Final Report

The Relationship Between Using Saxon Elementary and Middle-School Math and Student Performance on California Statewide Assessments

January 2007

Prepared By
Miriam Resendez, MA, Senior Researcher
Mariam Azin, PhD, President

PRES Associates, Inc.
For more information:
(307) 733-3255
www.presassociates.com

SAXON™
A Harcourt Education Company

The Relationship Between Using Saxon Elementary and Middle-School Math and Student Performance on California Statewide Assessments

Table of Contents

Figures and Tables

Figures

Tables

Appendix: Tables of Statistical Results

Executive Summary

PRES Associates, an external, independent educational research firm with over 15 years of experience in applied educational research and evaluation, conducted analyses using existing California state assessment data to examine the relationship between math performance and *Saxon Math* programs at elementary and middle school grade levels. The purpose of this report is to present the results of statistical analyses conducted in order to examine how well the *Saxon Math* program helps California elementary and middle-school students attain critical math skills.

Major findings arranged by evaluation questions include the following:

1. **Are there significant changes in Saxon students' math performance?**

   - Math performance on the *Stanford Achievement Test, Ninth Edition (Stanford 9)* and *California Achievement Test* (CAT 6) among Saxon elementary and middle-school students increased significantly as they progressed through grade levels.
   - Generally, the percentage of students meeting California math standards in Saxon elementary and middle schools increased over time from 2002 to 2006. However, this seemed to be more pronounced among elementary students.
   - Changes in math performance among Saxon schools on the *California Standards Test* (CST) are not dependent on how long a school has used the program. Therefore, schools that had only implemented the Saxon program for 1 year showed similar rates of change as schools that had implemented the program for 4 years.

2. **Does achievement across Saxon students vary depending on the type of student?**

   - In general, there were significant increasing trends in math performance among all subgroups of students, including males and females, minorities and nonminorities, economically disadvantaged and non–economically disadvantaged students, English language learners (ELLs) and non-ELLs, and students with disabilities and students without disabilities.
   - Generally, improvement in math performance among subgroups of students was found consistently at both the elementary and middle-school level and among all statewide math assessments.

3. **How does math performance differ between students in Saxon and non–Saxon schools? Are there differences between subgroups of students in Saxon and non–Saxon schools?**

   - Examination of differences over time (i.e., cross-sectional analyses) showed that, overall, both groups (Saxon and non-Saxon) generally showed improvement in performance. In addition, although on most measures and years, the performance of Saxon students was higher than those of non–Saxon students, given the small effect sizes (d = .01 to .18), which provide an indication on the importance of findings, the focus should be on the positive changes themselves and not necessarily on differences between the groups.
   - Results of similar groups of students followed over time (i.e., cohort analyses) show that in general, Saxon and non–Saxon students showed similar increases in math performance. While at times Saxon students outperformed non–Saxon students (and vice versa), patterns of changes between groups were not consistent as to allow for more conclusive comments to be made about differences between groups.
   - School-level analyses controlling for pre–Saxon differences revealed that Saxon elementary schools show similar levels of math performance as non–Saxon elementary schools when averaged across all years. However, results from the CAT 6 and CST also suggest that Saxon schools may need some time using Saxon before there is differentiation in performances between Saxon schools and schools using other math curricula. More specifically, while Saxon schools started out at a lower level in math performance compared to non–Saxon schools, Saxon schools subsequently surpassed non–Saxon schools.
   - Differences among subgroups of students were observed. In particular, use of *Saxon Math*

was associated with greater math performance among students in certain subpopulations, including Whites, Hispanics, ELLs, non–economically disadvantaged students, and students with disabilities. In contrast, non–Saxon students who were African American, non-ELL, and did not have disabilities performed better than Saxon students.

In summary, the results of this study using California state assessment data provides some support for a positive relationship between the *Saxon Math* program in elementary and middle-school levels and math performance. However, stronger (and more conclusive) findings have been obtained in other research on the *Saxon Math* curriculum. Therefore, further research is needed to more fully explore the effectiveness of the *Saxon Math* program.

Project Background

"In my experience, competency in mathematics—both in numerical manipulation and in understanding its conceptual foundations—enhances a person's ability to handle the more ambiguous and qualitative relationships that dominate our day-to-day decision-making."

—Federal Reserve Chairman
Alan Greenspan

A strong foundation in math skills early on is critical to students' future participation in higher level math courses as well as to their academic and career success (Glenn, 2000; National Research Council, 2001). Unfortunately, research continues to show that U.S. students are not being sufficiently prepared to meet the demands of future careers, including advanced skills in critical thinking and mathematics. While the latest results from the National Assessment of Educational Progress (NAEP; 2005) points to improvements in the math performance of fourth and eighth graders, international comparisons have shown that U.S. students are falling behind in math as compared to students of other countries. On the most recent Program for International Student Assessment, U.S. 15-year-olds performed below the international average in mathematics literacy and problem solving (U.S. Department of Education, 2006). In addition, results from the Third International Mathematics and Science Study found that eighth-grade students' achievement in math is below average internationally and is lower than students in many countries that are economic competitors to the United States (Mullis, Martin, & Foy, 2005). In order to adequately prepare students to be competitive in a global economy, it is imperative that the mathematics skills and knowledge of U.S. students be improved upon.

In an effort to improve mathematical understanding of students, John Saxon developed the *Saxon Math* program in the 1980s. Based on several research-based strategies to promote student success, the program uses incremental development and continual review to teach students math concepts. Saxon's instructional approach breaks complex concepts into related increments, with the idea that smaller pieces of information are easier to teach and easier to learn. Thus, the incremental approach provides students with time to solidify prerequisite concepts and skills before they are introduced to the next step of instruction. Through continual review, previously taught concepts are practiced frequently and extensively over the year. The goal is to help students build knowledge of math concepts over time, and through repetitive practice, reinforce those concepts.

Given how important math skills are to the future success of children, programs that can help in the development of these skills need to be looked at carefully to determine the extent to which they help students attain critical math skills. Indeed, the No Child Left Behind Act of 2001(NCLB) mandates that educational materials purchased with public funds be proven by scientific research to improve student achievement in the classroom. In an effort to examine the effectiveness of the Saxon Math program in the state of California, Planning, Research, and Evaluation Services (PRES Associates),[1] conducted analyses using California state assessment data to examine the relationship between math performance and use of the *Saxon Math* program among elementary and middle-school students.

Project Overview

The overarching purpose of this report is to present the results of statistical analyses conducted on existing California state assessment data in order to examine how well the *Saxon Math* program helps California elementary and middle-school students attain vital math skills. Specifically, the analyses are designed to address the following key evaluation questions:

1. **Are there significant changes in Saxon students' math performance?**
2. **Does achievement across Saxon students vary depending on the type of student?**

[1] PRES Associates is an external, independent, educational research firm with more than 15 years of experience in applied educational research and evaluation. For more information, please visit www.presassociates.com.

3. **How does math performance differ between students in Saxon and non–Saxon schools? Are there differences between subgroups of students in Saxon and non–Saxon schools?**

The remainder of this report includes a description of the methods employed, measures, sample, curricula, and results of the analyses performed. In addition, where appropriate, results from the present analyses are triangulated with results from prior archival studies conducted in the states of Georgia and Texas as well as with a recent randomized control trial.[2]

Design and Methodology

Archival California assessment data were used to evaluate the *Saxon Math* program in elementary and middle schools. The California Department of Education (CDE) was first contacted to determine what data were available and at what level[3] (school or student). Based on this feedback, evaluation questions and an analyses plan were developed.

Data for students from schools using Saxon and matched comparison schools were requested from the CDE. It should be noted that, per state policy, the CDE could only release unidentifiable student-level data. That is, all student and school identifiers[4] were excluded. This eliminated the possibility of conducting student-level longitudinal growth analyses. A detailed description of the measures and samples used follows.

Measures

The Stanford 9, CAT 6, and CST are the three statewide exams that have been used by California to assess student learning during the spring over the past 8 years. The Stanford 9 was used from 1998 to 2002 and was administered to Grades 2 through 8. In spring 2003, the Stanford 9 was replaced by the CAT 6. This test was administered to Grades 2 through 8 in spring 2003 and 2004, but in spring 2005–2006 it was administered only to Grades 3 and 7. In 1998, the state of California also began testing via the CST. However, information obtained from the CDE indicated that CST data are only available for spring 2002 to 2006. Table 1 displays the data available for each assessment.

Table 1. California State Assessments (and Sample) by Year Administered and Grades Tested

Test Year	Sample 1: STANFORD	Sample 2: CAT6	Sample 3: CST
1998	2, 3, 4, 5, 6, 7, 8		
1999	2, 3, 4, 5, 6, 7, 8		
2000	2, 3, 4, 5, 6, 7, 8		
2001	2, 3, 4, 5, 6, 7, 8		
2002	2, 3, 4, 5, 6, 7, 8		2, 3, 4, 5, 6, 7, 8
2003		2, 3, 4, 5, 6, 7, 8	2, 3, 4, 5, 6, 7, 8
2004		2, 3, 4, 5, 6, 7, 8	2, 3, 4, 5, 6, 7, 8
2005		3, 7	2, 3, 4, 5, 6, 7, 8
2006		3, 7	2, 3, 4, 5, 6, 7, 8

Note. Numbers within cells represent grade levels.

Data from each of these three assessments are comparable across years (i.e., have not changed) and therefore support trend analyses. However, information obtained from the California Department of Education indicated that the tests are not comparable. Therefore, separate analyses are conducted for each test. Both student- and school-level[5] data were obtained for the three state assessments. However, although student-level data were obtained, as previously noted, no identifiers were provided. As such, individual student growth analyses could not be conducted. Instead, most analyses involve comparisons of *groups* of

[2] In particular, two studies using Texas and Georgia statewide assessment data were conducted previously by PRES Associates (Resendez, Fahmy, & Manley, 2004; Resendez, Sridharan, & Azin, 2005), in addition to a recently completed randomized control trial (Resendez & Azin, 2006). For more information on these studies, the reader is referred to http://saxonpublishers.harcourtachieve.com, or contact PRES Associates at info@presassociates.com.

[3] This took several months to finalize due to state policy changes.

[4] Instead of school identifiers, codes were sent to CDE to allow PRES Associates to identify different *groups* of schools (e.g., Saxon and non–Saxon schools, Saxon elementary schools that began using the program in the 1999–2000 school year, etc.).

[5] School-level data were downloaded from the CDE Web site (http://star.cde.ca.gov/).

students over time (i.e., cross-sectional analyses; e.g., comparing 2001 elementary students with 2002 elementary students). In addition, analyses of similar groups of students followed over grade levels were also performed (i.e., cohort analyses; e.g., comparing second graders in 2001 with third graders in 2002). School-level data were obtained to supplement the student-level data because this dataset allows researchers to examine changes over time *within* schools. This is because, as opposed to the student-level data received from the CDE, schools could be readily identified, and therefore researchers could match schools' math performance over time.

The Stanford 9 and CAT 6 are norm-referenced tests and consist of multiple-choice items. According to the CDE, these exams are valid and reliable for the population of California public school students.[6] The analyses presented in this report use the total math scale score[7] for each of these tests as outcome measures. It should be noted that because these are development scale scores, which increase from the lowest to highest grades tested, Stanford 9 and CAT 6 data were analyzed to measure both changes over the years and grade levels.

The CST is a criterion-referenced test that measures how well students attain identified state-adopted content standards. Performance levels establish points at which students have demonstrated sufficient knowledge and skills to be regarded as performing at a particular achievement level. The identified performance levels on the CST are: (1) far below basic, (2) below basic, (3) basic, (4) proficient, and (5) advanced.

Scores on *each* math objective were not provided from the CDE. However, overall math performance level and the CST scale score were provided. These are used as math outcomes for analyses pertaining to the CST. It is important to note that the CST scale score is horizontally equated, meaning that cross-sectional analyses (i.e., examining scores of students in the same grades over time) is supported. However, scale scores are not vertically equated. That is, unlike with the Stanford 9 and CAT 6 tests, these are not developmental scale scores, which are designed to increase with each grade level. Rather, they are on the same scale, grade to grade and year to year (range is 150–600). Furthermore, statewide, there tends to be a downward shift in scale-score performance between the elementary and middle school grade levels, suggestive of the more varying and difficult levels of math (e.g., prealgebra, algebra) students are expected to know at the middle school level. Given these considerations, for the CST sample, analyses were performed to examine changes over time (and not over grade levels or cohort analyses) among Saxon elementary and middle-school students.

Sample

California schools using the *Saxon Elementary* and *Middle School Math* program in the second through eighth grades between 1998 and 2005 were selected for inclusion in this study ($n = 64$[8]). Control sites[9] ($n = 64$) were randomly selected from a list of *similar schools* for each Saxon school. Schools with similar characteristics are determined by the state of California via the School Characteristics Index (SCI). The SCI is a composite of the demographic characteristics of a school derived through multiple linear regression. This technique yields a single composite index based on important school background characteristics, including

- enrollment,
- ethnicity distribution,
- average parent educational level,
- free or reduced-price lunch participation,
- fully credentialed teachers,
- teachers with emergency permits,

6 For more information on the validity of the these tests, see (1) Harcourt Assessment, Inc.'s, *Stanford 9 Technical Manual*; (2) CTB/McGraw-Hill's *CAT 6 Technical Manual*; and (3) the *2004 CST Technical Manual*, available online (http://www.cde.ca.gov/ta/tg/sr/resources.asp).

7 Scores on math objectives were unavailable from the CDE.

8 Note that only schools *confirmed* to be Saxon users through contact with the school by an independent call center were included in this study. These schools had to have used Saxon Math in 75% or more of their math classes.

9 Similarly, only schools *confirmed* to be non-Saxon users by an independent call center, during the years of interest (1999-2006) were included in this study.

Table 2. Sample Sizes by Year and Grade Level

Sample	Grade	1998		1999		2000		2001		2002		2003		2004		2005		2006	
		non-Saxon	Pre-Saxon	non-Saxon	Pre-Saxon	non-Saxon	Saxon	non-Saxon	Saxon	non-Saxon	Saxon	non-Saxon	Saxon	non-Saxon	Saxon	non-Saxon	Saxon	non-Saxon	Saxon
Stanford 9	2	5586	2098	5589	2279	5890	2230	6109	2253	6405	2350								
	3	5306	2183	5696	2251	6022	2418	6232	2368	6327	2299								
	4	5281	2257	5189	2191	5708	2282	5948	2498	6078	2452								
	5	5198	2245	5263	2173	5349	2229	5802	2372	5967	2423								
	6	6105	3531	6603	3077	6680	3407	6972	3520	7490	3765								
	7	4828	3942	5865	3751	6161	3721	6451	3971	6700	3987								
	8	4739	3599	5543	3806	5816	3765	6118	3818	6413	3921								
CAT 6	2											4614	3819	4687	3901				
	3											4704	3879	4602	3921	4574	3907	4370	3510
	4											4490	3913	4375	4089				
	5											4262	4038	4393	3879				
	6											7427	4027	7103	4159				
	7											7334	4283	7083	4547	7022	4473	7246	4287
	8											6816	3897	7339	4323				
CST	2									4675	1661	4615	1570	4692	1589	4405	1437	4332	1376
	3									4681	1611	4703	1573	4600	1583	4571	1513	4378	1336
	4									4380	1727	4500	1601	4370	1614	4264	1512	4030	1387
	5									4329	1672	4271	1668	4400	1555	4335	1545	4138	1354
	6									7557	3747	7440	4022	7111	4166	7174	3981	6872	3978
	7									6734	3957	7340	4263	7095	4550	7028	4468	7263	4313
	8									6128	3864	6707	3775	6947	4191	7126	4263	6914	4206

Note. Figures represent students with valid scores on each assessment. Black/dark gray areas in the table highlight the elementary cohorts examined in the study. Blue areas highlight the middle school cohorts examined in the study.

[a] 44 Saxon schools began using *Saxon Math* in the 1999–2000 school year. The remaining 20 schools began using *Saxon Math* in the 2002–2003 school year, all of which were elementary schools.

- English language learners,
- average class size k–3 and 4–6, and
- multitrack year-round program.

According to the CDE, schools with nearly identical SCIs will be "similar" with respect to the overall educational challenge and opportunity presented by their respective constellations of background factors. For more detailed information on this procedure, the reader is referred to the CDE's report *Construction of California's 1999 School Characteristics Index and Similar Schools' Ranks.*[10]

Data were obtained for all students in the selected Saxon and non–Saxon California schools between the 1997–1998 and 2005–2006 school years. The total sample includes 48 Saxon elementary schools, 45 non–Saxon elementary schools, 16 Saxon middle schools, and 19 non–Saxon middle schools. As shown on Table 2, the samples are defined by the assessment used. Specifically, the Stanford 9 sample consists of all second through eighth graders from the 1997–1998 to 2001–2002 school years. The CAT 6 sample consists of second through eighth graders from 2002–2003 to 2003–2004 and third and seventh graders from 2004–2005 to 2005–2006. The CST sample consists of second through eighth graders from the 2001–2002 to 2005–2006 school years.

Table 3 shows the average site and statewide characteristics for elementary and middle schools in 2005–2006.[11] Results of the comparability of the Saxon and control sites showed that the schools were equivalent on most of the measured demographic variables. Significant differences were observed for the following variables: (a) total enrollment, $t(126) = 2.78$, $p = .006$, (b) percentage of African Americans, $t(126) = 5.99$, $p < .001$, and (c) percentage of Hispanics, $t(126) = 2.76$, $p = .007$. In general, there was a higher enrollment and percentage of Hispanics in non–Saxon schools compared to Saxon schools. In addition, there were a higher percentage of African Americans in Saxon schools compared to non–Saxon schools. These results indicate that it is important to control for demographic differences in analyses involving comparisons between Saxon and non–Saxon schools. It should also be noted that this sample consisted of a higher minority population (heavily Hispanic), and a higher proportion of English language learners and socioeconomically disadvantaged students than that was found statewide. As such, comparisons between the study groups and the State of California are not made.

Table 3. Sample and Statewide Average Demographic Characteristics (2005–2006) for Elementary and Middle Schools

Sites	Average Enrollment	% White	% Hispanic	% African American	% Female	% Limited English	% with Disabilities	% Socioeconomically Disadvantaged
Saxon	595	21	58	16	50	33	9	63
Non-Saxon	744	17	71	4	49	41	10	69
CA (K–8)	**614**	**33**	**45**	**7**	**47**	**26**	**10**	**49**

Settings

Figure 1 shows the geographical location of the sites used in this study. Schools came from a mixture of urban, suburban, and rural communities. For confidentiality purposes, the names and exact location of the schools are excluded.

Curricula

Saxon Math

In the early 1980s, John Saxon developed a theoretically based and distributed approach to mathematics instruction, practice, and assessment that has evolved to include a textbook series and a comprehensive approach for K–12 students. At the foundation of the Saxon program is the

[10] This document is available at http://www.cde.ca.gov/ta/ac/ap/documents/tdgreport0400.pdf

[11] Schools demographic characteristics were somewhat consistent over the course of the 6 years in which demographic data were available (2000–2006).

Figure 1. Map of study schools

Note. Number within blue boxes note the number of identified Saxon schools in a particular county. Number within white boxes note the number of identified non-Saxon schools in a particular county.

premise that students learn best if (a) instruction is incremental and explicit; (b) they can continually review previously learned concepts; and (c) assessment is frequent and cumulative. In *Saxon Math,* new increments of instruction are regularly introduced while students continually review previously introduced math concepts. Such an approach to learning ensures that students truly integrate and retain math concepts rather than forget them as soon as they are no longer exposed to them.

Confirmation phone calls were made to all schools that were identified as current or former *Saxon Math* users. Data collected from these confirmation calls included (a) verification of periods of use of the *Saxon Math* program (b) the *Saxon Math* program used at different grade levels, and (c) the proportion of students within schools that used this curriculum.[12] Results showed that, in elementary grades, schools used the *Saxon Math* program recommended for each grade level. As such, first graders used *Saxon Math 1,* second graders used *Saxon Math 2,* and so forth. However, in the middle schools, there was more variability. This is because different *Saxon Math* programs can be used depending on the ability levels of the students. For example, advanced seventh-grade students can use *Saxon Algebra 1* instead of *Saxon Math 87.* Table 4 shows the average and range of the percentage of students using the Saxon texts at the middle-school level. Typically, schools that did not use only a single textbook at each grade level tended to use the next level above and/or below of the Saxon text for remaining students (e.g., a school used *Saxon Algebra 1/2* with 80% of its seventh graders, and the remaining 20% used *Saxon Algebra 1*).

[12]Schools had to use this program with at least 75% of their math classes to be included in the Saxon sample.

[13]Note that analyses could not be performed to examine if there were differences between the various types of control curricula and the *Saxon Math* program. This is because the required coding could potentially enable the identification of schools and students, which the CDE's privacy policy could not allow. Thus, the data had to be requested (and released) without these school identifiers.

Table 4. Percentage of Students Using Saxon Textbooks by Middle-School Grade Level

Grade	Program	Average	Range
6	Saxon 76	29%	0–100%
6	Saxon 87	39%	0–100%
6	Saxon Alg ½	24%	0–100%
6	Saxon Alg 1	8%	0–25%
7	Saxon 76	26%	0–25%
7	Saxon 87	40%	0–75%
7	Saxon Alg ½	27%	0–100%
7	Saxon Alg 1	7%	0–100%
8	Saxon 76	11%	0–25%
8	Saxon 87	23%	0–75%
8	Saxon Alg ½	22%	0–50%
8	Saxon Alg 1	44%	0–100%

Note. Range refers to the percentage of students noted by schools as using the indicated text, from the lowest to highest percentage.

Non–Saxon Site Curricula

The majority of non–Saxon schools (75%) used core basal math curricula. These curricula typically consist of a chapter-based approach to math instruction. Five schools (9%) use an investigative approach, with an emphasis on purposeful, inquiry-based math instruction involving integration across various mathematical topics and content areas. The remaining 16% used a mix of basal, investigative, computer-based, and/or used other printed material (non–textbook based).[13]

Summary of Findings

Major findings included the following:

1. **Are there significant changes in Saxon students' math performance?**
   - Math performance on the Stanford 9 and CAT 6 among Saxon elementary and middle-school students increased significantly as they progressed through grade levels.
   - Generally, the percentage of students meeting California math standards in Saxon elementary and middle schools increased over time from 2002 to 2006. However, this seemed to be more

pronounced among elementary students.

- Changes in math performance among Saxon schools on the CST are not dependent on how long a school has used the program. Therefore, schools that had implemented the Saxon program for only 1 year showed similar rates of change as schools that had implemented the program for 4 years.

2. Does achievement across Saxon students vary, depending on the type of student?

- In general, there were significant increasing trends in math performance among all subgroups of students, including males and females, minorities and nonminorities, economically disadvantaged and non–economically disadvantaged students, English language learners and non-ELLs, and students with disabilities and students without disabilities.
- Generally, improvement in math performance among subgroups of students was found consistently at both the elementary and middle-school level and among all statewide math assessments.

3. How does math performance differ between students in Saxon and non–Saxon schools? Are there differences between subgroups of students in Saxon and non–Saxon schools?

- Examination of differences over time (i.e., cross-sectional analyses) showed that overall, both groups (Saxon and non-Saxon) generally showed improvement in performance. In addition, although on most measures and years, the performance of Saxon students was higher than those of non-Saxon students, given the small effect sizes (d = .01 to .18), which provide an indication on the importance of findings, the focus should be on the positive changes themselves and not necessarily on differences between the groups.
- Results of similar groups of students followed over time (i.e., cohort analyses) show that in general, Saxon and non–Saxon students showed similar increases in math performance. While at times Saxon students outperformed non–Saxon students (and vice versa), patterns of changes between groups were not consistent as to allow for more conclusive comments to be made about differences between groups.
- School-level analyses controlling for pre-Saxon differences revealed that Saxon elementary schools show similar levels of math performance as non–Saxon elementary schools when averaged across all years. However, results from the CAT 6 and CST also suggest that Saxon schools may need some time using Saxon before there is differentiation in performances between Saxon schools and schools using other math curricula. More specifically, while Saxon schools started out at a lower level in math performance compared to non–Saxon schools, Saxon schools subsequently surpassed non–Saxon schools.
- Differences among subgroups of students were observed. In particular, use of *Saxon Math* was associated with greater math performance among students in certain subpopulations, including Whites, Hispanics, ELLs, non–economically disadvantaged students, and students with disabilities. In contrast, non–Saxon students who were African American, non-ELL, and did not have disabilities performed better than did Saxon students.

What follows is a detailed account of the findings, organized by the evaluation questions. For detailed statistical tables, the reader is referred to the Appendix.

Results

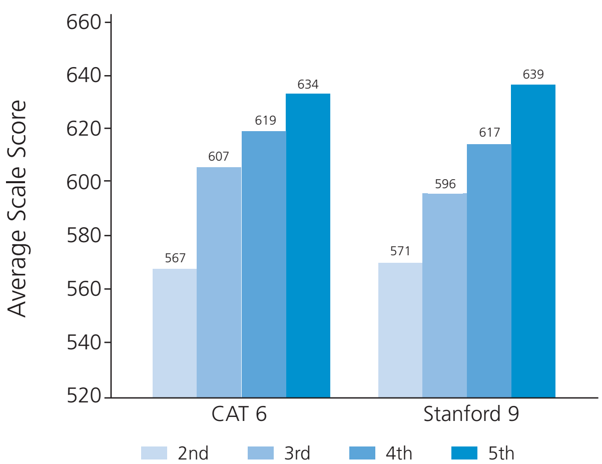
1. Are there significant changes in Saxon students' math performance?

Separate analyses were conducted on the Stanford 9, CAT 6, and CST data to examine changes among grade levels and/or over time. As previously noted, the Stanford 9 (1998–2002) was replaced by the CAT 6 in 2003. In addition, the CAT 6 was administered in Grades 2 through 8 in the spring of 2003 and 2004 and in Grades 3 and 7 only in the spring of 2005 and 2006. Data from the CST were available from students in Grades 2 through 8 from spring 2002 to 2006.

Given the characteristics of the different sets of samples, the first set of analyses involved cross-sectional analyses of the CAT 6 and Stanford 9 samples. Specifically, comparisons were made between different grade levels across time as follows: (a) examination of differences between elementary students across all years in which assessment data is available (i.e., second versus third versus fourth versus fifth graders on average performance from 2000 to 2002[14] for the Stanford 9 and 2003 to 2006 for the CAT 6[15]); and (b) examination of differences between middle school students across all years in which assessment data is available (i.e., sixth versus seventh versus eighth graders on average performance from 2000 to 2002 for the Stanford 9 sample and 2003 to 2006 for the CAT 6 sample). It is important to note that scores from these norm-referenced tests are not comparable and, therefore, results should be examined separately.

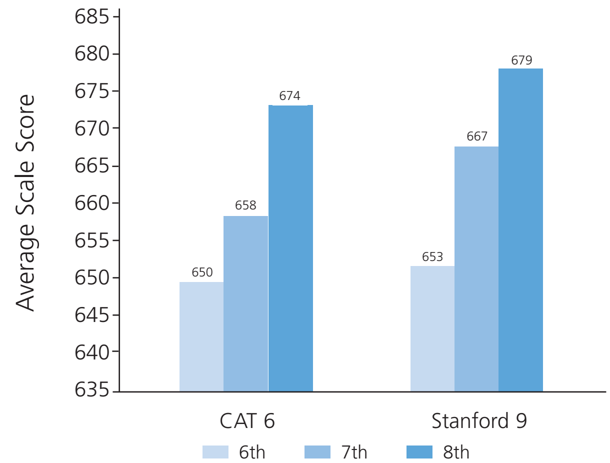
Results of the Stanford 9 and CAT 6 data revealed that Saxon exposure was related to differences in math performance over grade levels. That is, as grade levels increased, so did math performance. These significant increases were observed in both the elementary grade level, F-Stanford 9 (3, 28170) = 3374.5, $p < .001$ and F-CAT 6 (3, 38852) = 2538.2, $p < .001$, and middle school grade level, F-Stanford 9 (2, 33872) = 1155.0, $p < .001$ and F-CAT 6 (2, 33988) = 418.4, $p < .001$ (see Figures 2 and 3).

Figure 2.
Saxon elementary students' Stanford 9 (2000–2002) and CAT 6 (2003–2006) math performance: Cross-sectional comparisons.



Note. Third-grade CAT 6 students include all third-grade students from 2003 to 2006. Second-, fourth-, and fifth-grade CAT 6 students include students at these grade levels for 2003 to 2004.

Figure 3.
Saxon middle-school students' Stanford 9 (2000–2002) and CAT 6 (2003–2006) math performance: Cross-sectional comparisons.



Note. Seventh-grade CAT 6 students include all seventh-grade students from 2003 to 2006. Sixth- and eighth-grade CAT 6 students include students at these grade levels for 2003 to 2004.

There were significant, positive changes in math performance from one grade level to the next in elementary and middle schools using *Saxon Math*. Specifically, cross-sectional analyses of students at different grade levels showed that as grade levels increased, so did math performance.

A more precise measure of change is provided when a similar group of students is followed over time (i.e., cohort analyses). For the Stanford 9 sample, this is accomplished by examining the performance of third graders in 2000 and comparing their performance to fourth graders in 2001 and fifth graders in 2002 (see black/dark gray highlighted groups in Table 2). For the CAT 6 sample, since only 2 years of data are available for Grades 2 through 8 (spring 2003 and 2004), researchers examined the performance of three similar groups (i.e., second graders in 2003 compared to third

[14] The Stanford 9 analyses include data only from the 2000 to 2002 school years because the earliest year that schools began using *Saxon Math* was the 1999–2000 school year.

[15] Note that only third and seventh graders had data available from 2003 and 2006. At the remaining grade levels, only 2003 to 2004 data were available due to changes in grade levels tested by the CDE.

graders in 2004, third graders in 2003 compared to fourth graders in 2004, and fourth graders in 2003 compared to fifth graders in 2004). A similar group was also examined at the middle-school level in both the Stanford 9 and CAT 6 sample (see blue highlighted groups in Table 2).

These cohort analyses, however, do not consist of longitudinal analyses; that is, these analyses do not measure growth within individual students. Such analyses could not be done because student identifiers were not provided per the current policy at the California Department of Education. Therefore, caution should be placed in interpreting these results as reflecting true individual change. Nevertheless, these analyses were conducted to obtain a closer approximation of actual change in math performance by following a similar group of students.

Results showed significant growth in the performance of elementary school students from third to fifth grades as measured by the Stanford 9,[16] $F(2, 7336) = 965.8$, $p < .001$, and of students from second to fifth grades as measured by the CAT 6 math tests,[17] $p < .05$ (see Figure 4). Similarly, analyses of middle school students also showed significant growth in performance from sixth to eighth grades on the Stanford 9, $F(2, 11296) = 628.4$, $p < .001$, and CAT 6 math tests, $p < .05$ (see Figure 5).

Figure 4.
Saxon students' Stanford 9 (2000–2002) and CAT 6 (2003–2004) math performance: Elementary cohorts.

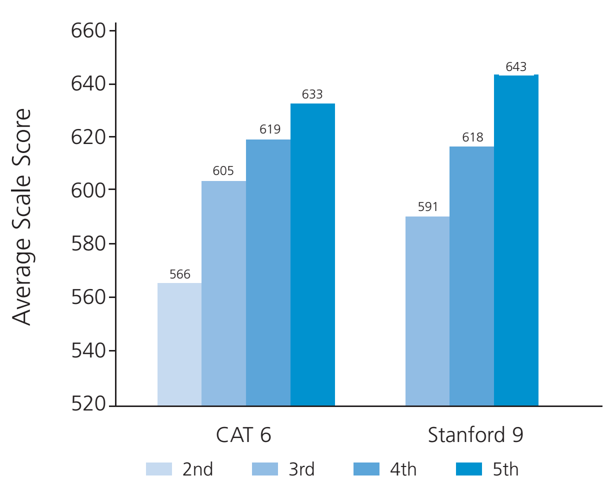
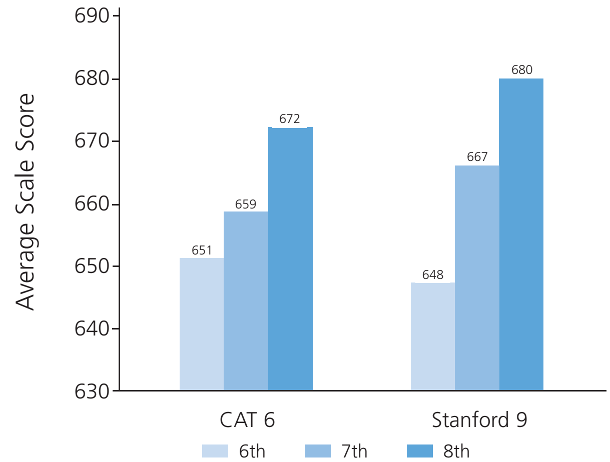


Figure 5.
Saxon students' Stanford 9 (2000–2002) and CAT 6 (2003–2004) math performance: Middle-school cohorts.



Cohort analyses on the Stanford 9 and CAT 6 revealed that elementary and middle-school Saxon students showed significant increases in math performance as they progressed from one grade level to the next.

Stanford 9 and CAT 6 math performance among Saxon elementary and middle school students increased significantly as they progressed through grade levels.

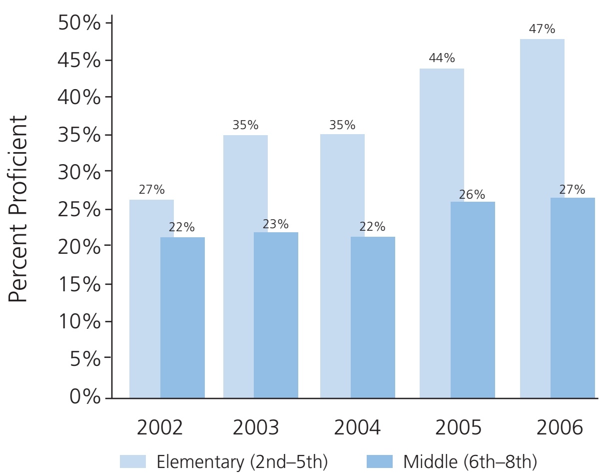
In addition, data on the criterion-referenced test, the CST, was analyzed to examine the extent to which Saxon students are meeting California math standards. As previously noted, as a result of the changing standards from grade to grade and the general decrease in performance from elementary and middle-school grade levels observed statewide, analyses consisted of examining changes in performance trends over time. As is shown in

[16] Because the earliest year in which schools used Saxon was the 1999-2000 school year, 2nd grade performance from 1999 is not included. These students were not exposed to Saxon Math. Analyses of pre-post differences in Saxon use are examined later in this report.

[17] Detailed statistical tables (A1-A2, pgs. 44-45) and how the cohorts in this sample are structured are available in the Appendix. Also note that the average mean is presented for 3rd grade (2003 & 2004) and 7th grade (2003 & 2004) in Figures 4 and 5.

Figure 6, *generally,* among both elementary and middle-school students, there was an increasing trend in the percentage of students meeting state math standards. Specifically, results showed a significant relationship such that, as the school year increased (from 2002 to 2006), so did the percentage of Saxon elementary students meeting the California math standards, $F(1, 70784) = 1309.6$, $p < .001$. In addition, with the exception of the slight drop in performance in 2004, results also showed an increasing trend among middle school Saxon students, $F(1, 61725) = 150.6$, $p < .001$.

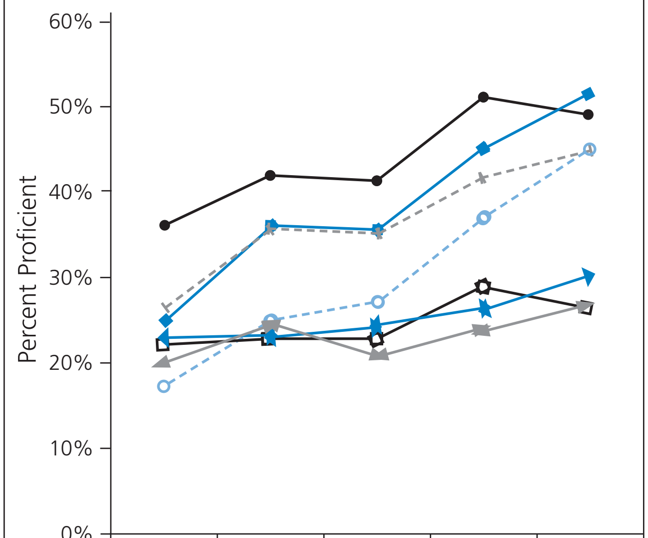
Figure 6.
Percentage of students proficient on the *California Standards Test* in math by school type.



Generally, the percentage of students meeting California math standards in Saxon elementary and middle schools increased over time.

For descriptive purposes and to better understand the observed trend in CST math performance, Figure 7 presents the percentage of students proficient for each grade level separately. As shown, it appears that this trend is strongest among the elementary-school grade levels (2–5) compared to middle-school grade levels (6–8).

Figure 7.
Percentage of students proficient on the *California Standards Test* in math by grade level.



	2002	2003	2004	2005	2006
2	36%	43%	42%	51%	50%
3	26%	36%	35%	46%	52%
4	27%	35%	35%	41%	45%
5	18%	25%	29%	37%	42%
6	22%	23%	23%	29%	27%
7	23%	23%	24%	27%	30%
8	20%	24%	21%	23%	26%

Across all grade levels (second to eighth), there tended to be an increase in the percentage of Saxon students meeting California math standards over the past 5 years. This pattern is strongest among students in elementary grade levels.

Are Changes in math performance related to the number of years a school has been using *Saxon Math*?

The degree of change in school-level math performance (from Spring 2002 to 2003[18]) as a function of the number of years a school had used Saxon (i.e., school exposure) was examined.[19] Saxon

[18] The Stanford 9 test and CAT 6 are excluded because these tests do not provide data at the two target years (2002–2003), where the effect of exposure can be readily assessed.

schools had used the program for either 1 (n = 20) or 4 (n = 24) years by Spring 2003. Since all schools that began to use the *Saxon Math* program in the 2002–2003 school year were elementary schools, analyses were conducted at this level only.

Results showed that number of years a school was exposed to Saxon was not significantly related to school growth in math performance, as measured by the percentage of students meeting math standards on the CST, $F(1, 42) = .01$, $p = .93$. This means that any effect the program has on math performance is *unlikely* to be dependent on how much time a school has used the program. For example, a school that had just begun implementing the program showed the same level of growth from one year to the next as a school that had used it for 4 years.

These findings are consistent with those found in our prior archival study, which examined the impact of exposure to *Saxon Math* in the states of Texas and Georgia. Namely, the amount of exposure had no relationship with growth in test scores. Together, these findings suggest that the Saxon program is fairly easy to learn and implement by teachers (i.e., there is a small learning curve) and as such, effects are likely to quickly manifest.

Growth in math performance among Saxon schools is not dependent on how long a school has used the program. That is, schools that have used Saxon Math *for shorter periods of time generally show the same amount of change as schools that have used the program for longer periods of time.*

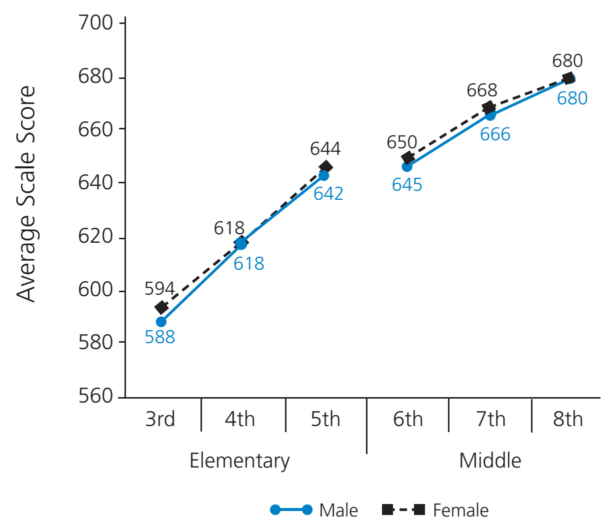
2. Does achievement across Saxon students vary depending on the type of student?

In order to obtain preliminary information on the performance of different types of Saxon students, analyses were conducted to examine if subgroups of students (defined by gender, English language learner status, disability status, economically disadvantaged status, and ethnicity) showed different patterns of math performance over time.

[19] Because school-level data allowed for the analyses of change over time *within* schools, analyses of the effects of exposure on changes in performance were conducted at the school level.

In addition, analyses examined whether within each subgroup there was significant change. For these analyses, cohorts consisting of a similar group of students followed over time, were examined (see black and blue highlighted groups in Table 2). For example, for the Stanford 9 sample, third graders in 2000 are compared to fourth graders in 2001 and fifth graders in 2002.[20] Similarly, sixth graders in 2000 are compared to seventh graders in 2001 and eighth graders in 2002. Figures 8 through 17 show the patterns in Stanford 9 and CAT 6[21] math performance for students in special populations.

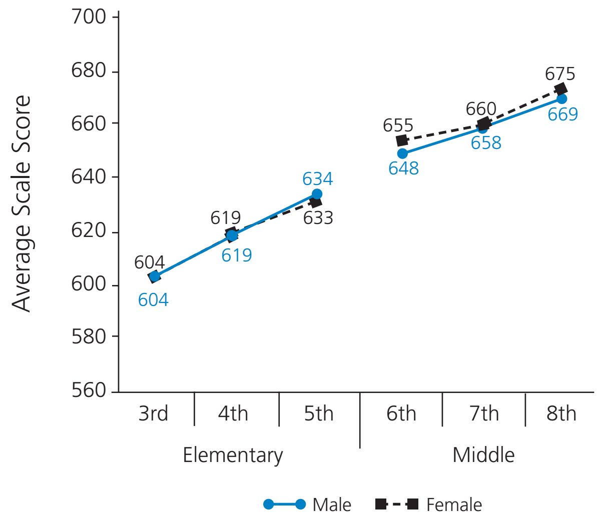
Figure 8.
Stanford 9 math performance by gender.



[20] Since researchers wanted to examine change after students were exposed to *Saxon Math*, second graders (in 1999 pre-Saxon) are excluded. Also note that because cohort analyses are not supported by CST data, this assessment is excluded in this section.

[21] To ease the presentation of results for the CAT 6 sample, the average fourth-grade scale score and average seventh-grade scale score between 2003 and 2004 are presented in the figures. For actual means, see the Appendix.

Figure 9.
CAT 6 math performance by gender.



On both the Stanford 9 and CAT 6, Saxon students who were females and males showed increasing patterns in math performance as they progressed from one grade level to the next, $p < .05$.

Figure 10.
Stanford 9 math performance by ethnicity.

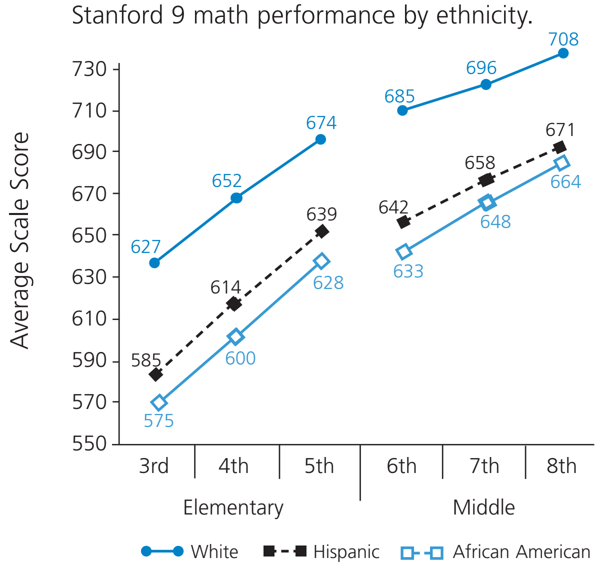
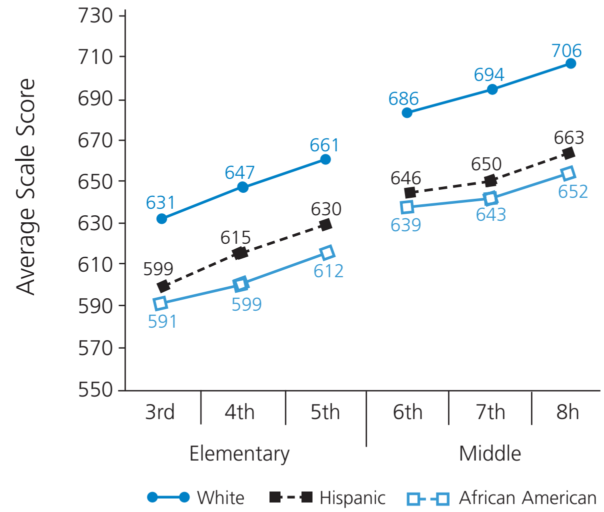


Figure 11.
CAT 6 math performance by ethnicity.



On the Stanford 9 and CAT 6, results show that Whites, Hispanics, and African Americans in elementary and middle-school grade levels showed increasing patterns in math performance as the progressed from one grade level to the next, $p < .05$. In general, increases in math performance did not depend on the ethnicity of Saxon students.

Figure 12.
Stanford 9 math performance by
English language learner status.

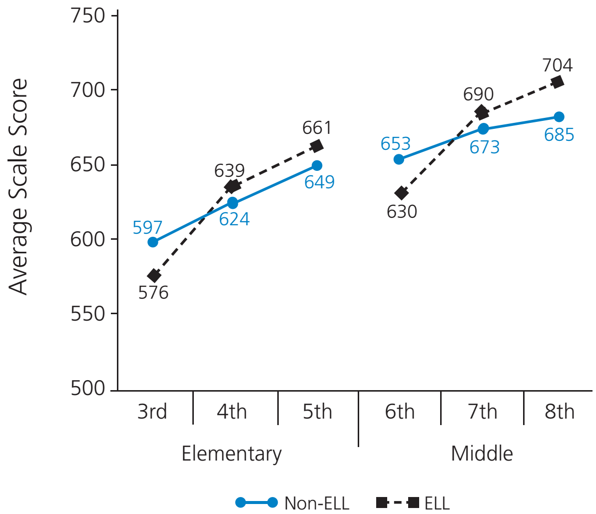
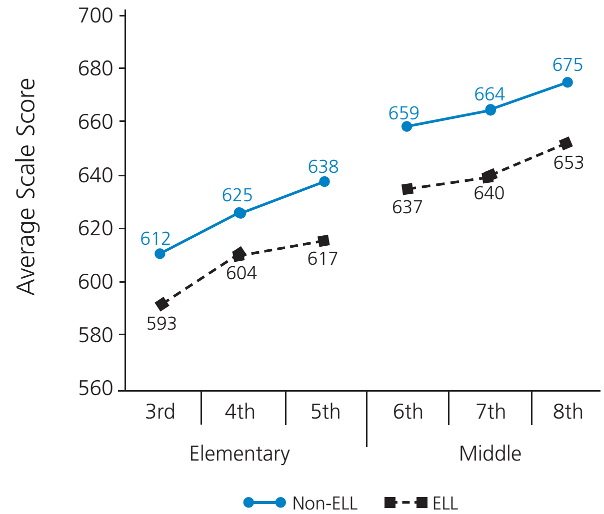


Figure 13.
CAT 6 math performance by
English language learner status.



On both measures, Saxon English language learners and non–Englishs language learners in elementary and middle-school grade levels showed significant upward progress in math performance as they progressed from one grade level to the next, $p < .05$. In addition, among the Stanford 9 sample only, there was an interaction between ELL status and grade such that Saxon elementary and middle-school ELL students started off with lower math performance and then subsequently outperformed non-ELLs,[22] $p < .05$.

[22] Note that this may be the result of the significant drop in students classified as ELLs from the third to fourth grade and sixth to seventh grade, $p < .05$. This may have influenced the findings. See Table A5 in the Appendix for sample sizes.

Figure 14.
Stanford 9 math performance by economic disadvantage status.

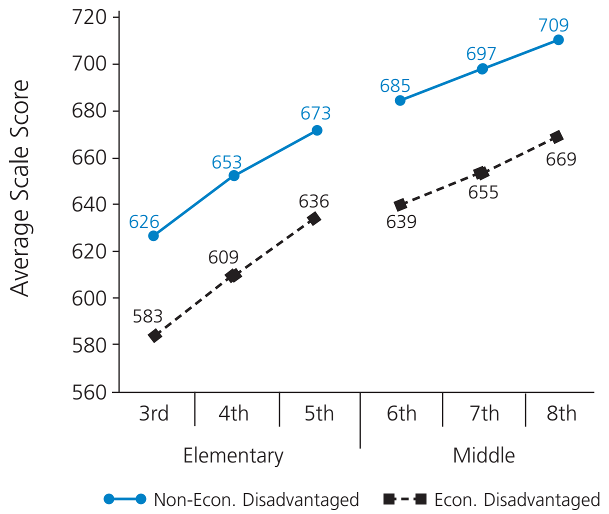
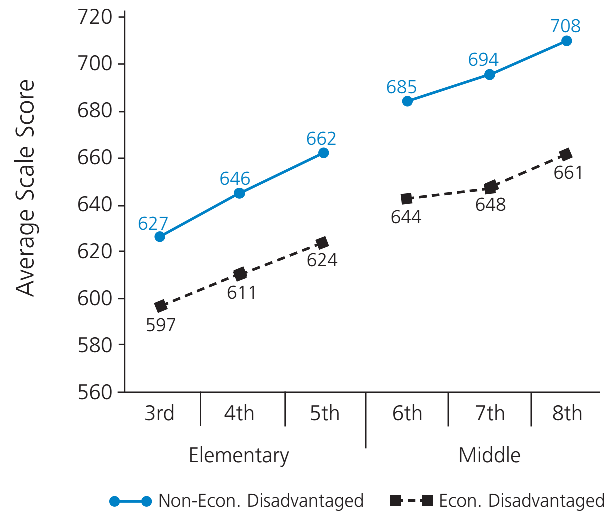


Figure 15.
CAT 6 math performance by economic disadvantage status.



On the Stanford 9, a significant interaction emerged such that there were greater positive changes among economically disadvantaged Saxon students compared to non-economically disadvantaged Saxon students, $p < .05$.
On both tests, economically disadvantaged and non-economically disadvantaged Saxon elementary and middle school students showed significant increases in math performance as they progressed from one grade level to the next, $p < .05$.

Accelerated rates of improvement were observed among ELL and economically disadvantaged Saxon students on the Stanford 9. That is, there were greater positive changes among ELL and economically disadvantaged students compared to non-ELLs and non–economically disadvantaged students, respectively.

Figure 16.
Stanford 9 math performance by disability status.

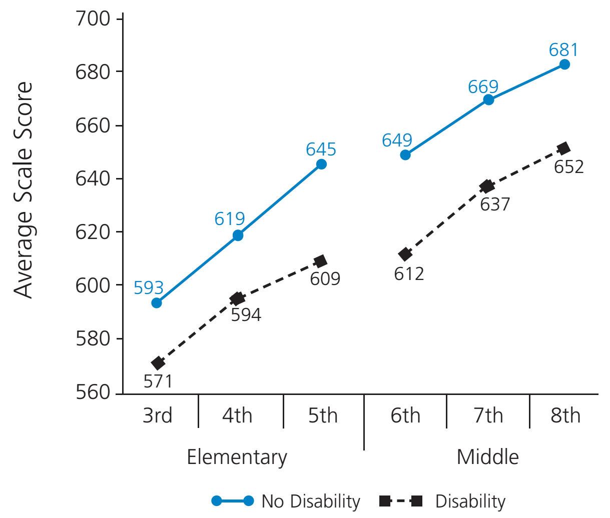
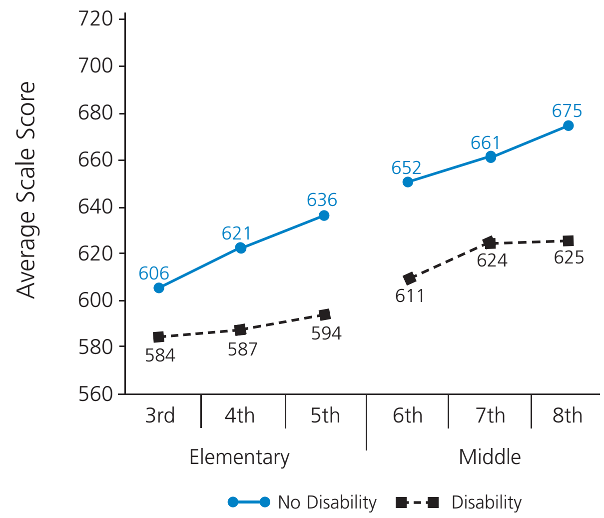


Figure 17.
CAT 6 math performance by disability status.



On the Stanford 9 sample, Saxon elementary and middle-school students with and without disabilities[23] showed improvement in math performance as they progressed from one grade level to the next, $p < .05$. However, on the CAT 6, elementary and middle-school students without disabilities showed significant improvement in test performance, $p < .05$, while students with disabilities showed no significant change (i.e., they performed at a stable pace from one grade level to the next), $p < .05$.

In summary, results showed significantly increasing trends in math performance among all subgroups of students and across all measures, $p < .05$,[24] with the exception of students with disabilities. Among students with disabilities, significant improvement was observed on the Stanford 9, and a steady rate of math performance was observed on the CAT 6. In general, Saxon elementary and middle-school students in special populations (i.e., females, minorities, economically disadvantaged students, ELL students, and Stanford 9 students

[23] This is defined as students with Individualized Education Programs (IEPs).

[24] Detailed statistics are provided in Tables A3 through A7 in the Appendix.

with disabilities) and those not in these special populations have consistently shown increases in math performance in both the elementary and middle-school level.

It should be noted that these results are consistent with those found in the analysis of Texas and Georgia statewide assessment data and a randomized control trial (Resendez & Azin, 2006). Among students who use *Saxon Math,* growth among all different types of students has been consistently observed.

Among Saxon students, there were consistent increasing patterns of Stanford 9 and CAT 6 math performance among females, males, Hispanics, African Americans, Whites, ELLs and non-ELLs, economically disadvantaged and non–economically disadvantaged, and students without disabilities.

3. How does math performance differ between Saxon and non–Saxon schools?

his set of analyses provides information on the relationship between *Saxon Math* and math performance relative to non–Saxon students. In order to address this question, analyses of covariance were performed on student- and school-level data.

Student-Level Analyses

For the student-level analyses, separate analyses were run for students at the elementary level (Grades 2–5) and middle-school level (Grades 6–8). In addition, the following variables were used as covariates in an effort to equate Saxon and non–Saxon students in terms of important demographic characteristics: [25]

- Gender
- Disability status
- White
- Hispanic
- African American
- Asian
- Migrant status
- Economically disadvantaged status

It should be noted that even when available,[26] pre–Saxon performance could not be used as a covariate to equate groups in the student-level analyses. Recall that student identifiers were not provided, and therefore no matching of student performance is possible. As such, it is not possible to control for math performance prior to the use of *Saxon Math.* Instead, the student-level analyses, when possible, examines whether there are significant differences between groups at different years (e.g., before adoption of *Saxon Math*). Lack of differences before the introduction of *Saxon Math* provides support that the groups are similar and that any differences observed afterwards are likely the result of the program. However, the existence of baseline differences indicates that the groups are not equivalent and that results should be interpreted with caution.

Cross-Sectional Analyses

Results are first presented for cross-sectional analyses. These analyses involved comparing average Saxon and non–Saxon math performance over time (e.g., elementary Saxon and non–Saxon students' math performance in 2000 vs. their performance in 2001).

Cross-Sectional Results for the Stanford 9 Sample

For the Stanford 9 elementary student sample, results showed a significant interaction between year and group, $F(4, 157720) = 36.46, p < .001$.[27] As is shown in Figure 18, Saxon elementary students had similar math scores in Spring 1999, prior to the introduction of *Saxon Math,* $p > .05$. However, during the spring following adoption of *Saxon Math*

[25] Due to extensive missing data (approximately 43,000) associated with the ELL-status variable, it is excluded as a covariate.

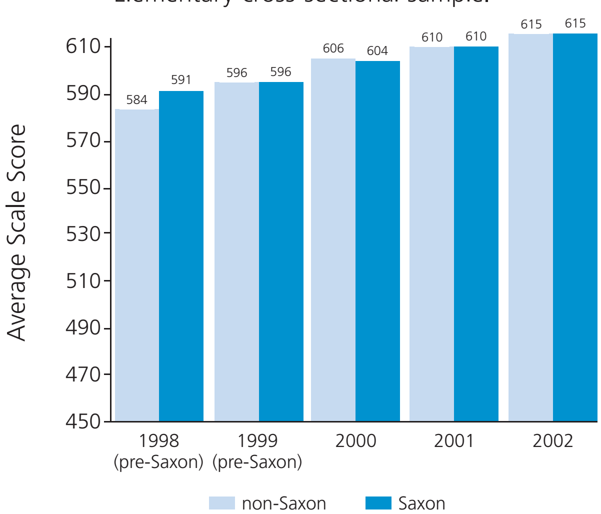
[26] In particular, pre-Saxon performance is available for the Stanford 9 and CST samples. Pre–post data is not available for the CAT 6 sample because this test started in 2003, and at this point all schools in the Saxon sample had begun to use *Saxon Math*.

[27] Detailed statistical tables of results in this section are shown in Tables A8 through A10 in the Appendix.

(2000), there was a significantly lower performance of Saxon students compared to non–Saxon students, $F(1, 31759) = 5.84$, $p = .02$, $d = .03$. Note that schools had not had a full school year of exposure at this point.[28] During spring 2001 to 2002, Saxon and non–Saxon students showed similar math performance.

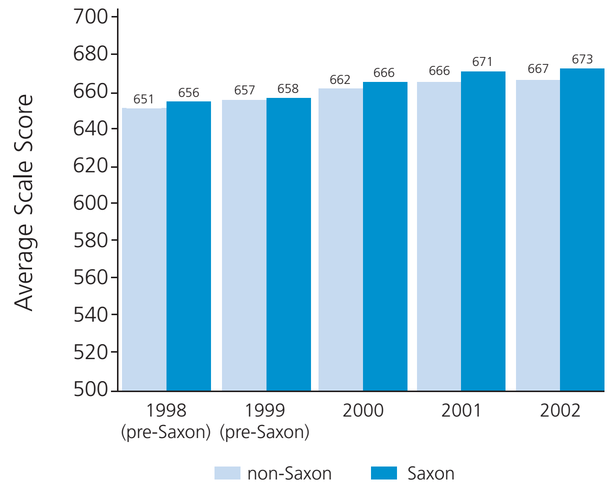
The Stanford 9 middle-school sample also showed a significant interaction between year and group, $F(4, 145308) = 17.74$, $p < .001$. In particular, there tended to be a higher discrepancy in math performance, in favor of Saxon schools, following adoption of *Saxon Math* (i.e., 2000–2002) (see Figure 19). That is, the differences between Saxon and non–Saxon students, in which Saxon students had higher math scores, were greater after 1999 compared to before they used *Saxon Math*. However, it is important to not ignore the finding that Saxon middle schools also had higher math performance prior to the use of *Saxon Math* (1998–1999), $p < .05$.

Figure 18.
Stanford 9 math performance by group and year: Elementary cross-sectional sample.



Note. Means adjusted for demographic covariates.

Figure 19.
Stanford 9 math performance by group and year: Middle school cross-sectional sample.



Note. Means adjusted for demographic covariates.

Cross-sectional analyses (i.e., examining changes over time) revealed that overall, there was no notable relationship between *Saxon Math* use and greater math performance relative to non-Saxon users among the elementary students tested via the Stanford 9. However, the results from the middle-school sample suggest an improvement in math performance following the introduction of *Saxon Math*.

Cross-Sectional Results for the CAT 6 Sample

It should be noted that the CAT 6 sample does not allow for the examination of baseline differences. This is because the CAT 6 was first administered in 2003, and at this point, all schools in the Saxon sample had begun to use *Saxon Math*. Thus, data involves post–*Saxon Math* performance.[29]

Results among the elementary sample showed no significant interaction between year and group, $F(3, 82913) = .79$, $p = .50$. This means that differences between Saxon and non–Saxon students were consistent over the years. Indeed, as is shown in Figure 20, Saxon elementary students outperformed

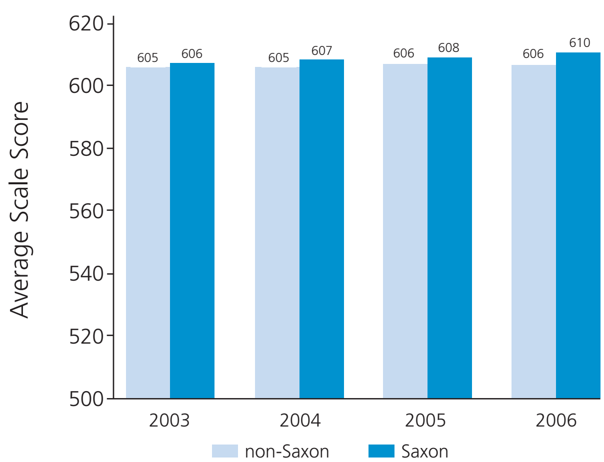
[28] The trivial effect size is also of note. Such a small effect size of $d = .03$ means that the difference is not really meaningful or important. More elaboration on this point follows.

[29] Without pre–Saxon data, it is not possible to rule out the possibility that differences observed may be the result of preexisting differences.

non–Saxon elementary students from 2003 to 2006, $F(1, 82913) = 18.74, p < .001, d = .03$.

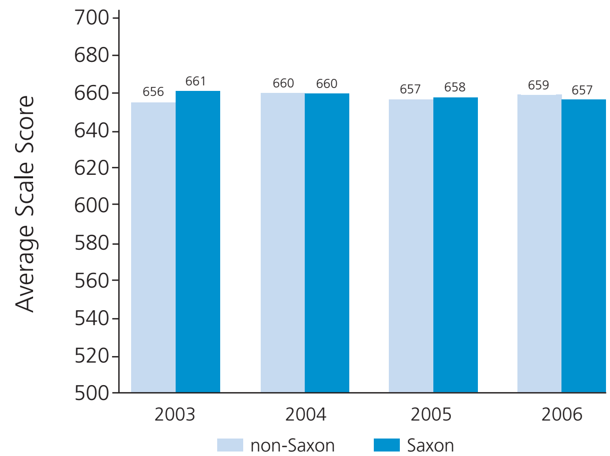
Among the middle-school students, results were not as clear (see Figure 21). In particular, while the performance of Saxon students in 2003 was significantly higher than that of non–Saxon students, $F(1, 33442) = 47.50, p < .001, d = .09$, there were no differences between groups in 2004 and 2005, $p > .05$. However, on the most recent assessment (2006), there were again significant differences, but this time, in favor of non–Saxon students, $F(1, 11441) = 4.99, p = .03, d = .04$.[30] When math performance is averaged across all school years, Saxon students outperformed non–Saxon students, $F(1, 90419) = 4.55, p = .03, d = .01$.

Figure 20.
CAT 6 math performance by group and year: Elementary cross-sectional sample.



Note. The 2005 and 2006 CAT 6 data consist of third graders only. For 2003 and 2004, data consists of second to fifth graders. Means are adjusted for covariates.

Figure 21.
CAT 6 math performance by group and year: Middle-school cross-sectional sample.



Note. The 2005 and 2006 CAT 6 data consist of seventh graders only. For 2003 and 2004, data consists of sixth to eighth graders. Means are adjusted for covariates.

Among elementary students tested via the CAT 6, Saxon students showed higher math performance than did non-Saxon students from 2003 to 2006. The relationship between math performance and *Saxon Math* in the middle-school sample is less clear; at times, Saxon students performed better, and at another time non-Saxon students did. Across all school years, however, Saxon students had higher math scores than did non-Saxon students.

Cross-Sectional Results for the CST Sample

A subset of Saxon elementary schools had pre–Saxon *California Standards Test*[31] data available. As is shown in Figure 22, there was a significant difference in baseline math performance, $F(1, 24356) = 109.79, p < .001, d = .13$. Specifically, Saxon students started out with higher test scores compared to non–Saxon students. Thus, although Saxon students consistently performed

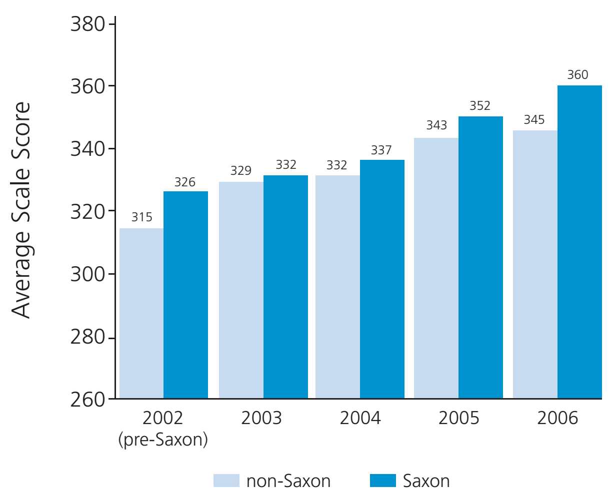
[30] An alternative explanation to these results is that they involve different samples (i.e., sixth through eighth graders in 2003–2004 and seventh graders only in 2005–2006). However, analyses were also conducted among seventh graders only from 2003–2006 and these results were consistent with those displayed previously.

[31] Specifically, analyses consisted of students in schools that began using *Saxon Math* in 2003, since these elementary schools had pre–post data, and all non–Saxon elementary schools to ensure comparability in terms of grade levels. In addition, instead of proficiency levels, analyses involving comparisons between Saxon and non–Saxon students included the CST scale score, as this is a more sensitive measure.

higher than non–Saxon students, these differences may be the result of preexisting differences in math performance. Nevertheless, the significant interaction between year and group, $F(4, 118499) = 21.06$, $p < .001$, suggests that the improvements in math performance were greater among Saxon students than it was among non–Saxon students.

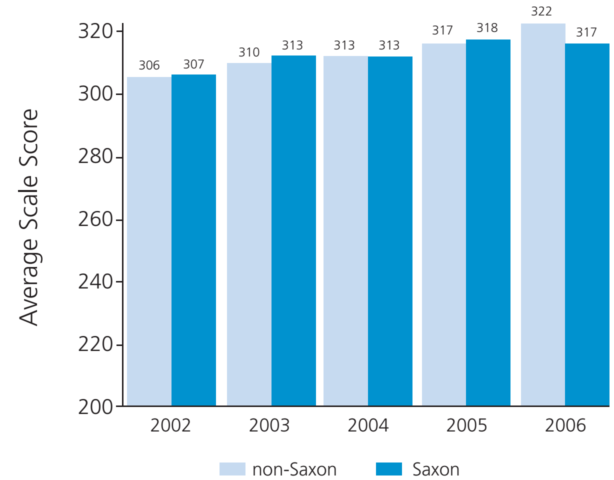
Because all middle schools in the Saxon sample began to use *Saxon Math* in the 1999–2000 school year, no pre–Saxon data is available for the middle school CST sample. Thus, analyses of the middle school students include only post–*Saxon Math* performance. Results among the middle-school sample were not consistent. In particular, while Saxon students showed higher performance in 2003, $F(1, 33209) = 12.51$, $p < .001$, $d = .09$, and 2005, $F(1, 33705) = 4.25$, $p = .04$, $d = .02$, non–Saxon students showed higher performance in 2006 $F(1, 33331) = 30.05$, $p < .001$, $d = .06$. In addition, across all school years, there was no significant difference between Saxon and non–Saxon middle school students, $F(1, 165447) = .01$, $p = .93$.

Figure 22.
California Standards Test math performance by group and year: Elementary cross-sectional sample.



Note. Means adjusted for demographic covariates.

Figure 23.
California Standards Test math performance by group and year: Middle school cross-sectional sample.



Note. Means adjusted for demographic covariates.

While there were significant differences on the CST between elementary students prior to using *Saxon Math* and non-Saxon students (in 2002), results suggested that improvements in math performance were greater among Saxon students than were among non-Saxon students. Overall, results among the middle-school students showed no significant relationship between *Saxon Math* use and math performance relative to non-Saxon students.

Given the large sample sizes involved, it is critical to examine effect sizes,[32] which represent a measure of the relative *importance* of differences observed. Although the large sample sizes involved increases the ability to detect differences, it also facilitates the detection of trivial or unimportant relationships. Examination of the effect sizes (refer to Tables A6–A8 in the Appendix) shows that the overall program effects were small ($d = .01$ to $.18$). One way to understand what these effect sizes mean is to examine the performance of Saxon students

[32] Effect size (ES) is commonly used as a measure of the magnitude of an effect of an intervention relative to a comparison group. It provides a measure of the relative position of one group to another. For example, with a moderate effect size of $d = .5$, we expect that about 69% of cases in Group 2 are above the mean of Group 1, whereas for a small effect of $d = .2$ this figure would be 58%, and for a large effect of $d = .8$ this would be 79%.

relative to non–Saxon students. With a small effect size of .18 (the largest effect size obtained), we could expect that about 57% of students using Saxon perform higher than the average of non–Saxon students. This is quite small and does not exceed the .25 value that Slavin (1986), a leader in educational research, notes as being *educationally significant*. Because the obtained effect sizes are below this threshold, the results between Saxon and non–Saxon students can be considered weak. In other words, both groups (Saxon and non-Saxon) generally showed increases in performance, and although at times the performance of Saxon students was higher than those of non-Saxon students (and vice versa), the focus should be on the positive changes themselves and not necessarily on differences between the groups. Note that small effect sizes are to be expected in any type of study that evaluates entire curricula against one another, given the similarities in content coverage. It must be emphasized that such overlap between curricula will reduce effect sizes. Typically, effect sizes found in comparisons of entire curricula are small to very small.

Overall, cross-sectional analyses showed that both groups (Saxon and non-Saxon) generally showed improvement in performance over time. While, on most measures and years, the performance of Saxon students was higher than that of non–Saxon students, given the small effect sizes observed, the focus should be on the positive changes themselves and not necessarily on differences between the groups.

Cohort Analyses

In order to better understand the relationship between math performance and *Saxon Math,* analyses were conducted to compare a similar group of Saxon and non–Saxon students over time (i.e., cohort analyses). To reiterate, these analyses have the strength of allowing for comparisons of changes over time within what should be similar groups of students. In addition, pre–Saxon data is included for the Stanford 9[33] sample in order to examine whether there are preexisting differences in this cohort sample. As previously noted, two cohorts are available in this dataset: one measured via the Stanford 9 (1999–2002) and the other measured by the CAT 6 (2003–2004). The latter is limited in that we must compare cohorts over only 2 years, from 2003 to 2004. Thus, five groups of CAT 6 students are examined: (a) second graders in 2003 and third graders in 2004, (b) third graders in 2003 and fourth graders in 2004, (c) fourth graders in 2003 and fifth graders in 2004, (d) sixth graders in 2003 and seventh graders in 2004, and (e) seventh graders in 2003 and eighth graders in 2004.

Cohort-Analyses' Results for the Stanford 9

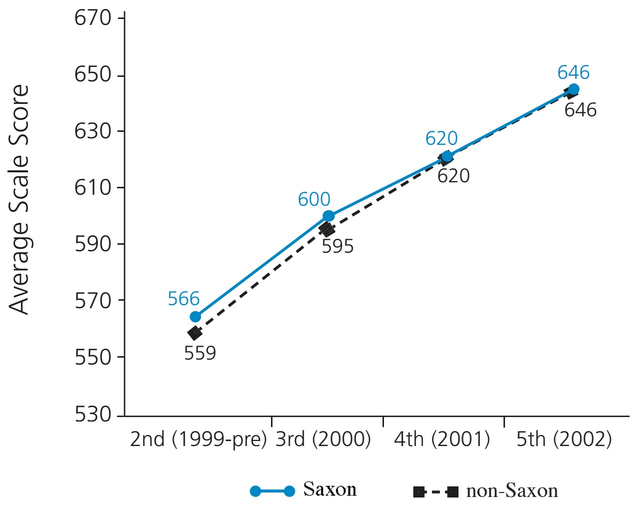
Sample

Results for the Stanford 9 elementary cohort sample showed a significant interaction between time and group, $F(3, 32724) = 29.52$, $p < .001$. Specifically, non–Saxon students tended to show greater rates of improvement than did Saxon students (see Figure 24). Across both groups, there were significant increases in math performance, $F(3, 32724) = 6396.65$, $p < .001$. Furthermore, across all years, there was no significant difference between Saxon and non–Saxon students, $F(1, 32724) = .74$, $p = .39$.

Results for the Stanford 9 middle school cohort sample showed a significant interaction between time and group as well, $F(2, 29033) = 46.47$, $p < .001$. However, the pattern of results was opposite to those found with the elementary sample (see Figure 25). Specifically, Saxon students had a more accelerated increase in test scores. Across all years, there was a significant difference between Saxon and non–Saxon middle-school students such that Saxon students had higher test scores than non–Saxon students, $F(1, 29033) = 40.62$, $p < .001$, $d = .06$, and across both groups there were significant increases in math performance, $F(2, 29033) = 2061.48$, $p < .001$.

[33] This is not possible with the CAT 6 since this test began in 2003, at which point all Saxon schools were using *Saxon Math*. For detailed statistics of these cohort analyses, see Appendix Tables A11 and A12.

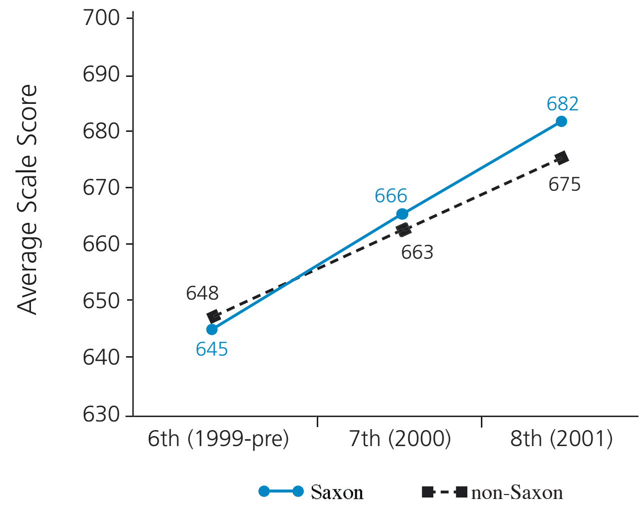
Figure 24.
Stanford 9 math performance by group and grade: Elementary cohort



Note. Means adjusted for covariates.

Although both groups showed improvement in math performance as students progressed from one grade to the next, non-Saxon students tended to show greater rates of improvement than did Saxon students.

Figure 25.
Stanford 9 math performance by group and grade: Middle-school cohort



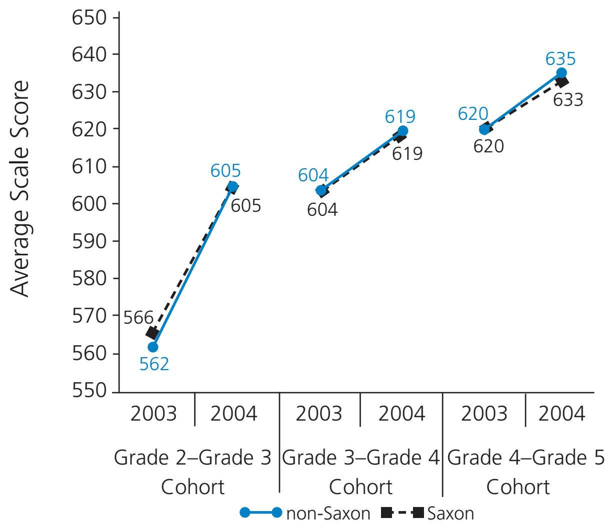
Note. Means adjusted for covariates.

While both groups showed increases in math performance as they progressed from one grade to the next, middle-school Saxon students showed more accelerated improvement in math performance.

Cohort Analyses' Results for the CAT 6 Sample

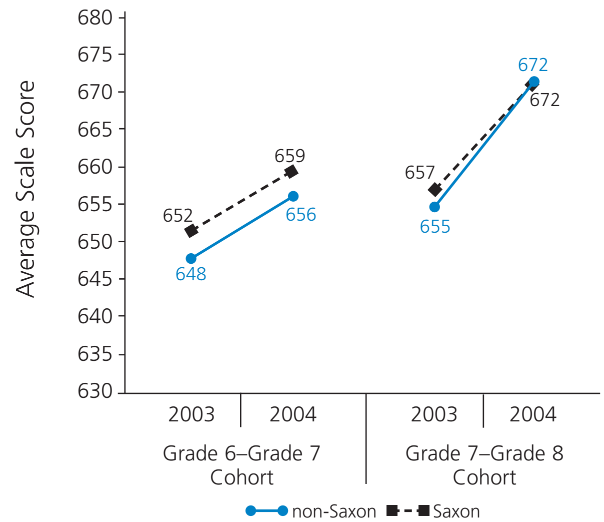
Analyses of the CAT 6 elementary cohort sample showed that, in general, there were significant increases in math performance among both groups of students. That is, consistent with the Stanford 9 elementary sample, there were no overall differences between Saxon and non–Saxon students in improvements in math performance, $p > .05$[34] (see Figure 26). The only exception was the Grade 2 to 3 cohort; there was greater change from second to third grade among non–Saxon students than among Saxon students. Results of the CAT 6 middle-school cohort sample showed that, in general, Saxon students had higher scores than non–Saxon students (see Figure 27). However, among the seventh- to eighth-grade cohort, non–Saxon students showed greater increases in math performance from seventh to eighth grades compared to Saxon students.[35] No such differences were observed among the sixth- to seventh-grade cohort.

Figure 26.
CAT 6 math performance by group and grade: Elementary cohorts



Note. Means adjusted for covariates.

Figure 27.
CAT 6 math performance by group and grade: Middle-school cohorts



Note. Means adjusted for covariates.

In general, results based on the CAT 6 indicate that both Saxon and non–Saxon elementary and middle-school students, followed from one grade level to the next, showed improvement in math performance. The only exception was the changes among the second-to-third and seventh-to-eighth–grade cohorts. In both cases, non–Saxon students showed greater improvement than did Saxon students.

Examination of the effect sizes obtained in these cohort analyses (refer to Tables A11–A12 in the Appendix) shows that the effects were small (d = .06 to d = .22). This means that, while increases in math performance exist, the differences between Saxon and non–Saxon students may not be meaningful.

Results from analyses from similar groups of students followed over time show that, in general, Saxon and non–Saxon students showed similar increases in math performance. Patterns of changes between groups were not consistent enough to allow for more conclusive inferences to be made about differences between groups.

[34] More specifically, for the Grade 2 to Grade 3 cohort, there was significant growth among both samples, but the growth among non–Saxon students was greater, $F(1, 16677) = 5.19$, $p = .02$. The Grade 3 to Grade 4 cohort showed that changes did not depend group, $F(1, 16809) = .12$, $p = .73$. In other words, both groups showed significant increases in math performance, $F(1, 16809) = 426.04$, $p < .001$. The Grade 4 to Grade 5 cohort also did not show a significant interaction, $F(1, 16522) = .16$, $p = .69$. Instead, both Saxon and non–Saxon students showed significant improvement in math performance, $F(1, 16522) = 376.99$, $p < .001$.

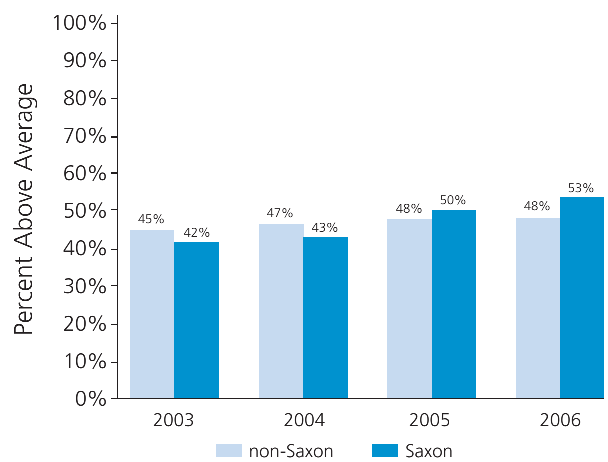
[35] Specifically, for the Grade 7 to Grade 8 cohort, there was significant growth among both samples, but the growth among non–Saxon students was greater than among Saxon students, $F(1, 22987) = 5.46$, $p = .02$. The Grade 6 to Grade 7 cohort showed that changes did not vary by group, $F(1, 22875) = 1.18$, $p = .28$. Instead, there was a significant and similar improvement among Saxon and non–Saxon students, $F(1, 22875) = 110.92$, $p < .001$.

School–Level Analyses

Thus far, analyses have focused on differences in the performance among Saxon and non–Saxon students. These analyses have the following limitations: (a) inability to control for pretest differences as a result of the structure of the dataset, and (b) very large sample sizes facilitate the attainment of significant yet trivial differences as evidenced by the effect sizes obtained ($d = .01$ to $d = .22$). Therefore, to supplement these student-level data, analyses were performed to examine differences at the school level. The advantage of this data is that researchers can control for preexisting differences on the CST and CAT 6 because schools can be readily identified and data across years can be matched to each school.[36] However, this meant selecting Saxon schools that began using *Saxon Math* in 2003, which were all *elementary* schools, controlling for 2002 pre–*Saxon Math* performance, and comparing these to non–Saxon elementary schools. In contrast, control for such preexisting differences was not possible for the Stanford 9 schools because of the lack of pre–Saxon data. Therefore, both elementary and middle schools are included in Stanford 9 analyses.[37] School-level measures included the percentage of students above average in comparison to the national sample of students who took the CAT 6 and Stanford 9, and the percentage of students proficient or advanced on California math standards (see Figures 28–30).

Results of school-level differences using all measures showed no significant interactions between time and group, $p > .05$, after equating groups on important demographic differences and baseline math performance (see Figures 28–30). This means that changes in math performance over the years did not differ significantly between Saxon and non–Saxon schools. However, note that this is school-level analysis, which means that the sample size being used is very small (e.g., 65 to 95 schools). This means that these analyses do not have much power to detect differences between groups that are potentially meaningful.[38] That is, even if noteworthy differences do exist between the Saxon and non–Saxon schools, the small sample size at the school level means that this particular analysis is not sensitive enough to detect such differences.[39] In fact, patterns of results found across all the different analyses conducted suggest that math performance may indeed differ between Saxon and non–Saxon schools; however, it may take time for such differences to emerge. Specifically, the consistent pattern of results obtained with the CAT 6 and CST samples, suggest that changes over time may vary by group. That is, while Saxon schools started out at a lower level in math performance compared to non–Saxon schools, Saxon schools subsequently surpassed non–Saxon schools. This suggests that schools may need some time using Saxon before there is differentiation in performances between Saxon schools and schools using other math curricula. Figures 28 and 29 illustrate these patterns.

Figure 28.
Percentage of elementary students above average on the CAT 6: School level.



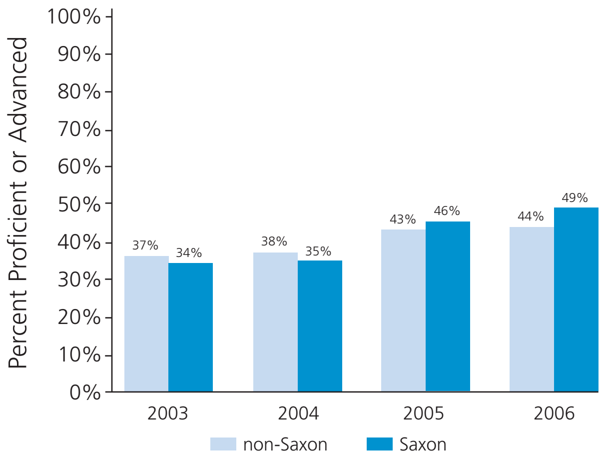
Note. Means adjusted for covariates. In 2005 and 2006, testing occurred in third grade only.

[36] In addition to controlling for pretest differences, other covariates used include enrollment, percentage Hispanics, percentage African Americans, percentage Asians, percentage Whites, percentage of students with disabilities, percentage of ELLs, and percentage of students who are economically disadvantaged. Note, however, that these analyses are limited in that the samples are small (Saxon = 65 schools and non–Saxon schools = 65). See the Appendix, Tables A13–A15 for detailed statistics.

[37] Note that given the small sample, analyses are not conducted for elementary and middle schools separately.

[38] Conversely, when you have very large sample sizes, it is possible to detect trivial effects (e.g., very small differences between groups) that are not educationally meaningful but are still statistically significant. According to Slavin (1986), an effect size of .25 is considered educationally meaningful.

Figure 29.
Percentage of students proficient or advanced on the *California Standards Test:* School level.

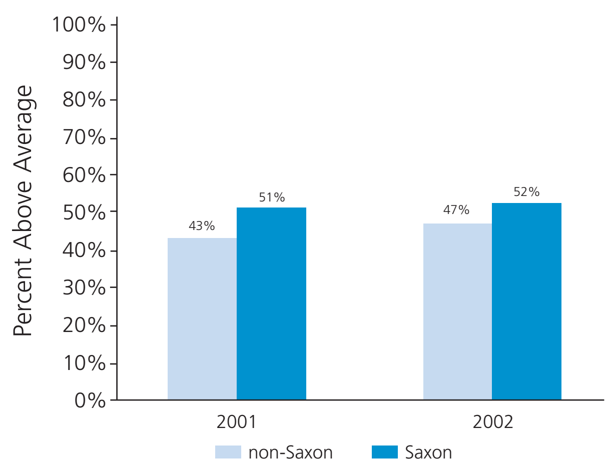


Note. Means adjusted for covariates.

On the CAT 6 and CST, results were similar. Across all school years, there was no significant difference between Saxon and non–Saxon elementary schools. However, an interesting pattern emerged. Specifically, the percentage of elementary students in Saxon schools that were above average tended to be initially lower compared to non–Saxon schools and after controlling for preexisting differences. However, following these 2 years, Saxon schools subsequently surpassed non–Saxon schools on both measures. This suggests that schools may need some time using Saxon before there is differentiation in performance between Saxon schools and schools using other math curricula.

Note that this interpretation does not contradict the earlier findings about the effect of years of *Saxon Math* exposure on math performance. Recall that the prior results showed that Saxon schools that had 1 year of exposure showed similar gains to Saxon schools that had been using the program for 3 years. This means that performance gains *among Saxon schools only* do not depend on the number of years a school has used *Saxon Math*. However, when this growth is compared to other, non–Saxon schools, differences between Saxon and non–Saxon schools may take time to manifest. It is not until schools have used *Saxon Math* for some time that differences between Saxon and non–Saxon schools become apparent. Note that this is also supported, in part, by the Stanford 9 results. As shown in Figure 30, these schools had been using *Saxon Math* for 2 years in 2001 and, at that point, the differences between *Saxon Math* schools and non–*Saxon Math* schools are significant and in favor of *Saxon Math* schools.

Figure 30.
Percentage of elementary students above average on the Stanford 9: School level.



Note. Means adjusted for covariates.

On the Stanford 9, analyses of each school year revealed that Saxon schools performed significantly better than non–Saxon schools in spring 2001.

[39] This refers to being able to detect *statistically significant* differences. Statistical significance is usually determined at the threshold of .05 level or below. "Significant" means that we can be 95% or more confident that the observed differences are real. If this value is greater than .05, it means that any observed differences are not statistically significant and may be interpreted as inconclusive. With small sample sizes, the differences between groups need to be larger in order to attain statistical significance. Thus, while smaller differences between groups may be educationally meaningful, they may still not be significant.

School-level analyses controlling for pre–Saxon differences revealed that Saxon elementary schools show similar levels of math performance as non–Saxon elementary schools when averaged across all years. However, results from the CAT 6 and CST also suggest that Saxon schools may need some time using Saxon before there is differentiation in performances between Saxon schools and schools using other math curricula. More specifically, while Saxon schools started out at a lower level in math performance compared to non–Saxon schools, Saxon schools subsequently surpassed non-Saxon schools.

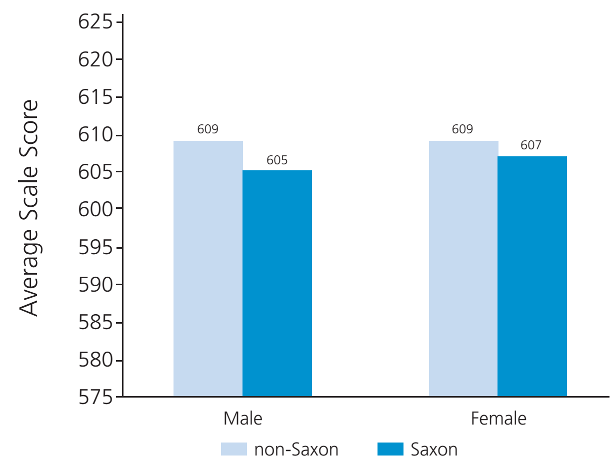
It should be noted that these findings are somewhat consistent with those obtained in prior analyses of Texas and Georgia assessment data, in addition to those obtained in a recent randomized control trial on the effects of *Saxon Math* in the middle school grades (Resendez & Azin, 2006). Specifically, in prior research, middle-school Saxon students tended to outperform non–Saxon students. In the current study, this finding was more evident with the Stanford 9 than with the CAT 6 and CST. In addition, as is consistent with prior research, the relationship between *Saxon Math* use and math performance among elementary students is inconclusive.

Are there differences between subgroups of students in Saxon and non–Saxon Schools?

Data on students in various subgroups (i.e., gender, ethnicity, economically disadvantaged status, ELL status, and students with disability status) were examined to determine whether there were significant differences between students in these subgroups who were in Saxon and non–Saxon schools. Analyses focused on students in active Saxon schools; that is, pre–Saxon data is excluded, and data is analyzed across all years of post–Saxon data. Furthermore, analyses were run separately for elementary students and middle-school students.

Analyses of the Stanford 9 by gender showed that the overall performance[40] of students differed significantly by group and gender among elementary students only, $F(1, 99824) = 12.50$, $p < .001$. Specifically, the difference between Saxon and non–Saxon elementary students, in favor of non-Saxon students, was greater among males than females (see Figure 31). This relationship was not observed at the middle school level, $p > .05$. On the CST and CAT 6, performance of female and male students did not depend on group (i.e., Saxon vs. non-Saxon), and this was found in both elementary and middle-school samples. Thus, overall, across all measures, there does not appear to be a notable interaction between gender and the math performance of Saxon and non–Saxon students.

Figure 31.
Stanford 9 math performance by group and gender: Elementary.



Note. All Saxon and non–Saxon differences are significant, $p = .05$.

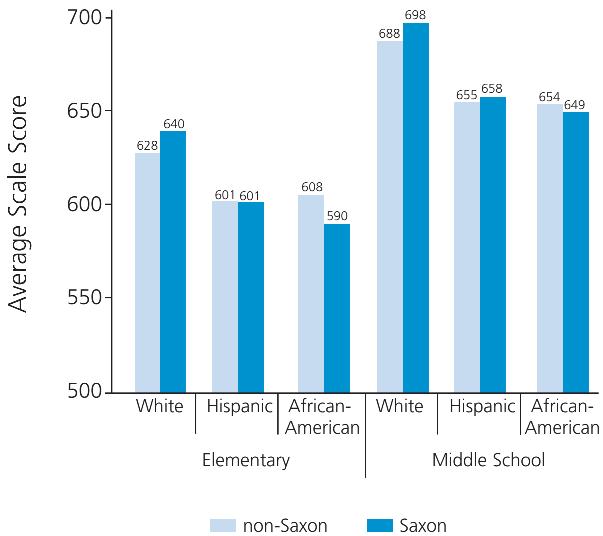
Overall, math performance differences between Saxon and non-Saxon students does not vary by gender.

Examination of differences by ethnicity revealed that among middle and elementary students, Saxon White and Hispanic students tended to perform better than non–Saxon White and Hispanic students, $p < .05$. However, among African American elementary and middle-school students, Saxon students had lower math performance than

[40] The appendix contains detailed statistical tables of all analyses presented in this section (see Tables A16–A20), including interaction and simple effects tests.

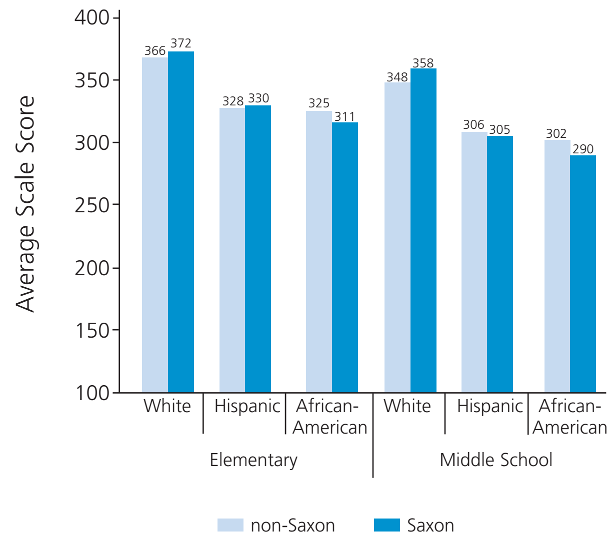
did non–Saxon students. These findings were found consistently across all measures (Stanford 9, CST, and CAT 6), with the exception of the lack of differences among Hispanic elementary students in the Stanford 9 sample (see Figures 32–34).

Figure 32.
Stanford 9 math performance by group and ethnicity: Elementary and middle school.



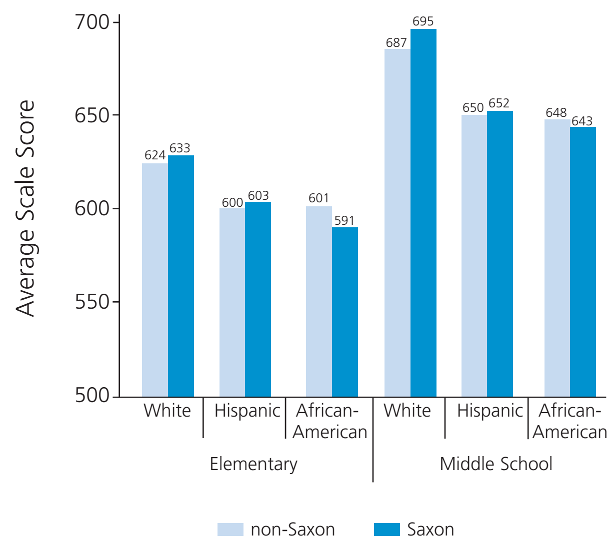
Note. With the exception of Hispanic elementary students, all Saxon and non–Saxon differences are significant, $p < .05$.

Figure 33.
California Standards Test math performance by group and ethnicity: Elementary and middle school.



Note. All Saxon and non–Saxon differences are significant, $p < .05$.

Figure 34.
CAT 6 math performance by group and ethnicity: Elementary and middle school.

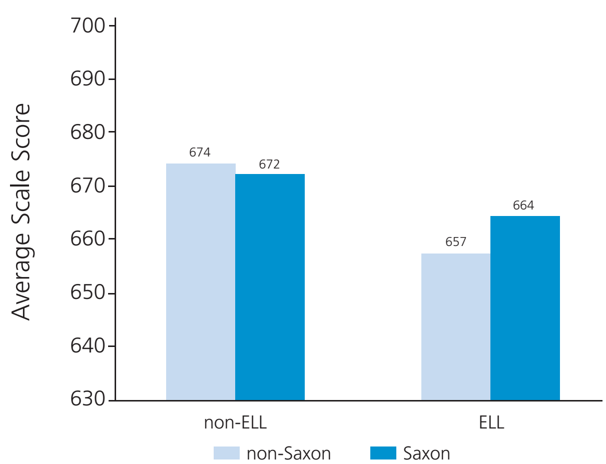


Note. All Saxon and non–Saxon differences are significant, $p < .05$.

> *Whites and Hispanics who attended Saxon schools showed higher math performance compared to these students in non–Saxon schools. However, among African Americans, non–Saxon students tended to show better performance than did Saxon students.*

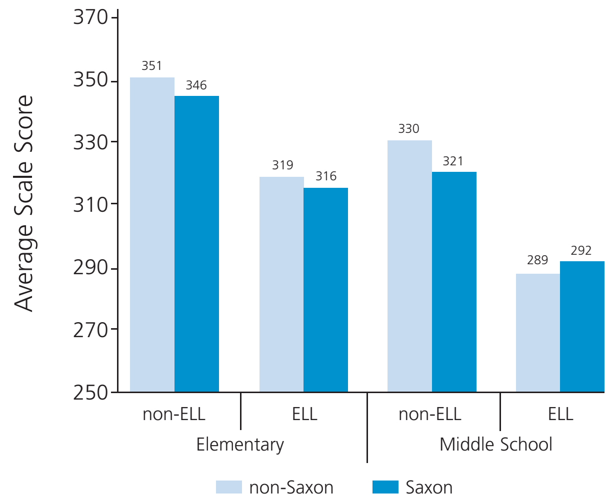
Overall, analyses by ELL status showed that math performance of students differed significantly by group and ELL status among elementary and middle-school students, $p < .05$. As is shown in Figures 35 through 37, among non-ELL students, non–Saxon students tended to show higher math performance than did Saxon students. In contrast, among ELL students, Saxon students had higher math performance than did non–Saxon students. There are two exceptions to this general relationship: (a) this interaction was not observed at the elementary level for the Stanford 9 measure; and (b) among CST elementary ELL students, non–Saxon students performed better than did Saxon students; however, the difference between Saxon and non–Saxon students tended to be smaller among ELL students than among non-ELL students.

Figure 35.
Stanford 9 math performance by group and English language learner status: Middle school.



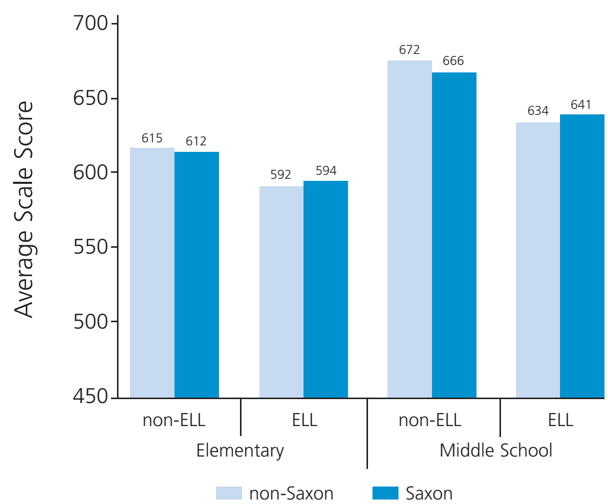
Note. All Saxon and non–Saxon differences are significant, $p < .05$.

Figure 36.
California Standards Test math performance by group and English language learner status: Elementary and middle school.



Note. All Saxon and non–Saxon differences are significant, $p < .05$.

Figure 37.
CAT 6 math performance by group and English language learner status: Elementary and middle school.

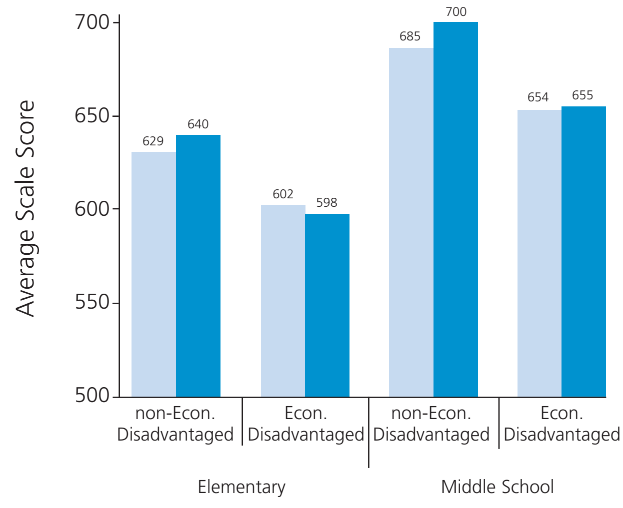


Note. All Saxon and non–Saxon differences are significant, $p < .05$.

Among ELL elementary and middle-school students, Saxon students generally performed better than do non–Saxon students. The opposite pattern was found for non-ELL students.

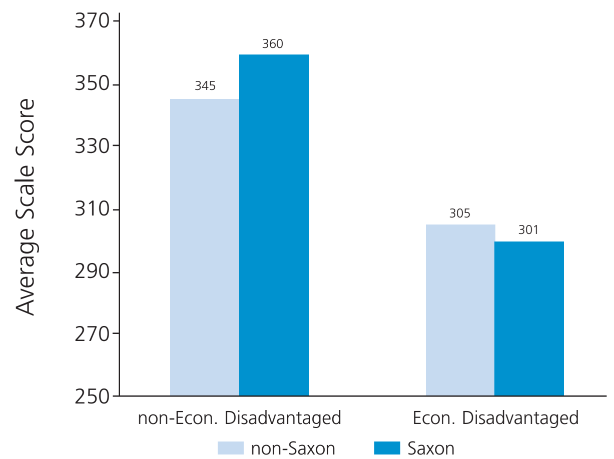
Examination of differences by economic disadvantage status showed significant interactions between group and economically disadvantaged status on math performance, with the exception of the CST elementary sample, $p < .05$. A consistent pattern among non–economically disadvantaged students was observed. Among these students, Saxon elementary and middle-school students tended to outperform non–Saxon students (see Figures 38–40). However, among economically disadvantaged students, results were inconsistent. For example, on the Stanford 9 and CAT 6, *middle-school* Saxon students had higher math performance than did non–Saxon students. However, on the CST, the opposite result was found, with non–Saxon students performing better than did Saxon students. At the *elementary* level, non–Saxon students had higher math performance than did Saxon students as measured on the Stanford 9, but differences were nonexistent as measured by the CAT 6.

Figure 38.
Stanford 9 math performance by group and economically disadvantaged status: Elementary and middle school.



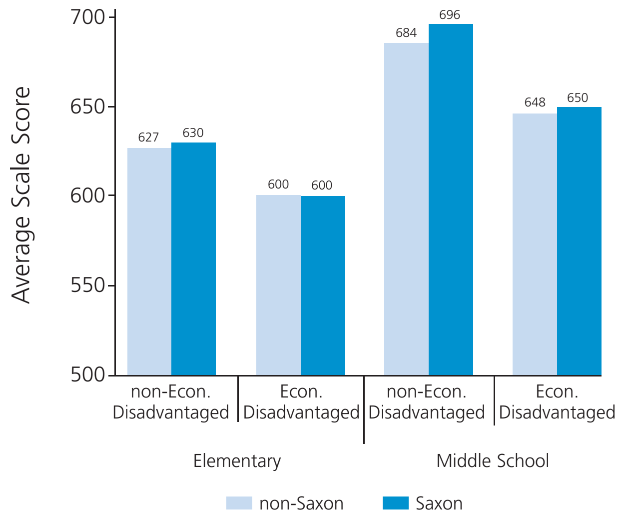
Note. All Saxon and non–Saxon differences are significant, $p < .05$.

Figure 39.
California Standards Test math performance by group and economically disadvantaged status: Middle school.



Note. All Saxon and non–Saxon differences are significant, $p < .05$.

Figure 40.
CAT 6 math performance by group and economically disadvantaged status: Elementary and middle school.

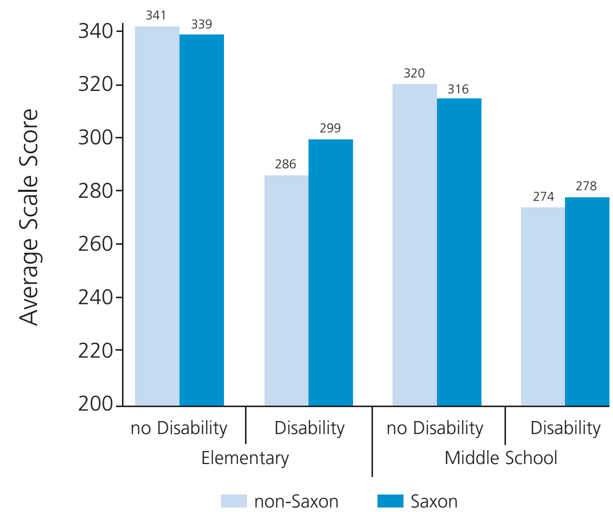


Note. With the exception of elementary disadvantaged students, all Saxon and non–Saxon differences are significant, $p < .05$.

> *Overall, it appears that* Saxon Math *is related to positive differences among non–economically disadvantaged students. However, among economically disadvantaged students, the relationship between* Saxon Math *and student performance relative to non–Saxon students, is unclear.*

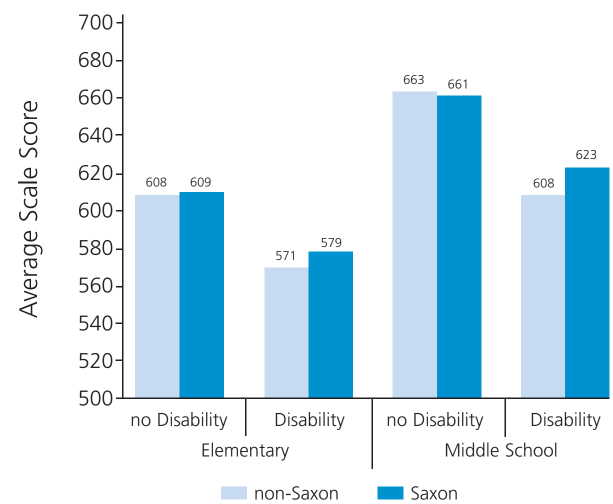
Results by disability status (i.e., students with Individualized Education Programs) revealed that there were significant interactions between disability status and group as measured by the CST and CAT 6, $p < .05$. On both of these measures and in elementary and middle-school samples, students with disabilities who also used *Saxon Math* showed better math performance than did non–Saxon students with disabilities. In contrast, among students without disabilities, non–Saxon students had higher math scores than did Saxon students. The only exception to this was elementary students without disabilities in the CAT 6 sample; for this group, differences between Saxon and non–Saxon students were not significant. In addition, on the Stanford 9, math performance between Saxon and non–Saxon students did not vary as a function of disability status, $p > .05$.

Figure 41.
California Standards Test math performance by group and disability status: Elementary and middle school.



Note. All Saxon and non–Saxon differences are significant, $p < .05$.

Figure 42.
CAT 6 math performance by group and disability status: Elementary and middle school.



Note. With the exception of elementary students without disabilities, all Saxon and non–Saxon differences are significant, $p < .05$.

In general, Saxon Math *students who had disabilities tended to outperform non–Saxon students with disabilities. However, among students without disabilities, non–Saxon students tended to show better math performance than did Saxon students.*

Overall, the findings of these subgroup analyses provide further support that *Saxon Math* is associated with greater math performance among students in certain subpopulations (i.e., Hispanics, ELLs, and students with disabilities). Prior research conducted on the *Saxon Math* curricula also shows significant differences between Saxon and non–Saxon users in special populations (e.g., minorities, economically disadvantaged students, special-education students, and students at risk of dropping out). These findings, along with those obtained in this study, suggest that Saxon may be particularly effective with students who are disadvantaged, as compared to other math curricula. However, given the exploratory, preliminary nature of these analyses, further research is needed to examine this claim more thoroughly.

Summary

Analyses of California statewide assessment data show that the *Saxon Math* program is associated with positive student outcomes. Specifically, significant positive changes in math performance were observed among Saxon elementary and middle-school students across years and grade levels. In addition, these increasing scores were observed among Hispanics, African Americans, Whites, females, males, economically disadvantaged and non–economically disadvantaged, ELL and non-ELLs, and students with disabilities and students without disabilities. In particular, there is some preliminary evidence (i.e., accelerated rates of change) that suggests that *Saxon Math* may work particularly well with ELL and economically disadvantaged students.

However, examination of differences between Saxon and non–Saxon students generally showed no consistent or meaningful differences. That is, both groups tended to show improvements in (or similar) math scores via cross-sectional, cohort, and school-level analyses. Furthermore, while at times Saxon students outperformed non–Saxon students (and vice versa), the small effect sizes obtained (d = .01 to .22), which provide an indication on the importance of findings, suggest that the focus should be on the positive changes themselves and not necessarily on differences between the groups. More consistent differences among students in special populations were observed. In particular, Saxon students who were White, Hispanic, English language learners, non–economically disadvantaged, and had disabilities tended to outperform non–Saxon students. In contrast, non–Saxon students who were African American, non-ELL, and did not have disabilities performed better than Saxon students.

In addition, findings were somewhat consistent with those found in prior Saxon Archival studies conducted using Texas and Georgia statewide assessment data as well as a randomized control trial. Like the results found in these studies, (a) there were positive changes in math performance among Saxon students, and (b) positive changes and differences between Saxon and non–Saxon students in special populations were observed. However, prior research has shown stronger relationships in math performance among Saxon students compared to non–Saxon students. Factors that may have influenced the lack of consistency in results across these various studies include (a) the current study used a number of outcome measures given at various periods of time, and (b) unlike prior research, baseline differences in math performance could not always be controlled. The increased variability evident in the current study is likely to yield more mixed results, making interpretation more challenging.

Limitations

There are several limitations to this study that readers should take into account when interpreting the study's results. First, this study relied on matching procedures employed by the California Department of Education and statistical controls in order to equate groups on important demographic characteristics. However, since it is not a true experiment with random assignment to conditions, there may still be other variables that have not been accounted for that may be producing differential effects, the most likely being preexisting differences in math performance. The only exception to this was the school-level analyses using the CAT 6 and CST samples; in these analyses, pre–*Saxon Math* performance was controlled for.

Secondly, teacher effects could not be examined. Research has shown that teacher quality has significant effects on student achievement (Mendro et al., 1998; Sanders & Rivers, 1996). Unfortunately, due to the retrospective nature of this study, it was not possible to gather information on teacher quality. Related to this, implementation information is not available. Therefore, it is not known *how* teachers implemented *Saxon Math* in their classrooms. Fidelity of implementation is an important construct to consider when examining the effects of interventions, because it gives an indication of whether the teachers are using the program as it was intended.

Third, although the large sample size increases our ability to detect differences, it also facilitates the detection of trivial or unimportant relationships. For this reason, it is important to consider the effect size associated with each analysis. As previously noted, the effect sizes found in this study could be classified as small (d = .01–d = .22). According to Slavin (1986), a leader in educational research, an effect size of .25 is considered educationally significant.

Fourth, generalizability is limited to sites with similar characteristics. This sample was heavily Hispanic and had a higher proportion of English language learners and socioeconomically disadvantaged students than found statewide.

In summary, the results of this study using California state assessment data provides some support for a positive relationship between the *Saxon Math* program in elementary and middle-school levels and math performance. However, stronger (and more conclusive) findings have been obtained in other research on the *Saxon Math* curriculum. Therefore, further research is needed to more fully explore the effectiveness of the *Saxon Math* program.

References

Federal Reserve Board. (2000, September 21). *Testimony of chairman Alan Greenspan: The economic importance of improving math-science education.* Retrieved February 28, 2007, from http://www.federalreserve.gov/boarddocs/testimony/2000/20000921.htm

Glenn, J. (2000). *Before it's too late: A report to the nation from the National Commission on Mathematics and Science Teaching for the 21st Century*. Washington, DC: U.S. Department of Education.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Mendro, R. L., Jordan, H. R., Gomez, E., Anderson, M. C., & Bembry, K. L. (1998). *An application of multiple linear regression in determining longitudinal teacher effectiveness.* Dallas, TX: Dallas Public Schools.

Mullis, V. S., Martin, M. O., & Foy, P. (2005). *TIMSS 2003 international report on achievement in the mathematics cognitive domains.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center. Retrieved September 25, 2006, from http://timss.bc.edu/PDF/t03_download/T03MCOGDRPT.pdf

National Assessment of Educational Progress. (2005). *The nation's report card.* Washington, DC: National Center for Education Statistics.

National Research Council. (2001). Adding it up: Helping children learn mathematics. In J. Kilpatrick, J. Swafford, & B. Findell (Eds.), *Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education.* Washington, DC: National Academy Press.

Resendez, M., & Azin, M. (2006). *Final report: Saxon Math randomized control trial*. Jackson, WY: PRES Associates.

Resendez, M., Fahmy, A., & Manley, M. (2004). *The relationship between using Saxon Math and student performance on Texas statewide assessments*. Jackson, WY: PRES Associates.

Resendez, M., Sridharan, S., & Azin, M. (2005). *The relationship between using Saxon Elementary and Middle School Math and student performance on Georgia statewide assessments*. Jackson, WY: PRES Associates.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student achievement.* Knoxville: University of Tennessee.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analysis and traditional reviews. *Educational Researcher*, *15*, 5–11.

U.S. Department of Education. (2006). *Math Now: Advancing math education in elementary and middle school.* Retrieved October 18, 2006, from http://www.ed.gov/about/inits/ed/competitiveness/math-now.html

Appendix: Tables of Statistical Results

The following tables display statistical results for ANCOVA, ANOVA, and repeated measures ANOVA.

For the majority of these analyses, a "significant" difference means that we can be 95% or more confident that the observed differences are real. If the significance level is less than or equal to .05, then the differences are considered statistically significant. If this value is greater than .05, then any observed differences are not statistically significant and may be interpreted as inconclusive. It is also important to point out that only analyses (and results) that were of interest in this study and determined a priori are included in these tables.

In some of the following tables, superscripts (in the form of letters) are provided to identify significant differences between different grade levels and/or years. To interpret these results, the reader should compare the letters next to each year/grade level (e.g., 2000^a). If the letters are different between years/grades, this means that the difference is statistically significant ($p < .05$). If they are the same letter, then the difference is not significant. In the example below, 1998 and 1999 test scores are significantly different from both 2000 and 2001 scores. The 2000 and 2001 scores are also significantly different from each other. However, the 1998 and 1999 test scores are not significantly different from each other.

Example:

Cohort	Mean
1998^a	590.7
1999^a	591.0
2001^b	617.5
2002^c	643.0

Cohort Analyses Among Saxon Math Students

Tables A1 and A2 summarize the ANOVA results of the cohort analyses *among Saxon Math students* for the Stanford 9 and CAT 6 scale-score measures for cohorts of similar students in elementary grades (2–5) and middle grades (6–8). For example, for the Stanford 9 sample, third graders in 2000 are compared to fourth graders in 2001 and fifth graders in 2002. Similarly, sixth graders in 2000 are compared to seventh graders in 2001 and eighth graders in 2002. For the CAT 6 sample, it becomes a bit more complicated. There are only 2 years in which similar groups of students (cohorts) can be compared. This is because the CAT 6 was administered in Grades 2 through 8 during spring of 2003 and 2004 only. As such, five cohorts were created and compared: (a) second graders in 2003 versus third graders in 2004, (b) third graders in 2003 versus fourth graders in 2004, (c) fourth graders in 2003 versus fifth graders in 2004, (d) sixth graders in 2003 versus seventh graders in 2004, and (e) seventh graders in 2003 versus eighth graders in 2004.

Table A1. Cohort Analyses Among Saxon Students: Stanford 9

Cohort	*M*	*SD*	*N*	*F*
3 (2000)[a]	590.80	43.40	2,418	$F(2, 7336) = 965.8$, $p < .001$
4 (2001)[b]	617.56	41.55	2,498	
5 (2002)[c]	643.10	39.15	2,423	
Cohort	***M***	***SD***	***N***	***F***
6 (2000)[a]	647.59	39.49	3,407	$F(2, 11296) = 628.4$, $p < .001$
7 (2001)[b]	666.82	39.52	3,971	
8 (2002)[c]	680.12	38.81	3,921	

Note. Different letters between grades (years) in cohort group represent significant differences in pairwise comparisons.

Table A2. Cohort Analyses Among Saxon Students: CAT 6

Cohort	*M*	*SD*	*N*	*F*
2 (2003)	565.61	48.99	3,819	$F(1, 7738) = 1354.8$, $p < .001$
3 (2004)	604.93	44.96	3,921	
Cohort	***M***	***SD***	***N***	***F***
3 (2003)	604.21	48.22	3,879	$F(1, 7966) = 164.0$, $p < .001$
4 (2004)	618.87	53.63	4,089	
Cohort	***M***	***SD***	***N***	***F***
4 (2003)	619.27	53.99	3,913	$F(1, 7790) = 135.8$, $p < .001$
5 (2004)	633.30	52.30	3,879	
Cohort	***M***	***SD***	***N***	***F***
6 (2003)	651.41	54.14	4,027	$F(1, 8572) = 58.5$, $p < .001$
7 (2004)	660.30	53.28	4,547	
Cohort	***M***	***SD***	***N***	***F***
7 (2003)	657.60	54.43	4,280	$F(1, 8601) = 148.2$, $p < .001$
8 (2004)	672.22	56.89	4,323	

Subgroup Differences Among Saxon Math Students

Tables A3 to A7 summarize the results of the subgroup analyses *among Saxon Math students* for the Stanford 9 and CAT 6 scale-score measures for cohorts of similar students in elementary grades (2–5) and middle grades (6–8); see prior page for description of these cohorts.

The ANOVA results for the interaction of each subgroup classification and group within the cohort (e.g., third graders in 2000 vs. fourth graders in 2001 vs. fifth graders in 2002) is first presented. Note that the structure of the data obtained from CDE prohibits repeated measures analyses. Because researchers were interested in examining whether there was significant change *within* the subgroups, the corresponding ANOVA examining differences between students at different grade levels (and years) for *each* subgroup are presented. A significant pattern of math performance (e.g., an increasing trend) would suggest that, within that subgroup (e.g., females), there are differences in performance over time. Results of pairwise comparison between the groups within the cohorts (in terms of whether significant or not) are noted as well. It is important to note that, because this is an observational study and these analyses *only* include Saxon students, these results should be viewed as descriptive and exploratory.

Table A3. Subgroup Differences Among Saxon Students: Gender Status

Elementary Cohort Results: Stanford 9 Scale Score

Subgroup	Cohort Group	*M*	*SD*	*N*	*F*-interaction between cohort group and subgroup	*F*-cohort group within subgroup
Male	3 (2000)[a]	587.99	43.85	1,242	$F(2, 7322) = 3.25$, $p = .04$	$F(2, 3699) = 509.30$, $p < .001$
	4 (2001)[b]	617.65	42.12	1,259		
	5 (2002)[c]	642.18	40.01	1,201		
Female	3 (2000)[a]	593.76	42.74	1,176		$F(2, 3623) = 460.11$, $p < .001$
	4 (2001)[b]	617.46	40.97	1,239		
	5 (2002)[c]	644.23	38.26	1,211		

Middle-School Cohort Results: Stanford 9 Scale Score

Subgroup	Cohort Group	*M*	*SD*	*N*	*F*-interaction between cohort group and subgroup	*F*-cohort group within subgroup
Male	6 (2000)[a]	644.77	39.87	1,677	$F(2, 11274) = 4.37$, $p = .01$	$F(2, 5547) = 346.98$, $p < .001$
	7 (2001)[b]	666.12	40.56	1,974		
	8 (2002)[c]	680.04	39.85	1,899		
Female	6 (2000)[a]	650.25	38.98	1,723		$F(2, 5727) = 284.52$, $p < .001$
	7 (2001)[b]	667.51	38.46	1,997		
	8 (2002)[c]	680.29	37.83	2,010		

Elementary Cohorts' Results: CAT 6 Scale Score

Subgroup	Cohort Group	*M*	*SD*	*N*	*F*-interaction between cohort group and subgroup	*F*-cohort group within subgroup
Male	3 (2003)[a]	604.06	50.25	1,962	$F(1, 7962) = .47$, $p = .49$	$F(1, 4040) = 67.52$, $p < .001$
	4 (2004)[b]	617.93	56.64	2,080		
Female	3 (2003)[a]	604.40	46.05	1,915		$F(1, 3952) = 100.29$, $p < .001$
	4 (2004)[b]	619.85	50.32	2,009		
Male	4 (2003)[a]	620.01	56.39	1,944	$F(1, 7788) = .23$, $p = .64$	$F(1, 3873) = 57.86$, $p < .001$
	5 (2004)[b]	633.47	53.72	1,931		
Female	4 (2003)[a]	618.53	51.52	1,969		$F(1, 3915) = 79.65$, $p < .001$
	5 (2004)[b]	633.13	50.87	1,948		

Middle-School Cohorts' Results: CAT 6 Scale Score

Subgroup	Cohort Group	*M*	*SD*	*N*	*F*-interaction between cohort group and subgroup	*F*-cohort group within subgroup
Male	6 (2003)[a]	648.29	55.88	2,023	$F(1, 8570) = 7.24$, $p = .007$	$F(1, 4314) = 50.24$, $p < .001$
	7 (2004)[b]	660.28	55.11	2,293		
Female	6 (2003)[a]	654.57	52.14	2,004		$F(1, 4256) = 13.08$, $p < .001$
	7 (2004)[b]	660.32	51.37	2,254		
Male	7 (2003)[a]	655.96	56.09	2,109	$F(1, 8597) = .93$, $p = .34$	$F(1, 4248) = 58.77$, $p < .001$
	8 (2004)[b]	669.41	58.30	2,141		
Female	7 (2003)[a]	659.20	52.73	2,171		$F(1, 4349) = 92.50$, $p < .001$
	8 (2004)[b]	674.97	55.38	2,180		

Note. Different letters between grades (years) in cohort group represent significant differences in pairwise comparisons.

Table A4. Subgroup Differences Among Saxon Students: Ethnicity Status

Subgroup	Cohort Group	*M*	*SD*	*N*	***F*-interaction between cohort group and subgroup**	***F*-cohort group within subgroup**
Elementary Cohort Results: Stanford 9 Scale Score						
White	3 (2000)[a]	627.33	40.34	379	$F(4, 6918) = 1.87$, $p = .11$	$F(2, 1191) = 140.84$, $p < .001$
	4 (2001)[b]	652.23	38.58	412		
	5 (2002)[c]	673.94	37.59	403		
Hispanic	3 (2000)[a]	585.45	39.08	1,344		$F(2, 4127) = 691.75$, $p < .001$
	4 (2001)[b]	614.00	37.60	1,416		
	5 (2002)[c]	639.15	36.17	1,370		
African-American	3 (2000)[a]	574.88	39.61	554		$F(2, 1600) = 276.91$, $p < .001$
	4 (2001)[b]	600.15	37.09	537		
	5 (2002)[c]	628.27	33.87	512		
Middle-School Cohort Results: Stanford 9 Scale Score						
Subgroup	**Cohort Group**	***M***	***SD***	***N***	***F*-interaction between cohort group and subgroup**	***F*-cohort group within subgroup**
White	6 (2000)[a]	684.57	39.40	573	$F(4, 10445) = 2.89$, $p = .02$	$F(2, 2051) = 53.74$, $p < .001$
	7 (2001)[b]	695.81	40.85	761		
	8 (2002)[c]	707.78	40.00	720		
Hispanic	6 (2000)[a]	642.26	33.76	1,841		$F(2, 5837) = 397.33$, $p < .001$
	7 (2001)[b]	658.12	30.73	1,999		
	8 (2002)[c]	671.35	31.45	2,000		
African-American	6 (2000)[a]	632.75	33.50	814		$F(2, 2557) = 203.83$, $p < .001$
	7 (2001)[b]	648.40	30.47	876		
	8 (2002)[c]	663.85	30.84	870		
Elementary Cohort Results: CAT 6 Scale Score						
Subgroup	**Cohort Group**	***M***	***SD***	***N***	***F*-interaction between cohort group and subgroup**	***F*-cohort group within subgroup**
White	3 (2003)[a]	631.02	48.17	682	$F(2, 7602) = 3.82$, $p = .02$	$F(1, 1420) = 35.58$, $p < .001$
	4 (2004)[b]	645.90	49.69	740		
Hispanic	3 (2003)[a]	598.80	45.83	2,293		$F(1, 4743) = 128.54$, $p < .001$
	4 (2004)[b]	614.75	50.75	2,452		
African-American	3 (2003)[a]	591.00	45.21	725		$F(1, 1439) = 8.55$, $p = .004$
	4 (2004)[b]	598.84	56.05	716		
White	4 (2003)[a]	648.09	47.37	715	$F(2, 7451) = .06$, $p = .94$	$F(1, 1442) = 27.09$, $p < .001$
	5 (2004)[b]	661.38	49.59	729		
Hispanic	4 (2003)[a]	615.82	50.83	2,267		$F(1, 4519) = 94.24$, $p < .001$
	5 (2004)[b]	630.00	47.30	2,254		
African-American	4 (2003)[a]	598.13	56.73	758		$F(1, 1490) = 12.13$, $p < .001$
	5 (2004)[b]	611.51	55.68	734		

(*continues*)

Middle-School Cohort Results: CAT 6 Scale Score						
Subgroup	**Cohort Group**	***M***	***SD***	***N***	***F*-interaction between cohort group and subgroup**	***F*-cohort group within subgroup**
White	6 (2003)[a]	685.90	46.30	630	$F(2, 8034) = .02$, $p = .99$	$F(1, 1322) = 5.81$, $p = .02$
	7 (2004)[b]	691.86	43.60	694		
Hispanic	6 (2003)[a]	646.13	51.52	2,433		$F(1, 5007) = 16.08$, $p < .001$
	7 (2004)[b]	651.96	51.35	2,576		
African-American	6 (2003)[a]	638.71	55.68	838		$F(1, 1705) = 6.13$, $p = .01$
	7 (2004)[b]	645.02	49.62	869		
White	7 (2003)[a]	695.77	50.67	666	$F(2, 7850) = 2.59$, $p = .08$	$F(1, 1353) = 13.59$, $p < .001$
	8 (2004)[b]	705.70	48.48	689		
Hispanic	7 (2003)[a]	647.32	50.08	2,321		$F(1, 4688) = 110.14$, $p < .001$
	8 (2004)[b]	663.34	54.31	2,369		
African-American	7 (2003)[a]	640.69	48.06	926		$F(1, 1809) = 23.54$, $p < .001$
	8 (2004)[b]	652.02	51.27	885		

Note. Different letters between grades (years) in cohort group represent significant differences in pairwise comparisons.

Table A5. Subgroup Differences Among Saxon Students: English Language Learner Status

Subgroup	Cohort Group	*M*	*SD*	*N*	*F*-interaction between cohort group and subgroup	*F*-cohort group within subgroup
Elementary Cohort Results: Stanford 9 Scale Score						
Non-ELL	3 (2000)[a]	597.37	44.86	3,556	$F(2, 5267) = 63.48$, $p < .001$	$F(2, 4203) = 496.68$, $p < .001$
	4 (2001)[b]	623.99	44.78	1,310		
	5 (2002)[c]	648.50	40.94	1,340		
ELL	3 (2000)[a]	576.10	36.86	729		$F(2, 1064) = 530.13$, $p < .001$
	4 (2001)[b]	539.02	33.59	149		
	5 (2002)[c]	660.60	31.45	189		
Middle-School Cohort Results: Stanford 9 Scale Score						
Non-ELL	6 (2000)[a]	652.73	41.48	2,369	$F(2, 8190) = 143.26$, $p < .001$	$F(2, 6791) = 347.31$, $p < .001$
	7 (2001)[b]	672.58	42.88	2,202		
	8 (2002)[c]	684.90	41.09	2,223		
ELL	6 (2000)[a]	629.96	28.90	782		$F(2, 1399) = 726.71$, $p < .001$
	7 (2001)[b]	689.78	39.05	282		
	8 (2002)[c]	703.55	36.78	338		
Elementary Cohort Results: CAT 6 Scale Score						
Non-ELL	3 (2003)[a]	611.62	48.60	2,218	$F(1, 7642) = .30$, $p = .58$	$F(1, 4552) = 74.36$, $p < .001$
	4 (2004)[b]	624.91	55.06	2,336		
ELL	3 (2003)[a]	592.50	45.40	1,568		$F(1, 3090) = 49.92$, $p < .001$
	4 (2004)[b]	604.51	49.14	1,524		
Non-ELL	4 (2003)[a]	624.91	54.77	2,303	$F(1, 7108) = .004$, $p = .95$	$F(1, 4559) = 66.22$, $p < .001$
	5 (2004)[b]	638.18	55.38	2,258		
ELL	4 (2003)[a]	603.76	50.68	1,306		$F(1, 2549) = 47.23$, $p < .001$
	5 (2004)[b]	616.86	45.30	1,245		
Middle-School Cohort Results: CAT 6 Scale Score						
Non-ELL	6 (2003)[a]	658.89	55.15	2,106	$F(1, 7730) = .19$, $p = .66$	$F(1, 4566) = 16.61$, $p < .001$
	7 (2004)[b]	665.48	53.84	2,462		
ELL	6 (2003)[a]	636.69	51.18	1,586		$F(1, 3164) = 9.53$, $p < .001$
	7 (2004)[b]	642.21	49.35	1,580		
Non-ELL	7 (2003)[a]	664.34	54.48	2,453	$F(1, 7825) = 3.16$, $p = .08$	$F(1, 5031) = 49.64$, $p < .001$
	8 (2004)[b]	675.40	56.72	2,580		
ELL	7 (2003)[a]	637.85	48.95	1,489		$F(1, 2794) = 64.86$, $p < .001$
	8 (2004)[b]	653.44	53.44	1,307		

Note. Different letters between grades (years) in cohort group represent significant differences in pairwise comparisons.

Table A6. Subgroup Differences Among Saxon Students: Economically Disadvantaged Status

Subgroup	Cohort Group	*M*	*SD*	*N*	*F*-interaction between cohort group and subgroup	*F*-cohort group within subgroup
Elementary Cohort Results: Stanford 9 Scale Score						
Non-Econ. Disadvantaged	3 (2000)[a]	625.67	42.38	447	$F(2, 7333) = 3.81$, $p = .02$	$F(2, 1383) = 159.62$, $p < .001$
	4 (2001)[b]	653.30	38.92	469		
	5 (2002)[c]	672.80	38.94	470		
Econ. Disadvantaged	3 (2000)[a]	582.89	39.57	1,971		$F(2, 5950) = 974.19$, $p < .001$
	4 (2001)[b]	609.29	37.58	2,029		
	5 (2002)[c]	635.95	35.69	1,953		
Middle-School Cohort Results: Stanford 9 Scale Score						
Subgroup	**Cohort Group**	***M***	***SD***	***N***	***F*-interaction between cohort group and subgroup**	***F*-cohort group within subgroup**
Non-Econ. Disadvantaged	6 (2000)[a]	684.57	40.34	615	$F(2, 11293) = 3.10$, $p = .05$	$F(2, 2864) = 73.73$, $p < .001$
	7 (2001)[b]	697.21	41.32	1,143		
	8 (2002)[c]	708.99	39.82	1,109		
Econ. Disadvantaged	6 (2000)[a]	639.44	34.31	2,792		$F(2, 8429) = 568.56$, $p < .001$
	7 (2001)[b]	654.54	31.29	2,828		
	8 (2002)[c]	668.73	31.88	2,812		
Elementary Cohort Results: CAT 6 Scale Score						
Subgroup	**Cohort Group**	***M***	***SD***	***N***	***F*-interaction between cohort group and subgroup**	***F*-cohort group within subgroup**
Non-Econ. Disadvantaged	3 (2003)[a]	626.76	47.31	932	$F(1, 7964) = 4.17$, $p = .04$	$F(1, 1864) = 75.90$, $p < .001$
	4 (2004)[b]	645.87	47.46	934		
Econ. Disadvantaged	3 (2003)[a]	597.08	46.28	2,947		$F(1, 6100) = 117.31$, $p < .001$
	4 (2004)[b]	610.88	52.74	3,155		
Non-Econ. Disadvantaged	4 (2003)[a]	645.60	48.25	970	$F(1, 7788) = 1.02$, $p = .31$	$F(1, 1891) = 52.03$, $p < .001$
	5 (2004)[b]	662.05	50.96	923		
Econ. Disadvantaged	4 (2003)[a]	610.59	52.97	2,943		$F(1, 5897) = 106.13$, $p < .001$
	5 (2004)[b]	624.32	49.41	2,956		
Middle-School Cohort Results: CAT 6 Scale Score						
Subgroup	**Cohort Group**	***M***	***SD***	***N***	***F*-interaction between cohort group and subgroup**	***F*-cohort group within subgroup**
Non-Econ. Disadvantaged	6 (2003)[a]	684.91	45.86	698	$F(1, 8570) = .05$, $p = .82$	$F(1, 1759) = 8.78$, $p = .003$
	7 (2004)[b]	691.83	49.25	1,063		
Econ. Disadvantaged	6 (2003)[a]	644.39	53.11	3,329		$F(1, 6811) = 24.99$, $p < .001$
	7 (2004)[b]	650.68	50.69	3,484		
Non-Econ. Disadvantaged	7 (2003)[a]	695.17	50.94	1,060	$F(1, 8599) = 1.41$, $p = .24$	$F(1, 2128) = 31.65$, $p < .001$
	8 (2004)[b]	707.49	50.11	1,070		
Econ. Disadvantaged	7 (2003)[a]	645.24	49.67	3,220		$F(1, 6471) = 141.73$, $p < .001$
	8 (2004)[b]	660.62	54.15	3,253		

Note. Different letters between grades (years) in cohort group represent significant differences in pairwise comparisons.

Table A7. Subgroup Differences Among Saxon Students: Disability Status

Subgroup	Cohort Group	*M*	*SD*	*N*	*F*-interaction between cohort group and subgroup	*F*-cohort group within subgroup
Elementary Cohort Results: Stanford 9 Scale Score						
No Disability	3 (2000)[a]	592.70	43.00	2,206	$F(2, 7333) = 4.85$, $p = .008$	$F(2, 6891) = 920.40$, $p < .001$
	4 (2001)[b]	618.61	41.41	2,393		
	5 (2002)[c]	645.02	38.21	2,295		
Disability	3 (2000)[a]	570.93	42.66	212		$F(2, 442) = 35.97$, $p < .001$
	4 (2001)[b]	593.55	37.55	105		
	5 (2002)[c]	608.69	40.00	128		
Middle-School Cohort Results: Stanford 9 Scale Score						
Subgroup	**Cohort Group**	***M***	***SD***	***N***	***F*-interaction between cohort group and subgroup**	***F*-cohort group within subgroup**
No Disability	6 (2000)[a]	649.39	39.05	3,245	$F(2, 11293) = 2.12$, $p = .12$	$F(2, 10742) = 584.24$, $p < .001$
	7 (2001)[b]	668.67	39.35	3,737		
	8 (2002)[c]	681.29	38.85	3,763		
Disability	6 (2000)[a]	611.59	30.05	162		$F(2, 551) = 85.36$, $p < .001$
	7 (2001)[b]	637.31	29.11	234		
	8 (2002)[c]	652.16	24.77	158		
Elementary Cohort Results: CAT 6 Scale Score						
Subgroup	**Cohort Group**	***M***	***SD***	***N***	***F*-interaction between cohort group and subgroup**	***F*-cohort group within subgroup**
No Disability	3 (2003)[a]	605.59	47.42	3,622	$F(1, 7911) = 4.04$, $p = .05$	$F(1, 7481) = 168.49$, $p < .001$
	4 (2004)[b]	620.60	52.28	3,861		
Disability	3 (2003)[a]	583.73	57.02	223		$F(1, 430) = .69$, $p = .41$
	4 (2004)[b]	588.66	65.87	209		
No Disability	4 (2003)[a]	621.93	51.67	3,612	$F(1, 7757) = 1.25$, $p = .26$	$F(1, 7231) = 136.97$, $p < .001$
	5 (2004)[b]	636.02	50.72	3,621		
Disability	4 (2003)[a]	584.93	68.66	279		$F(1, 526) = 2.47$, $p = .12$
	5 (2004)[b]	593.75	59.26	249		
Middle-School Cohort Results: Stanford 9 Scale Score						
Subgroup	**Cohort Group**	***M***	***SD***	***N***	***F*-interaction between cohort group and subgroup**	***F*-cohort group within subgroup**
No Disability	6 (2003)[a]	652.40	53.61	3,928	$F(1, 8559) = .39$, $p = .56$	$F(1, 8289) = 64.73$, $p < .001$
	7 (2004)[b]	661.81	52.83	4,363		
Disability	6 (2003)[a]	610.47	62.27	91		$F(1, 270) = 3.60$, $p = .06$
	7 (2004)[b]	623.92	51.26	181		
No Disability	7 (2003)[a]	659.10	54.17	4,104	$F(1, 8579) = 6.35$, $p = .01$	$F(1, 8209) = 162.26$, $p < .001$
	8 (2004)[b]	674.52	55.51	4,107		
Disability	7 (2003)[a]	624.12	46.96	172		$F(1, 370) = .01$, $p = .91$
	8 (2004)[b]	624.82	64.79	200		

Note. Different letters between grades (years) in cohort group represent significant differences in pairwise comparisons.

Saxon Versus Non–Saxon Comparisons

Cross-Sectional Analyses: Differences by Year

Tables A8 to A10 in the following pages summarize the results for differences between *Saxon Math* and non–*Saxon Math* users for the Stanford 9, CST, and CAT 6 scale-score measures and for students in elementary grades (2–5) and middle grades (6–8) separately. The ANOVA results for the interaction of year(time) with group (Saxon vs. non-Saxon) are first presented. In addition, ANOVA conducted for differences between group at *each year* for which data are available is also presented. It is important to note that, given that this is an observational study and schools were not randomized into the treatment groups, there may be preexisting differences. Indeed, when available, differences between groups *before* Saxon schools were using *Saxon Math* are analyzed (i.e., for the Stanford elementary and middle-school sample, this would consist of the 1998 and 1999 data; for the CST elementary sample only, this would consist of the 2002 data).

Effect sizes are also presented. Eta^2 [i.e., proportion of variance accounted for (PV)] obtained from SPSS 14.0 was converted to Cohen's *d*. This was done to ease interpretation. The following formula was used for this conversion (Lipsey, 1990):

$$ES = \sqrt{\frac{4\ (PV)}{1\text{-}(PV)}}$$

Table A8. Saxon vs. Non-Saxon by Time: Stanford 9

Elementary Results: Stanford 9 Scale Score							
Year	**Group**	**Adjusted *M*[a]**	**Unadjusted *SD***	***N***	***F***	***F*-group within year**	**Effect size (*d*)**
1998 (pre)	Non-Saxon	583.63	46.77	20,732	*F* interaction (4, 157720) = 36.46, $p < .001$ *F* group (1, 157720) = 11.40, $p = .001$ *F* time (4, 157720) = 1418.73, $p < .001$	*F* (1, 28971) = 91.98, $p < .001$	.11
	Saxon	590.46	48.56	8,249			
1999 (pre)	Non-Saxon	596.13	45.41	21,521		*F* (1, 29924) = 1.05, $p = .31$	na
	Saxon	595.57	47.34	8,413			
2000	Non-Saxon	605.67	46.28	22,750		*F* (1, 31759) = 5.84, $p = .02$	.03
	Saxon	603.79	48.17	9,019			
2001	Non-Saxon	610.08	46.60	23,964		*F* (1, 33399) = 1.36, $p = .24$	na
	Saxon	610.40	47.96	9,445			
2002	Non-Saxon	615.07	46.24	24,584		*F* (1, 33635) = .48, $p = .49$	na
	Saxon	614.60	47.92	9,061			
Middle-School Results: Stanford 9 Scale Score							
Year	**Group**	**Adjusted *M*[b]**	**Unadjusted *SD***	***N***	***F***	***F*-group within year**	**Effect size (*d*)**
1998 (pre)	Non-Saxon	651.41	38.16	15,158	*F* interaction (4, 145308) = 17.74, $p < .001$ *F* group (1, 145308) = 457.13, $p < .001$ *F* time (4, 145308) = 1009.27, $p < .001$	*F* (1, 25775) = 59.70, $p < .001$	.09
	Saxon	656.19	39.74	10,627			
1999 (pre)	Non-Saxon	656.92	37.21	17,705		*F* (1, 27898) = 54.18, $p < .31$	.09
	Saxon	658.03	39.72	10,203			
2000	Non-Saxon	661.63	37.21	18,389		*F* (1, 29078) = 113.77, $p < .001$	.13
	Saxon	665.47	40.89	10,699			
2001	Non-Saxon	665.47	38.46	19,461		*F* (1, 30732) = 161.76, $p < .001$	.14
	Saxon	670.58	40.94	11,281			
2002	Non-Saxon	667.16	38.47	20,517		*F* (1, 31794) = 180.86, $p < .001$	.16
	Saxon	672.57	41.86	11,287			

Note. na = not applicable.
[a]Covariates appearing in the model are evaluated at the following values: Gender = .49, Disability Status = .06, White = .19, Hispanic = .62, African American = .13, Asian = .04, Migrant Status =.03, Economically Disadvantaged Status =.64.
[b]Covariates appearing in the model are evaluated at the following values: Gender = .50, Disability Status = .07, White = .21, Hispanic = .61, African American = .11, Asian = .05, Migrant Status =.06, Economically Disadvantaged Status =.61.

Table A9. Saxon vs. Non-Saxon by Time: CAT 6

Year	Group	Adjusted M^a	Unadjusted *SD*	*N*	*F*-interaction between group and year	*F*-group within year	Effect size (*d*)
Elementary Results: CAT 6 Scale Score							
2003	Non-Saxon	604.57	53.95	17,895	F interaction (3, 82913) = .79, p = .50 F group (1, 82913) = 18.74, p < .001, d = .03 F time (3, 82913) = 6.27, p < .001	F (1, 33303) = 8.85, p = .003	.03
	Saxon	606.44	56.51	15,418			
2004	Non-Saxon	605.16	55.67	17,873		F (1, 33377) = 6.66, p = .01	.03
	Saxon	606.72	55.97	15,514			
2005	Non-Saxon	606.35	44.81	4,531		F (1, 8381) = 2.43, p = .12	na
	Saxon	607.50	46.33	3,860			
2006	Non-Saxon	606.41	45.71	4,337		F (1, 7829) = 11.38, p = .001	.06
	Saxon	609.77	46.48	3,501			
Middle-School Results: CAT 6 Scale Score							
Year	**Group**	**Adjusted M^b**	**Unadjusted *SD***	***N***	***F*-interaction between group and year**	***F*-group within year**	**Effect size (*d*)**
2003	Non-Saxon	656.08	55.35	21,357	F interaction (3, 90419) = 17.06, p < .001 F group (1, 90419) = 4.55, p = .03, d = .01 F time (3, 90419) = 13.27, p < .001	F (1, 33442) = 47.50, p < .001	.09
	Saxon	660.57	54.90	12,095			
2004	Non-Saxon	660.04	54.07	21,304		F (1, 34169) = .47, p = .49	na
	Saxon	660.42	56.38	12,875			
2005	Non-Saxon	657.15	49.63	6,919		F (1, 11344) = 2.62, p = .11	na
	Saxon	658.15	53.01	4,435			
2006	Non-Saxon	659.14	51.51	7,215		F (1, 11441) = 4.99, p = .03	.04
	Saxon	656.66	53.42	4,235			

Note. na = not applicable.
[a]Covariates appearing in the model are evaluated at the following values: Gender = .49, Disability Status = .07, White = .16, Hispanic = .70, African American = .11, Asian = .04, Migrant Status =.03, Economically Disadvantaged Status =.79.
[b]Covariates appearing in the model are evaluated at the following values: Gender = .49, Disability Status = .07, White = .17, Hispanic = .68, African American = .10, Asian = .04, Migrant Status =.06, Economically Disadvantaged Status =.75.

Note that the CAT 6 was administered in Grades 2 through 8 in 2003 and 2004. In 2005 and 2006, it was administered to their and seventh graders only. Hence, there is a decrease in sample size.

Examination of third- and seventh-grade data only across these years showed a similar pattern of results.

Table A10. Saxon vs. Non-Saxon by Time: CST

Year	Group	Adjusted M[a]	Unadjusted *SD*	*N*	*F*	*F*-group within year	Effect size (*d*)
Elementary Results: CST Scale Score							
2002 (pre)	Non-Saxon	315.23	63.42	17,938	*F* interaction (4, 118499) = 21.06, $p < .001$ *F* group (1, 118499) = 348.82, $p < .001$ *F* time (4, 118499) = 699.59, $p < .001$	*F* (1, 24536) = 109.79, $p < .001$	.13
	Saxon	325.58	63.20	6,608			
2003	Non-Saxon	328.77	68.06	17,939		*F* (1, 24229) = 7.80, $p = .005$	.01
	Saxon	332.00	66.56	6,300			
2004	Non-Saxon	331.99	68.37	17,918		*F* (1, 24149) = 28.13, $p < .001$	.06
	Saxon	336.99	68.09	6,241			
2005	Non-Saxon	342.79	75.24	17,429		*F* (1, 23334) = 89.66, $p < .001$	.13
	Saxon	352.36	74.62	5,915			
2006	Non-Saxon	345.31	78.03	16,780		*F* (1, 22220) = 171.94, $p < .001$	.18
	Saxon	360.04	80.29	5,449			
Middle-School Results: CST Scale Score							
Year	**Group**	**Adjusted M[b]**	**Unadjusted *SD***	***N***	***F***	***F*-group within year**	**Effect size (*d*)**
2002	Non-Saxon	306.37	53.36	20,328	*F* interaction (4, 165447) = 19.13, $p < .001$ *F* group (1, 165447) = .01, $p = .93$ *F* time (4, 165447) = 279.52, $p < .001$	*F* (1, 31491) = 1.30, $p = .25$	na
	Saxon	307.31	58.18	11,173			
2003	Non-Saxon	310.22	54.10	21,269		*F* (1, 33209) = 12.51, $p < .001$	.09
	Saxon	312.62	55.85	11,950			
2004	Non-Saxon	312.73	53.94	20,936		*F* (1, 33680) = .20, $p = .66$	na
	Saxon	312.97	59.11	12,754			
2005	Non-Saxon	316.89	59.17	21,114		*F* (1, 33705) = 4.25, $p = .04$	.02
	Saxon	317.73	62.23	12,601			
2006	Non-Saxon	321.51	61.77	20,952		*F* (1, 33331) = 30.05, $p < .001$	.06
	Saxon	317.21	64.35	12,388			

Note. na = not applicable.
[a]Covariates appearing in the model are evaluated at the following values: Gender = .49, Disability Status = .09, White = .17, Hispanic = .70, African American = .08, Asian = .05, Migrant Status =.04, Economically Disadvantaged Status =.77.
[b]Covariates appearing in the model are evaluated at the following values: Gender = .49, Disability Status = .07, White = .17, Hispanic = .67, African American = .10, Asian = .04, Migrant Status =.05, Economically Disadvantaged Status =.75.

For this dataset, the elementary sample only includes students in Saxon schools that began using Saxon in 2003. This allows for comparisons between elementary Saxon and non–Saxon students at baseline (before a school started using Saxon). The middle-school sample includes *all* students in Saxon middle schools, all of which began using the program in the 1999–2000 school year. Thus, for the middle-school sample, there is no pre–Saxon data available.

Cohort Analyses

Tables A11 to A12 in the following pages summarize the results of cohort analyses between *Saxon Math* and non–*Saxon Math* students for the Stanford 9 and CAT 6 scale-score measures. For the Stanford 9 sample, second graders in 1999 are compared to third graders in 2000, fourth graders in 2001, and fifth graders in 2002. Similarly, sixth graders in 1999 are compared to seventh graders in 2000 and eighth graders in 2001. Note that these analyses include pre-Saxon data. In spring 1999 (i.e., second and sixth grade, respectively), students were *not* using *Saxon Math*. Exposure to *Saxon Math* occurred in fall of the 1999–2000 school year, and thus, the first year of post–Saxon data is Spring 2000 (i.e., third and seventh grade, respectively).

For the CAT 6 sample, it becomes a bit more complicated. There are only 2 years in which similar groups of students (cohorts) can be compared. This is because the CAT 6 was administered in Grades 2 through 8 during spring of 2003 and 2004 only. As such, five cohorts were created and compared: (a) second graders in 2003 versus third graders in 2004, (b) third graders in 2003 versus fourth graders in 2004, (c) fourth graders in 2003 versus fifth graders in 2004, (d) sixth graders in 2003 versus seventh graders in 2004, and (e) seventh graders in 2003 versus eighth graders in 2004. For this dataset, no pre–Saxon data is available because, at this point in time, all schools were actively using *Saxon Math*.

The ANOVA results for the interaction of group and grade level is presented. In addition, results of pairwise comparison between the groups within the grade level are noted. It is important to note that given that this is an observational study and that students were not randomized to conditions, these results should be viewed as preliminary. Effect sizes are also presented using the formula previously noted.

Table A11. Saxon vs. Non-Saxon by Grade: Stanford 9

Elementary Results: Stanford 9 Scale Score							
Year	**Group**	**Adjusted *M*[a]**	**Unadjusted *SD***	***N***	***F* test**	***F*-group within grade**	**Effect size (*d*)**
2 (pre-1999)	Non-Saxon	559.05	40.85	3,999	*F* interaction (3, 32724) = 29.52, $p < .001$ *F* group (1, 32724) = .74, $p = .39$ *F* cohort (3, 32724) = 6396.65, $p < .001$	*F* (1, 7659) = 67.29, $p < .001$	.19
	Saxon	565.66	41.68	3,670			
3 (2000)	Non-Saxon	599.64	41.64	5,955		*F* (1, 8334) = 27.34, $p < .001$	.11
	Saxon	594.65	43.45	2,389			
4 (2001)	Non-Saxon	619.92	39.46	5,926		*F* (1, 8406) = .70, $p = .40$	na
	Saxon	620.11	41.52	2,490			
5 (2002)	Non-Saxon	646.20	38.48	5,927		*F* (1, 8301) = .17, $p = .68$	na
	Saxon	645.96	39.19	2,384			
Middle-School Results: Stanford 9 Scale Score							
Year	**Group**	**Adjusted *M*[b]**	**Unadjusted *SD***	***N***	***F*-interaction between group and year**	***F*-group within year**	**Effect size (*d*)**
6 (pre-1999)	Non-Saxon	647.51	38.67	6,478	*F* interaction (2, 29033) = 46.47, $p < .001$ *F* group (1, 29033) = 40.62, $p < .001$ *F* time (2, 29033) = 2061.48, $p < .001$	*F* (1, 9423) = .01, $p < .19$	na
	Saxon	645.38	37.35	2,955			
7 (2000)	Non-Saxon	663.14	35.60	6,053		*F* (1, 9706) = 20.63, $p < .001$	.09
	Saxon	665.72	39.38	3,663			
8 (2001)	Non-Saxon	674.64	34.66	6,088		*F* (1, 9888) = 119.75, $p < .001$	.22
	Saxon	681.96	38.38	3,810			

Note. na = not applicable.
[a]Covariates appearing in the model are evaluated at the following values: Gender = .50, Disability Status = .06, White = .19, Hispanic = .62, African American = .13, Asian = .04, Migrant Status =.03, Economically Disadvantaged Status =.72.
[b]Covariates appearing in the model are evaluated at the following values: Gender = .50, Disability Status = .06, White = .21, Hispanic = .61, African American = .11, Asian = .05, Migrant Status =.07, Economically Disadvantaged Status =.62.

Table A12. Saxon vs. Non-Saxon by Grade: CAT 6

Cohort	Year	Group	Adjusted M^a	Unadjusted SD	N	F test	F-group within grade	Effect size (d)
Elementary Results: CAT 6 Scale Score								
1	2 (2003)	Non-Saxon	562.33	48.13	4,553	F interaction (1, 16677) = 5.19, $p = .02$ F group (1, 16677) = 3547.6, $p < .001$ F grade (1, 16677) = 8.50, $p = .004$	F (1, 8301) = 12.98, $p < .001$	.09
		Saxon	565.96	49.04	3,758			
	3 (2004)	Non-Saxon	604.56	44.80	4,542		F (1, 8368) = .14, $p = .71$	na
		Saxon	605.08	44.90	3,836			
2	3 (2003)	Non-Saxon	604.34	43.57	4,641	F interaction (1, 16809) = .12, $p = .73$ F group (1, 16809) = .13, $p = .72$ F grade (1, 16809) = 426.04, $p < .001$	F (1, 8438) = .12, $p = .73$	na
		Saxon	604.32	48.38	3,807			
	4 (2004)	Non-Saxon	619.27	51.29	4,345		F (1, 8363) = .07, $p = .79$	na
		Saxon	618.76	53.66	4,028			
3	4 (2003)	Non-Saxon	620.28	46.80	4,462	F interaction (1, 16522) = .16, $p = .69$ F group (1, 16522) = 2.15, $p = .14$ F grade (1, 16522) = 376.99, $p < .001$	F (1, 8317) = .78, $p = .38$	na
		Saxon	619.46	53.94	3,865			
	5 (2004)	Non-Saxon	634.71	50.25	4,364		F (1, 8197) = 1.49, $p = .22$	na
		Saxon	633.31	52.40	3,843			
Middle-School Results: CAT 6 Scale Score								
Cohort	**Year**	**Group**	**Adjusted M^a**	**Unadjusted SD**	**N**	**F test**	**F-group within grade**	**Effect size (d)**
4	6 (2003)	Non-Saxon	647.65	55.51	7,378	F interaction (1, 22875) = 1.18, $p = .28$ F group (1, 22875) = 29.78, $p < .001$ F grade (1, 22875) = 110.92, $p < .001$	F (1, 11364) = 17.35, $p < .001$	.06
		Saxon	652.21	54.17	3,996			
	7 (2004)	Non-Saxon	655.45	52.52	7,022		F (1, 11503) = 1.68, $p = .001$	.09
		Saxon	658.55	53.19	4,491			
5	7 (2003)	Non-Saxon	654.47	53.42	7,244	F interaction (1, 22987) = 5.46, $p = .02$ F group (1, 22987) = 3.55, $p = .06$ F grade (1, 22987) = 611.85, $p < .001$	F (1, 11476) = 7.38, $p = .007$	.06
		Saxon	657.29	54.39	4,242			
	8 (2004)	Non-Saxon	672.21	52.95	7,251		F (1, 11503) = .00, $p = .99$	na
		Saxon	671.97	56.83	4,262			

Note. na = not applicable.

[a]Covariates appearing in the model for cohort 1: Gender = .49, Disability Status = .06, White = .16, Hispanic = .68, African American = .11, Asian = .04, Migrant Status =.04, Economically Disadvantaged Status =.79.

Covariates appearing in the model for cohort 2: Gender = .49, Disability Status = .07, White = .17, Hispanic = .68, African American = .11, Asian = .04, Migrant Status =.03, Economically Disadvantaged Status =.79.

Covariates appearing in the model for cohort 3: Gender = .49, Disability Status = .08, White = .17, Hispanic = .67, African American = .11, Asian = .04, Migrant Status =.04, Economically Disadvantaged Status =.78.

Covariates appearing in the model for cohort 4: Gender = .50, Disability Status = .07, White = .16, Hispanic = .69, African American = .10, Asian = .04, Migrant Status =.07, Economically Disadvantaged Status =.77.

Covariates appearing in the model for cohort 5: Gender = .49, Disability Status = .07, White = .18, Hispanic = .65, African American = .11, Asian = .04, Migrant Status =.06, Economically Disadvantaged Status =.72.

School-Level Analyses

Tables A13 to A15 in the following pages summarize the results of school-level analyses between *Saxon Math* and non–*Saxon Math* students for the Stanford 9, CAT 6, and CST measures. The advantage of this data is that researchers can control for preexisting differences on the CST and CAT 6 because schools can be readily identified and data across years can be matched to each school. Note that for the Stanford 9 sample, school level data is only available from Spring 2001 to Spring 2002, and all Saxon schools in the Stanford 9 sample had been using *Saxon Math* for 2 years in 2001. Thus, controlling for differences in 2001 may eliminate potential Saxon effects. Therefore, analyses were conducted both controlling and not controlling for 2001 math performance. The outcome measure is the percentage of students (elementary and middle school) who were above average relative to the norm sample of the Stanford 9.

The CAT 6 and CST school-level analyses included only *elementary* Saxon and non–Saxon schools. This is because researchers wanted to control for preexisting differences prior to the use of *Saxon Math* so as not to control for any potential Saxon effects. For Saxon schools, this meant selecting Saxon schools that began using *Saxon Math* in 2003 (which happened to be all elementary schools), and controlling for 2002 pre–*Saxon Math* performance and comparing these to non-Saxon elementary schools. For the CAT 6 sample, the outcome measure is the percentage of elementary students who were above average relative to the norm sample of the CAT 6. For the CST, the outcome measure is the percentage of elementary students who were proficient or advanced relative to California math standards.

The repeated measures ANOVA results for the interaction of group and time is presented, along with main effects tests for time and group. In addition, results of pairwise comparison between the groups within each year are noted. Effect sizes are also presented, using the formula previously noted.

Table A13. Saxon vs. Non-Saxon Elementary and Middle Schools by Year: Stanford 9

Elementary Results: Stanford 9 Percentage of Students Above Average							
Year	**Group**	**Adjusted *M*[a]**	**Unadjusted *SD***	***N***	**Repeated Measures *F***	***F*-group within year**	**Effect size (*d*)**
2001	Non-Saxon	43.26	17.35	63	*F* interaction (1, 95) = 3.11, *p* = .08 *F* group (1, 95) = 4.09, *p* = .05, *d* = .41 *F* time (1, 95) = .04, *p* = .83	*F* (1, 95) = 6.75, *p* = .01	.55
	Saxon	50.71	22.82	42			
2002	Non-Saxon	47.21	16.68	63		*F* (1, 96) = 2.30, *p* = .13	.29
	Saxon	51.83	23.64	42			
Analyses Below Control for 2001 Stanford 9 Math (Percentage Above Average)							
2002[b]	Non-Saxon	50.20	16.68	63	na	*F* (1, 94) = 2.90, *p* = .09	.35
	Saxon	47.35	23.64	42			

[a]Covariates appearing in the model are evaluated at the following values: Total = 707.23, per_aa = 8.31, per_as = 2.87, per_hi = 66.06, per_wh = 18.23, per_sd = 65.26, per_el = 38.74, per_di = 8.77.
[b]Covariates appearing in the model are evaluated at the following values: Percentage of students above average in Stanford9-2001 = 46.24, total = 707.23, per_aa = 8.31, per_as = 2.87, per_hi = 66.06, per_wh = 18.23, per_sd = 65.26, per_el = 38.74, per_di = 8.77.

Table A14. Saxon vs. Non-Saxon Elementary Schools by Year: CAT 6

Elementary Results: CAT 6 Percentage of Students Above Average							
Year	**Group**	**Adjusted *M*[a]**	**Unadjusted *SD***	***N***	**Repeated Measures *F***	***F*-group within year**	**Effect size (*d*)**
2003	Non-Saxon	45.39	13.04	45	*F* interaction (3, 52) = 2.02, *p* = .12 *F* group (1, 54) = .003, *p* = .96 *F* time (3, 52) = .72, *p* = .55	*F* (1, 54) = 1.95, *p* = .17	.41
	Saxon	42.12	8.47	20			
2004	Non-Saxon	47.07	12.36	45		*F* (1, 54) = 2.33, *p* = .13	.41
	Saxon	43.44	8.35	20			
2005	Non-Saxon	47.45	13.32	45		*F* (1, 54) = .56, *p* = .46	.20
	Saxon	49.88	9.26	20			
2006	Non-Saxon	47.89	12.42	45		*F* (1, 54) = 1.56, *p* = .22	.35
	Saxon	52.80	11.38	20			

[a]Covariates appearing in the model are evaluated at the following values: Total = 597.09, per_aa = 8.39, per_as = 2.84, per_hi = 66.21, per_wh = 17.60, per_sd = 69.46, per_el = 37.00, per_di = 11.48. Percentage of students above average in Stanford9-2002 = 53.65.

Table A15. Saxon vs. Non-Saxon Elementary Schools by Year: CST

Elementary Results: CST Percentage of Students Meeting Math Standards							
Year	**Group**	**Adjusted *M*[a]**	**Unadjusted *SD***	***N***	**Repeated Measures *F***	***F*-group within year**	**Effect size (*d*)**
2003	Non-Saxon	36.75	14.05	45	*F* interaction (3, 52) = 2.58, *p* = .06 *F* group (1, 54) = .04, *p* = .85 *F* time (3, 52) = 2.60, *p* = .06	*F* (1, 54) = 1.78, *p* = .19	.35
	Saxon	33.61	8.95	20			
2004	Non-Saxon	37.89	12.98	45		*F* (1, 54) = 1.57, *p* = .22	.35
	Saxon	34.79	9.38	20			
2005	Non-Saxon	42.93	12.55	45		*F* (1, 54) = 1.07, *p* = .31	.29
	Saxon	46.01	8.74	20			
2006	Non-Saxon	44.01	12.48	45		*F* (1, 54) = 2.12, *p* = .15	.41
	Saxon	48.89	10.27	20			

[a]Covariates appearing in the model are evaluated at the following values: Percentage of students proficient/advanced in 2002 CST = 28.95, total = 597.09, per_aa = 8.39, per_as = 2.84, per_hi = 66.21, per_wh = 17.60, per_sd = 69.46, per_el = 37.00, per_di = 11.48.

Subgroup Differences

Tables A16 to A20 in the following pages summarize the results of the subgroup analyses for the Stanford 9, CST, and CAT 6 scale-score measures for students in elementary grades (2–5) and middle grades (6–8) separately. Note that analyses only included Saxon students in a school *actively* using *Saxon Math*. The ANOVA results for the interaction of each subgroup classification with group (Saxon vs. non-Saxon) are first presented. When this interaction is significant, the corresponding simple effects *t*-test results are presented. It is important to note that given that this is an observational study and data were not randomized into the treatment groups and that there may be preexisting differences, and also in the absence of a strong theory, it is important to view the pattern of subgroup results as a primarily, exploratory exercise.

Table A16. Subgroup Differences: Gender Status

Subgroup	Group	*M*	*SD*	*N*	*F*-interaction	*t* test
Elementary Level Results: Stanford 9 Scale Score						
Male	Non-Saxon	608.62	46.85	36,480	$F(1, 99824) = 12.50$, $p < .001$	$t(24505) = 7.80$, $p < .001$
	Saxon	604.88	48.84	14,018		
Female	Non-Saxon	608.62	46.18	35,346		$t(24995) = 2.98$, $p = .003$
	Saxon	607.22	47.54	13,984		
Middle-School Level Results: Stanford 9 Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Male	Non-Saxon	662.50	39.53	29,745	$F(1, 92618) = .02$, $p = .90$	na
	Saxon	665.66	42.45	16,896		
Female	Non-Saxon	664.27	36.59	29,047		na
	Saxon	667.51	40.00	16,934		
Elementary Level Results: CST Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Male	Non-Saxon	337.06	75.07	36,246	$F(1, 131872) = .34$, $p = .56$	na
	Saxon	336.49	76.50	30,825		
Female	Non-Saxon	335.91	70.24	34,351		na
	Saxon	335.81	71.47	30,454		
Middle-School Level Results: CST Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Male	Non-Saxon	314.27	59.31	43,239	$F(1, 135166) = .64$, $p = .43$	na
	Saxon	312.41	61.84	25,079		
Female	Non-Saxon	316.83	55.47	41,766		na
	Saxon	315.49	59.22	25,086		
Elementary Level Results: CAT 6 Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Male	Non-Saxon	605.15	54.42	23,148	$F(1, 83909) = .05$, $p = .83$	na
	Saxon	606.91	56.00	19,504		
Female	Non-Saxon	604.93	51.75	21,913		na
	Saxon	606.85	52.94	19,348		
Middle-School Level Results: CAT 6 Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Male	Non-Saxon	656.58	56.46	29,240	$F(1, 91350) = .15$, $p = .70$	na
	Saxon	658.46	57.08	17,027		
Female	Non-Saxon	659.55	50.77	28,127		na
	Saxon	661.15	53.05	16,960		

Note. na = not applicable.

Table A17. Subgroup Differences: Ethnic Status

Elementary Level Results: Stanford 9 Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
White	Non-Saxon	627.88	47.12	13,521	F (2, 93629) = 350.40, $p < .001$	t (18038) = 14.91, $p < .001$
	Saxon	639.98	47.59	4,519		
Hispanic	Non-Saxon	601.29	43.92	47,303		t (27331) = 1.08, $p = .28$
	Saxon	600.85	45.02	16,137		
African American	Non-Saxon	607.94	44.68	6,588		t (12153) = 21.69, $p < .001$
	Saxon	590.44	43.97	5,567		
Middle-School Level Results: Stanford 9 Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
White	Non-Saxon	687.92	41.25	12,266	F (2, 86350) = 139.68, $p < .001$	t (13108) = 16.24, $p < .001$
	Saxon	698.09	40.07	6,326		
Hispanic	Non-Saxon	654.80	32.41	40,194		t (31125) = 11.18, $p < .001$
	Saxon	658.24	34.44	17,329		
African American	Non-Saxon	654.22	31.70	2,898		t (5536) = 7.38, $p < .001$
	Saxon	649.02	33.18	7,343		
Elementary Level Results: CST Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
White	Non-Saxon	365.85	76.70	10,081	F (2, 124386) = 64.82, $p < .001$	t (21073) = 6.25, $p < .001$
	Saxon	372.35	75.35	11,424		
Hispanic	Non-Saxon	327.61	68.50	52,781		t (78027) = 4.39, $p < .001$
	Saxon	329.67	69.57	36,645		
African American	Non-Saxon	324.88	67.33	2,736		t (13459) = 9.37, $p < .001$
	Saxon	311.27	67.89	10,725		
Middle-School Level Results: CST Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
White	Non-Saxon	348.28	62.78	14,532	F (2, 127308) = 150.31, $p < .001$	t (22226) = 10.51, $p < .001$
	Saxon	357.59	62.79	7,696		
Hispanic	Non-Saxon	305.67	50.63	62,399		t (55519) = 2.96, $p = .003$
	Saxon	304.58	52.62	29,350		
African American	Non-Saxon	302.11	50.20	3,692		t (6290) = 12.69, $p < .001$
	Saxon	290.01	46.81	9,645		
Elementary Level Results: CAT 6 Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
White	Non-Saxon	623.97	53.54	6,555	F (2, 78980) = 68.08, $p < .001$	t (13532) = 10.27, $p < .001$
	Saxon	633.37	52.85	6,979		
Hispanic	Non-Saxon	599.62	51.36	33,559		t (56555) = 6.80, $p < .001$
	Saxon	602.61	51.45	22,998		
African American	Non-Saxon	600.46	52.91	1,754		t (2764) = 6.91, $p < .001$
	Saxon	590.64	55.12	7,141		
Middle-School Level Results: CAT 6 Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
White	Non-Saxon	686.54	51.40	10,185	F (2, 85980) = 47.99, $p < .001$	t (11214) = 10.46, $p < .001$
	Saxon	695.26	47.67	5,209		

Hispanic	Non-Saxon	649.61	50.85	41,695		*t* (37467) = 4.43, $p < .001$
	Saxon	651.59	51.96	19,535		
African American	Non-Saxon	648.10	53.88	2,619		*t* (4573) = 4.30, $p < .001$
	Saxon	642.83	51.39	6,743		

Table A18. Subgroup Differences: English Language Learner Status

Subgroup	Group	*M*	*SD*	*N*	*F*-interaction	*t* test
Elementary Level Results: Stanford 9 Scale Score						
Non-ELL	Non-Saxon	617.27	47.27	38,222	F (1, 70143) = 3.15, p = .08	na
	Saxon	612.84	49.79	16,035		
ELL	Non-Saxon	602.61	46.27	12,170		na
	Saxon	596.40	47.00	3,720		
Middle-School Level Results: Stanford 9 Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Non-ELL	Non-Saxon	674.34	40.66	29,534	F (1, 65868) = 138.36, $p < .001$	t (43273) = 6.92, $p < .001$
	Saxon	671.71	43.18	20,892		
ELL	Non-Saxon	656.94	33.59	11,490		t (5797) = 9.68, $p < .001$
	Saxon	664.07	41.95	3,956		
Elementary Level Results: CST Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Non-ELL	Non-Saxon	351.22	75.52	32,152	F (1, 124263) = 14.14, $p < .001$	t (66671) = 9.66, $p < .001$
	Saxon	345.52	77.05	34,741		
ELL	Non-Saxon	318.90	65.73	34,543		t (49282) = 4.68, $p < .001$
	Saxon	316.29	64.94	22,831		
Middle-School Level Results: CST Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Non-ELL	Non-Saxon	330.23	61.87	39,444	F (1, 116697) = 326.64, $p < .001$	t (55997) = 19.48, $p < .001$
	Saxon	320.53	63.23	26,504		
ELL	Non-Saxon	289.40	41.71	33,008		t (33621) = 5.90, $p < .001$
	Saxon	291.83	45.63	17,745		
Elementary Level Results: CAT 6 Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Non-ELL	Non-Saxon	614.91	52.99	20,495	F (1, 79379) = 39.21, $p < .001$	t (43078) = 5.70, $p < .001$
	Saxon	611.90	56.41	22,645		
ELL	Non-Saxon	592.36	50.24	22,167		t (30521) = 3.29, $p < .001$
	Saxon	594.11	48.97	14,076		
Middle-School Level Results: CAT 6 Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Non-ELL	Non-Saxon	671.89	53.16	27,009	F (1, 79846) = 284.74, $p < .001$	t (39220) = 11.92, $p < .001$
	Saxon	665.72	55.33	18,724		
ELL	Non-Saxon	633.58	49.54	22,464		t (23007) = 12.08, $p < .001$
	Saxon	640.54	50.96	11,653		

Note. na = not applicable.

Table A19. Subgroup Differences: Economic Disadvantage Status

Subgroup	Group	*M*	*SD*	*N*	*F*-interaction	*t* test
Elementary Level Results: Stanford 9 Scale Score						
Non-Econ. Disadvantaged	Non-Saxon	629.06	47.60	18,221	$F(1, 100007) = 331.83$, $p < .001$	$t(23274) = 14.82$, $p < .001$
	Saxon	640.28	47.69	5,055		
Econ. Disadvantaged	Non-Saxon	601.67	44.04	53,616		$t(76733) = 9.63$, $p < .001$
	Saxon	598.31	45.00	23,119		
Middle-School Level Results: Stanford 9 Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Non-Econ. Disadvantaged	Non-Saxon	684.53	40.91	18,598	$F(1, 92672) = 670.77$, $p < .001$	$t(27424) = 29.04$, $p < .001$
	Saxon	699.84	40.57	8,828		
Econ. Disadvantaged	Non-Saxon	653.57	32.37	40,203		$t(50534) = 4.66$, $p < .001$
	Saxon	654.84	34.53	25,047		
Elementary Level Results: CST Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Non-Econ. Disadvantaged	Non-Saxon	372.36	76.96	14,100	$F(1, 131880) = 1.44$, $p = .23$	na
	Saxon	371.32	77.67	14,372		
Econ. Disadvantaged	Non-Saxon	327.55	68.82	56,504		na
	Saxon	325.37	69.42	46,908		
Middle-School Level Results: CST Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Non-Econ. Disadvantaged	Non-Saxon	345.49	64.28	22,625	$F(1, 135184) = 601.84$, $p < .001$	$t(20288) = 19.20$, $p < .001$
	Saxon	360.31	66.42	10,663		
Econ. Disadvantaged	Non-Saxon	304.66	50.58	62,392		$t(82015) = 9.71$, $p < .001$
	Saxon	301.43	52.25	39,508		
Elementary Level Results: CAT 6 Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Non-Econ. Disadvantaged	Non-Saxon	626.59	54.09	9,012	$F(1, 83923) = 19.49$, $p < .001$	$t(18206) = 4.75$, $p < .001$
	Saxon	630.40	54.34	9,196		
Econ. Disadvantaged	Non-Saxon	599.65	51.52	36,059		$t(65717) = .17$, $p = .87$
	Saxon	599.58	52.45	29,660		
Middle-School Level Results: CAT 6 Scale Score						
Subgroup	**Group**	***M***	***SD***	***N***	***F*-interaction**	***t* test**
Non-Econ. Disadvantaged	Non-Saxon	683.87	51.95	15,609	$F(1, 91357) = 171.90$, $p < .001$	$t(15820) = 17.12$, $p < .001$
	Saxon	695.83	48.80	7,558		
Econ. Disadvantaged	Non-Saxon	648.39	51.18	41,761		$t(55219) = 2.72$, $p = .006$
	Saxon	649.50	52.43	26,433		

Note. na = not applicable.

Table A20. Subgroup Differences: Disability Status

Subgroup	Group	*M*	*SD*	*N*	*F*-interaction	*t* test
Elementary Level Results: Stanford 9 Scale Score						
No Disability	Non-Saxon	610.41	46.26	66,533	F (1, 100007) = .01, p = .91	na
	Saxon	607.12	48.02	26,681		
Disability	Non-Saxon	586.07	43.80	5,304		na
	Saxon	582.93	46.72	1,493		
Middle-School Level Results: Stanford 9 Scale Score						
No Disability	Non-Saxon	665.85	37.68	54,117	F (1, 92672) = 1.61, p = .21	na
	Saxon	668.16	41.16	32,064		
Disability	Non-Saxon	634.62	30.58	4,684		na
	Saxon	638.32	31.33	1,811		
Elementary Level Results: CST Scale Score						
No Disability	Non-Saxon	341.08	71.40	64,608	F (1, 131499) = 93.86, $p < .001$	t (118909) = 5.39, $p < .001$
	Saxon	338.84	73.54	57,091		
Disability	Non-Saxon	286.31	68.70	5,820		t (8330) = 8.77, $p < .001$
	Saxon	299.01	71.50	3,984		
Middle-School Level Results: CST Scale Score						
No Disability	Non-Saxon	319.61	57.11	77,354	F (1, 134821) = 31.82, $p < .001$	t (96840) = 11.56, $p < .001$
	Saxon	315.68	60.58	47,858		
Disability	Non-Saxon	273.50	42.76	7,376		t (3398) = 3.67, $p < .001$
	Saxon	277.63	47.69	2,237		
Elementary Level Results: CAT 6 Scale Score						
No Disability	Non-Saxon	608.13	51.08	41,258	F (1, 83462) = 29.32, $p < .001$	t (75194) = 1.84, p = .07
	Saxon	608.82	53.45	36,239		
Disability	Non-Saxon	570.82	62.64	3,606		t (5967) = 5.20, $p < .001$
	Saxon	579.38	61.59	2,363		
Middle-School Level Results: CAT 6 Scale Score						
No Disability	Non-Saxon	663.04	50.61	52,011	F (1, 91073) = 106.24, $p < .001$	t (65262) = 4.38, $p < .001$
	Saxon	661.40	54.47	32,538		
Disability	Non-Saxon	608.21	59.03	5,127		t (6526) = 8.57, $p < .001$
	Saxon	623.37	57.40	1,401		

Note. na = not applicable.

SAXON™
A Harcourt Education Company

saxonpublishers.harcourtachieve.com
1.800.531.5015