Black and Smith, 2004). Indeed, in the next section we show that the OLS estimates overstate the precision with which we can compare the curricula in Florida.

VI. Results & Falsification Tests

Results

Table 4 reports our primary results for each grade-3 cohort. For brevity we present kernel-matching estimates only (using the Epanechnikov kernel).[21] We compare the effect of the Harcourt curriculum to the effect of the composite alternative on overall math test scores, and on scores in each subtopic of the FCAT. We also report OLS estimates for the overall effects, which we obtain by regressing test score outcomes on an indicator for using Harcourt and the covariates used in the propensity score model. The standard errors for the matching and OLS estimates are clustered at the district level and the matching-estimator standard errors are bootstrapped using 250 repetitions. All of the matching estimators impose the common support condition.[22]

Focusing first on our most-compelling estimates for the cohorts that were exposed to the curricula of interest in all three grades (2007-09), the estimated effect of Harcourt on overall achievement (relative to the composite alternative) is consistently positive. The OLS and matching estimates are similar in magnitude; however, consistent with the preceding discussion, the OLS estimates are always statistically significant, whereas the matching estimates are too imprecise to make definitive claims. Still, the matching estimates are suggestive of an economically meaningful effect of the Harcourt curriculum relative to the alternative on the whole. For instance, the results for the 2007 cohort indicate that the average effect of using Harcourt instead of the composite

[21] In unreported results we show that our findings are qualitatively robust to alternative estimators such as local-linear-regression and radius matching estimators (using various radii).

[22] The estimates in the table are reported in standard deviations of the student-level distribution of test scores. We convert the estimates from the school-level distribution by multiplying by the ratio of the school-level standard deviation to the student-level standard deviation (using data on students' overall FCAT scores) as in Bhatt and Koedel (forthcoming). In Florida, the scaling factor is approximately 0.40.

alternative in grades 1, 2 and 3 was 0.077 standard deviations of test scores.[23]

Turning to the estimates of primary interest from the topical analysis, we find that the Harcourt curriculum is not of uniform quality. The estimates in Table 4 indicate that Harcourt outperformed the composite alternative in the areas of data analysis and geometry, but was indistinguishable from the composite alternative in the other three subtopics of the test. Guided by this pattern in the point estimates in Table 4, we test the following null hypothesis for the uniformity of curriculum effects:

$$H_0 : \frac{\theta_{NS} + \theta_M + \theta_{AT}}{3} = \frac{\theta_{DA} + \theta_G}{2} \qquad (6)$$

In (6), θ_{NS}, θ_M, θ_{AT}, θ_{DA} and θ_G represent the effects of the Harcourt curriculum (relative to the composite alternative) on numbers sense, measurement, algebraic thinking, data analysis and geometry, respectively.[24] We report the test statistic in the last row of Table 4 for each year-cohort. We reject the null hypothesis at the 10 percent level or better for all three fully-exposed cohorts, confirming that the Harcourt curriculum is non-uniform in its effects across the subtopics of the FCAT exam.[25]

---

[23] We attempted to improve power by averaging the data across years for the fully-exposed and partially-exposed cohorts to reduce sampling variability. The point estimates from the averaged model are similar to what we report in Table 4 for both the matching and OLS estimates, with a modest improvement in precision. Also, recall from Section II that there may be some measurement error in the treatment designations in the data, which we expect to attenuate our estimates to some degree (see footnote 6). This further strengthens the suggestion from Table 4 that Harcourt outperformed the composite alternative. Finally, note that the general similarity between the matching and OLS results is not surprising (also see Lamo and Messina, 2010; Mueser et al., 2007; Reynolds, 2012; Smith and Todd, 2004). Matching is best viewed as a refinement of OLS – the same underlying assumption (selection on observables) is required for identification for both estimators.

[24] Defining $A = \frac{\theta_{NS} + \theta_M + \theta_{AT}}{3}$ and $B = \frac{\theta_{DA} + \theta_G}{2}$, we construct the test statistic as: $\frac{\hat{A} - \hat{B}}{\sqrt{\text{var}(\hat{A}) + \text{var}(\hat{B}) - 2 * \text{cov}(\hat{A}, \hat{B})}}$. We estimate the terms in the denominator by bootstrapping using 250 repetitions (our findings are not sensitive to increasing the number of bootstrap repetitions to as high as 500).

[25] We also considered several other variants of the uniformity test. For example, we constructed a similar test for $H_0 : \frac{\theta_{NS} + \theta_{AT}}{2} = \frac{\theta_{DA} + \theta_G}{2}$. The construction of this test is again guided by our point estimates in Table 4. In particular, note that the estimated Harcourt effects for the measurement subtopic are nominally larger than for numbers sense and algebraic thinking. Consistent with the results reported in Table 4 for our main test, we reject the null hypothesis at the 5-percent level or better for each of the 2007-2009 cohorts using this alternative test. Finally, we also constructed tests

Our topical results add to the literature on the Harcourt curriculum, and raise an important issue with curriculum evaluation more generally. In their 2007 study, for example, Nelson et al. (2007) compared the total achievement effects of the Harcourt elementary mathematics curriculum to Everyday Mathematics and found no difference in overall mathematics achievement across users of the different curricula. While such a result might lead educational administrators to be indifferent between adopting either curriculum, our findings suggest that it would be worthwhile to also consider curriculum performance within specific content areas. In sum, our results indicate that important differences in performance can exist across curricula *within* subjects, and these differences may be concealed in studies that only evaluate overall achievement.[26]

Next we briefly turn to the estimates for the partially exposed cohorts. Recall that because other curricula were used by the 2006 cohort in grade 1 and the 2005 cohort in grades 1 and 2, these results are likely to be attenuated. The degree of attenuation will depend on (i) the extent to which curriculum quality is correlated across adoption cycles within districts, and (ii) the relative importance of curriculum quality in grades 1, 2 and 3 in determining grade-3 test scores. Unfortunately we can only speculate about these issues, but our findings are consistent with some attenuation. Still, the results for the partially exposed cohorts are at least nominally consistent with our main findings for the fully-exposed cohorts.[27]

Finally, recall from above that some of the comparison curricula have more in common with

---

for pairwise comparisons between the coefficients in Table 4. The general pattern of results from those tests is that the comparisons that pit data analysis or geometry against numbers sense or algebraic thinking reject the null (although there are a handful of exceptions from year-to-year). The test statistics reported in Table 4 are a parsimonious representation of the flavor of the findings from the pairwise tests.

[26] We do not mean to suggest that it is not important to evaluate overall differences in curriculum performance; indeed, given the sparseness of the empirical literature on curricular effectiveness this would be a worthwhile first step (Bhatt and Koedel, forthcoming; Chingos and Whitehurst, 2012). Nonetheless, our findings suggest that the question of curriculum quality is more nuanced, and additional information of direct relevance to school districts can be gained by comparing curriculum performance by subtopic.

[27] A final note about the results in Table 4 relates to the issue of "curriculum familiarity", by which we mean the extent to which teachers are familiar with the curricula being used. Our data are poorly suited to speak to the familiarity issue because familiarity effects in our data are confounded by differences in exposure to the curricula across cohorts.

the Harcourt curriculum than others. Although it is an open empirical question as to whether the differences between the curricula that we outline in Section II affect student performance, one might hypothesize that curricula that are similar along these dimensions are more likely to generate similar outcomes for students. Although the Florida data do not permit a thorough investigation of this issue because there are only 17 comparison districts, we provide an exploratory analysis in Appendix C. The appendix shows that when we compare Harcourt adopters to non-adopters that use what appear to be similar curricula, the differences between the Harcourt and non-Harcourt adopters nominally shrink. This is true in terms of the overall point estimates, and the subtopic point estimates as well. These results point to the possibility for even larger differences in curriculum effects, overall and by subtopic, across curricula that differ more in pedagogy and topical organization. If more data become available, this is an obvious area for future research.

Falsification Tests

As discussed above, the matching estimates can only be interpreted causally if the conditional independence assumption holds. Recall that a number of potential issues could lead to a violation of conditional independence, including differences between Harcourt and non-Harcourt adopters in terms of unobserved student characteristics, administrator quality and/or differences in how districts emphasize test-score performance.[28] We use falsification tests to determine whether our findings in Table 4 are likely to be biased by unobserved selection into the different curricula. Our falsification tests estimate a series of curriculum "effects" that, in the absence of bias from unobserved selection, should be zero or near-zero. We perform two types of falsification tests. First, we estimate curriculum "effects" for the two cohorts of students who were never exposed to the

[28] With regard to the latter, a concern would be that districts that place more emphasis on test-score performance may be drawn to the Harcourt curriculum (perhaps because of its emphasis on specific problem-solving methods and drill-based learning) , in which case these districts may also take other actions to improve performance that would bias our findings. The lagged test-score controls are surely helpful with this issue in the matching model, but like other possible confounders, we rely primarily on evidence from the falsification tests to alert us to the possibility that unobserved correlates of curriculum adoptions that are not otherwise accounted for in the matching model are biasing our findings.

curricula that we study. Second, we estimate the "effects" of the math curricula on reading test scores for cohorts of students who did and did not use the curricula. In both cases, if our matching procedure is effectively mitigating selection bias, we should estimate zero or near-zero curriculum "effects."[29]

The falsification estimates are reported in Table 5. The first two columns show estimates from the pre-adoption years in math, overall and by subtopic, and in reading. Recall that we do not observe the curricula to which students from these cohorts were exposed during the 1998 adoption cycle. Curriculum quality is likely to be correlated across adoption cycles to some degree.[30] If differences in curriculum quality from the previous adoption cycle are not captured by the matching model (most likely by the lagged test-score controls) our falsification estimates will be biased away from zero. However, the confounding effects will be small if curriculum quality is not strongly correlated across adoption cycles, and/or if the matching procedure appropriately captures pre-2004 differences in curriculum quality. Indeed, the confounding influence of prior curriculum adoptions on our estimates must be small. We do not observe any statistically significant "effects" in the pre-adoption years, and all the point estimates for the falsification cohorts are close to zero.[31]

Table 5 also reports estimated curriculum "effects" on reading test scores over the course of the entire data panel. The reading estimates do not uncover any evidence to suggest that our results in Table 4 are biased – they are all small and statistically indistinguishable from zero. Overall, the

[29] *Ex ante*, small spillover effects of the math curricula on reading scores cannot be ruled out.

[30] The most-obvious scenario where curriculum quality will be correlated across adoption cycles is when schools and districts select the same publisher across cycles. Publishers do update and change their curriculum offerings; however, it is reasonable to hypothesize that curricula from the same publisher over time will be similar. Harcourt did offer a curriculum in Florida during the 1998 adoption cycle, called Math Advantage. The 2004 Harcourt offering was Harcourt Math. Unfortunately, data on which districts used which curricula in Florida during the 1998 adoption cycle are not available.

[31] In addition to the differences between the subtopic-score point estimates in Tables 4 and 5 being visually apparent, we can also formally test for whether the pattern of subtopic-score coefficients in Table 5 is the same as in Table 4. In 2001 the test suggests that, if anything, the average of the data analysis and geometry coefficients is statistically *smaller* than for the other three subtopics. In 2002 the test for equality between the average for data analysis and geometry and the average across the other three subtopics has a p-value that exceeds 0.95.

falsification estimates in Table 5 provide no evidence to suggest that our primary findings are driven by unobserved differences between Harcourt and non-Harcourt adopters.

In summary, our analysis provides suggestive evidence that the Harcourt curriculum was more effective than the composite alternative curriculum in terms of raising student test scores during the 2004 adoption cycle in Florida, which is encouraging because the Harcourt curriculum was clearly the most popular curriculum in the state.[32] On the issue of non-uniform curriculum effects the evidence in Table 4 is compelling – the Harcourt curriculum clearly excelled in two of the five subtopics of the math exam, but was indistinguishable from the composite alternative in the other three subtopics. We find no evidence to suggest that our findings are driven by unobserved selection into the different curricula.

VII. Conclusion

We present evidence showing that curriculum quality can be non-uniform. The most popular elementary-mathematics curriculum in the state of Florida during the 2004 adoption cycle (Harcourt) was more effective than a composite alternative in two of the five math topics on the FCAT, but indistinguishable from the composite alternative in the other three topics. In this particular case the Harcourt curriculum did not perform worse than the composite in any subtopic; however, in other contexts the non-uniformity of curriculum quality may result in gains for some students at the direct expense of others. Our results raise general concerns about the potential for curriculum quality to be non-uniform, and suggest that educational administrators and other decision makers should be cognizant of this issue when making curriculum-related choices.

Local education agencies, particularly agencies that serve diverse student populations, should carefully consider their options for improving alignment between students and curricula. These

[32] Bhatt and Koedel (forthcoming) find that in Indiana the most commonly-adopted curriculum is the *least* effective in terms of raising test scores. It may not be a coincidence that Florida school districts are more likely to adopt a curriculum that positively influences test scores given the state's strong focus on test-based accountability.

options include moving from the single-curriculum model to a model that uses multiple curricula, lobbying for more customized curriculum material from publishers, and/or supplementing existing curricula with additional instructional materials. But such responses may be premature without more evidence on the uniformity in quality of different curriculum packages. This point leads directly to the larger issue that much remains to be learned about curricular effectiveness. The evidence we provide here comes from the only state in the country where this type of analysis is possible (that is, Florida is the only state where data on curriculum adoptions and topical test scores are available). It would be of great value to know the extent to which curriculum effects are non-uniform for other curricula, and *how* they are non-uniform. More studies along these lines of what we present here are clearly needed. As discussed by Bhatt and Koedel (forthcoming), Chingos and Whitehurst (2012) and Slavin and Lake (2007), the thinness of the empirical literature on curricular effectiveness is striking. An important impediment to the expansion of research is the lack of data – Chingos and Whitehurst (2012) report that Florida is currently the *only* state that centrally collects curriculum adoption information.[33] State longitudinal data systems are well-suited for evaluations like ours if only states would begin to track which curriculum materials are being used in which schools and districts. Without curriculum-adoption data linking schools and districts to the materials that students use in classrooms, an evidence base cannot be assembled.

[33] Bhatt and Koedel (forthcoming) use data from Indiana, but Indiana recently stopped collecting curriculum-adoption data from school districts.

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.

Bhatt, Rachana and Cory Koedel (forthcoming). "A Non-Experimental Evaluation of Curricular Effectiveness in Math," *Educational Evaluation and Policy Analysis.*

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.

Burstein, Leigh. 1980. "The Analysis of Multilevel Data in Educational Research and Evaluation," *Review of Research in Education*," 8, 158-233.

Cambium Learning Publishers. 2012. Accessed online at: http://www.voyagerlearning.com/index.jsp

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

Florida Department of Education. 2012. Statewide Uses of the FCAT. Accessed online at: http://fcat.fldoe.org/pdf/FCATStatewideUses.pdf

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

Hanushek, Eric and Steven Rivkin and Lori Taylor. 1996. "Aggregation and the Estimated Effects of School Resources," *Review of Economics and Statistics* 78 (4) 611-627.

Harcourt Publishers. 2012. Accessed online at: http://www.harcourtschool.com/menus/math2004/math2004_gr3.html

Idaho State Department of Education. 2006. Elementary Math Adoption Guide. Accessed online at: http://www.sde.idaho.gov/adoptionguide/docs/math/Math%20Grade%201%20-%20approved%20listing.pdf

Kullback, Solomon and Richard Leibler. 1951. "On Information and Sufficiency," *Annals of Mathematical Statistics* 22 (1), 79-86.

Lamo, Ana and Julian Messina. 2010. "Formal education, mismatch and wages after transition: Assessing the impact of unobserved heterogeneity using matching estimators," *Economics of Education Review*, 29 (1), 1086-1099. (See Tables 3-5)

Levin, Jesse. 2006. Why Do Some Schools Do Better? Elementary School Curriculum Program and API: A More Detailed Examination. Additional Findings, Curriculum, Ed Source. Accessed online at: http://www.edsource.org/assets/files/OCFindings4-19-06.pdf.

MacMillan/McGraw Hill. 2012. Accessed online at:
http://www.mhschool.com/math/2003/student/index.html

Millimet, Daniel L. and Rusty Tchernis. 2009. "On the Specification of Propensity Scores, with Applications to the Analysis of Trade Policies," *Journal of Business & Economic Statistics* 27 (3), 397-415.

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 Association of State Textbook Administrators. 2012. Accessed online: http://nasta.org/

Nelson, Catherine and Julia Kaufman and Kevin Booker and Brian Gill. 2007. Elementary-Grade Math Programs in Pittsburgh Public Schools: A Comparison of Everyday Mathematics and Harcourt Math, Policy Report, Mathematica Policy Research.

Padilla, Christine and Katrina Woodworth and Andrea Lash and Patrick M. Shields and Katrina G. Laguarda. 2005. Evaluation of Title I Accountability Systems and School Improvement Efforts: Findings From 2002-03, U.S. Department of Education, Office of Planning, Evaluation and Policy Development.

Raudenbush, Stephen W. 1988. "Educational Applications of Hierarchical Linear Models: A Review," *Journal of Education Statistics*, 13(2), 85-116.

Raudenbush, Stephen W. and Anthony S. Bryk. 1986. "A Hierarchical Model for Studying School Effects," *Sociology of Education*, 59(1), 1-17.

Reed Elsevier. 2004. Annual Review and Summary Financial Statements. Accessed online at: http://www.reedelsevier.com/investorcentre/reports%202007/Pages/2004.aspx

Resendez, M. and Manley. M.A. 2005. The relationship between using Saxon Elementary and Middle School Math and student performance on Georgia Statewide assessments, Harcourt Publishers.

Reynolds, Lockwood. 2012. "Where to attend? Estimating the effects of beginning college at a two-year institution," *Economics of Education Review*, 31 (4), 345-362.

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

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

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

Slavin, Robert and Nancy Karweit. 1985. "Effects of Whole Class, Ability Grouped, and Individualized Instruction on Mathematics Achievement" American Education Research Journal. 22 (3), pp. 351-367.

Slavin, Robert. 2005. Evidence Based Reform: Advancing the Education of Students at Risk. Report Prepared for Renewing Our Schools, Securing our Future A National Taksforce on Public Education. Policy Report. Center for American Progress.

Slavin, Robert and Cynthia Lake. 2007. Effective Programs in Elementary Mathematics: A Best-Evidence Synthesis, Best Evidence Encyclopedia.

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

Srikantaiah, Deepa. 2009. How State and Federal Accountability Policies Have Influenced Curriculum and Instruction in Three States, Center on Education Policy.

State of Florida. 2001. Florida's Instructional Materials Specifications for the 2003-2004 Adoption Grades K-8. Florida Department of Education, Office of Instructional Materials.

Tyson, Harriet. 1997. Overcoming Structural Barriers to Good Textbooks. Report for the National Education Goals Panel.

What Works Clearinghouse. 2007. Houghton Mifflin Mathematics. WWC Intervention Report for Elementary School Math.

What Works Clearinghouse. 2009. Investigations in Number, Data and Space. WWC Intervention Report for Elementary School Math.

What Works Clearinghouse. 2010. Scott-Foresman Addison Wesley. WWC Intervention Report for Elementary School Math.

Xin, Yan Ping and Jia Liu and Xiaoning Zheng. 2011. "A Cross-Cultural Lesson Comparison on Teaching the Connection Between Multiplication and Division," *Social Science and Mathematics*. 111 (7), 354-367.

Figure 1. Probability Density Functions over the Common Support for Estimated Propensity Scores, by Treatment Status (Treatment Density = Solid, Control Density = Dashed).

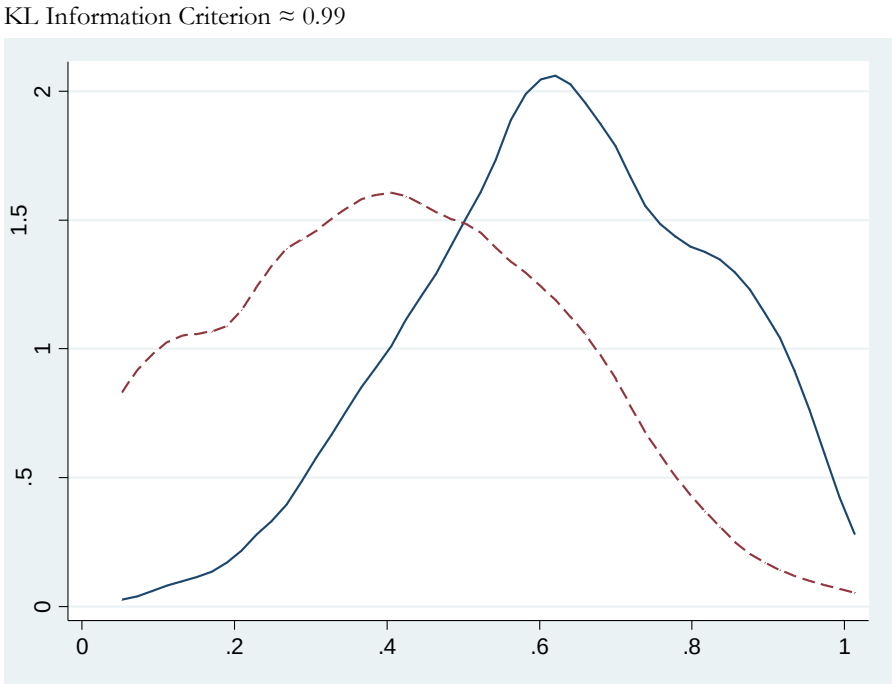


Table 1. Numbers of Schools and Districts that Adopted Each Curriculum in our Sample.

	Schools	Districts
Full Sample	1205	45
Harcourt	636	28
Composite Alternative (All)	569	17
Scott Foresman Addison Wesley	141	4
SRA/McGraw Hill	119	2
Cambium*	108	1
Scott Foresman Investigations	103	1
MacMillan/McGraw Hill	86	4
Houghton Mifflin	12	5

* The Cambium curriculum was not on the DOE's approved list for the 2004 adoption.

Table 2. Average Characteristics of Schools and Districts, by Curriculum Group (2003 values).

	Sample Mean	Harcourt	Composite Alternative
School-Level Outcomes			
Grade-3 Math FCAT Score	305.5	307.4	303.3**
Grade-3 Language FCAT Score	297.8	298.5	297.1
School-Level Characteristics			
Percent Free/Reduced Lunch	54.9	52.8	57.2**
Percent Limited English Proficient	6.3	6.5	6.1
Percent White	52.7	53.8	51.4
Percent Black	27.2	26.5	27.9
Percent Asian	1.9	1.9	1.9
Percent Hispanic	15.1	14.4	15.7
Percent Migrant	2.0	2.0	1.9
Percent Disabled	0.7	0.9	0.5**
Number of Test Takers (logs)	5.7	5.8	5.7**
District-Level Outcomes			
Grade-3 Math FCAT Score	308.4	310.4	306.1**
Grade-3 Language FCAT Score	300.2	301.1	299.1**
District-Level Characteristics			
Number of Test Takers (logs)	10.5	10.4	10.6**
Total Per-Pupil Revenue (logs)	8.9	9.0	8.9**
Census Information (School Level)			
Median Household Income (logs)	10.6	10.6	10.5
Share of Population with Low Education	20.4	19.4	21.6**
N (Schools)	1205	636	569
N (Districts)	45	28	17

Note: 2004 values are included in the propensity score specification for the italicized covariates.
* Indicates statistically significant difference at the 10 percent level or better
** Indicates statistically significant difference at the 5 percent level or better.

Table 3. Balancing Details for the 28 Covariates Included in the Propensity-Score Specification, by Year.

	2001	2002	2005	2006	2007	2008	2009
Average p-value from regression tests	0.54	0.54	0.46	0.48	0.48	0.47	0.48
Number of unbalanced covariates (5/10 percent level)	2/4	3/4	3/6	2/4	2/5	3/5	2/4
Mean Standardized Bias	3.3	3.7	6.1	5.7	5.5	5.1	3.7
Median Standardized Bias	1.9	2.4	3.5	4.3	4.0	3.1	2.1
N (Districts)	45	45	45	45	45	45	45
N (Schools)	1129	1155	1197	1197	1195	1194	1182

Table 4. Average Treatment Effects of Harcourt Relative to the Composite Alternative, Overall and by Math Topic. Effects are Measured in Standard Deviations of Student-Level Test-Score Distribution.

	2005	2006	**2007**	**2008**	**2009**
Overall Score (OLS)	0.057	0.061	**0.093**	**0.085**	**0.066**
	(0.022)**	(0.019)**	**(0.023)****	**(0.020)****	**(0.019)****
Overall Score (Matching)	0.059	0.040	**0.077**	**0.091**	**0.058**
	(0.054)	(0.064)	**(0.058)**	**(0.058)**	**(0.048)**
Topical Scores (Matching)					
Numbers Sense	0.053	0.024	**0.024**	**0.052**	**0.017**
	(0.050)	(0.053)	**(0.067)**	**(0.062)**	**(0.051)**
Measurement	0.020	0.013	**0.097**	**0.048**	**0.058**
	(0.049)	(0.058)	**(0.071)**	**(0.060)**	**(0.054)**
Algebraic Thinking	0.064	0.043	**0.031**	**0.034**	**0.006**
	(0.049)	(0.055)	**(0.049)**	**(0.054)**	**(0.048)**
Data Analysis	0.117	0.085	**0.092**	**0.114**	**0.115**
	(0.048)**	(0.062)	**(0.069)**	**(0.061)***	**(0.046)****
Geometry	0.078	0.047	**0.126**	**0.113**	**0.108**
	(0.053)	(0.058)	**(0.060)****	**(0.057)****	**(0.043)****
|Test Statistic| for Null Hypothesis of Uniformity	2.215**	1.237	**1.824***	**2.478****	**2.037****

Notes: Bolded columns are for the fully exposed cohorts of students. Standard errors are clustered at the district level, and for the matching estimators, bootstrapped using 250 repetitions. The uniformity test is for the null hypothesis $H_0: \frac{\theta_{NS}+\theta_M+\theta_{AT}}{3} = \frac{\theta_{DA}+\theta_G}{2}$ as described in the text.

* Indicates statistical significance at the 10 percent level or better.

** Indicates statistical significance at the 5 percent level or better.

Table 5. Falsification Tests. All Estimates are Matching Estimates. Estimates are Measured in Standard Deviations of Student-Level Test-Score Distribution.

	2001	2002	2005	2006	2007	2008	2009
Overall Math Score	0.014	0.024					
	(0.023)	(0.025)					
Topical Math Scores							
Numbers Sense	0.036	0.026					
	(0.061)	(0.065)					
Measurement	0.017	0.028					
	(0.068)	(0.069)					
Algebraic Thinking	0.036	-0.016					
	(0.047)	(0.062)					
Data Analysis	0.000	-0.028					
	(0.065)	(0.072)					
Geometry	-0.029	0.056					
	(0.070)	(0.057)					
Overall Reading Score	0.015	0.035	-0.002	-0.034	0.003	0.008	0.008
	(0.066)	(0.073)	(0.074)	(0.088)	(0.061)	(0.060)	(0.076)

Notes: Standard errors are clustered at the district level and bootstrapped using 250 repetitions.

Appendix A
Supplementary Tables

Appendix Table A.1. Data Sample Details.

	Schools	% of Universe	Districts	% of Universe
Universe*	1725		68	
Missing Information:				
Curriculum adoptions	14	0.8	7	10.2
District Test Scores (2003)	14	0.8	2	2.9
School Test Scores (2003)	23	1.4	0	0.0
Census Information	57	3.2	0	0.0
Multiple Curriculum Adopters	412	23.9	14	20.6
Final Sample	*1205*	*69.9*	*45*	*66.2*

Notes: The universe consists of those schools and districts for which any information was reported in 2003, and at least one grade-3 math test score was reported for an exposed cohort (2005-2009). The issue with the Census information was that some schools did not have a valid zip code. In an omitted analysis we assigned these schools Census characteristics based on the zip code assigned to their districts so that they could be included in our analysis. The inclusion of these schools does not qualitatively affect our findings.

Appendix Table A.2. Full Balancing Details for a Single Year (2007).

	Regression P-Value	Standardized Difference
2003 School Level Variables		
Math Test Score	0.99	3.55
Reading Test Score	0.73	2.47
Percent Free Lunch	0.98	3.69
Percent Students with Disabilities	0.33	8.23
Limited English Proficiency	0.44	1.97
Percent Immigrant	0.66	10.33
Percent Black	0.07	2.30
Percent Asian	0.81	4.83
Percent Hispanic	0.48	4.55
Percent White	0.65	0.07
Test Takers (log)	0.01	15.02
2004 School Level Variables		
Percent Free Lunch	0.98	3.25
Percent Students with Disabilities	0.07	15.67
Limited English Proficiency	0.36	2.33
Percent Immigrant	0.74	5.92
Percent Black	0.10	1.93
Percent Asian	0.65	3.81
Percent Hispanic	0.68	3.78
Percent White	0.67	0.11
Test Takers (log)	0.01	14.80
2003 District Level Variables		
Math Test Score	0.70	2.21
Reading Test Score	0.12	4.48
Test Takers (log)	0.18	9.71
Total Per-Pupil Revenue (log)	0.16	4.24
2004 District Level Variables		
Test Takers (log)	0.21	9.57
Total Per-Pupil Revenue (log)	0.58	0.92
Census Variables		
Median Household Income (log)	0.88	6.91
Share of Population with Low Education	0.31	6.14
Average	0.56	5.46

Note: We show full balancing results for just this single year for brevity. The results from other years are similar and available from the authors upon request.

Appendix B
Bandwidth Selection

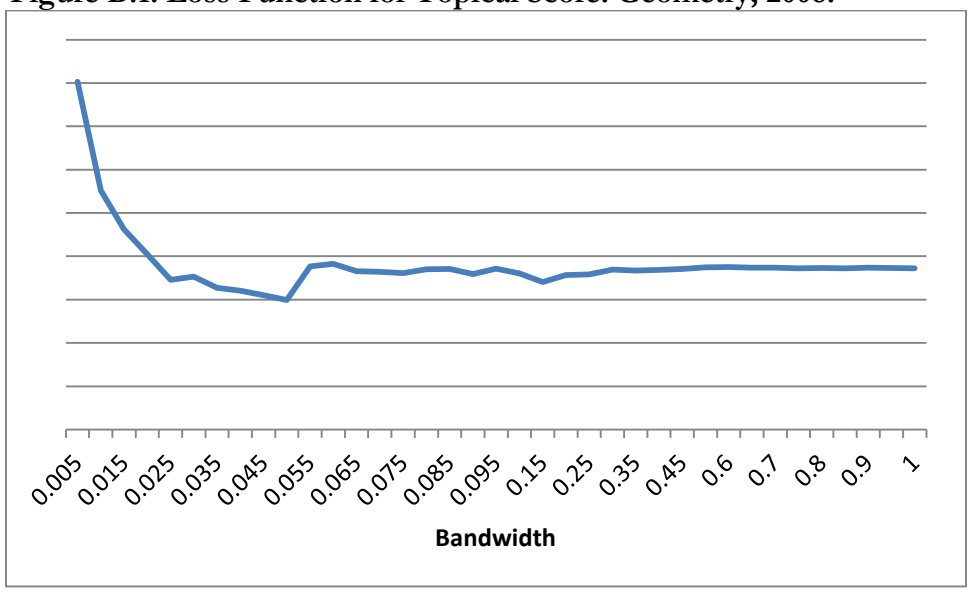
We use standard leave-one-out cross validation (C-V) to obtain fixed bandwidths for the kernel matching estimators. The grid search is over the range (0.005, 2.0). Using Frölich's (2004) notation, the C-V approach selects the optimal bandwidth, h_{CV}, by solving the following minimization problem using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

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

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.

Figure B.1 illustrates one of the many loss functions generated by equation (B.1) for our analysis. This particular example is for our evaluation of geometry scores in 2008, where we use the bandwidth 0.05 for our estimator.

Figure B.1. Loss Function for Topical Score: Geometry, 2008.



Note: The figure stops at h=1.0 although our grid search extends to h=2.0. The loss function is flat beyond h=1.0.

Appendix C
Alternative Composite Comparison

The curriculum descriptions in Section II indicate that some of the curricula in the composite comparison group have features similar to Harcourt. An interesting extension of our analysis would be to formally compare Harcourt to the other curricula in Florida that are the most distinct observationally. However, a practical limitation is that only four districts used a curriculum that is distinct from Harcourt along the two key dimensions of pedagogy and content organization (SRA/McGraw Hill, Cambium and Scott Foresman Investigations), which makes such an analysis infeasible empirically. We can, however, compare the Harcourt curriculum to the three *most-similar* alternatives used in Florida (Scott Foresman Addison Wesley, MacMillan/McGraw Hill and Houghton Mifflin), which were used by 13 of the 17 comparison districts. An expectation from this comparison is that the performance differences between Harcourt and the composite alternative will be smaller than what we report in the text.

Appendix Table C.1 shows our results using the alternative composite-comparison curriculum. Consistent with our expectation, the performance differences between Harcourt and the composite alternative are smaller relative to what we report in Table 4. When we run the formal tests for non-uniformity, the differences across subtopics are no longer statistically significant.

We also show balancing results for the restricted comparison in Appendix Table C.2 (standardized differences). Although the standardized-bias calculations across years look reasonable (e.g., they are much lower than the benchmark of 20 suggested by Rosenbaum and Rubin, 1985), the estimates are notably larger when we use the smaller composite-alternative comparison group. The fact that balance weakens in the data as the comparison group shrinks is not surprising, but it does suggest some caution in interpreting our estimates with the restricted comparison sample.

Although the findings reported in this appendix are best-viewed as suggestive due to data limitations, they are consistent with the hypothesis that our main results in the text are muted by the general similarity between Harcourt and the curricula that comprise most of the composite alternative in our main analysis. If we had access to more data for more schools and districts where curricula with substantive differences from Harcourt were used, a reasonable expectation is that we would find even more-pronounced differences between curricula. Again, the key limitation preventing further investigation is the lack of data for empirical research on curricular effectiveness. In this case, data from a state with more school districts (i.e., more unique sources of variation in adoptions) would likely be required for an informative study. If and when more data become available, extensions of this line of inquiry seem warranted.

Appendix Table C.1. Average Treatment Effects of Harcourt Relative to the Composite Alternative of Three Traditional and Massed Curricula, Overall and by Math Topic. Effects are Measured in Standard Deviations of Student-Level Test-Score Distribution.

	2005	2006	**2007**	**2008**	**2009**
Overall Score (OLS)	0.045	0.039	**0.076**	**0.055**	**0.052**
	(0.026)*	(0.028)	**(0.026)****	**(0.024)****	**(0.018)****
Overall Score (Matching)	0.026	0.008	**0.068**	**0.023**	**0.017**
	(0.063)	(0.065)	(**0.058**)	(**0.077**)	(**0.047**)
Topical Scores (Matching)					
Numbers Sense	0.019	0.008	**0.026**	**0.020**	**-0.014**
	(0.061)	(0.060)	(**0.065**)	(**0.067**)	(**0.055**)
Measurement	0.010	-0.012	**0.092**	**0.017**	**0.043**
	(0.057)	(0.082)	**(0.062)**	**(0.066)**	**(0.063)**
Algebraic Thinking	0.024	0.039	**0.043**	**-0.019**	**0.031**
	(0.077)	(0.066)	**(0.053)**	**(0.056)**	**(0.055)**
Data Analysis	0.080	0.042	**0.094**	**0.057**	**0.055**
	(0.051)	(0.072)	**(0.054)***	**(0.071)**	**(0.051)**
Geometry	0.050	-0.027	**0.076**	**0.030**	**0.048**
	(0.048)	(0.079)	**(0.085)**	**(0.054)**	**(0.054)**
|Test Statistic| for Null Hypothesis of Uniformity	0.941	0.109	**0.895**	**1.211**	**0.755**

Notes: Bolded columns are for the fully exposed cohorts of students. Standard errors are clustered at the district level, and for the matching estimators, bootstrapped using 250 repetitions.

The uniformity test is for the null hypothesis $H_0: \frac{\theta_{NS}+\theta_M+\theta_{AT}}{3} = \frac{\theta_{DA}+\theta_G}{2}$ as described in the text.

* Indicates statistical significance at the 10 percent level or better.

** Indicates statistical significance at the 5 percent level or better.

Appendix Table C.2. Balancing Details for the 28 Covariates Included in the Propensity-Score Specification, by Year. Comparison Between Harcourt and the Composite Alternative Consisting of the Three Other Traditional/Massed Curricula.

	2005	2006	2007	2008	2009
Mean Standardized Bias	8.1	7.8	7.8	10.5	8.2
Median Standardized Bias	7.0	4.7	5.0	6.6	6.2
N (Districts)	41	41	41	41	41
N (Schools)	867	869	870	868	860