Appendix A: Technical Appendix

Density Divergence

We calculate the divergence between the densities of the estimated propensity scores for treated and control units in each comparison. Frölich (2004) measures density divergence using the Kullback-Leibler (KL) information criterion; we follow his approach here, using kernel-density plots based on the Epanechnikov kernel. We estimate the divergence between the densities of ρ_{jm} for treatment and control schools as:

$$KL = \int \ln\left(\frac{f_{p|D=j}(\rho_{jm})}{f_{p|D=m}(\rho_{jm})}\right) f_{p|D=j}(\rho_{jm}) d\rho_{jm} + \int \ln\left(\frac{f_{p|D=m}(\rho_{jm})}{f_{p|D=j}(\rho_{jm})}\right) f_{p|D=m}(\rho_{jm}) d\rho_{jm} \qquad \text{(A.1)}$$

In (A.1), $f_{p|D=j}(\rho_{jm})$ is the density function of ρ_{jm} among schools treated with *j*, and $f_{p|D=m}(\rho_{jm})$ is the density function for schools that used *m*. Intuitively, density divergence will affect the precision of the estimates. A KL-information-criterion measure of zero indicates that the densities are identical, and the measure increases with density divergence. When the parameters of interest are *ATT*s, researchers use a unidirectional version of the KL information criterion (Frölich, 2004). Because we estimate *ATE*s, we use the bidirectional version originally suggested by Kullback and Leibler (1951).

Figure A.1 plots the estimated density functions of propensity scores for treatment and control schools in each comparison, and Table A.1 reports the corresponding KL information criteria. Similarly to the balancing tests, the density-divergence measures suggest that the data conditions are most favorable in our comparison between SBG and Saxon. Density divergence is largest in our comparison between SFAW and Saxon.[1]

[1] Frölich (2004) uses unidirectional density divergence measures in his study. Although the one and two-sided measures are not directly comparable; roughly speaking, our comparison of SBG and Saxon corresponds to his most favorable design, SFAW and SBG to his middle, and SFAW and Saxon to his least favorable design. This is purely by coincidence.

Bandwidth Selection

We use standard leave-one-out cross validation (C-V) to obtain fixed bandwidths for the kernel and LLR matching estimators. 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 for control observations:[2]

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

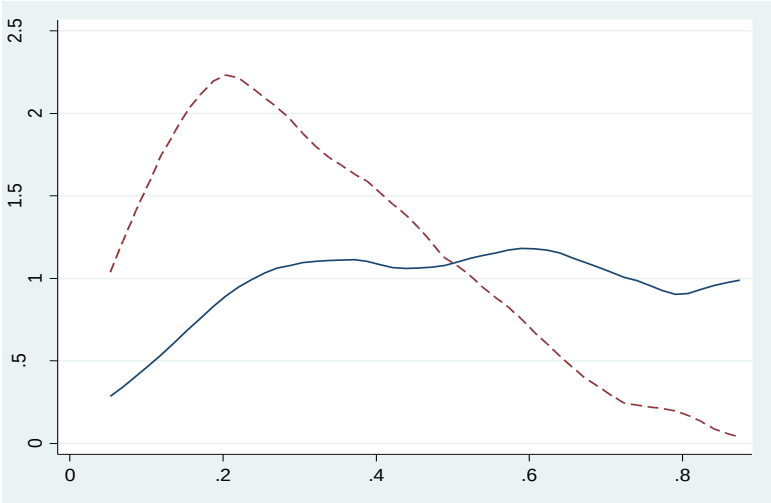
where q indexes the sample of control units, Y is the outcome (test score) and $\hat{m}_{-q}(p_q)$ is the estimate of the mean outcome among the control observations, excluding observation q, conditional on the estimated propensity score for unit q.

Figure A.2 shows two examples of loss functions generated by equation (A.2). In the first example the C-V procedure identifies a clear bandwidth choice. In the second example there is a flat region in the loss function as has been encountered in other contexts (e.g., Ludwig and Miller, 2007), and the C-V procedure suggests an optimal bandwidth at the edge of the grid search. In cases like this one we use a combination of conventional C-V and "visual inspection" to identify the appropriate bandwidth. That is, by visual inspection, we can see that there is very little difference in the loss function between the optimal bandwidth as determined by the mechanical C-V procedure and a much narrower bandwidth selected after the initial drop in the loss function. We 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.

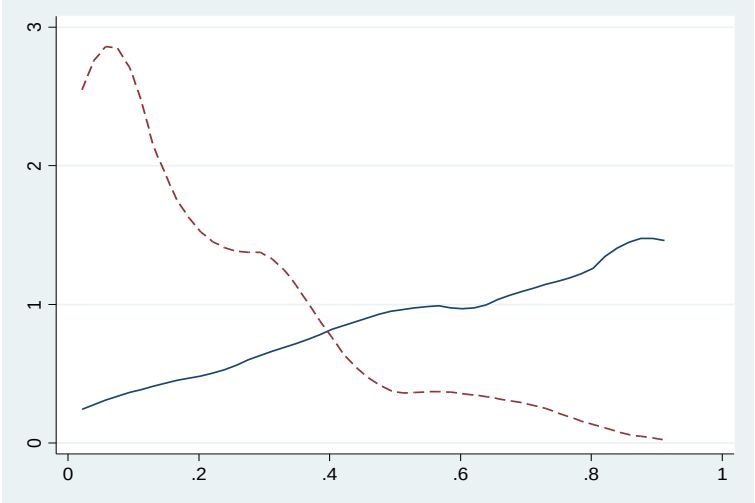
[2] In our case the definition of "treatment" and "control" is arbitrary and therefore, we could use either group. We use the largest group in each comparison as the control group.

Figure A.1 Probability Density Functions for Estimated Propensity Scores for Treatment and Control Units on the Common Support in Each Comparison Using 2001 Data (Solid Lines are Treatment Densities, Dashed Lines are Control Densities).

Treatment: SBG Control: Saxon



Treatment: SFAW Control: Saxon



Treatment: SFAW Control: SBG

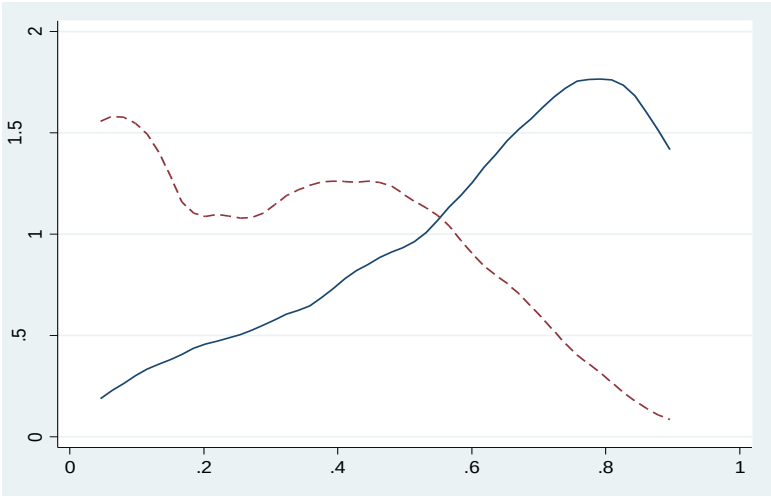
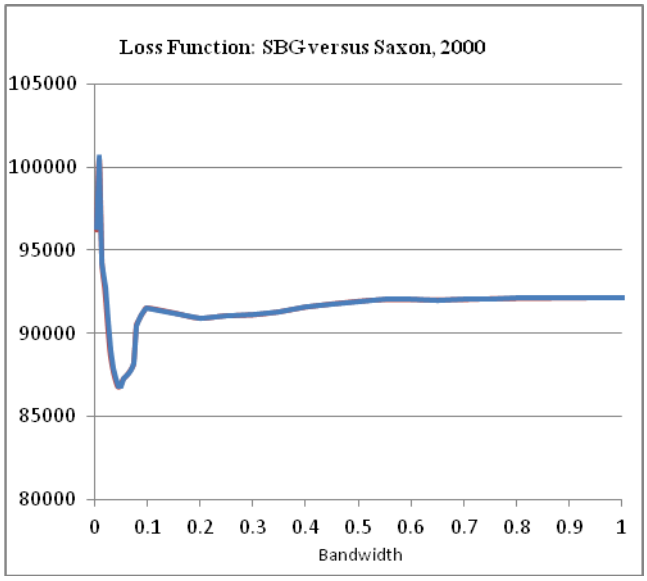


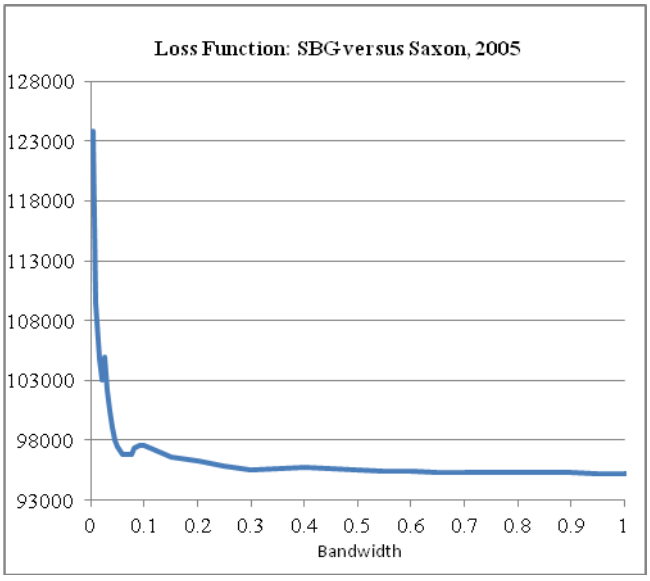
Table A.1 Kullback-Leibler Information Criteria by Curriculum Comparison

Comparison	KL Information Criterion
SBG and Saxon	0.63
SFAW and Saxon	1.58
SFAW and SBG	0.91

Note: Based on 2001 sample of schools.

Figure A.2. Loss Function Examples.





Appendix B: Supplementary Analyses

School Consolidations

Our data panel spans the course of 17 years. Over this time frame, we observe a series of school consolidations. As we discuss in Section V of the text, we match schools based on their static characteristics from the 1996-1997 and 1997-1998 school years. Consequently, school consolidations may alter the populations of students served by the schools that remain in our data over time, reducing the quality of our matches and potentially introducing bias into our estimates.

In order for the school consolidations to bias our estimates they must be correlated with curriculum 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 curriculum adoptions are independent of whether a district experiences a school closing (p-value ≈ 0.40). Additional evidence that our results are unlikely to be biased by school consolidations is provided in the paper in Table 2, where we evaluate covariate balance across matched treatment and control schools over the entire data panel. If schools that drop out of our sample over time systematically adopted specific curricula, we should find that our sample becomes less balanced as we move away from the matching years. We find little evidence of this, suggesting that school closings are not correlated with curriculum adoptions.[3]

Finally, in an omitted analysis (available upon request) 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 (an analogous procedure was done for schools that came into existence between 1991-1992 and 1996-1997). This alternative approach produces estimates that are qualitatively similar to what we report in the text, but comes at the cost of reduced efficiency.

[3] This use of balancing will only catch non-random attrition if it is correlated with observables.

Covariate Balance

In Section V we examine covariate balance using the regression-based test proposed by Smith and Todd (2005) and standardized differences in observables across matched treated and control schools (Rosenbaum and Rubin, 1985). We display the results from our balancing tests in Table 2 for each year of our data panel. For brevity, we only report the average p-value across F-tests and the average standardized difference across all covariates in the table. To provide more insight on covariate balance, in Table B.1 we report the p-value and standardized difference for each covariate separately, across all curriculum comparisons, for the year 2002.

Ultimately, as noted in the text, our most-compelling comparison is between SBG and Saxon. Table B.1 confirms the consistency of the balance across schools in that comparison. In the comparison between SFAW and Saxon, a notable result from Table B.1 is that the district-level characteristics do not balance well, generally speaking, relative to the school-level characteristics. The less-favorable balance in the district-level characteristics supports a cautious interpretation of the estimates from that comparison; however, the extent to which the balancing results in the table are problematic is unclear. For example, despite the apparent lack of balance in the district-level characteristics more generally, district-level achievement appears to be well balanced, and only three district-level covariates have a standardized difference in excess of 15. Additionally, the falsification tests in Section VII provide no indication of systematic bias in the comparison between SFAW and Saxon.

Grade 6 Falsification

In Section VII we provide the results from falsification tests where we estimate curriculum "effects" on the grade-3 math achievement for students who were not exposed to the curricula of interest, as well as the effects on grade-3 reading achievement for students who were and were not exposed (Tables 5 and 6). Here we show an additional set of falsification estimates for cohorts of grade-6 students who were not exposed to the curricula of interest (cohorts from 1993-2001).

For these falsification tests we use the same matching procedure to predict the same treatments (the uniform adoption of curriculum Saxon, SBG, or SFAW in grades one, two and three), only we match schools that have grade-6 classrooms and estimate the "effects" of the curricula on grade-6 achievement. Unfortunately, our grade-6 analysis is limited by sample-size issues. Specifically, because many districts teach grade six in middle school, and multiple elementary schools generally feed into a single middle school, our grade-6 sample of schools is much smaller than our grade-3 sample. The number of grade-6 schools that use SFAW is particularly small (roughly 80, on average, across the data panel), and we cannot balance treatment and control schools in either of our comparisons involving this curriculum. Because the unbalanced comparisons will not be informative, we present grade-6 falsification estimates only for our strongest comparison, between SBG and Saxon.[4] These estimates are reported in Table B.2. We estimate one non-zero "effect" in 1993 but otherwise, the point estimates are generally small and statistically indistinguishable from zero.

[4]For example, taking the average p-values from the Smith and Todd (2005) balancing regressions across years for the comparisons involving SFAW, they fall from roughly 0.50 in the grade-3 analysis (Table 2), to roughly 0.20 in the grade-6 analysis. In contrast, the average p-value falls to just below 0.50 in our grade-6 SBG to Saxon comparison.

Appendix Table B.1. Detailed Results from Balancing Tests in 2002.
SBG and Saxon

	Regression P-Value	Standardized Difference
1997 School Level Variables		
Math Test Score	0.19	1.64
Reading Test Score	0.56	3.93
Attendance	0.44	4.49
Percent Free Lunch	0.17	6.69
Percent Reduced Lunch	0.80	9.33
Percent Not Fluent in English	0.34	1.21
Percent Language Minority	0.91	2.65
Percent Black	0.13	5.69
Percent Asian	0.29	0.30
Percent Hispanic	0.98	1.79
Percent American Indian	0.91	1.49
Enrollment (log)	0.67	4.17
1998 School Level Variables		
Percent Free Lunch	0.36	8.41
Percent Reduced Lunch	0.83	7.11
Percent Not Fluent in English	0.24	0.15
Percent Language Minority	0.41	1.77
Percent Black	0.09	7.55
Percent Asian	0.63	0.26
Percent Hispanic	0.98	1.06
Percent American Indian	0.10	3.22
Enrollment (log)	0.83	4.97
1997 District Level Variables		
Math Test Score	0.64	3.69
Reading Test Score	0.38	5.24
Attendance	0.97	10.50
Enrollment (log)	0.18	1.11
Total Per-Pupil Revenue (log)	0.78	0.50
Local Per-Pupil Revenue (log)	0.98	4.14
1998 District Level Variables		
Enrollment (log)	0.16	1.04
Total Per-Pupil Revenue (log)	0.97	0.09
Local Per-Pupil Revenue (log)	1.00	3.49
Census Variables		
Median Household Income (log)	0.71	2.71
Share of Population with Low Education	0.14	0.08
Average	0.56	3.45

Appendix Table B.1 (continued).
SFAW and Saxon

	Regression P-Value	Standardized Difference
1997 School Level Variables		
Math Test Score	0.66	1.39
Reading Test Score	0.99	1.32
Attendance	0.62	2.36
Percent Free Lunch	0.60	0.22
Percent Reduced Lunch	0.88	2.67
Percent Not Fluent in English	0.33	3.55
Percent Language Minority	0.69	4.57
Percent Black	0.34	1.57
Percent Asian	0.76	4.95
Percent Hispanic	0.55	5.34
Percent American Indian	0.47	0.71
Enrollment (log)	0.06	3.39
1998 School Level Variables		
Percent Free Lunch	0.48	0.44
Percent Reduced Lunch	0.40	5.31
Percent Not Fluent in English	0.33	2.46
Percent Language Minority	0.77	3.31
Percent Black	0.34	0.59
Percent Asian	0.19	4.62
Percent Hispanic	0.83	2.94
Percent American Indian	0.02	3.94
Enrollment (log)	0.08	4.48
1997 District Level Variables		
Math Test Score	0.79	4.51
Reading Test Score	0.80	3.52
Attendance	0.38	10.57
Enrollment (log)	0.46	12.32
Total Per-Pupil Revenue (log)	0.49	3.97
Local Per-Pupil Revenue (log)	0.02	18.41
1998 District Level Variables		
Enrollment (log)	0.44	12.32
Total Per-Pupil Revenue (log)	0.09	26.17
Local Per-Pupil Revenue (log)	0.01	17.50
Census Variables		
Median Household Income (log)	0.92	9.57
Share of Population with Low Education	0.81	11.67
Average	0.49	5.96

Appendix Table B.1 (continued).
SFAW and SBG

	Regression P-Value	Standardized Difference
1997 School Level Variables		
Math Test Score	0.08	1.86
Reading Test Score	0.15	0.74
Attendance	0.20	2.53
Percent Free Lunch	0.34	3.56
Percent Reduced Lunch	0.88	17.06
Percent Not Fluent in English	0.42	6.48
Percent Language Minority	0.68	0.27
Percent Black	0.67	16.06
Percent Asian	0.39	8.03
Percent Hispanic	0.75	16.04
Percent American Indian	0.11	0.52
Enrollment (log)	0.61	7.38
1998 School Level Variables		
Percent Free Lunch	0.22	3.67
Percent Reduced Lunch	0.76	22.57
Percent Not Fluent in English	0.89	0.86
Percent Language Minority	0.68	5.17
Percent Black	0.56	23.10
Percent Asian	0.31	6.92
Percent Hispanic	0.57	18.04
Percent American Indian	0.04	5.70
Enrollment (log)		14.35
	0.82	
1997 District Level Variables		
Math Test Score	0.91	5.27
Reading Test Score	0.84	5.09
Attendance	0.09	12.00
Enrollment (log)	0.45	19.63
Total Per-Pupil Revenue (log)	0.06	32.74
Local Per-Pupil Revenue (log)	0.63	8.17
1998 District Level Variables		
Enrollment (log)	0.44	19.83
Total Per-Pupil Revenue (log)	0.58	6.16
Local Per-Pupil Revenue (log)	0.54	4.41
Census Variables		
Median Household Income (log)	0.82	14.61
Share of Population with Low Education	0.62	5.15
Average	0.50	9.81

Appendix Table B.2. Falsification Estimates of Math Curricular Effectiveness, Estimated Using Math Test Scores for Grade-6 Cohorts who were Never Exposed to the Curricula of Interest. Comparison of SBG and Saxon only.

	1992	1993	1994	1995	1996	1999	2000	2001
Treatment: SBG Control: Saxon								
Kernel Matching	-0.042	-0.096	-0.018	-0.045	0.015	0.006	-0.063	-0.034
	(0.052)	(0.055)†	(0.052)	(0.046)	(0.047)	(0.059)	(0.050)	(0.043)
N(Saxon)	205	208	213	213	218	212	205	204
N(SBG)	117	118	122	125	127	122	120	120

Notes: Matching estimators impose the common support restriction. Standard errors in parentheses 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 C: Data Appendix

Appendix Table C.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 (1997)	2	0.2	2	0.7
School outcome variables (1997)	23	2.2	1	0.3
District finance/enrollment data (1997, 1998)	2	0.2	1	0.3
School enrollment/demographic data (1997, 1998)	82	7.3	12	4.0
Did not use one of the primary curricula in grades 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
Final 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 C.2. Determination of 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

Note: In the years prior to 2000 the scaling factor is assumed to be 0.33. Falsification estimates with reading scores and grade-6 math scores were scaled similarly – in all grades/subjects/years, the scaling factor that converts estimates based on the school-level score distributions into student-level standard deviations is roughly one-third.