The ARC Center Tri-State Student Achievement Study

Executive Summary

As Ball and Cohen (1996) point out, there are good reasons to believe that changing curriculum materials can alter classroom instruction. Curriculum materials provide the activities that shape the daily interactions between teachers and students. If they are written with sufficient specificity, curriculum materials can help teachers translate research findings and authoritative recommendations into classroom reality. Because they can be easily disseminated, curriculum materials have the potential to help large numbers of teachers transform their classroom instruction.

Recognizing this potential, the National Science Foundation (NSF) in the early 1990s funded three projects to create comprehensive elementary mathematics curricula that would be aligned with the vision for school mathematics in the NCTM's *Curriculum and Evaluation Standards for School Mathematics* (1989). The three projects were The University of Chicago School Mathematics Project, the TIMS Project at the University of Illinois at Chicago, and TERC in Cambridge, Massachusetts. The curricula those projects produced are *Everyday Mathematics; Math Trailblazers;* and *Investigations in Number, Data, and Space,* respectively.

In 2000, the ARC Center at COMAP in Lexington, Massachusetts, received funding from NSF to carry out a large-scale study of the effects of *Everyday Mathematics* (EM), *Investigations in Number, Data, and Space* (IN), and *Math Trailblazers* (MT) on student performance on state-mandated standardized tests in Massachusetts, Illinois, and Washington State.

Method

The study combined survey data from schools using the three curricula with publicly available data from state-mandated tests in the three states. The combined data set made it possible to compare the achievement of students studying the reform curricula with matched comparison students not using any of the three curricula.

The ARC Center study focused on Illinois, Massachusetts, and Washington State for two reasons: First, the reform programs EM, MT, and IN were represented by substantial numbers of users in these states, and second, the five different state-mandated standardized tests used in these states permitted analysis across a variety of instruments. The five tests were:

- *Illinois Standards Achievement Test* (ISAT), grade 3
- *Illinois Standards Achievement Test* (ISAT), grade 5
- *Massachusetts Comprehensive Assessment System* (MCAS), grade 4
- *Iowa Test of Basic Skills* (ITBS), grade 3 (Washington State)
- *Washington Assessment of Student Learning* (WASL), grade 4

Each project conducted a telephone survey of districts and schools in the three states that were known to use, or were suspected of using, its curriculum. These districts and schools were identified principally through customer lists provided by the respective program publishers. Schools designated to be surveyed included approximately 90% of all students in the three states using the three curricula. The coverage rate for the survey—the ratio of total students in all schools responding to the survey to total students in all schools designated to be surveyed—was calculated separately by program and state, and was generally at the 90% level or higher. The schools in the study, therefore, include a near census of all students using these three curricula in these three states.

The survey collected 1999–2000 school year data for the grades for which state test data were available. To gauge the extent and length of implementation, data were also collected for previous grades, and in some cases for the following grade. Survey respondents included district math supervisors, principals, or other knowledgeable persons.

The primary reason for conducting the implementation survey was to verify usage of the reform curricula and to determine eligibility for each grade in each school. A grade in a school was considered eligible for inclusion in the analysis provided:

- The grade was one for which student test data is available (grades 3 and 5 in Illinois, grade 4 in Massachusetts, and grades 3 and 4 in Washington).
- The school-grade reported full implementation of EM, IN, or MT during the 1999–2000 school year.
- The program had been implemented in the previous grade within the school for at least two years (1998–2000), so that students in the given grade would have had at least a two-year exposure to the program.

The coded implementation survey file contained 1,058 school-grade records for which student test data were available. Of these, 742 (70%) were classified as eligible and were subsequently matched to comparison schools. Failure to meet the two-year implementation requirement was the most common reason for classifying a school-grade case as ineligible.

Matching Reform and Comparison School-Grade Combinations

A matching routine was carried out for each of the five state-grade combinations in order to identify a set of comparison schools that had not implemented any one of the three reform programs, but that were similar in how they would be expected to perform on the respective statewide test. The matching procedure selected one matched comparison school for each of the 742 eligible reform school-grade cases used in the analysis.

Within each state-grade combination, schools known to use, or suspected of using, any one of the three reform curricula were excluded as possible matched comparison schools. All remaining schools appearing on that state's public education data files formed the pool of schools eligible for selection as comparison schools.

Separate school-level regression analyses for the different state-grade combinations provided information concerning the strongest predictors of the average school mathematics score for each state-grade test. Reading score and income variables consistently accounted for the greatest percentage of total variance. These variables were given greater weight in the matching process. Other variables—such as percent white, school mobility rate, and percent with limited English proficiency (LEP)—accounted for little of the total variance, but were typically significant. These variables were given less weight in the matching process.

The actual matching routine was carried out separately for each of the five state-grade combinations. The variables used in matching for the different state-grade combinations were as follows:

- Illinois:
  School averages for reading score, low-income %, white %, LEP %, and mobility %.
- Massachusetts:
  School averages for reading score, free/reduced lunch %, and white %
- Washington:
  School averages for reading score, TitleI Mathematics %, and white %.

(Additionally, the school variable TitleIS identifies TitleI schools and was used as a stratification variable: A reform school and its matched/comparison school were required to have the same TitleIS designation.)

For each reform school-grade case, the matching routine identified a comparison school that resembled the reform school with respect to the matching variables.

Table 1 shows the matching variable averages for students in the 742 eligible reform school-grade cases and their comparison school-grades. There is generally close agreement between the reform-student and comparison-student averages for the variables used in matching, but differences do exist, and such differences could bias any subsequent comparisons. Therefore, the comparison-student averages for all test variables were adjusted first—before any tabulated comparisons were made. Adjustment ensured that any bias ensuing from the matching procedure to select comparison schools was minimized.

Exclusions, Missing Data, and Weights

Before any comparisons of the performance of reform and comparison students were tabulated, various student-record exclusions and imputations were made, and a set of case weights for comparison students was calculated.

All reform and comparison student records for IEP, "mathematically disabled", and "special education" students were deleted from the analysis and excluded from the tabulated comparisons. These deleted records represent approximately 10% of all student records. Fewer than 3% of the student records included missing or incomplete math test data or reading scores. All such records were deleted from the analysis and excluded from the tabulated comparisons.

Table 2 shows, for each state-grade combination, the number of student records for reform and comparison students that were in fact used for tabulated comparisons and all subsequent analysis. In all, more than 100,000 student records are represented, with approximately equal numbers of reform-student and comparison-student records.

The near equality in numbers of reform-student and comparison-student records shown in Table 2 does not apply, however, at the individual school level. The difference between the number of students in a given reform school-grade and its matched school-grade was highly variable and sometimes substantial. Weighting was therefore necessary, and case weights were constructed for all comparison-student records. Use of case weights for all tabulations ensured that comparison schools contributed to overall statistics with the same proportions as their reform-school counterparts.

Results

Tabulations of differences between reform and comparison student scores were made separately for each of the five state-grade combinations; these results were also pooled to yield overall tabulations. For disaggregated comparisons by race/ethnicity and income, results from all state-grade tabulations were pooled.

The mathematics test variables used for all tabulations are student-level variables. The overall mathematics test score variables are "math" and "total". "Math" is the scaled test score; "total" is the percent of total possible points on the test. Each of the variables "computation", "measurement", "geometry", "prob/stat", and "algebra" denotes the percent of total possible points for the corresponding strand of test items.

Each set of tabulated comparisons for a state-grade combination compares averages for reform students and comparison students within that state-grade combination. These differences between averages were not calculated simply by subtracting the observed comparison-student average from the observed reform-student average for each test variable. The observed comparison-student average for each test variable

was instead adjusted prior to subtraction. The adjustment procedure was based on regression analyses and ensured that any bias ensuing from imperfect matching of reform and comparison schools was minimized.

Having calculated an adjusted difference of average scores between reform students and their comparison students, the effect size for that difference was then calculated by dividing the adjusted difference by the standard deviation of the comparison student scores.

For this study, an effect size can be thought of as the percentile standing of the average reform student relative to the average comparison student. An effect size of 0.10 (the approximate 3-state weighted average effect size for both the "math" and "total" test scores) indicates that the mean of the reform-student group is at the 54th percentile of the comparison group. This, in turn, implies a change in percentile standing of 4 percentile points. Tables 3, 5, and 6 all use the label "percentile change" to denote the change in percentile standing of the average reform student relative to the average comparison student, as determined by the effect size.

Comparisons by State-Grade Combination.

Table 3 shows comparisons for all test variables, by state-grade combination. The effect sizes for "math" and "total" are approximately the same for all state-grade combinations—as expected, since these are the overall test score variables. The combined state-grade effect sizes for "math" and "total" are virtually identical and correspond to a percentile change of about +4% (favoring the reform students).

Counting "math" and "total" as a single comparison within each state-grade combination, 34 different comparisons are represented within Table 3, of which 28 favor the reform students, six show no statistically significant difference, and none favor the comparison students. The combined state-grade effect sizes are highly significant ($p < 0.001$) for all mathematics strands; and they are fairly consistent across strands, with probability and statistics as the single exception.

The comparisons shown in Table 3 are differences and convey no information about the level of student performance. Table 4, however, does show the actual levels of student performance corresponding to the differences in Table 3. For comparative purposes, Table 4 also shows the levels of performance for all non-reform students. "Non-reform students" include all students within a state-grade combination that do not attend any of the eligible reform schools, and that are not identified as IEP, mathematically-disabled, or special-education students. In particular, non-reform students include all comparison students. The side-by-side bar graph in Figure 1 summarizes information from Table 4. The differences reported in Table 4 correspond to the differences in adjacent bar heights in Figure 1.

Including averages for all non-reform students in Table 4 and Figure 1 highlights the impact of the matching procedure used to select comparison schools. The reform schools have, relative to other schools in their state, higher average reading scores, higher percentages of white students, and lower percentages of low-income students, all of which are associated with higher mathematics achievement. Direct comparison of reform- and non-reform-student performance would, therefore, largely reflect the differences in these factors between the two student groups and would not furnish valid measures of the effects of the reform curricula. The matching procedure, however, selected comparison schools with comparable values for these factors, and the differences between reform- and comparison-student performance do furnish valid measures of program effects.

Comparisons by Race/Ethnicity and by Income

Table 5 and Figure 2 show comparisons by race/ethnicity that combine results from the individual state-grade tabulations. The effect sizes are remarkably similar for blacks and whites, and these effect sizes very nearly duplicate the combined state-grade effect sizes. The effect sizes for Asians are generally at the same level or higher than those for blacks and whites. With the exception of probability and statistics,

virtually all of the effect sizes for Asians, blacks, and whites are highly significant and favor the reform students.

The results for Hispanics, however, are quite different. None of the effect sizes for "math", "total", computation, algebra, and probability and statistics are statistically significant. The effect sizes for measurement and geometry are both positive and significant; each, however, is much smaller than the corresponding effect size for Asians, blacks, or whites.

The effect sizes for probability and statistics are exceptional within Table 5, just as they are within Table 3. They are small and generally favor the reform students, but are not statistically significant except for whites. The combined state-grade effect size for probability and statistics is highly significant, but so small (0.025) that the result lacks practical significance.

Table 6 shows comparisons by student family income that combine results from the individual state-grade tabulations. Comparisons are reported by TitleIS status for Washington State, and by SES categories for combined Illinois/Massachusetts data.

The effect sizes shown by Table 6 are quite similar across SES and TitleIS categories for the overall test score variables "math" and "total". All such effect sizes are highly significant and all favor the reform students. Effect sizes for low-SES and top-SES students are at the same level, and marginally higher than those for middle-SES students. Effect sizes for TitleIS students are marginally higher than those for non-TitleIS students, and marginally lower than those for low-SES students.

Conclusion

This study examined achievement test data from three states for a near census of students in schools using NSF-funded comprehensive elementary mathematics curricula. These students' test results were compared to those of students in non-using schools carefully matched by reading score, SES, and other variables. Possible bias due to imperfect matching was controlled by adjustments based on regression studies. The principal finding of the study is that the students in the NSF-funded reform curricula consistently outperformed the comparison students: All significant differences favored the reform students; no significant difference favored the comparison students. This result held across all tests, all grade levels, and all strands, regardless of SES and racial/ethnic identity. The data from this study show that these curricula improve student performance in all areas of elementary mathematics, including both basic skills and higher-level processes. Use of these curricula results in higher test scores.

References

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is--or might be--the role of curricular materials in teacher learning and instructional reform? *Educational Researcher*, 25 (9), pp. 6–8.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Table 1: Matching variable averages for students in eligible reform school/grades and their comparison school/grades*.

State-Grade Level	Matching Variable				
		Average percentage of:			
Illinois-Grade 3	reading score	white	low-income	mobility	LEP
Reform	165.75	74%	18%	13%	6%
Comparison	165.85	77%	18%	13%	5%
Illinois-Grade 5	reading score	white	low-income	mobility	LEP
Reform	164.46	76%	17%	11%	5%
Comparison	164.2	81%	18%	12%	4%
Massachusetts-Grade 4	reading score	white	free/reduced lunch	mobility**	LEP**
Reform	236.99	77%	12%	16%	2%
Comparison	236.28	80%	14%	11%	1%
Washington-Grade 3	reading score	white	TitleI Math	TitleIS	
Reform	192.07	80%	2%	16%	
Comparison	191.97	82%	2%	16%	
Washington-Grade 4	reading score	white	TitleI Math	Title1S	
Reform	412.46	80%	3%	6%	
Comparison	412.36	82%	3%	6%	

* Averages for comparison students in matched schools are weighted averages.

** Variable was tracked but not used in matching .

Table 2: Number of student records used for tabulated comparisons, by state, grade level, school status, and curriculum.

			Reform Program			
State	**Grade Level**	**School Status**	**Everyday Math**	**Investigations**	**Math Trailblazers**	**Total**
Illinois	3rd	Reform	13,840	0	1,035	14,875
		Comparison	13,216	0	901	14,117
	5th	Reform	12,988	0	832	13,820
		Comparison	13,098	0	563	13,661
Massachusetts	4th	Reform	3,962	2,917	0	6,879
		Comparison	4,181	3,337	0	7,518
Washington	3rd	Reform	4,412	916	2,485	7,813
		Comparison	3,923	783	2,150	6,856
	4th	Reform	4,499	920	2,534	7,953
		Comparison	4,063	907	2,413	7,383
Totals		Reform	39,701	4,753	6,886	51,340
		Comparison	38,481	5,027	6,027	49,535
			78,182	9,780	12,913	100,875

Table 3: Average differences and effect sizes, by state/grade combination.

		math	total	computation	measurement	geometry	prob/stat	algebra				
IL grade3	difference	1.39***	1.82***	2.78***	3.84***	0.76***	0.06	1.44***				
(n=14,875)	effect size	0.098	0.099	0.141	0.164	0.038	0.003	0.073				
IL grade5	difference	1.82***	2.20***	2.29***	3.02***	3.26***	1.55***	1.44***				
(n=13,820)	effect size	0.121	0.116	0.117	0.132	0.165	0.079	0.067				
									open response	short answer	multiple choice	
MA grade 4	difference	1.34***	1.33***	2.36***		-0.19	-0.16	3.07***	2.46***	-0.62	0.89***	
(n=6,879)	effect size	0.087	0.078	0.127		-0.010	-0.008	0.137	0.119	-0.024	0.053	
									problem solving	concepts/ estimation		
WA grade3	difference	1.34***	1.27***	0.74**					0.76**	1.86***		
(n=7,813)	effect size	0.073	0.078	0.039					0.036	0.108		
										logic	commun- icating	making connections
WA grade4	difference	3.02***	1.77***	1.02***	3.43***	1.61***	0.00	2.99***	2.51***	1.17***	-0.03	3.55***
(n=7,953)	effect size	0.093	0.093	0.041	0.120	0.078	0.000	0.112	0.090	0.040	-0.001	0.116
Combined	effect size	0.098***	0.097***	0.102***	0.142***	0.078***	0.025***	0.088***				
(n=51,340)												
	percentile change	+3.92%	+3.88%	+4.08%	+5.68%	+3.12%	+1.00%	+3.52%				

"Math" is scaled test score; "total" and remaining strand scores are percent of total possible points on entire test or appropriate strand portion of test.

The record counts in column one are the numbers of reform-student records used for tabulations. For any given tabulation, the weighted number of comparison-student records used is equal to the number of reform-student records used.

Two-sided significance levels are defined as follows: *** is $p < 0.001$, ** is $p < 0.01$, * is $p < 0.025$.

Table 4: Averages, by state/grade combination and reform status

State- Grade Level	Reform Status	Student Records	math	total	computation	measurement	geometry	prob/stat	algebra				
IL grade3	reform	14,875	167.18	70.22	68.95	67.78	73.26	74.64	66.44				
	comparison	14,117	165.79	68.40	66.17	63.94	72.50	74.58	65.00				
	non-reform	115,053	160.84	61.62	59.47	56.42	65.58	67.27	59.38				
IL grade5	reform	13,820	170.10	66.29	65.08	61.03	73.01	68.49	63.85				
	comparison	13,661	168.28	64.09	62.79	58.01	69.75	66.94	62.41				
	non-reform	116,316	162.19	56.34	55.03	50.60	62.33	58.82	54.56				
										open response	short answer	multiple choice	
MA grade 4	reform	6,879	244.13	66.88	70.13		62.05	72.16	61.50	60.32	59.65	72.65	
	comparison	7,518	242.79	65.55	67.77		62.24	72.32	58.43	57.86	60.27	71.76	
	non-reform	58,716	236.52	57.37	*		*	*	*	*	*	*	
										problem solving	concepts/ estimation		
WA grade3	reform	7,813	194.00	71.17	72.53					66.52	72.42		
	comparison	6,856	192.66	69.90	71.79					65.74	70.56		
	non-reform	59,021	189.58	67.18	69.74					61.34	67.83		
											logic	commun- icating	making connections
WA grade4	reform	7,953	401.60	56.62	55.22	67.26	60.58	62.64	69.50	38.82	57.62	45.03	61.02
	comparison	7,383	398.58	54.85	54.20	63.83	58.97	62.64	66.51	36.31	56.45	45.06	57.47
	non-reform	60,656	393.57	51.93	49.93	58.80	55.17	59.25	61.60	32.22	51.34	40.27	52.86

"Math" is scaled test score; "total" and remaining strand scores are percent of total possible points on entire test or appropriate strand portion of test.

Averages for non-reform students exclude all IEP, mathematically-disabled, and special-education students.

* Data for individual strands and test item categories was available only for reform and comparison schools.

Table 5: Average effect sizes and percentile changes, by student race/ethnicity

		math	total	computation	measurement	geometry	prob/stat	algebra
Asian (n=3,071)	effect size	0.106***	0.115***	0.097***	0.175***	0.162***	0.043	0.086***
	percentile change	+4.24%	+4.60%	+3.88%	+7.00%	+6.48%	+1.72%	+3.44%
Black (n=3,509)	effect size	0.092***	0.101***	0.109***	0.129***	0.081***	0.029	0.087***
	percentile change	+3.68%	+4.04%	+4.36%	+5.16%	+3.24%	+1.16%	+3.48%
Hispanic (n=3,002)	effect size	0.021	0.031	0.017	0.094***	0.049*	-0.005	0.035
	percentile change	+0.84%	+1.24%	+0.68%	+3.76%	+1.96%	-0.20%	+1.40%
White (n=37,609)	effect size	0.100***	0.100***	0.106***	0.144***	0.070***	0.020***	0.091***
	percentile change	+4.00%	+4.00%	+4.24%	+5.76%	+2.80%	+0.80%	+3.64%
Combined (n=51,340)	effect size	0.098***	0.097***	0.102***	0.142***	0.078***	0.025***	0.088***
	percentile change	+3.92%	+3.88%	+4.08%	+5.68%	+3.12%	+1.00%	+3.52%

"Math" is scaled test score; "total" and remaining strand scores are percent of total possible points on entire test or appropriate stand portion of test

The record counts in column one are the numbers of reform-student records used for tabulations. For any given tabulation, the weighted number of comparison-student records used is roughly equal to the number of reform-student records used.

The record count for "Combined" also includes Native Americans, "mixed" and "other" race categories, and all records for which race was missing and subsequently imputed.

The tabulations for geometry, prob/stat and algebra are based on 5-10% fewer student records, because these strands are not separately scored in the WA (grade 3) test.

The tabulations for measurement are based on 10-20% fewer student records, because this strand is scored separately by neither the MA (grade 4) test nor the WA (grade 3) test.

Two-sided significance levels are defined as follows: *** is $p < 0.001$, ** is $p < 0.01$, * is $p < 0.025$

Table 6: Average effect sizes and percentile changes, by school SES and Title1S status

			math	total	computation	measurement	geometry	prob/stat	algebra
Illinois and Massachusetts	SES low (n=9,723)	effect size	0.114***	0.102***	0.103***	0.144***	0.152***	0.027*	0.072***
		percentile change	+4.56%	+4.08%	+4.12%	+5.76%	+6.08%	+1.08%	+2.88%
	SES middle (n=8,476)	effect size	0.077***	0.078***	0.124***	0.110***	0.000	0.004	0.081***
		percentile change	+3.08%	+3.12%	+4.96%	+4.40%	+0.00%	+0.16%	+3.24%
	SES top (n=17,375)	effect size	0.101***	0.108***	0.151***	0.150***	0.046***	0.039***	0.081***
		percentile change	+4.04%	+4.32%	+6.04%	+6.00%	+1.84%	+1.56%	+3.24%
Washington	TitleIS (n=1,793)	effect size	0.096***	0.094***	0.046	0.065 #	0.032 #	-0.026 #	0.087 #
		percentile change	+3.84%	+3.76%	+1.84%	+2.60%	+1.28%	-1.04%	+3.48%
	non-TitleIS (n=13,973)	effect size	0.083***	0.085***	0.039***	0.125***	0.082***	0.003	0.116***
		percentile change	+3.32%	+3.40%	+1.56%	+5.00%	+3.28%	+0.12%	+4.64%
Combined	(n=51,340)	effect size	0.098***	0.097***	0.102***	0.142***	0.078***	0.025***	0.088***
		percentile change	+3.92%	+3.88%	+4.08%	+5.68%	+3.12%	+1.00%	+3.52%

Math is scaled test score; "total" and remaining strand scores are percent of total possible points on entire test or appropriate stand portion of test.

The record counts in column two are the numbers of reform-student records used for tabulations. For any given tabulation, the weighted number of comparison-student records used is equal to the number of reform-student records used.

For IL and MA, the tabulations for measurement are based on approximately 20% fewer student records because this strand is not scored separately by MA.

Statistics are based on only n=507 grade 4 reform-student records and an equal number of comparison-student records..

For WA, the tabulations for measurement, geometry, prob/stat, and algebra are based only on grade 4 student records. Thus, for these strands, the tabulations use only 507 reform-student records for TitleIS schools, and 7,446 reform-student records for non-TitleIS schools.

Two-sided significance levels are defined as follows: *** is $p < 0.001$, ** is $p < 0.01$, * is $p < 0.025$

Figure 1: Averages for the overall test score variable "total", by state/grade and reform status.

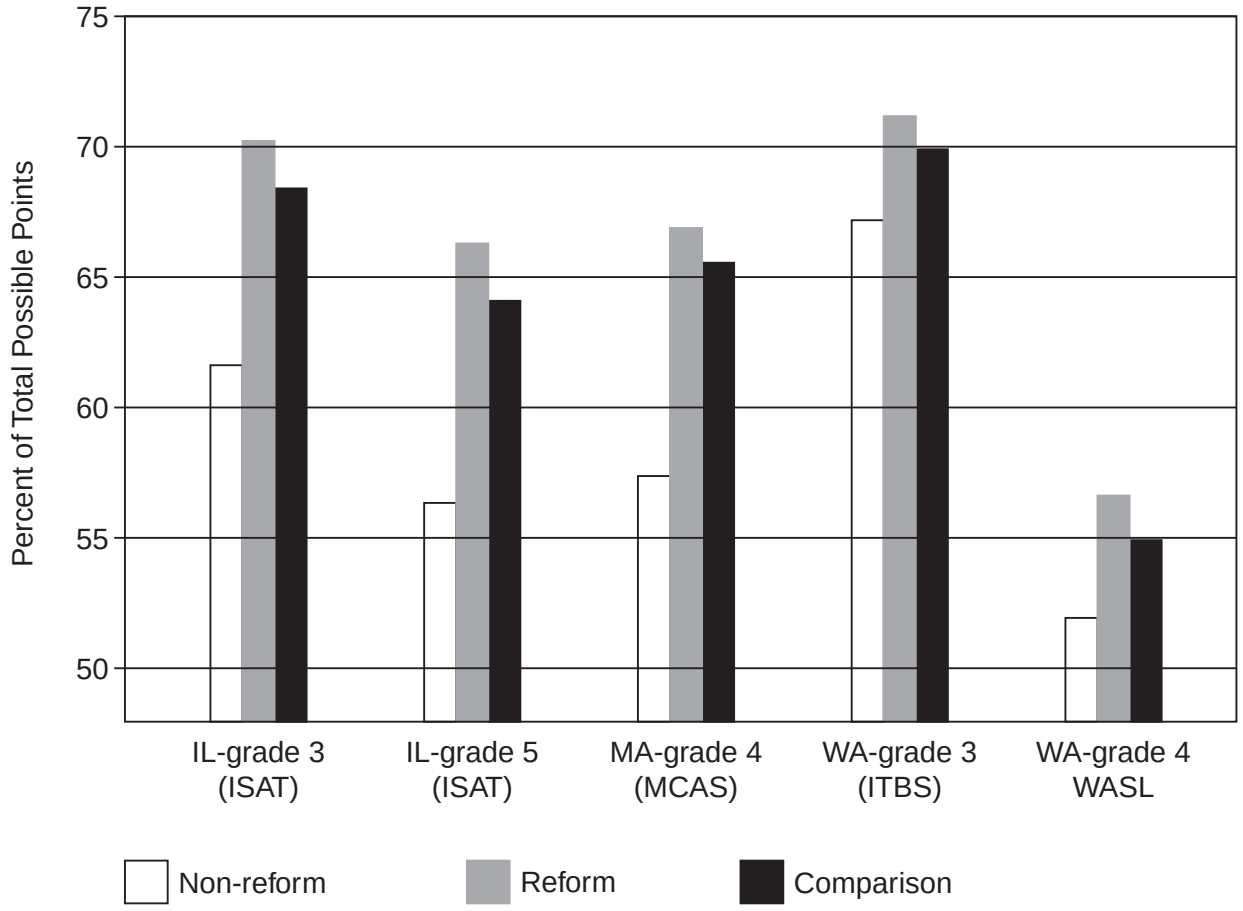
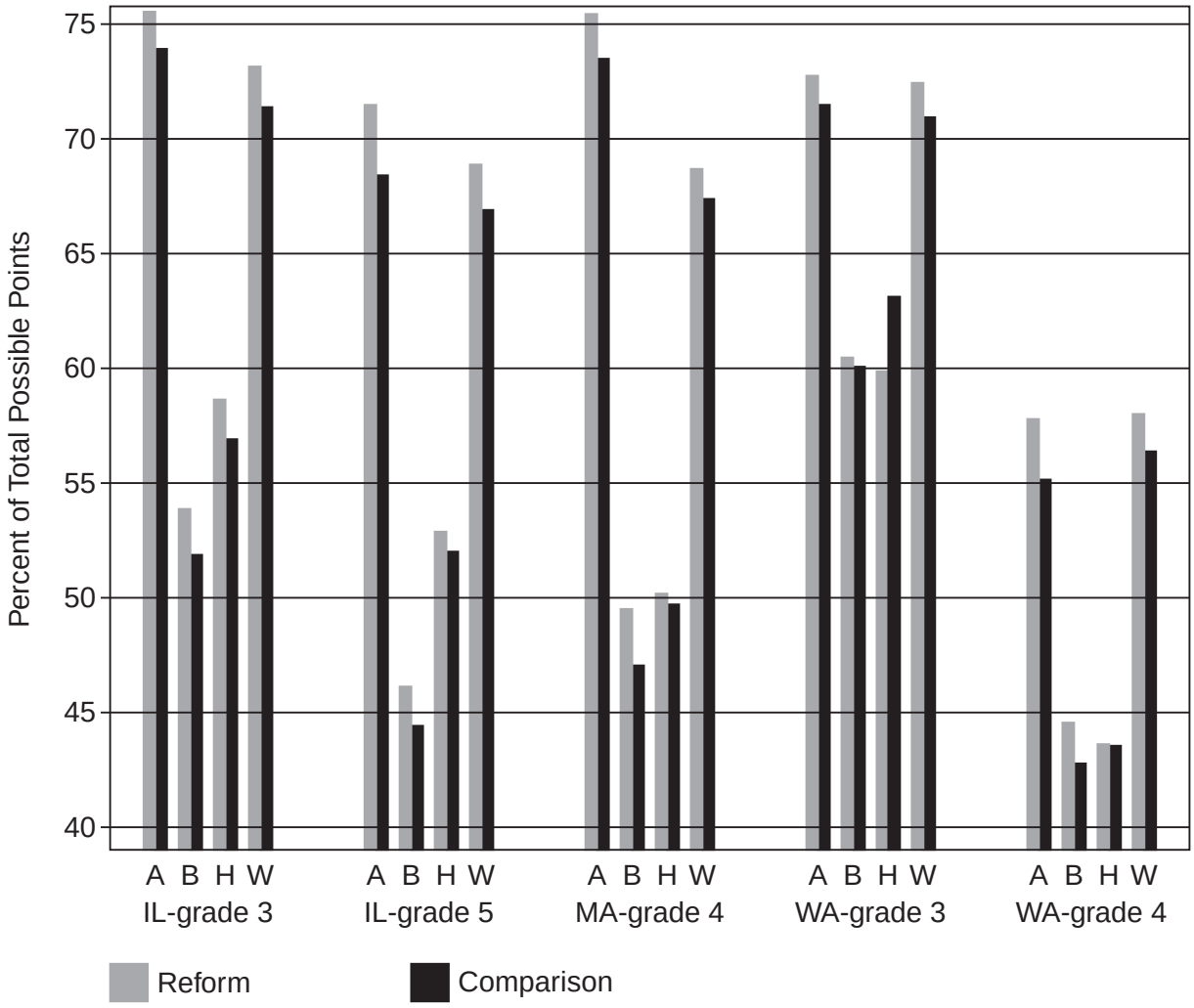


Figure 2: Averages for the overall test score variable "total", by state/grade, race/ethnicity, and reform status.



Race/ethnicity categories are defined as follows:
A = Asian, B = black, H = Hispanic, and W = white.