Standards-based Mathematics Curricula and Middle Grade Students' Performance on Standardized Achievement Tests

(A Report Submitted to the Division of Elementary, Secondary, and Informal Education, within the Directorate for Education and Human Resources, a division of the National Science Foundation, April 2004.)

Thomas R. Post
Arnie Cutler
Jon D. Davis
Edwin Andersen
Michael R. Harwell
Yukiko Maeda
Jeremy A. Kahan

University of Minnesota

NSF 9618741

The research was supported by the National Science Foundation under Grant No. NSF 9618741. Any opinions, findings, and conclusions expressed are those the authors and do not necessarily reflect the views of the National Science Foundation .

Abstract

This study extends our understanding of the relationship between the use of Standards based middle grades curricula and the learning of traditional mathematics topics as measured by two widely used standardized tests, the SAT-9 and the New Standards Mathematics Reference Exam. Sixteen hundred middle grades students in 43 classrooms from five districts with varying demographic profiles participated. These districts had participated in a Local Systemic Change (LSC) project supported by NSF. Students had used either the *Connected Mathematics Project* (CMP) or the *MATH Thematics* (STEM) program for three years.

Achievement related information was accumulated and used to detect patterns and to estimate their magnitude on the Open Ended, Problem Solving and Procedures subtests of the SAT-9. Hierarchical Linear Modeling (HLM) was used to analyze subtest results following methods described by Raudenbush and Bryk (2002). HLM permitted us to model variation in mathematics proficiency by controlling for other factors such as student prior achievement and classroom ethnic composition. It also allowed us to examine student and classroom patterns of achievement simultaneously. Student and classroom level predictive models were developed and fitted to predict student scores for various subgroups and to compare the predicted values against national Normal Curve Equivalence (NCE) means. A group of "value added" students were tested on three separate occasions over a two-year period of time to permit examination of achievement trends and achievement gaps among the various subpopulations.

In the within classrooms HLM analysis, students' socio-economic status (SES) and prior mathematics knowledge were significant predictors of achievement. A number of different factors such as SES, prior knowledge, concentrations of Asian students, percent special education and school district were significant predictors of between classroom variation.

Students' performance on the Open Ended and Problem Solving subtests were above national means, despite the fact that many began below this level. Their progress on the Procedures subtest was less stellar with districts ending below national means. This is perhaps not surprising since these curricula do not focus on paper and pencil calculation to the same extent as more traditional curricula.

Results suggest that Standards based students do learn traditional mathematics topics, at least those topics assessed by the two tests used. This study continues to raise the question "What mathematics is most valued and what type of a curriculum is most likely to provide it?

Introduction

This study examined achievement patterns of middle school students enrolled in *Standards*-based curricula. The focus was on traditional topics in mathematics as measured by two nationally normed achievement tests. By *Standards*-based curricula we are referring to those curricula which were funded from a solicitation of proposals through the National Science Foundation (NSF) in the early 1990's (Senk & Thompson, 2003).

This study builds on and extends our existing understanding of student achievement in *Standards*-based programs in four significant ways. First, we examine the impact of published editions of these curricula on student understanding as contrasted to earlier studies using field-test versions. Second, this research concerns the use of *Standards*-based curricula as part of district-wide curricula adoptions. By examining adopted versions of these curricula we study the achievement of students whose teachers are required to, and have not necessarily volunteered to teach *Standards*-based curricula. This provides a more accurate overall picture of expected student achievement since many earlier field-test teachers were volunteers and therefore may not be typical of all teachers that will eventually be expected to implement a *Standards*-based curriculum. Third, we employ hierarchical linear modeling (HLM) to account for the wide variability between classrooms and the inter-dependency of students within the same classroom (Kilpatrick, 2003). Thus, HLM allows us to consider student and classroom results simultaneously. And lastly, we examine achievement gaps over a two year period for

students whose prior mathematics achievement was categorized as high or low, and for high and low levels of socio-economic status (SES) as determined by eligibility for free or reduced lunch.

The present study adds yet another brushstroke to the emerging picture of mathematics achievement in classrooms using curricula directly and fundamentally influenced by the *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989) and similar documents (National Research Council, 1989). Schoenfeld (2002) reviews this emerging body of work and concludes that there is growing support for the success of such programs in terms of problem solving or other in-depth measures. This characterization is largely consistent with research on *Everyday Mathematics* (Briars & Resnick, 2000; Carroll, 1997; Riordan and Noyce 2001), *Connected Mathematics* (Riordan & Noyce, 2001; Ridgway, Zawojewski, Hoover, & Lambdin, 2003; Reys, Reys, Lapan, Holliday, & Wasman, 2003), *MATH Thematics* (Billstein, 1998; Reys, Reys, Lapan, Holliday, & Wasman, 2003), *Contemporary Mathematics in Context* (CMIC) (Schoen & Hirsch, 2003), *Interactive Mathematics Project* (Webb, 2003), University of Chicago School Mathematics Project (Thompson & Senk, 2001, who note it is *Standards*-based and NSF-funded), and *Mathematics: Modeling Our World* (Abeille & Hurley, 2001). Kilpatrick (2003), when referring to the 13 chapters in the Senk and Thompson book, concludes that "[the] studies reported in this volume offer the best evidence we have that *Standards*-based reform works" (pg. 487).

The research of student achievement in *Standards*-based curricula with regard to facility with both arithmetic and symbolic manipulation procedures is mixed over the short and long terms. For instance, Ridgway, Zawojewski, Hoover, & Lambdin (2003) found that 6th grade CMP students started one year behind non-CMP students on the Iowa Test of Basic Skills in the fall and at the end of grade 6 were 1.5 years behind the other group. The CMP Students who started .52 SD behind non-CMP students were .61 SD behind after one year. At the end of 8th grade, however (3 years), CMP students were .32 SD ahead (p. 207). In related investigations, the authors conclude that there is no immediate short term advantage but that the longer view is promising with CMP students making large gains on a broad range of curriculum topics and processes when compared to non-CMP students (pg. 215). Mokros (2003) reported that there were no differences between students studying from *Investigations in Number, Data, and Space* (TERC, 1998) and students studying from traditional curricula on mastery of basic facts. At the high school level, students studying from CMIC (Core-Plus) did less well on symbolic manipulation in abstract contexts without graphing calculators than traditional students (Huntley, Rasmussen, Villarubi, Sangtong, & Fey, 2000). Schoen & Hirsch (2003) found that students who studied Algebra I or Accelerated Geometry outperformed students studying from the first year of the CMIC curriculum on algebraic procedures. However, students who were in their second year of CMIC performed at statistically the same level as students enrolled in algebra or geometry. Only students in Accelerated Algebra 2, the third year of a traditional curriculum, outperformed CMIC Course 2 students on algebraic procedures.

There is research that the benefits from a *Standards*-based curriculum extend beyond increases in mathematics achievement on open ended problem solving. Billstein & Williamson (2003) found that students who used STEM improved in their attitudes towards mathematics at the middle school level. Cichon & Ellis (2003) support this finding among students at the secondary level. Billstein & Williamson (2003) found that students who used STEM at the middle-school level also had higher scores on the language achievement subtest of the Iowa Test of Basic Skills than a comparable group of students studying from other mathematics curricula.

Research suggests that curriculum is only one of the factors that influences student achievement. "Whereas improved curriculum materials can provide rich activities that support students' mathematical investigations, in and of themselves such materials may not be sufficient enablers of instruction that affords pursuit of conceptual issues." (Gearhart, Saxe, Seltzer, Schlackman, Ching, Nasir, et al., 1999, p. 309; c.f. Ball & Cohen, 1996). Schoen, Cebulla, Finn, & Fi (2003) found that fidelity of implementation of CMIC by teachers was "positively related to growth in student achievement" (p. 228).

Briars & Resnick (2000) looked at fidelity of implementation in the *Everyday Mathematics* program in the Pittsburgh schools. They found that schools with high fidelity of implementation scored two to five times higher on skills, problem solving, and concepts on the New Standards Mathematics Reference Examination. McCaffrey et al. (2001) also found that *Standards*-based teaching was positively related with student achievement, but only made a significant impact when a *Standards*-based curriculum was

also in place. Weiss, Banilower, Overstreet & Soar (2002) found that classrooms using a *Standards*-based curriculum were rated higher on a scale measuring inquiry-oriented teaching practices when compared to classrooms with traditional mathematics curricula. These findings suggest that a *Standards*-based curriculum alone can positively influence teacher pedagogy. However, the results are especially promising if combined with high fidelity of implementation and effective instruction of these new materials.

Another aspect of this complex interaction is the students themselves. The ways students react to and interface with the curriculum in the classroom can affect implementation of the curriculum (Cooney, 1985; Henningsen & Stein, 1997). In addition, the characteristics that students bring with them to the classroom also help to shape their achievement. For instance, in the SMSG study Begle (1973) stated,

"Even a casual inspection of the results of this study of predictors reveals two clear generalizations. The first of these is that the best predictor of mathematics achievement is previous mathematics achievement. … What we do say is that these nonmathematics scales do not affect later mathematics achievement nearly as much as previous achievement in mathematics does. The second generalization is this: The best predictors of computational skill at the end of the school year are generally computational skills at the beginning of the school year. On the other hand, the best predictors of performance at the high cognitive levels of understanding, application, and analysis seldom include computational skills." (p. 213-214).

Student socio-economic status (SES) has also been shown to play a role in how students interact with *Standards*-based curricula (Lubienski, 2000). Past research on student achievement in *Standards*-based classrooms has used SES either as a variable in matching groups for comparison purposes (Riordan & Noyce, 2001; Reys, Reys, Lapan, Holliday, & Wasman, 2003) or as a predictor of student achievement in regression analyses (Schoen, Cebulla, Finn, & Fi, 2003).

School environment also affects successful implementation of any curriculum (c.f., Eisenhart et al., 1993; Cohen & Ball, 2001). Schoen, Cebulla, Finn, & Fi (2003) found that professional development that is related to the curriculum is positively correlated with student achievement.

Methodology

Selection of the Districts.

In the mid to late 1990's the National Science Foundation, in an attempt to provide much needed professional development for school districts that adopted one or more of the new NSF funded *Standards*-based curricula, created the Local Systemic Change through Teacher Enhancement Initiatives (LSCs). The 47 funded LSCs (NSF 95-145) were "designed to engage entire school districts in the reform of science, mathematics and technology education, … to provide 47,000 teachers with professional development and … reach 1.6 million students in 240 school districts nationally" (p. 1 http://www.nsf.gov/pubs/1997/nsf97145/intro.htm.)

The LSC discussed here, "Minneapolis and St Paul Merging to Achieve Standards Project," $(MASP)^2$, was one of these 47 Projects. $(MASP)^2$ provided professional development (PD) targeted to new NSF funded *Standards*-based curricula to over 1100 middle grades and secondary teachers in 21 districts between 1997 and 2000. These teachers then provided *Standards*-based mathematics instruction to over 74,000 students in the 2000-2001 school year, and slightly larger numbers of students in the years since. Of these twenty-one school districts, five were invited to participate at the middle school level in the study of student achievement reported here. These five districts were chosen due to their geographic variety (urban and suburban), use of a variety of *Standards*-based curricula, and differing school contexts. In some cases there were high degrees of implementation while in other districts the middle-school and/or secondary mathematics faculty actively tried to undermine efforts to use *Standards*-based curricula.

The five districts included in our sample used either the *Connected Mathematics Project* (*CMP*) (Lappan, Fey, Fitzgerald, Friel, & Phillips, 1997) or *MATH Thematics* (*MT*) (Billstein & Williamson, 1998). These curricula differ from each other in various ways; the length of the units, the reality of the contexts, and the emphasized content are a few examples. Because both curricula responded to the same NSF Request for Proposals, they share many similarities such as recurring integration of topics within grade levels and a decreased emphasis on procedural knowledge. Although we recognize the danger of combining similar curricula (Davis, 1990) our research question referred to students in broadly defined *Standards*-based mathematics classrooms, not those studying from a specific *Standards*-based curricula.

The location, curriculum, rationale and assessment for each of the five districts are shown in Table 1. There were sharp differences among some of the districts in geographic location, student enrollment, and student characteristics. One school district had substantially greater enrollment than the others and was the only one located in a purely urban setting. Another district was located on an urban-suburban boundary, while the remaining districts were located in suburban settings. The largest district also showed the greatest diversity in student ethnicity, eligibility for a free or reduced price lunch, percentage of English language learners, and special education status. In contrast, the remaining four districts had only modest variation on student demographic variables and their students were predominately white native English speakers who were not eligible for free or reduced lunch.

Table 1: Location, Curriculum, Rationale, and Assessment for Choice of Districts

District	Geographic Location	Middle Grades Curriculum Assessed in this Study	Rationale for Choice of District	Assessment Used
A	Urban-Suburban Boundary	*MT*	High fidelity of implementation with much parent support	SAT-9
B	Urban	*CMP*	Large Urban Population – Variable Implementation	SAT-9
C	Suburban	*CMP*	Wholesale adoption sabotaged by a few faculty dissenters and parents	SAT-9
D	Suburban	*MT*	Wholesale adoption with district authorized supplements	SAT-9
E	Suburban	*CMP*	Enthusiastic Adoption	SAT-9 & New Standards

Data Collection

Two types of data were gathered. We refer to them as "one-shot" and "value-added". "One-shot" cross-sectional data consisted of two groups of 8th grade middle school students, one tested in the Spring of 2001, and the other in the Spring 2002. These students had been studying from either *CMP* or *MT* for a total of three years. The Spring 2002 group also contained 8th grade "value-added" students. There were no theoretical reasons to consider the cohorts of eighth grade students tested as separate. Similarly, the results of analyses such as ANOVA and HLM in which cohort served as a predictor indicated that there were no empirical reasons to treat the groups as separate. As a result, they were combined into a single group for analysis purposes. The sample for this study consists of approximately 1600 *Standards*-based students who took either the Stanford 9 or the New Standards Test.

In the "value-added" data, students who had studied from either *CMP* or *MT* as sixth-graders and continued through 7th and 8th grades in the curriculum were tested three times: in the beginning of 7th grade (Fall 2000), the end of 7th grade (Spring 2001), and at the end of 8th grade (Spring 2002). The purpose of this phase of the assessment was to provide a longitudinal description of student achievement over two successive school years.

In both the value-added and one-shot data, 43 selected mathematics classes were identified by the districts as being typical of classes using these materials.

Design

A non-experimental design with clustering was used. Students were considered to be clustered (nested) within classrooms, which in turn were clustered within school districts. Information was obtained for each level of clustering in the sample, but the focus was on students and classrooms. The lack of experimental manipulation means that study results support inferences about relationships among variables and their magnitude, but do not generally support strong causal inferences.

Instruments

This research project began with questions from districts, teachers, and parents concerning achievement of students enrolled in *Standards*-based classes in schools for which the $(MASP)^2$ LSC provided professional staff development for the teachers and staff. Our partner school districts wanted testing instruments with national norms. After reviewing several instruments from national publishers and consulting with our partner districts, we narrowed the list to two; the Stanford Achievement Test, Ninth Edition (SAT-9) developed by Harcourt Brace Educational Measurement and the New Standards Reference Examination in Mathematics developed by the Learning Research and Development Center of the University of Pittsburgh and the National Center on Education and the Economy. Both sets of tests were distributed and scored by Harcourt Brace.

The mathematics portion of the SAT-9 has three subtests. The Problem Solving subtest contains 30 multiple-choice problems which require students to solve problems set within

real world and mathematical contexts. The Procedures subtest has 20 multiple choice questions which require students to perform one of the four basic arithmetic operations with whole numbers, integers, and fractions. The Open Ended subtest is designed to "assess the concepts and skills of mathematics within the context of realistic and engaging problems" (Stanford 9 Technical Data Report). Each of the SAT-9 subtests test the content areas of number, measurement, geometry, algebra, functions, statistics and probability as deemed appropriate for each grade level. The two multiple choice subtests combined and the Open Ended subtest each take two 50-minute school periods to administer. Calculators are allowed on all subtests except for Procedures. Sample items for the SAT-9 Procedures, Problem Solving and Open Ended subtests are located in Appendix A.

The mathematics portion of the New Standards Reference Examination consists of three parts. The first consists of 20 multiple choice questions which are a subset of SAT-9 multiple-choice test items. In addition, this first section contains short tasks. These SAT-9 questions enable this portion of the test to be compared to national norms. The student has 20 minutes to complete the multiple-choice questions and 35 minutes for the short tasks. Students then spend 55 minutes on the second section which is made up of long and medium length tasks. The third section requires 55 minutes to complete and covers both short and long tasks. Short tasks are constructed response items while the medium and long tasks require detailed answers and are considered extended response items.

The New Standards Reference Examinations are criterion-referenced tests. These tests set levels of constructed response performance in three areas: Skills, Concepts and Problem Solving. The performance levels are derived from national Standards developed by a conglomerate of assessment-related organizations (Wiley and Resnick, 1998). The content and process areas assessed include number and operations, geometry and measurement, algebra and function, mathematics skills, problem solving and reasoning and mathematical communication. Sample items from the New Standards Reference Examination can be found in Appendix B.

The SAT-9 reports both scale scores and normal curve equivalents (NCEs). By definition, NCEs are "Normalized standard scores with a mean of 50 and a standard deviation of 21.06. The standard deviation of 21.06 was chosen so that NCEs of 1 and 99 are equivalent to percentiles of 1 and 99. There are approximately 11 NCEs to each stanine." (http://www.hemweb.com/library/glossary.htm#n). It is important to emphasize that NCEs are monotonically related to, but are not identical to, percentiles. Most of the data analysis results are reported in NCEs because of their familiarity and interpretability.

The New Standards Mathematics test reports student scores in a number of different ways, but we chose to examine student performance levels in the areas of skills, concepts, and problem solving. There are five student performance levels: achieved with honors, achieved the standard, nearly achieved, below standard and little achievement.

Student and Classroom Samples

Students in 43 *Standards*-based classrooms were tested. The teachers in these classrooms had participated in professional development provided through the $(MASP)^2$ LSC to varying degrees. Three types of professional development were provided. First, teachers participated in two weeks (80 hours) of summer training related to a particular *Standards*-based curriculum. This usually entailed working through activities of the curriculum while teaching strategies were modeled by experienced $(MASP)^2$ staff members. Second, during the school year teachers participated in sessions (30 hours) focused on more general topics such as facilitating cooperative learning in mathematics classrooms, current research on the brain and its implications for mathematics classroom instruction, meetings with teachers and administrators to discuss administrative issues and with their counselors relating to the scheduling of students. Lastly, $(MASP)^2$ employed district personnel experienced in the curriculum to serve during the school year as mentors to teachers newly implementing *Standards*-based curricula. The 20-hour mentoring component consisted primarily of classroom observations followed up with one-on-one debriefings and in some cases demonstration lessons. Middle grades teachers in this study had completed an average of 162 professional development hours over a three-year period.

The teachers whose students were tested were selected by administrators from each district. We requested that the students included should be representative of the entire spectrum of students enrolled in each school. Students who were part of the data set referred to earlier as "value added" came from classrooms that district personnel

identified as showing a high fidelity of implementation with the particular *Standards*-based curriculum being used.

Some schools had a difficult time getting teachers to agree to test their students on three different occasions and students did not always remain in the same group or the same school over the two-year period, and the sample showed some attrition. A greater source of attrition was the failure of students to sit for all three test administrations, with some tested twice while others were only tested once. The number of students tested three times as part of the "value added" data collection (i.e., provided three sets of scores) varied from a high of 92 in the district located in the urban-suburban boundary to a low of 20 in one of the suburban districts. The largest attrition occurred during the third testing period due to heavy demands on teachers and students for state and local district testing.

Data Analyses

We used five methods to analyze student achievement patterns. First, descriptive analyses were used to compare the achievement of *Standards*-based students against the national norms for the two instruments used. Descriptive statistics and graphs were used to detect patterns and estimate their magnitude. Second, we used hierarchical linear modeling (HLM) to model variation in mathematics proficiency using within-classroom factors such as student prior achievement. Between-classroom predictors such as classroom ethnic composition, SES, as well as predictors capturing school district membership were used. Where possible, the contribution of two-way interactions in accounting for variation in the outcomes was examined. Third, the fitted HLM models

were used to predict student mathematics performance and the predicted values were compared to the score known to reflect average performance to provide information about student mathematics performance controlling for various factors related to achievement. Fourth, a sub-sample of students within each district was tested three times to provide a value-added dimension to our analysis. These students were tested at the beginning of their 7th grade year, at the end of 7th grade, and at the end of their 8th grade, and served as their own baseline. We were, in this phase of the analysis, interested in examining achievement trends. Fifth, we examined patterns of changes in achievement gaps over the three test administrations.

Initial analyses focused on exploring patterns within and among districts and were followed by fitting regression models to try to account for variation in mathematics performance. Approximately 1,600 *Standards*-based students were treated as clustered within 43 classrooms (too few districts were tested to include this level of clustering). All students in this group had experienced three years of a *Standards*-based curriculum.

Since districts were concerned about how their students' scores compared to nationally normed groups, we used national NCE scores and compared each district's subtest mean NCE (Open Ended, Problem Solving, and Procedures) to the national mean NCE of 50. In the district that administered the New Standards Mathematics test we compared the percent of students at each of the five performance levels with the corresponding percent of the national sample at each performance level.

As other studies have noted (Riordan & Noyce, 2001; Reys, Reys, Lapan, Holliday, & Wasman, 2003, and Begle 1973), prior achievement in mathematics is an important predictor of student achievement, and, accordingly, we needed a measure of prior achievement for use in several statistical analyses. As is often the case, different districts administered different mathematics tests to students. At the middle school level the Northwest Achievement Level Test (NALT), Minnesota Comprehensive Assessment (MCA), Metropolitan Achievement Test (MAT7), and Terra Nova were used. In three of the districts a sub-sample of students had scores on two of these mathematics tests.

Because we wished to have a single (common) prior mathematics achievement score for each student, we began by examining the objectives, content, format, etc., of these tests and concluded that they assessed approximately the same construct of mathematics proficiency. Next, we empirically examined the effects of combining the various measures of prior mathematics knowledge. We first correlated the two sets of student's prior mathematics test scores that were available in three districts. The correlation between the MCA and NALT was .79 (N = 172), between MAT7 and MCA was .50 (108), and between MCA and NALT (302) was .69. These correlations provide support for the conclusion that these tests are assessing a common construct of mathematics proficiency. We also fitted multiple regression models to the SAT-9 student data within each district using the available prior mathematics measure as a predictor, along with other student-level predictors like gender, attendance, SES, and native versus nonnative English speaker. The results of these analyses produced similar percentages of variance

explained attributable to prior mathematics achievement with the effects of the other predictors held constant.

The above logical and empirical analyses led us to treat the different prior achievement measures as commensurable. We then created a combined, across-district prior achievement measure by treating the NCEs associated with these varied measures as equivalent. For example, students with an NCE of 70 on any of these tests were assumed to possess approximately the same mathematics proficiency. A plausible criticism of this assumption is that it suggests more precision than is justified. That is, students with the same NCE score from different mathematics tests probably have similar prior knowledge, but perhaps less than that implied by having the same NCE score.

To examine the effect of using the NCE metric of 1-99 for the combined measure versus another representation of this metric, some of the statistical analyses reported below were also performed using a polytomized form of the NCEs. Normal Curve Equivalent scores were replaced by a value indicating student membership in a particular decile of NCE performance. For example, the NCE performance of students in the sixth decile exceeded approximately 60% of the remaining students but was lower than approximately 30%. The similarity of results in using the combined prior mathematics scores in their original NCE metric of 1-99, versus replacing these scores with a value reflecting a student's decile membership, suggests that our findings are not overly dependent on the metric of the combined prior mathematics achievement variable.

Student demographic data such as attendance, student eligibility for a free or reduced price lunch, English language learner status, special education status, and gender were also gathered from each district. Because of missing data most statistical analyses were based on fewer than 1600 students, but there was no evidence that omitted students and/or classrooms differed systematically on our variables from those providing complete data, and hence no clear evidence of bias. Still, we acknowledge the difficulty of identifying the presence of such effects, and it would be prudent to interpret our findings in light of this.

Results

Method 1: Descriptive Summaries of District Performance

As shown in Table 2, students across all five districts performed above the national norm on the Problem Solving subtest. Only the large urban district had a mean below 50 on the Open Ended subtest. On the Procedures subtest, four of the five districts scored below the national mean. Recall that this subtest of the SAT-9 covers the four basic operations with whole numbers, integers, and fractions within purely symbolic settings and sometimes with one-step word problems solved without calculators.

Table 2: SAT-9 Open Ended, Problem Solving, and Procedures Sample Sizes, Means and Standard Deviations by District

	Subtest								
	Open Ended			Problem Solving			Procedures		
District	N	Mean	*SD*	N	Mean	*SD*	N	Mean	*SD*
A	579	58.7	18.3	584	62.9	19.3	565	37.1	16.2
B	385	47.2	24.8	399	52.6	23.1	386	36.7	20.3
C	113	57.3	17.7	120	60.3	15.8	120	49.9	18.0
D	161	63.4	16.1	162	63.3	20.0	158	40.1	15.4
E	128	76.5	16.8	123	84.9	14.8	123	59.2	17.8

The results for District E, the district that also administered the Mathematics portion of the New Standards Reference examination, are in Table 3. Recall also that the mathematical skills subtest of the New Standards is a sub-sample of the SAT-9 test. The results below show that 86% (57+29) of the students tested in District E achieved or exceeded the mathematical skills standard, while 33% (11+22) of the students at the national level performed at this same level. The above average performance of District E students who used CMP, extends also to the Mathematical Concepts and Mathematical Problem Solving subtests. Within Mathematical Concepts 71% of students achieved or exceeded the standard as compared to 20% nationally. On Mathematical Problem Solving 44% achieved or exceeded the standard while only 11% did so nationally.

Although these results come from an advantaged suburban district it should be kept in mind that the norming group for the New Standards instrument comes from the Northeastern part of the United States. This is an area of the United States that typically has higher scores than other geographical areas of the United States. (http://nces.ed.gov/nationsreportcard/mathematics/results2003/stateachieve-g8.asp)

Table 3: *Percentage and Number of Students Meeting New Standards Achievement Levels - District E*

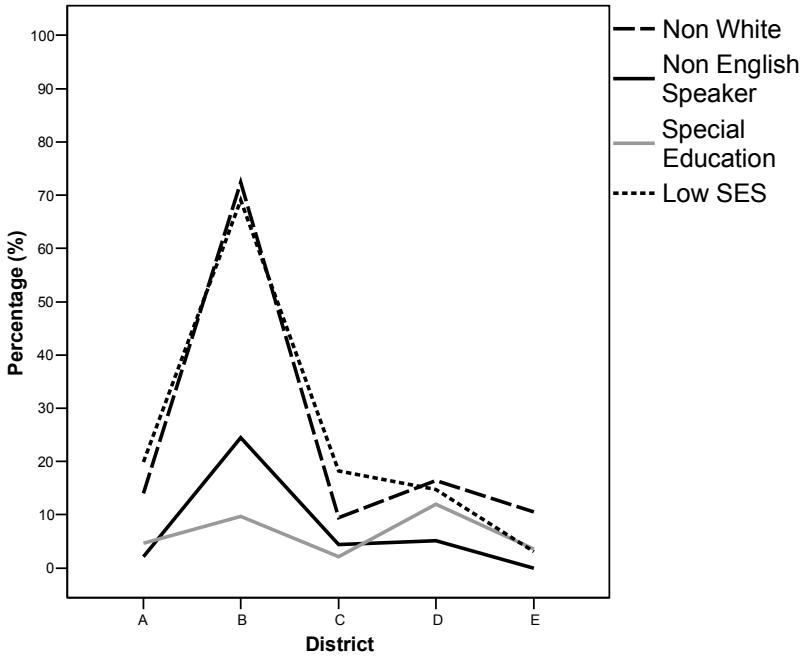
	Mathematical Skills		Mathematical Concepts		Mathematical Problem Solving	
Achievement Levels	Percent of Students (N)	National Norm %	Percent of Students (N)	National Norm %	Percent of Students (N)	National Norm %
Achieved the Standard with Honors	57% (137)	11%	29% (69)	6%	2% (5)	0%
Achieved the Standard	29% (70)	22%	42% (101)	14%	42% (101)	11%
Nearly Achieved the Standard	9% (21)	24%	19% (45)	17%	17% (41)	14%
Below the Standard	4% (10)	24%	6% (15)	26%	30% (73)	27%
Little Evidence of Achievement	1% (2)	19%	4% (10)	37%	8% (20)	48%

There were sharp differences among some of the districts in student characteristics. Figure 1 shows that District B had just under 70% non-white students and the remaining districts approximately 20% or less. Similarly, one district had approximately 21% of its middle school students classified as nonnative English speakers, while the remaining districts had values ranging from 0% - 6%. The percentage of Special Education students varied from 2% - 12%, with the highest value attached to a suburban district.

Socio-economic status showed comparatively more variability across all districts. One district had more than 60% of its middle school students eligible for a free or reduced

price lunch (low SES on Figure 1), two districts had 18% - 20% eligible, another about 15%, and in one district 3% of the middle school students were eligible.

Figure 1: District Demographic Data



Average performances for the SAT-9 mathematics subtests and prior mathematics achievement are displayed by district in Figure 2 and show considerable variability. The outcome showing the greatest variability was Problem Solving, with 27 NCE points separating the highest and lowest performing school districts. The Open Ended subtest produced almost as much variability (25), followed by Procedures (15) and Prior Mathematics Achievement (18). Collectively, this variation suggests that there are large differences in mathematics proficiency across the districts and somewhat smaller differences in prior mathematics knowledge. However, in the analysis to follow these differences shrink when various demographic variables are statistically partialled out.

Figure 2: District Achievement Data

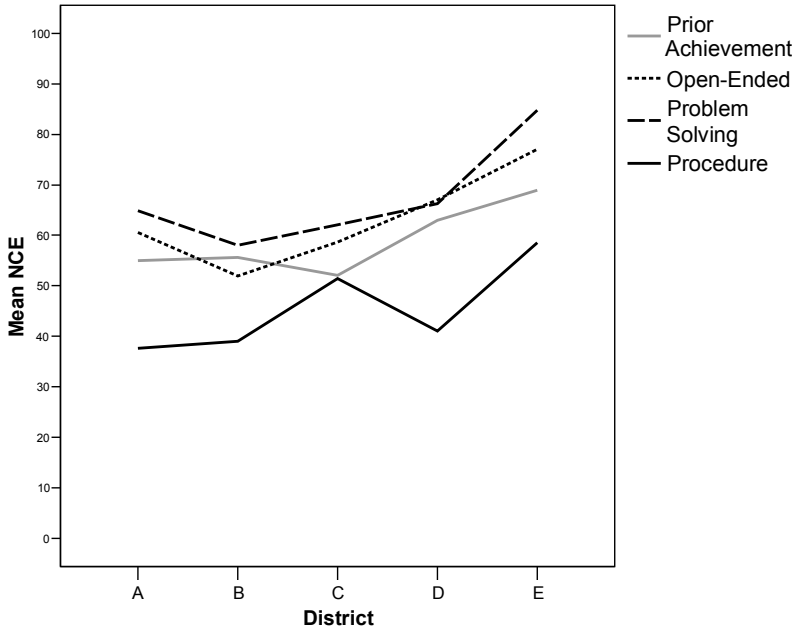
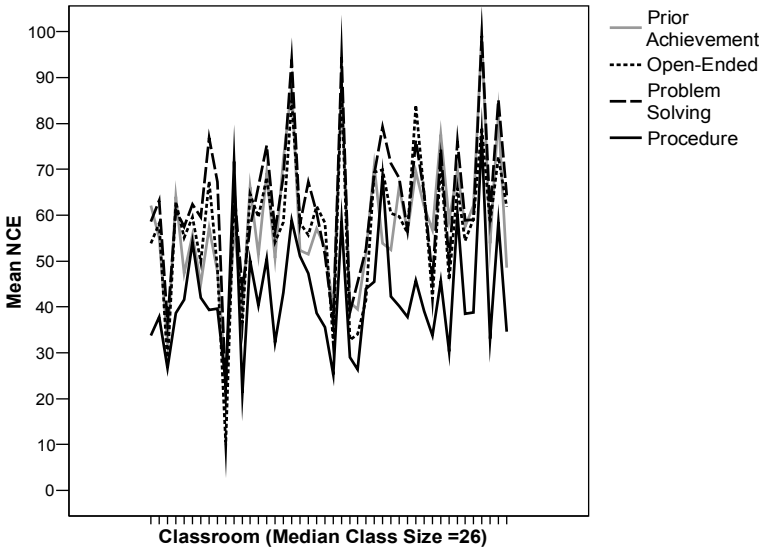


Figure 3: Mean Classroom Scores



There was also evidence of substantial variability in average mathematics performance at the classroom level. Figure 3 shows the classroom means for the SAT-9 subtests. Similar classroom performance on these subtests would produce approximately a straight

line, and it is apparent from Figure 3 that there is substantial variation in classroom performance on the SAT-9 subtests.

There was also variation in average mathematics performance across SES and English language (native vs. nonnative speaker). Overall, students not eligible for a free or reduced price lunch (high SES) scored on average 17, 17, and 7 NCE points higher than those eligible (low SES) on the Open Ended, Problem Solving, and Procedures subtests, respectively. Similarly, native English speakers scored 26, 23, and 11 NCE points higher than nonnative speakers on these subtests. Based on the descriptive statistics, English speaker status had a greater impact on mathematics performance than SES.

The role of subgroups generated by these variables (e.g., high SES/native English speaker) for each ethnic group is displayed in Figures 4-7 and shows that their combination has differential effects on mathematics performance. In general, native speakers outperform nonnative speakers in the same SES group. With the exception of Asian American students in the Procedures subtest, subtest scores largely mimic prior achievement scores in all subgroups. Low SES, nonnative white students performed at the lowest level (note small N).

Figure 4: Achievement for African American Students

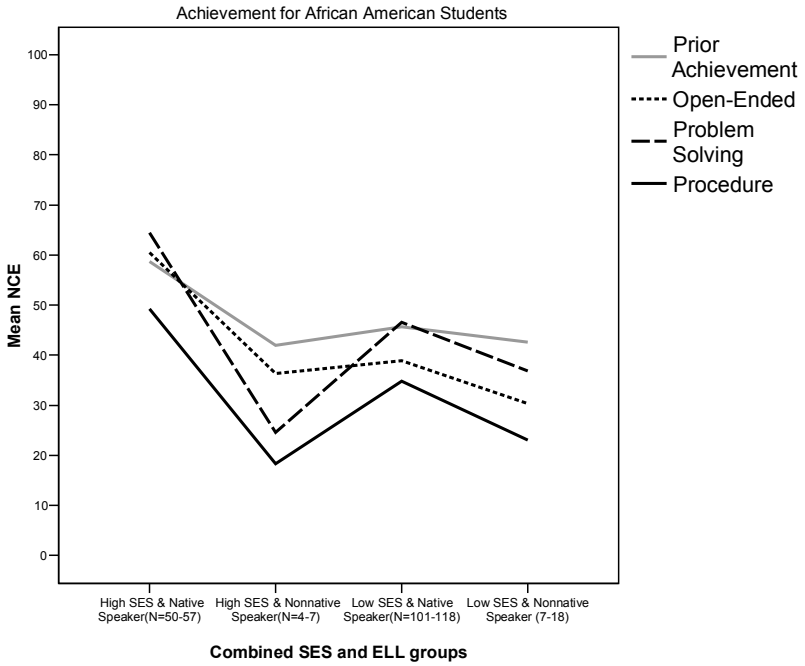


Figure 5: Achievement for Asian American Students

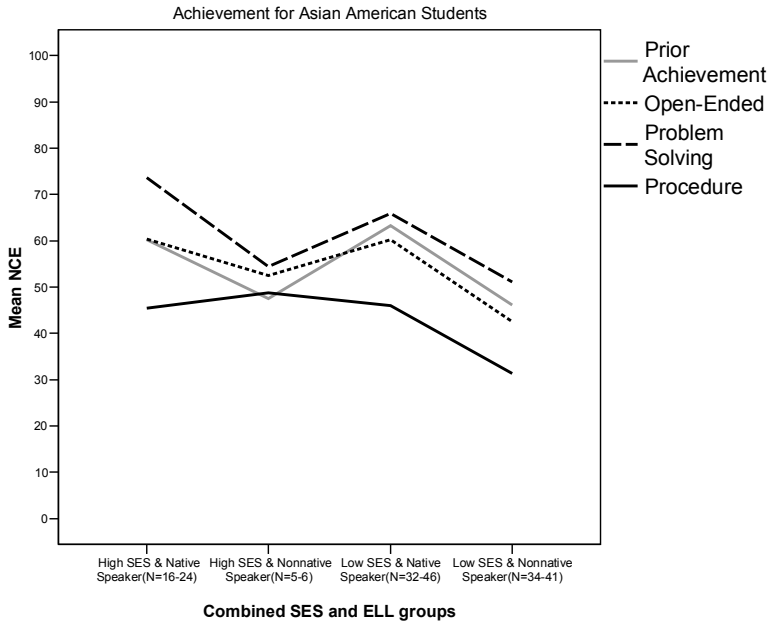


Figure 6: Achievement for Hispanic Students

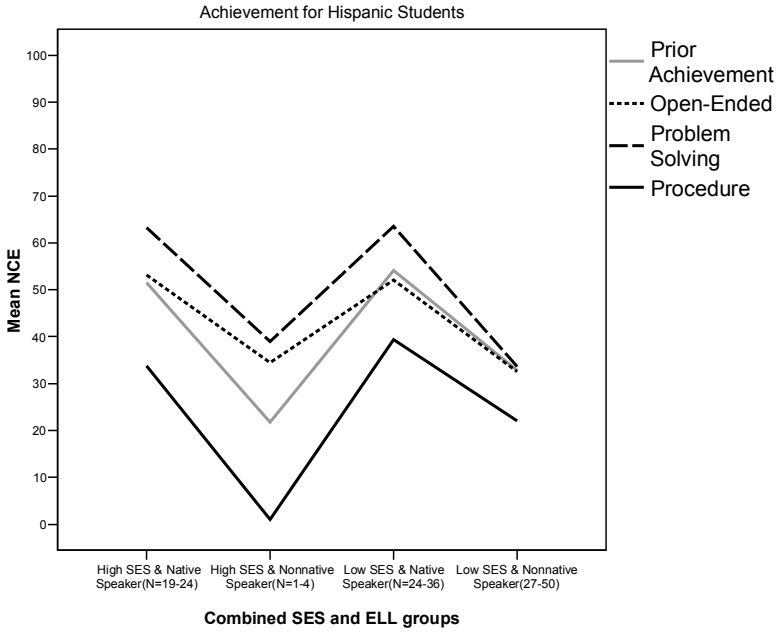
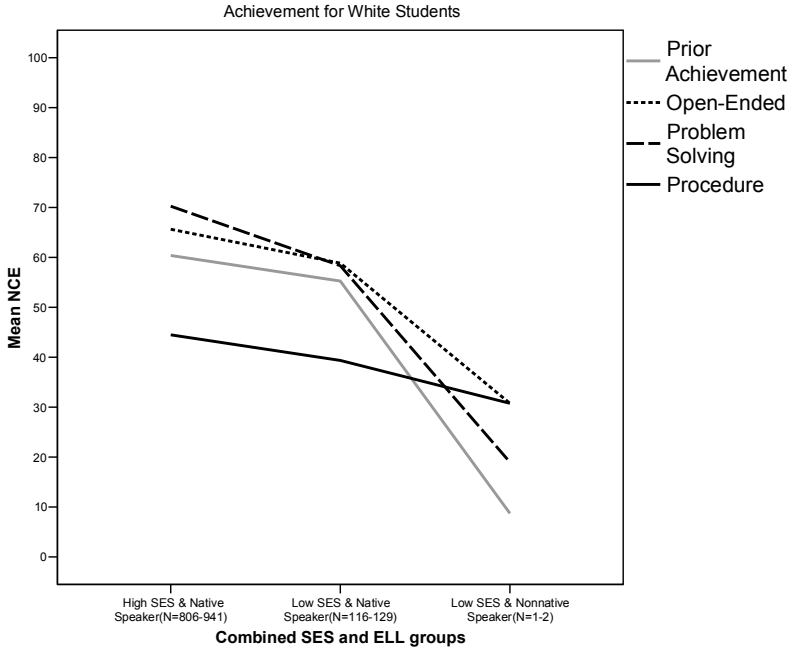


Figure 7: Achievement for White Students



Additional descriptive statistics for middle school students including effect sizes appear in Table 4. The effect sizes help to quantify differences apparent in Figures 4-7. Following Hedges and Olkin (1985, p. 78), effect sizes were computed as the difference in two means divided by the estimated pooled standard deviation of the difference. For example, low SES students scored on average .88 standard deviations lower on the Open Ended subtest than high SES students.

Several patterns are apparent among the effect sizes. One is the lower average performance for low SES, non-white, urban, nonnative English speakers, and special education students. These effect sizes ranged from -.12 to -1.24 standard deviation units. The only demographic variables not showing a statistically significant effect were gender (not reported) and Asians vs. whites. With one exception, each of the remaining 23 effect sizes was statistically significant. Another pattern apparent in Table 4 is that Problem Solving subtest NCE means were higher than those for Open Ended subtest across all subgroups. The Procedures subtest produced the lowest NCEs for all subgroups.

Table 4: Descriptive Data and Effect Sizes for Various Subgroups

Group		Open Ended				Problem Solving				Procedures				Prior Achievement			
		N	Mean	SD	Effect size	N	Mean	SD	Effect size	N	Mean	SD	Effect size	N	Mean	SD	Effect size
SES	**High**	936	62.9	18.5		938	67.1	20.0		919	43.0	18.8		1044	59.8	17.9	
	Low	419	45.3	23.0	-0.81*	439	50.1	20.9	-0.84*	422	35.1	18.5	-0.42*	342	48.5	19.2	-0.62*
Ethnicity	**White**	962	63.3	18.2		967	66.4	19.9		948	42.5	18.3		1068	59.6	17.9	
	African American	186	42.5	24.4	-1.06*	191	49.3	22.0	-0.84*	184	35.3	20.0	-0.38*	169	47.6	20.4	-0.66*
	Asian American	115	50.6	18.5	-0.65*	117	59.2	17.8	-0.36*	112	40.2	18.4	-0.12	87	53.9	17.1	-0.31*
	Hispanic	103	39.0	23.0	-1.24*	113	45.2	23.1	-1.04*	108	31.8	19.7	-0.57*	71	44.5	17.2	-0.83*
location	**Suburban**	981	61.6	18.7		989	65.4	20.0		966	42.0	18.3		1033	58.3	17.7	
	Urban	385	47.2	24.8	-0.28*	399	52.6	23.1	-0.23*	386	36.7	20.3	-0.62*	362	53.5	21.3	-0.58*
Language status	**Native English**	1249	59.8	20.4		1261	63.8	20.7		1230	41.6	18.7		1315	58.1	18.6	
	Non Native English	117	33.7	20.1	-0.91*	127	40.6	19.5	-0.90*	122	29.1	18.3	-0.68*	79	39.9	13.8	-1.03*
Special Education	**Non Special ED**	1288	58.5	21.1		1306	62.9	21.3		1273	41.2	19.0		1310	58.1	18.2	
	Special ED	78	41.3	23.8	-0.73*	82	42.7	19.2	-0.95*	79	28.8	14.7	-0.83*	84	39.8	20.4	-1.00*

*$p<0.05$

In sum, there is ample evidence of variability among the school districts in student demographic characteristics and in average mathematics performance. The suburban districts, which included the district on the urban-suburban boundary, tended to have far smaller percentages of non-white, low SES, and nonnative English speakers. The distribution of special education students showed little relationship with the suburban or urban location of districts.

Method 2: HLM Analyses of Student and Classroom Data

The SAT-9 mathematics subtests data were analyzed with HLM following the methods described in Raudenbush and Bryk (2002). Treating students as clustered within classrooms permitted within-classroom dependency among student mathematics test scores to be modeled, and allowed both student- and classroom-level questions to be answered simultaneously. This in turn helped to ensure more credible statistical test results than would ordinarily be possible with traditional regression modeling.

Student-level regression models containing prior mathematics achievement, attendance, SES, and gender were fitted to each middle school classroom's data. Because of missing data the total number of students was reduced to approximately 1050 – 1200, depending on the outcome. For each outcome (Open Ended, Problem Solving, Procedures subtest scores) three models were fitted. First an unconditional model of the form

$$Y_{ij} = \beta_{0j} + r_{ij} \qquad (1)$$

$$\beta_{0j} = \gamma_{00} + u_{0j},$$

was fitted, where Y_{ij} is the mathematics score of the ith student in the jth classroom, β_{0j} is the average mathematics score (intercept) for the jth classroom, γ_{00} is the average mathematics performance across classrooms, r_{ij} is a student-level residual, and u_{0j} represents the unique effect of the jth classroom. The unconditional model results tell us whether average outcomes differ across classrooms. Next we fitted a student-level model of the form

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\text{attendance}_{1ij} - \overline{X}_{1j}) + \beta_{2j}(\text{SES}_{2ij} - \overline{X}_{2j}) + \beta_{3j}(\text{gender}_{3ij} - \overline{X}_{3j}) + \beta_{4j}(\text{prior}_{4ij} - \overline{X}_{4j}) + r_{ij}, \quad (2)$$

where β_{1j} is the student level slope capturing the effect of attendance on mathematics (with other predictors held constant), $\overline{X}_{1j}$ is the mean attendance in the jth classroom, and prior_{4ij} is the prior mathematics knowledge predictor. We also tested whether slopes for the predictors varied across classrooms.

Classroom-level predictive models for intercepts (average mathematics achievement), and, where appropriate, slopes, were then developed. That is, for instances when the effect of a student-level predictor like SES on a SAT-9 subtest varied across classrooms, we constructed a predictive model to try to explain variation in these slopes with classroom-level predictors. Key middle school classroom predictors included Class SES (percentage of students eligible for a free or reduced price lunch in a classroom) and average prior mathematics knowledge in a classroom. Other classroom predictors which we examined were the effect of different concentrations of various ethnic groups, nonnative English speakers, special education students, and female students in a

classroom. Average classroom attendance and predictors capturing school district membership were also used. Preliminary analyses showed that average classroom attendance and the percentage of female students in a classroom could be removed because they did not contribute anything to explaining variation in classroom mathematics means (intercepts) or slopes. These analyses also indicated that differences across the five districts could be captured by a single predictor indicating whether or not the classroom was in the urban district.

The classroom model for intercepts fitted in most analyses was

$$\begin{aligned}\beta_{0j} = {} & \gamma_{00} + \gamma_{01}(\text{Class SES}_{2j} - \overline{W}_1) + \gamma_{02}(\text{Class African American}_{2j} - \overline{W}_2) + \\ & \gamma_{03}(\text{Class Asian}_{3j} - \overline{W}_3) + \gamma_{04}(\text{Class Hispanic}_{5j} - \overline{W}_4) + \gamma_{05}(\text{Class nonnative} \\ & \text{English speakers}_{5j} - \overline{W}_5) + \gamma_{06}(\text{Class special educ}_{6j} - \overline{W}_6) + \gamma_{07}(\text{district}_{7j} - \overline{W}_7) + \\ & \gamma_{08}(\text{prof devel}_{8j} - \overline{W}_8) + \gamma_{09}(\text{prior}_{9j} - \overline{W}_9) + u_{0j} \qquad (3)\end{aligned}$$

where γ_{01} is the classroom level slope capturing the effect of class SES (percentage of low SES students) on average mathematics performance, $\overline{W}_1$ is class SES averaged across classrooms, Class African American is the percentage of African American students in a classroom, and so on. The percentage of White students in a classroom was not used as a predictor because doing so would have introduced a dependency among the ethnicity predictors.

In a few cases, student-level slopes varied randomly across classrooms, and models similar to those for intercepts tried to account for this variation. The deviance test

described in Raudenbush and Bryk (2002, pp. 59-61) was used to test for model fit, allowing us to discriminate among models with more or less explanatory power. Model-fitting was followed by extensive model-checking to help to ensure validity of inferences. Cases in which normality, homoscedasticity, or linearity appeared to be suspect were examined in detail, and various remedies (e.g., modeling unequal classroom variances) employed. The analyses reported below are based on fitted models in which these assumptions appeared to be at least approximately satisfied.

An initial difficulty with several of the classroom-level predictor variables, such as the percentage of nonnative English speakers in a classroom, was their ragged and discontinuous nature. For example, about 40% of the classrooms had less than 3% nonnative English speakers, another 25% of the classrooms had values between 5% - 7%, and, at the other end of the distribution, ten of the classrooms had values ranging between 14% - 96%. We explored various transformations of these variables with the goal of representing their variation in a more succinct form, and polytomized the distributions into quartiles as follows:

Table 5: Definitions of Classroom Level Predictors

Variable	Quartile	N	Range (R)
Class SES	1	10	$R \leq 15.38\%$
	2	12	$15.38\% < R \leq 23.53\%$
	3	11	$23.53\% < R \leq 69.57\%$
	4	10	$R > 69.57\%$
Class English Language Status	1	15	$R = 0\%$
	2	7	$0\% < R \leq 5.36\%$
	3	11	$5.36\% < R \leq 14.29\%$
	4	10	$R > 14.29\%$
Class Spec Ed	1	15	$R = 0\%$
	2	7	$0\% < R \leq 4.17\%$
	3	11	$4.17\% < R \leq 10.26\%$
	4	10	$R > 10.26\%$
Class African American	1	19	$R \leq 5.27\%$
	2	9	$5.27\% < R \leq 10.71\%$
	3	11	$10.71\% < R \leq 33.3\%$
	4	11	$R > 33.3\%$
Class Asian	1	15	$R = 0\%$
	2	7	$0\% < R \leq 3.85\%$
	3	11	$3.85\% < R \leq 8.7\%$
	4	10	$R > 8.7\%$
Class Hispanic	1	17	$R = 0\%$
	2	3	$0 < R \leq 3.33\%$
	3	12	$3.33\% < R \leq 9.52\%$
	4	11	$R > 9.52\%$
Professional Development Hours	1	6	$R = 65$
	2	8	$65 < R \leq 130$
	3	15	$130 < R \leq 158$
	4	14	$R > 158$

Thus, each of the above classroom predictors was transformed to a scale in which each was represented with four values corresponding to the above quartiles.

The HLM cross-sectional results are summarized in Table 6. All statistical tests used a Type I error rate of $\alpha = .05$. Several general findings emerged across the SAT-9 subtests. First, there was substantial between-classroom variation in the Open Ended, Problem Solving, and Procedures subtest scores with classroom means of 34%, 38%, and 34%, respectively.

Second, at the student level (level 1), prior mathematics knowledge was a statistically significant predictor in every model, although its effect expressed in NCE units tended to be modest (< 1). Student-level SES was statistically significant in models for the Open Ended and Problem Solving subtests and demonstrated a moderate effect on mathematics performance. Gender was never a statistically significant student-level predictor, and student attendance was only occasionally a significant (and weak) predictor of mathematics performance.

Third, results for the Procedures subtest were somewhat different from those for the others, in that there were fewer significant effects.

Fourth, there was evidence of differences in average classroom performance (level 2) between the large urban district and the remaining districts even when demographic

variables (e.g., SES and prior mathematics achievement) were held constant for the Open Ended and Problem Solving subtest analyses.

Fifth, there was no evidence of contextual effects for SES or prior mathematics knowledge (Raudenbush & Bryk, 2002, pp. 149-141), meaning, for example, that the effect of SES on mathematics achievement was the same at the student-level and classroom-level.

Other results were specific to particular subtests. For the Open Ended subtest, student-level SES was a significant predictor with an average slope of -3.36, meaning that, with the other predictors held constant, students eligible for a free or reduced price lunch tended to score slightly more than 3 NCE points below those not eligible. There was a strong district effect in classroom means of -7.8, meaning that urban classrooms tended to score on average about 8 NCE points lower than suburban classrooms. Other classroom effects were for SES (-3.14), prior mathematics knowledge (.72), special education (-1.57), and the concentration of Asian students (1.28). The latter finding means that, with other predictors held constant, increasing the concentration of Asian students by one quartile in a classroom was associated with average increases in Open Ended means of approximately 1.28 NCE points.

For Problem Solving, SES again had a pronounced effect at the student level (-5.1), while at the classroom level there was an even larger district effect favoring suburban classrooms (-11.7). Another statistically significant classroom effect was for the

concentration of Asian students in a classroom (2.58), meaning that a one quartile increase in this variable was associated with an increase in classroom Problem Solving means of 2.58 NCE points.

Three small cross-level interaction effects for Problem Solving also emerged. The slope capturing the effect of special education on the prior knowledge slopes was .09, meaning that increases in the concentration of special education students in a classroom tended to be associated with a (slightly) greater impact of prior mathematics knowledge on Problem Solving scores. Second, increasing concentrations of students eligible for a free or reduced price lunch in a class had a significant effect on attendance slopes (.79), meaning that increases in this variable tended to be associated with classrooms in which lower attendance was associated with lower Problem Solving scores. Third, increasing concentrations of nonnative English speakers in a classroom was also a significant predictor of attendance slopes (-.38), meaning that increasing numbers of nonnative speakers in a classroom tended to exacerbate the effect of attendance on Problem Solving scores.

For Procedures, only prior mathematics knowledge was a significant predictor at the student level (.46), while at the classroom level special education students (-3.9) and prior mathematics knowledge (.53) were significant predictors of classroom means. There were no between-district effects for Procedures.

In sum, the HLM cross-sectional results for the Open Ended and Problem Solving subtests were quite similar. Student's prior mathematics knowledge was a consistent predictor of mathematics proficiency, although its effect was modest, but student SES was a stronger predictor. Strong differences between the urban and suburban classrooms also emerged for the Open Ended and Problem Solving subtests (-7.8 and -11.7 points, respectively), along with other, smaller, classroom effects such as the concentration of Asian students and students eligible for a free or reduced price lunch. For the Procedures subtest, only prior mathematics knowledge and the concentration of special education students in a classroom were significant predictors.

Table 6: Results of HLM

Dependent variable	HLM Results
Open Ended	1. There was significant between-classroom variation in means (34%). 2. There were significant within-classroom student effects for SES (-3.38), prior mathematics knowledge (.72), and attendance (.27). Classroom Open Ended means and prior knowledge slopes varied significantly across classrooms. 3. There was a significant difference between urban and suburban districts (-7.8) favoring suburban districts. This effect accounted for approximately 30% of the variance in classroom Open Ended means. 4. Concentrations of low SES students, expressed in quartiles, was a significant predictor of classroom Open Ended means (-3.14), meaning that shifting from the first quartile ($\leq$ 15.38% eligible for a free or reduced price lunch) to the second quartile ($15.38\% < R \leq 23.53\%$) produces an approximate decline of 3 points in classroom Open Ended means. Classroom prior mathematics knowledge (.72) was also a significant predictor of classroom means, along with special education (-1.57) and Asian students (1.28). 5. Significant variation in classroom open ended means remained unexplained.
Problem Solving	1. There was significant between-classroom variation in means (38%). 2. There were significant within-classroom effects for SES (-5.1) and prior mathematics knowledge (.63). Intercepts and slopes for all student level predictors varied across classrooms. 3. There was a strong district effect (-11.4) for classroom means, accounting for approximately 19% of the variance in classroom means. 4. Concentrations of Asian students in a class (2.58), expressed through quartiles, was also significant. Classroom prior mathematics knowledge (.81) was a significant predictor of classroom Problem Solving means. 5. The slope capturing the effect of concentrations of special education students on prior mathematics knowledge slopes was .09, meaning that increases in the percentages of these students (expressed through quartiles) in a classroom was associated with a slightly greater impact of prior mathematics knowledge on Problem Solving scores. Second, concentrations of low SES students had a significant effect on attendance slopes (.79), meaning that increases in this variable tended to be associated with classrooms in which the effect of attendance on Problem Solving scores was weaker. Increasing concentrations of nonnative speakers in a classroom was also a significant predictor of attendance slopes (-.38), meaning that increases in nonnative speakers exacerbated the effect of attendance on Problem Solving scores. 6. Significant variation in classroom Problem Solving and the gender, SES, and prior mathematics knowledge slopes remained unexplained.
Procedures	1. There was significant between-classroom variation in means (34%). 2. The only significant within-classroom predictor was prior mathematics knowledge (.46). Procedures means showed significant variation within classrooms. 3. The largest classroom effect for classroom means was the concentration of special education students (-3.9), which accounted for approximately 44% of the variability in classroom Procedures means. Classroom prior mathematics knowledge (.53) was also a significant predictor. Classroom means continued to show significant variability. There were no significant between-district effects.

Method 3: HLM-Based Predicted NCE Scores

Another way to examine the impact of participating in a *Standards*-based curriculum for three years is to use the fitted HLM models for the one-shot data to predict SAT-9 scores for each of the classrooms in our sample. Specifically, we examined patterns in the model-predicted (empirical Bayes estimated) classroom NCE means produced by the HLM5 software (Raudenbush, et al, 2000). Comparing the predicted scores against the NCE mean of 50 (average performance) provides information about the expected performance of the classrooms, taking into account other factors such as prior mathematics knowledge and SES. A summary of this information for the Open Ended, Problem Solving, and Procedures subtests appears in Table 7 for key classroom variables such as whether the classroom was urban or suburban, and concentrations of various subgroups. For display purposes in Table 7 prior mathematics knowledge was polytomized into quartiles. To try to avoid over-interpreting such results we report the average model-predicted classroom mean values simply as above or below the NCE average of 50. We emphasize that these results have not been cross-validated.

Table 7: *HLM Predicted Scores*

HLM-Based Open Ended Predicted NCE Scores ***

		Class Prior achievement		Class SES		Class ELL		Class Special Ed.	
District		1*	4**	1	4	1	4	1	4
Urban		↑	↓		↑	↑	↓	↑	↓
Suburban		↑	↑	↑	↑	↑	↑	↑	↑

HLM-Based Problem Solving Predicted NCE Scores

		Class Prior achievement		Class SES		Class ELL		Class Special Ed.	
District		1*	4**	1	4	1	4	1	4
Urban		↑	↓		↓	↑	↓	↑	↓
Suburban		↑	↑	↑	↑	↑	↑	↑	↑

HLM-Based Procedures Predicted NCE Scores

		Class Prior achievement		Class SES		Class ELL		Class Special Ed.	
District		1*	4**	1	4	1	4	1	4
Urban		↓	↓		↓	↓	↓	↓	↓
Suburban		↑	↓	↓	↓	↓	↓	↑	↓

*1st Quartile
**4th Quartile
↑ model-predicted NCE score above 50
↓ model-predicted NCE score below 50
*** See Table 5 for a description of the quartiles for the classroom predictors.
ELL = English language status.

Overall, 57% of the classrooms in the urban district were predicted to score above 50 on the Open Ended subtest. The average predicted Open Ended score for urban classrooms classified in the 1st (highest prior mathematics knowledge) quartile was above 50, whereas the average of those in the 4th (lowest prior mathematics knowledge) was below 50. Similarly, the average predicted Open Ended score of urban classrooms with the

lowest concentration of non-native English speakers was above 50 and those with the highest concentration were below 50. All 22 suburban classrooms (100%) were predicted to score above 50 on this subtest. Because none of the urban classrooms appeared in the first quartile of Class SES (lowest concentration of students eligible for a free or reduced price lunch), average model-predicted values are not reported.

For Problem Solving, 71% of the urban classrooms and 100% of the suburban classrooms were predicted to score above 50.. The average predicted score for urban classrooms with the highest concentrations of non-native speakers, students eligible for a free or reduced price lunch, and special education students, or the lowest prior mathematics knowledge, was below 50.

For Procedures, only 9% of the urban classrooms were predicted to score above 50, and only 27% of the suburban classrooms. The average predicted scores of urban classrooms with the highest or lowest prior mathematics achievement and various concentrations of students eligible for a free or reduced price lunch, non-native speakers, or special education students were all below 50. Only those suburban classrooms with the highest prior mathematics knowledge and the lowest concentrations of special education students had an average predicted score above 50.

In sum, the general pattern in Table 7 for the Open Ended subtest showed that the average predicted score of most classrooms was above 50, whereas for Problem Solving sharper differences favoring suburban classrooms emerged. Procedures showed a different pattern, with most average predicted classroom scores below 50 regardless of whether a classroom was urban or suburban. This result is not entirely surprising because

Procedures was the only SAT-9 subtest in which every district except one scored on average below 50 (see Table 2).

Method 4: Student Achievement of Standards-based Value-Added Students (Repeat Testers)

Sub-samples of students sat for the SAT-9 test up to three times over a two academic year period. Analyzing these data provides a way to assess and interpret change over the indicated time span. Scaled scores were used in the longitudinal analyses because they more adequately capture change. However, scaled scores are calibrated such that they increase from one year to the next if a student has made expected progress. For the SAT-9 subtests, information supplied by the test publisher that takes into account the ages tested and the testing period indicates that a student scoring at the 50th percentile on the Open Ended subtest would have scores of 638, 646, and 654, corresponding to the Fall, 2000, Spring 2001, and Spring 2002 testing, respectively. For Problem Solving these values are 653, 663, and 670, and for Procedures the values are 671, 685, and 695. Looking at the resulting data, comparisons can be made of the average growth of student scores over the three test periods vis-à-vis the expected growth as determined by the test publisher. Table 8 highlights the performance of the repeat testers over time for each district.

Table 8. Average SAT-9 Scaled Scores of the Repeat Testers

Subtest	OE1 Oct, 2000	OE2 Apr, 2001	OE3 Apr, 2002	OE Growth 1st-3rd	PS1 Oct, 2000	PS2 Apr, 2001	PS3 Apr, 2002	PS Growth 1st-3rd	PR1 Oct, 2000	PR2 Apr, 2001	PR3 Apr, 2002	PR Growth 1st-3rd
Publishers Number	638	646	654	+16	653	663	670	+17	671	685	695	+24
Dist A (N=23-25)	636 (-2)	638 (-8)	670 (+16)	+34	678 (+25)	678 (+15)	670 (0)	-8	674 (+3)	670 (-15)	674 (-21)	0
Dist B (125-144)	635 (-3)	631 (-15)	655 (+1)	+20	659 (+6)	673 (+10)	684 (+14)	+25	659 (-12)	678 (-7)	671 (-24)	+12
Dist C (32-49)	637 (-1)	646 (0)	658 (+4)	+21	667 (+14)	678 (+15)	687 (+17)	+20	648 (-23)	705 (+20)	679 (-16)	+31
Dist E (47-52)	716 (+78)	670 (+24)	715 (+61)	-1	747 (+94)	772 (+94)	771 (+101)	+24	NV	NV	NV	NV

All available data for repeat testers were used to compute the values in this table. Values in parentheses represent the range of sample sizes across the three testings. District D had too few repeat tester students to warrant inclusion in this Table.

OE1 = Open Ended Testing in October, 2000
OE2 = Open Ended Testing in April, 2001
OE3 = Open Ended Testing in April 2002
PS = Problem Solving
PR = Procedures
NV = No data are reported for District E for Procedures because students were inadvertently allowed to use calculators during the October, 2000 testing and their scores are not valid.
( ) = Deviation from publisher's 50th percentile scores
N = number of students

Open Ended: The 638 reported in Table 8 for the publisher's number for the first Open Ended testing (October, 2000) is the scaled score reflecting 50th percentile performance for this subtest. Over time, maintaining 50th percentile performance for this subscale requires students to score at least 646 in April, 2001 and 654 in April, 2002. Students in district A scored on average two points below this value in October, 2000, 8 points below in April, 2001, but by April, 2002 scored on average 16 points above the value indicating adequate progress. Table 8 also reports the average change in scaled scores from the first (where students started) to the last (where students ended approximately two years later)

testing. District A showed an average gain of 34 scaled score points over the two-year period, meaning that by April, 2002 District A students were on average making adequate progress (and then some) on the concepts measured by the Open Ended subtest. In fact, all four districts ended the eighth grade year with means above the scale score associated with the 50th percentile. The district that did not show positive mean growth (District E) likely experienced a ceiling effect since their mean score ranged from the 97th to the 99th percentile rank. It should be observed that 3 of the 4 districts, including the urban district, began the seventh grade year with means below the national expected scale score.

Problem Solving: All four districts ended the eighth grade year with means at or above the 50th percentile scale score. Three of the four districts exceeded the publisher's expected growth over the span.

Procedures: One district did not produce valid scores for this subtest since calculators were erroneously used on the first administration of this subtest. The results of the remaining three districts are mixed. Two of the three districts began below the expected scale score and all three ended below the expected scale score. One district exceeded expected growth and two fell short of expected growth on the Procedures subtest.

Nine out of the 11 average growth values in Table 8 are positive, one is 0, and two are negative. This means that in 8 of the 11 cases, students' actual growth exceeded publisher's expected growth over the two year time frame.

In addition to the descriptive statistics reported in Table 8, hierarchical linear modeling was used to assess change over time within each school district. The advantages of an HLM approach with repeated measures data are that the measurement occasions do not have to be equidistant in time and students need not provide data for every occasion to be used in the analysis (Raudenbush & Bryk, 2001, pp. 160-176). These models also have the advantage of estimating a growth trajectory (e.g., linear) for each student. For each SAT-9 subtest we fitted a within-student model designed to estimate the linear growth rate, and a between-student model with prior mathematics achievement as the sole predictor. In most cases the sample size used in the model-fitting was somewhat small (e.g., 25).

Two key findings emerged from the HLM for the repeat testers that provide additional guidance in interpreting the results in Table 8. First, there was substantial variability in the magnitude and direction of change within-students over time. For example, one district showed positive mean change overall between the Fall, 2000 and Spring, 2002 testings. However, approximately 45% of the repeat testers in this district showed a negative change over time. In some cases a student scored at or below average initially and then declined over time, but in other cases their initial scores were quite high and the decline quite modest. There was also evidence that these patterns appeared both within and between particular subgroups, for example, high and low SES students. These findings remind us that the evidence of adequate progress provided by the summary statistics in Table 8 does not apply to all students within a given district. There remains a significant number of low achieving students in these districts.

Second, the HLM results (through tests of model-data fit) provided evidence that additional models and predictors should be examined. For example, a model with prior mathematics achievement as a between-student predictor fitted the data better than a model without this predictor. However, there were typically too few students to construct potentially more powerful predictive models. The results in Table 8, which do not take other predictors into account, should be interpreted accordingly.

Method 5: Assessing the Achievement Gap

The Elementary and Secondary Education Act of 2001 has as its focus the elimination of achievement gaps for various subgroups of students such as those of low SES and various ethnic groups. Growth over time was not examined for districts D and E because of very small numbers of students. With some schools' waning interest in having their students tested a third time (an additional three hours of testing time for the SAT 9 added to testing requirements for the state assessment program), large numbers of our originally identified students had to be omitted from this analysis.

We also examined the growth over time of students with high and low prior achievement by dichotomizing this variable at its median. These data are summarized by district in Tables 9 and 10 for the Open Ended, Problem Solving, and Procedures subtests of the SAT-9.

Table 9: Average SAT-9 Scaled Scores Over Time for High and Low Prior Achievement by District

District A

Subtest (Scale Score)	Prior Achievement	N	Test Administration: Fall 2000 Mean	Fall 2000 *SD*	Fall 2000 Gap	Spring 2002 Mean	Spring 2002 *SD*	Spring 2002 Gap	Change
Open Ended	High	13	654.2	21.4	46.4*	677.2	10.6	26.4*	-20.0
	Low	4	607.8	9.0		650.8	14.5		
Problem Solving	High	12	697.2	27.1	52.4*	711.4	21.4	46.4*	-6.0
	Low	5	644.8	6.0		665.0	23.5		
Procedures	High	13	694.6	23.5	66.6*	691.5	25.7	45.3*	-21.3
	Low	6	628.0	37.4		646.2	21.9		

** p<0.05*

District B

Subtest (Scale Score)	Prior Achievement	N	Test Administration: Fall 2000 Mean	Fall 2000 *SD*	Fall 2000 Gap	Spring 2002 Mean	Spring 2002 *SD*	Spring 2002 Gap	Change
Open Ended	High	61	656.6	30.4	52.4*	673.67	24.4	43.7*	-8.7
	Low	32	604.3	28.5		630	31.9		
Problem Solving	High	63	687.3	34.6	66.5*	707.06	35.4	57.2*	-9.3
	Low	35	620.8	32.4		649.86	25.8		
Procedures	High	62	684.4	45.9	64.4*	687.11	42.2	47.5*	-17.0*
	Low	33	620.0	32.4		639.67	29.1		

** p<0.05*

District C

Subtest (Scale Score)	Prior Achievement	N	Test Administration: Fall 2000 Mean	Fall 2000 *SD*	Fall 2000 Gap	Spring 2002 Mean	Spring 2002 *SD*	Spring 2002 Gap	Change
Open Ended	High	13	646.2	15.6	22.6*	671.2	17.2	21.7*	-0.9
	Low	14	623.6	17.5		649.4	21.8		
Problem Solving	High	13	687.8	23.7	42.9*	699.7	19.7	30.0*	-12.9
	Low	15	644.9	21.8		669.7	29.5		
Procedures	High	13	660.4	38.3	24.6*	698.3	22.3	39.7*	15.1
	Low	15	635.8	38.6		658.6	27.8		

*** p<0.05*

Table 10: Average SAT-9 Scaled Scores Over Time for High and Low SES by District

District A

Subtest (Scale Score)	SES	N	Test Administration: Fall 2000 Mean	Fall 2000 *SD*	Fall 2000 Gap	Spring 2002 Mean	Spring 2002 *SD*	Spring 2002 Gap	Change
Open Ended	High	18	643.2	24.8	11.0	670.1	16.1	-1.2	-12.2
	Low	4	632.2	27.5		671.3	16.0		
Problem Solving	High	17	685.5	33.7	20.7	702.5	22.5	15.5	-5.3
	Low	5	664.8	19.6		687.0	42.3		
Procedures	High	18	686.9	26.9	48.9*	679.9	28.0	19.9	-28.9
	Low	6	638.0	47.9		660.0	31.7		

*$p<0.05$

District B

Subtest (Scale Score)	SES	N	Test Administration: Fall 2000 Mean	Fall 2000 *SD*	Fall 2000 Gap	Spring 2002 Mean	Spring 2002 *SD*	Spring 2002 Gap	Change
Open Ended	High	42	658.9	28.3	35.6	671.83	28.26	24.9	-10.7
	Low	59	623.4	38.0		646.95	34.83		
Problem Solving	High	42	684.3	46.8	33.7	705.40	41.28	32.9*	-0.8
	Low	65	650.6	40.69		672.51	38.59		
Procedures	High	42	677.9	50.28	24.8*	682.86	45.91	19.1	-5.7
	Low	62	653.0	51.88		663.76	45.40		

* $p<0.05$

District C

Subtest (Scale Score)	SES	N	Test Administration: Fall 2000 Mean	Fall 2000 *SD*	Fall 2000 Gap	Spring 2002 Mean	Spring 2002 *SD*	Spring 2002 Gap	Change
Open Ended	High	26	639.1	16.5	20.4*	659.5	20.9	10.0	-10.5
	Low	6	618.7	20.1		649.5	30.1		
Problem Solving	High	29	668.6	31.8	19.4	690.4	25.6	19.6	0.1
	Low	6	649.2	26.2		670.8	41.8		
Procedures	High	29	647.7	45.0	11.2	681.8	31.5	18.7	7.5
	Low	6	636.5	25.5		663.2	32.0		

* $p<0.05$

An asterisk means that the mean difference is statistically significant at the .05 level. An examination of the Gap columns indicates that every difference between the prior achievement groups was statistically significant at both points in time. For example, the gap between the High and Low achievement groups in District A in Fall 2000 on the Open Ended subtest was statistically significant (46.4), and persisted in Spring 2002 (26.4). Although the patterns among the sample means suggest that the gap favoring high prior achievement students over low prior achievement students shrank between Fall, 2000 and Spring, 2002, there was no statistical significance indicating that the gaps between groups changed over time. For example, the Change value of -20 for District A on the Open Ended test indicates that while the mean difference between the High and Low prior achievement groups decreased by 20 points, they were statistically the same in Spring, 2002 as they were in Fall, 2000.

Only one value in the Change column was statistically significant, High and Low prior achievement groups in District B for Procedures. Seven of the remaining 8 Change values were not statistically significant but were negative, providing descriptive evidence that the process of narrowing the gap among High and Low prior mathematics achievement groups might be underway. Patterns for high and low SES students in Table 10 were similar to those reported in Table 9 for high and low prior achievement.

It is important to point out that listwise deletion (which requires that the same subjects provide scores at both time points) was used in producing the results in Tables 9 and 10. This allowed us to test the change between groups over time. This decision also lowered

the sample sizes. More importantly, this raises the possibility of bias. To investigate this empirically, the analyses comparing the High and Low prior achievement and SES groups at each time were repeated using listwise deletion versus using all available data. The sample sizes for listwise deletion versus using all available data tended to be most different for larger samples, although on the whole the change was not dramatic (e.g., 12 versus 13, 61 versus 76). Similarly, the means and standard deviations of the scaled scores with and without listwise deletion were generally quite similar. Only one of the 18 t-tests comparing High and Low prior achievement groups for Fall, 2000 and Spring, 2002 produced a statistically different result from listwise deletion. Likewise for SES, only one of the possible 18 t-tests produced a different result. We chose to use listwise deletion in generating Tables 9-10 because the two sets of results (with and without listwise deletion) in Fall 2000 and Spring 2002 were quite similar, and because this allowed for change to be tested statistically. However, we acknowledge the possibility of some bias in testing the change values.

Discussion

Section 1: Descriptive Data

The descriptive results reported in Table 2 show that on the SAT-9 tests designed to measure traditional content, students enrolled in *Standards*-based mathematics curricula performed above the NCE national mean of 50 on the Open Ended and Problem Solving subtests. Students were generally below the NCE mean of 50 on the Procedures subtest.

These results suggest that students are learning traditional topics but are also lacking in paper and pencil procedural skills. This result parallels the findings of other studies of a similar nature. (Schoen, et. al., 2003 and Senk and Thompson, 2003). Since these curricula consciously spend less time and effort developing student skills in paper and pencil calculations, these procedural subtest results may be largely a matter of reduced time-on-task in this area.

Although students were, on the whole, performing at or above expectations on two of the three SAT-9 subtests, the performance of various subgroups differed sharply. Students in the high SES and native English speaker groups on average scored substantially higher than those in the low SES and nonnative English speaker groups. Among the ethnic groups represented in the sample, White students uniformly produced the highest averages, with African-American and Hispanic students scoring significantly lower. There were no differences among male and female students on any of the SAT-9 subtests. There was substantial variation in average SAT-9 performance among classrooms, however.

One set of district results is particularly noteworthy. The results reported in Table 3 show that students enrolled in the *Standards*-based mathematics curricula in one of our districts far outperformed national norms. The New Standards Reference Examination, which is more closely aligned to the NCTM Standards, purports to measure conceptual development, problem solving and traditional skills and admits the use of calculator

technology. District E was the only one that elected to use the New Standards Test and was the highest achieving of our districts.

The often heard criticism from opponents of *Standards*-based curricula that high achieving students will be "held back" is soundly refuted in these related results from one of our districts. District E was the most affluent and high achieving district, and had implemented *Standards*-based curricula for all of its students at the elementary (Everyday Mathematics), middle grades (CMP) and high school (Core Plus) levels. This district in Spring 2003 reported that the number of high school students taking the AP Calculus exam (BC) jumped from 50 to 67 between 1999, the last year of traditional students, and 2003. The passing rate with a score of 3 or better during this period increased from 64% to 87%. For AB calculus, the corresponding numbers were: the number of students increased from 10 to 16 and there was an increase in the percentage with a grade of 3 or more from 30% to 81%. Similarly, the number of students taking the AP statistics exam increased from 31 to 71 between 1999 and 2003. The passing rate, with a score of 3 or higher, during this period also increased slightly from 74% to 76%. In 2003, all high school students had been exposed only to the CMP and Core Plus programs since grade 6 (District E, 2003).

Such information, relating to District E, certainly undercuts the viability of the premature and irrational claims of the "mathematically correct" and other anti-reform organizations related to lack of success and college readiness. District administrators need to understand that Standards-based curricula do not impede the mathematical performance

and development of high achieving students when they make curricula adoption decisions.

Section 2: HLM Across-District Results Discussion

The HLM results for the Open Ended and Problem Solving subtests indicated that prior mathematics knowledge was a consistent predictor of mathematics performance on these subtests at the student and classroom levels, although the effect was modest in size. Socio-economic status was a stronger predictor of mathematics performance at both the student and classroom levels, with higher SES linked to higher performance. Whether a classroom was in an urban or suburban school also impacted achievement, with strong differences favoring suburban classrooms emerging. Higher student and classroom prior mathematics scores were also associated with higher Procedures scores.

The effect of prior achievement on assessments of student understanding has been well documented elsewhere in terms of traditional mathematics instruction and alternative mathematics programs (e.g., Begle, 1973). Our analysis suggests that prior achievement is a significant predictor (though small, less than 1 NCE point on average) of student achievement on all three SAT-9 subtests. This finding underscores the position that achievement gaps need to be addressed early in students' academic careers.

Boaler (2003) suggests that gaps between low and high SES decrease over time when students are involved in an Open Ended project-based curriculum. The present study found results favoring the high SES group on both the Open Ended and Problem Solving

subtests of the SAT-9. This effect was moderate, consisting of 3.38 NCE points on the Open Ended subtest and 5.1 NCE points on the Problem Solving subtest. Student achievement on the Procedures subtest, however, was not significantly associated with student level SES. Our explanation for this finding is that the more context-bound and language intensive Problem Solving and Open Ended assessments may be more difficult for students of low SES for reasons described by Lubienski (2000).

High percentages of nonnative speakers, low SES students, and high percentages of minority students are commonly associated with urban schools (Grant and Tate, 2001). When these independent variables were accounted for in the HLM model employed, large significant differences between urban and suburban classrooms on the Open Ended and Problem Solving subtests remained. There apparently are other student or classroom predictors associated with urban and suburban schools that were not accounted for in our model (e.g., class size, degree of parental involvement, education level of parents, etc.).

Much research has been conducted on the importance of school and classroom culture when considering factors that affect student achievement (Finnan, 2000; Pang, 2003). Examining the effect of classroom level predictors on student achievement enables us to describe this classroom culture quantitatively. For example, as the percent of special education students increases in the classroom this tends to increase the association between prior mathematics knowledge and Problem Solving. That is, as the percent of special education students in the classroom increases, prior knowledge plays a bigger role in predicting student achievement on SAT-9 subtest scores.

Section 3: Predicted Classroom NCE Scores

The fitted HLM models were used to predict the performance of classrooms on the SAT-9 subtests with statistical control of other variables, such as SES and English speaker status. It's important to remember that the students in these classrooms had been in a *Standards*-based curriculum for three years.

The results for the Open Ended subtest showed that the average predicted classroom scores generally exceeded the test publisher's cutoff of satisfactory performance for the two year period, despite varying prior mathematics knowledge and classroom composition. These patterns somewhat favored suburban over urban classrooms. For the Problem Solving subtest sharper differences in average predicted scores favoring suburban classrooms emerged. For the Procedures subtest, none of the urban classrooms, and few of the suburban classrooms, had an average predicted score above 50.

Section 4: Value-Added Component

The Value-Added (repeat testers) component was designed to evaluate achievement patterns for students in middle school reform curricula. In value-added situations each student serves as their own control, allowing patterns of actual growth to be quantified and compared with patterns of expected growth. Table 8 reported average growth over time for these students. It should be noted that considerable variation occurred in student growth patterns.

All four districts with repeat tester data ended the eighth grade year with Open Ended and Problem Solving means exceeding the publisher's cutoff for satisfactory performance, and three of the four districts showed growth over two years that exceeded the publisher's expectations. The results were mixed for Procedures, with one district exceeding expected growth and the remaining districts falling short.

On the Open Ended subtest, the observed mean growth in scale scores exceeded the expectations of the test publishers in all cases except for the high SES suburban district. This district likely experienced a ceiling effect since their mean scores ranged from the 97^{th} to the 99^{th} percentile. It should be observed that 3 of the 4 districts, including the urban district, began the seventh grade year with means below the national average (50^{th} percentile) expected scale score. All districts ended the eighth grade year with means above the national average expected scale score, suggesting progress in students' ability to set up and solve problems that are open ended in nature.

On the Problem Solving subtest, all four districts ended the eighth grade year with means at or above the national expected scale score, also calibrated at the 50^{th} percentile. Three of the four districts exceeded the publisher's expected growth over the span. The district that did not exceed the expected growth finished the eighth grade year at the publisher's expected performance level. Students performed satisfactorily on the SAT-9 Problem Solving subscale. Problem solving is an important focus of all reform curricula and a major thrust in both the NCTM standards documents (1989 and 2000).

On the Procedures subtest, two of the three districts were initially (October 2000) below the expected Procedures scale score and all three ended (April 2002) below the 50th percentile scale score. This would lead one to believe that the issue of hand calculation has its roots in early grades and continues on into middle school. Due to the manner in which *Standards*-based curricula have been constructed, it is probable that classroom teachers in this study did not focus heavily on developing procedural skills. This might also be a logical extension of their observation and belief that the routine use of calculators and spreadsheets more realistically reflects real world situations. From this perspective, the curricula do not value procedural knowledge as highly as problem solving ability. It seems that students are indeed learning what they are being taught.

These results provide an opportunity for teachers, researchers and the public to discuss what exactly is valued in a middle school math curriculum. Can schools do it all? For all? What computational skills are basic? Who decides what society values in the area of computation? What mathematical abilities does the average citizen need to function well in our society? Is there a separate set of computational skills that future math-oriented students need for success? If so, what are they? Can society afford to implement a curriculum whose primary purpose is to benefit the 4% - 5%* who pursue math related careers?

> * "Between 4 and 5 percent of an age group will major in Mathematics, Science, or Engineering, one of the traditional mathematics-intensive disciplines. This percentage has been fairly constant since the 1950's through both mathematics reform and "back-to-basics" movements. Majors grow or shrink by reapportioning students in this 5 percent group." (Mathematics and the Mathematical Sciences in 2010: What Should Students Know? MAA, 2000)

Section 5: Achievement Gaps

With respect to high prior achievement vs. low prior achievement, Table 9 shows some evidence of a narrowing of the achievement gap between students with high and low prior achievement levels in the three districts for whom we had a value added (repeat testers) component. We examined the gaps as relating to the low prior achievement vs. high prior achievement as applicable to the SAT-9 Open Ended, Problem Solving and Procedures NCE subtest scores measured over the three testing periods between Fall 2000 and Spring 2002. Our results provide little statistical evidence of a narrowing of the achievement gap between students with high and low prior achievement levels or in the high and low SES groups. On the other hand, there was no evidence that the gaps were widening.

We don't want to attach too much weight to the descriptive statistics, but it is striking that the achievement gap between high and low levels of prior achievement decreased in 8 of the 9 comparisons. Only one of these was statistically significant, however. Likewise, in 7 of the 9 contrasts between high and low levels of SES there was descriptive evidence that the achievement gap decreased. In the case of SES, none of the comparisons were statistically significant.

One can argue that it is easier to improve from an NCE score of 40 than it is to improve with an initial score of 60 or higher given that there is much more room for improvement in the lower achieving case. Thus the descriptive evidence that the gaps may be narrowing is not unexpected given comparable curriculum, instruction and time on task. It is unlikely that the achievement gap can be entirely eliminated as some national initiatives have suggested (NCLB). One has to consider what it is that the higher achieving students are doing while the lower achieving students are busy closing the gap. The answer of course is that they are continuing to exhibit the behavior that resulted in their becoming the higher achieving students in the first place, and they are most likely continuing to benefit from the extra-class types of support normally associated with higher achieving students. The narrowing of the gap is of course a viable educational goal, one that seems more likely to be achieved now with the emergence of *Standards* based curricula. All students can now be realistically and consistently exposed to significant and powerful mathematical ideas with the use of these curricula.

The *Standards* -based curricula offer exceptional promise in this regard, as they were in every case designed for the vast majority (80% – 90%) of school students. This is a non-trivial distinction. In the not too distant past lower achieving students were, routinely redirected away from powerful mathematical ideas and placed in remedial situations. Here they were 'one more time' re-exposed to arithmetic algorithms and other numerically oriented basic skills. Such skills rarely were intended to lead to higher level coursework where other and perhaps more important mathematical ideas are developed. Mathematics had become a filter of students. Most contend today that

mathematics must become a pump (e.g., Steen, 1987), by increasing the number of students who will be involved in higher levels of educational achievement, and involving a greater range of students in the mathematical enterprise for a longer period of time. It is plausible that a good portion of the routinely observed achievement gaps in mathematics are, in a significant way, a result of uneven exposure to powerful mathematical ideas. As all students become accustomed to, and comfortable with, continuous exposure to important and powerful concepts, it is likely that the gaps will decrease.

Two things seem clear from these results. First, the achievement gaps are real. They are large and they persist between different populations of students over time. In many cases the gap amounted to a full standard deviation. Secondly, in this study, prior achievement was repeatedly a significant predictor of future achievement. It is important to "get it right" early in the child's educational career. It follows that increased attention must be paid to every student's mathematical (and other) development at the pre-school and early primary education levels. This is where students begin to be sorted out, and sort themselves out, by their attention, motivation, achievement levels, and often, by the expectations of their teachers. Additional commitment must be extended into each of the elementary grades to ensure that when students reach the junior high level, achievement gaps will be as small as possible and certainly less than presently observed. This assumes of course that at each level all students are exposed to significant mathematics, and regularly participate in activity promoting high levels of mathematical thinking. This is not the case at present

Broader Implications

These results indicate that middle grades students in the five districts discussed here who have been involved with either the MathThematics (STEM Project) or the Connected Mathematics Project (CMP) *Standards*-based mathematics curricula for three consecutive years demonstrated achievement patterns on the Stanford 9 that in general exceeded the means of the national samples upon which the SAT-9 was normed. The results are also promising on the New Standards exam, with roughly two and one half to four times the percentage of students in our sample meeting or exceeding the standards relating to Mathematical Skills, Mathematical Concepts and Mathematical Problem Solving when compared to the national norming group. Results on the SAT-9 Procedures subtest showed that four out of five of the districts scored below the mean of the national norming group. Given the decreased amount of attention to algorithmic development in these curricula, these results may reflect a time-on-task result.

The Stanford 9 and the New Standards tests, although not the most "conservative" measures available, are primarily attuned to traditionally oriented content. There are topics which *Standards*-based students have studied that are either inadequately assessed or are absent altogether from the two assessments used here. At the middle grades level quadratic and exponential growth, topics in transformational geometry, rational numbers, probability and statistics, and others fall into this category. Instructional approaches which focus on complex and extended problems whose multi-faceted solutions may require several days rather than several minutes of investigation, are unique to the *Standards*-based curricula. Such competencies are also not addressed in the two

assessments used here. This study was therefore not able to address important questions related to complex problem solving nor to the assessment of many of the non-traditional topics such as those mentioned above.

Having said that, recall that this study was motivated by a different set of concerns. Participating $(MASP)^2$ school district administrators were faced with parent concerns that their students were not learning the sets of skills contained in traditional coursework and therefore would not be prepared for college mathematics, especially the calculus. This concern was a bit premature since those concerned parents referred to here had children who were only in the middle grades. This concern was encouraged by several mathematicians from our university mathematics department who have, for the past four years, visited many of the $(MASP)^2$ districts with the message that students in *Standards*-based curricula will be unprepared for college calculus. Although their major focus was at the high school level their message nevertheless percolated down to the middle grades parents. To our knowledge there are no *published studies* to corroborate their concerns, although we know of several web circulated documents.

We now provide a glimpse into the district politics relating to the implementation of *Standards*-based curricula. As will be seen, the waters have not always been calm. This despite the fact that there is published evidence from other studies suggesting that when a *Standards*-based curriculum is fully implemented with fidelity, students achieve at a rate which is significantly higher than in classrooms where teachers regularly select, supplement and significantly modify the *Standards* based mathematics curriculum.

(Briars 2000, Senk & Thompson, 2003) With a dozen or so exceptions, we simply do not know the extent to which teachers in this study have abided by program directives relating to instruction in both content and method. There is anecdotal evidence that some teachers supplemented with skill directed worksheets. We can say that teachers in these classrooms averaged 165 hours of professional development with at least 100 of those hours targeted specifically to the curriculum they were teaching. This professional development included 20 personalized in-class contact hours with a mentor teacher whose purpose was to support and help develop instructional proficiency with the curriculum in question. Every teacher in the study had at least 65 hours of professional development.

The anecdotal evidence which we do have on this issue of fidelity of implementation suggests that there is considerable variation in teachers usage patterns, and, that in some of our districts teachers actually worked to discredit the *Standards* based programs altogether. This is counterintuitive since traditionally oriented standardized tests were used in this study, and lower achieving districts tended to be those with overt attempts to subvert the *Standards*-based curricula by re-emphasizing more traditional topics. One would think such classrooms would do well on these standardized measures. The evidence for such subversion, however, is sporadic and anecdotal, and was not systematically collected.

The HLM results document that when prior knowledge and several other demographic variables were taken into account there continued to be significant achievement

differences between the urban and suburban districts. There really are no surprises in our data relating to the impact of demographic variables on student achievement. That is, urban, low SES, nonnative speakers and low levels of prior achievement are all associated with lower achievement levels on the standardized tests.

In conclusion, we find that when *Standards*-based students' achievement patterns on these two standardized instruments are analyzed, traditional topics are learned, although the evidence here is that students' achievement levels on the Open Ended and Problem Solving subtests are greater than those on the Procedures subtest. This is in addition to whatever benefits might accrue from the use of the broader scoped, the more contextually based, and the increased emphasis on extended problems of the *Standards*-based curricula. This finding is consistent with results documented in many of the studies reported in Senk and Thompson (2003), and other sources. This study was not designed to evaluate those additional projected benefits.

Valid and reliable instruments that adequately measure the new content and processes inherent in *Standards*-based curricula are challenging to develop, will be cumbersome to administer and time consuming to score. The variety of instruments to be used in future NCLB assessments hold no promise in this regard, as they will of necessity focus on low level factual knowledge and procedural skills. The next step in ongoing research efforts in this area should paint a portrait of the content and processes that students in *Standards*-based curricula learn that are above and beyond traditional mathematical topics considered at the grade levels of interest. In a parallel effort, it will be important to

conduct studies that document the related situation, or what traditional students are learning that *Standards*-based students are not. It will then be possible to ask and answer the question "What kind of student mathematical outcomes do you value, and which type of programs are most likely to produce them?"

In conclusion, this study suggests that Standards-based middle grades students do learn traditional mathematical topics, but do not develop high levels of procedural skills.

References

Abeille, A., & Hurley, N. (2001). Final evaluation report: Mathematics modeling our world. Accessed on October 18, 2003 from http://www.comap.com/highschool/projects/mmow/FinalReport.pdf

Balanced Assessment. (1999). *Middle grades assessment package I*. Palo Alto, CA: Dale Seymour.

Begle, E. G. (1973). Some lessons learned by SMSG. *Mathematics Teacher*, *66*, 207-214.

Billstein, R. (1998). Middle Grades Math Thematics: The STEM Project. In L. Leutzinger (Ed.), *Mathematics in the Middle* (93-106). Reston, VA: NCTM.

Billstein, R., & Williamson, J. (1998). *Middle grades MATH Thematics* (Book 1, 2, and 3). Evanston, IL: McDougal Littell.

Billstein, R. & Williamson, J. (2003). *Middle grades MATH Thematics:* The *Stem Project*. In S. L. Senk and D. R. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 251-284). Mahwah, NJ: Lawrence Erlbaum Associates.

Boaler, J. (n.d.). Stanford University Mathematics Teaching and learning study: Initial report – A comparison of IMP1 and Algebra 1 at Greendale school. Accessed October 18, 2003 from
http://www.stanford.edu/~joboaler/Initial_report_Greendale.doc

Briars, D. & Resnick, L. (2000) *Standards, Assessments – and What Else? The essential elements of Standards-based school improvement.* Unpublished manuscript. Retrieved October 15, 2003, from http://www.cse.ucla.edu/CRESST/Reports/TECH528.pdf

Carroll, W.M. (1997). Results of third-grade students in a *Standards*-based curriculum on the Illinois state mathematics test. *Journal for research in Mathematics Education, 28,* 237-242.

Carroll, W. M. & Isaacs, A. (2003). Achievement of students using the University of Chicago School Mathematics Project's *Everyday Mathematics*. In S. L. Senk and D. R. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 79-108). Mahwah, NJ: Lawrence Erlbaum Associates.

Carter, A., Beissinger, J. S., Cirulis, A., Gartzman, M., Kelso, C. R., Wagreich, P. (2003). Student learning and achievement with *Math Trailblazers*. In S. L. Senk and D. R.

Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 45-78). Mahwah, NJ: Lawrence Erlbaum Associates.

Cichon, D., & Ellis, J. G. (2003). The effects of Math *Connections* on student achievement, confidence, and perception. In S. L. Senk and D. R. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 345-374). Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, D. K., & Ball, D. L. (2001). Making change: Instruction and its improvement. *Phi Delta Kappan, 83*, 73-77.

Cooney, T. J. (1985). A beginning teacher's view of problem solving. *Journal for Research in Mathematics Education*, *16*, 324-336.

Coxford, A. F., Fey, J. T., Hirsch, C. R., Schoen, H. L., Burrill, G., Hart, E. W., & Watkins, A. E. (1998). *Contemporary mathematics in context: A unified approach (Courses 1-4)*. New York: Glencoe/McGraw-Hill.

Davis, R.B. (1990) "Discovery Learning and Constructivism." *Journal for Research in Mathematics Education*, Monograph Number 4. Reston, VA: National Council of Teachers of Mathematics. Pp 93-106.

District E. (2003). "Mathematics Program Evaluation." Fall, 2003.

English, Lynn D. (2002) Handbook of International Research in Mathematics Education. Mahwah, NJ: Lawrence Erlbaum Associates.

Fendel, D., Resek, D., Alper, L., & Fraser, S. (1997). *Interactive mathematics program (Years 1-4)*. Berkeley, CA: Key Curriculum Press.

Finnan, C. (2000) "Implementing School Reform Models: Why is it So Hard for Some Schools and Easy for Others?" Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, Louisiana, April 24-28, 2000).

Fuson, K. C., Carroll, W. C., & Drueck, J. V. (2000). Achievement results for second and third graders using the *Standards*-based curriculum *Everyday Mathematics*. *Journal for Research in Mathematics Education*, *31*, 277-295.

Garfunkel, S., Godbold, L., & Pollak, H. (1998). *Mathematics: Modeling our world (Courses 1-4)*. Cincinnati, OH: South-Western Educational Publishing.

Gearhart, Maryl. , Saxe, Geoffrey B., Seltzer, Michael, Schlackman, Jonah, Ching, Cynthia C., Nasir, Na'ilah, Fall, Randy, Bennett, Tom, Rhine, Steven, & Sloan, Tina F. (1999). "Opportunities to Learn Fractions in Elementary Mathematics Classrooms." *Journal for Research in Mathematics Education, 30*, 286-315.

Grant, C., Tate, W. (2001) "Multicultural Education Through the Lens of the Multicultural Education Research Literature." *Handbook of Research on Multicultural Education*. J. Banks, Editor. San Francisco: Jossey-Bass Publishers.

Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test Ninth Edition*. San Antonio, TX: Harcourt Brace.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Henningsen, M., & Stein, M. K. (1997). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for Research in Mathematics Education*, *28*, 524-549.

Huntley, M. A., Rasmussen, C. L., Villarubi, R. S., Sangtong, J., & Fey, J. T. (2000). Effects of *Standards*-based mathematics education: a study of the Core-Plus Mathematics Project algebra and functions strand. *Journal for Research in Mathematics Education. 31*, 328-361.

Kilpatrick, J. (2003). What Works? In S.L. Senk and D.R. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 471-487). Mahwah, NJ: Lawrence Erlbaum Associates.

Lappan, G, Fey, James T., Phillip, E.D., and Anderson, C. (Eds.) (1997). *Connected Mathematics Series*. NJ: Prentice Hall, 1997.

Learning Research and Development Center of the University of Pittsburgh and the National Center on Education and the Economy (1998). *New Standards Mathematics Reference Examination*, Form C. San Antonio, TX: Harcourt Brace Educational Measurement.

Lubienski, S. T. (2000). Problem solving as a means toward mathematics for all: An exploratory look through a class lens. *Journal for Research in Mathematics Education*, *31*, 454-482.

Martin, T. S., Hunt, C. A., Lannin, J., Leonard, Jr., W., Marshall, G. L., & Wares, A. (2001). How *Standards*-based secondary mathematics textbooks stack up against NCTM's *Principles and Standards*. *Mathematics Teacher*, *94*, 540-589.

Mokros, J. (2003). Learning to reason numerically: The impact of *Investigations*. In S.L. Senk & D.R. Thompson (Eds.), *Standards-based school mathematics curricula: what are they? What do students learn?* (pp. 109-131). Mahwah, NJ: Lawrence Erlbaum Associates.

National Center for Education for Education Statistics. (2003). Percentage of students within each mathematics achievement level, grade 8 public schools: By state, 2003.

Retrieved November 21, 2003 from http://nces.ed.gov/nationsreportcard/mathematics/results2003/stateachieve-g8.asp

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

National Research Council. (1989). *Everybody counts: A report to the nation on the future of mathematics education.* Washington, DC: Author.

National Science Foundation. (1989). *Materials for middle school mathematics instruction: Program solicitation*. Washington, DC: Author.

Pang, J. (2003) "Understanding the Culture of Elementary Mathematics Classrooms in Transition." Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).

Raudenbush, S. & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd Ed.). Thousand Oaks, CA: Sage.

Raudenbush, S., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2000). HLM 5: Hierarchical Linear and Nonlinear Modeling. Lincolnwood, IL: Scientific Software International.

Reys, R., Reys, B., Lapan, R., Holliday, G., & Wasman, D. (2003). Assessing the impact of *Standards*-based middle grades mathematics curriculum material on student achievement. *Journal for Research in Mathematics Education*, *34*, 74-95.

Ridgway, J. E., Zawojewski, J. S., Hoover, M. N., Lambdin, D. V. (2003). Student attainment in the *Connected Mathematics* Curriculum. In S. L. Senk and D. R. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 193-224). Mahwah, NJ: Lawrence Erlbaum Associates.

Riordan, J. E., & Noyce, P. E. (2001). The Impact of two *Standards*-based mathematics curricula on student achievement in Massachusetts. *Journal for Research in Mathematics Education*, *32*, 368-398.

Schoen, H. L., Cebulla, K. J., Finn, K. F., & Fi, C. (2003). Teacher variables that relate to student achievement when using a *Standards*-based curriculum. *Journal for Research in Mathematics Education*, *34*, 228-259.

Schoen, H. L., & Hirsch, C. R. (2003). The Core-Plus Mathematics Project: Perspectives and student achievement. In S. L. Senk & D. R. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 311-343). Mahwah, NJ: Lawrence Erlbaum Associates.

Schoenfeld, A.H. (2002). Making mathematics work for all children: Issues of Standards, testing and equity. *Educational Researcher, 31*, 13-25.

Senk, S. L., & Thompson, D. R. (2003). Middle school mathematics curriculum reform. In S. L. Senk and D. R. Thompson (Eds.), *Standards-based school mathematics curricula? What are they? What do students learn?* (pp. 181-191). Mahwah, NJ: Lawrence Erlbaum Associates.

Steen, L. A. (1987). *Calculus for a New Century: a pump, not a filter, a national Colloquium, October 28-29*. Washington, DC: Mathematical Association of America, c1988.

TERC. (1998). *Investigations in Number, Data, and Space* (Grades K-5). White Plain, NY: Dale Seymour Publications.

Thompson, D. R., & Senk, S. L. (2001). The effects of curriculum on achievement in second-year algebra: The example of the University of Chicago School Mathematics Project. *Journal for Research in Mathematics Education*, *32*, 58-84.

Weiss, I. R., Banilower, E. R., Overstreet, C. M., Soar, E. H. (2002). *Local systemic change through teacher enhancement: Year seven cross-site report.* Chapel Hill, NC: Horizon Research.

Webb, N. L (2003). The Impact of the *Interactive Mathematics Program* on Student Learning. In S.L. Senk & D.R. Thompson(Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 375-398). Mahwah, NJ: Lawrence Erlbaum Associates.

Wiley D. E. & Resnick, L. B. (1998) *The new standards reference examination standards-referenced scoring system* . Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Graduate School of Education & Information Studies, University of California, Los Angeles.

Appendix A – Sample Test Items from the SAT-9

Procedures
Multiple Choice Items;
Read each question and select the best answer.

1) 5 1/2
+ 2 1/4
_____

A 7 1/2 C 7 1/6
B 7 3/4 D 7 1/8
E 8 3/4

Problem Solving

2) On a totem pole, the eagle was above the bear. The beaver was under the thunderbird. The thunderbird was above the eagle. Which animal was on the top of the pole?
F Thunderbird G Bear H Beaver J Eagle .

SAT- 9 High School Edited Sample items

1) When Mr. Tillen meets a client, the probability that he will make a sale is 1/4. How many sales can he expect if he meets 144 clients?
A 145 B 36 C 24 D 12

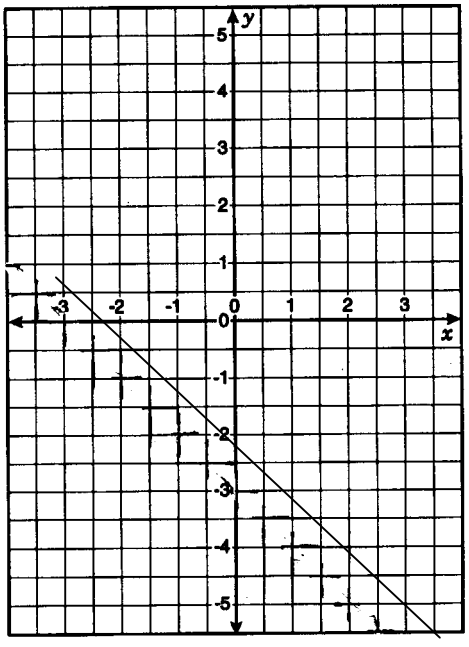
2) After t seconds, the velocity, v, of a basketball thrown upward at 256 feet per second is given by the equation

$$v = 128 - 32t$$

After how many seconds will the velocity equal zero?

F 4, G 6, H 8, J 12

3) A point at (1, -1) is reflected across the line shown on the grid.



What will the coordinates after reflection?
F (-1,1), B (0,0), C (-2, -4), (-4,-2)

Appendix B – Sample Test Items from the New Standards Reference Examination

Middle School Medium Task

How Fast?

In this task, your job is to show a method for figuring a person's biking speed in miles per hour and explain how to use the method for any biking speed.

Henry wants to use the number of pedal rotations he takes each minute to figure out his approximate biking speed in miles per hour. He knows that

a. the distance he travels in one pedal rotation, called "rotation distance" is very close to 5 feet;
b. he takes about 60 pedal rotations each minute;
c. there are 5,280 feet in a mile; and
d. there are 60 minutes in an hour.

State a rule or formula that can be used to approximate biking speed in miles per hour (mph) based on the number of pedal rotation taken in one minute.

Be sure to explain your rule or formula clearly enough that it can be used by any biker.

High School Short Task

In this task you are asked to explore if the melted ice cream will fit into the cone or if it will spill over.

The ice cream and cone below are drawn accurately and at full size.
If the scoop of ice cream is placed on top of the cone and its melts down into the cone, will all the ice cream fit inside the cone.

Show whether the melted ice cream will fit inside the cone.
For full credit, you must show your calculations and reasoning.

Picture of a sphere here. Picture of a cone here.