Standards-based Mathematics Curricula and Middle-Grades Students’ Performance on Standardized Achievement Tests

Thomas R. Post and Michael R. Harwell
University of Minnesota

Jon D. Davis
Western Michigan University

Yukiko Maeda
Michigan State University

Arnie Cutler and Edwin Andersen
University of Minnesota

Jeremy A. Kahan
Ida Crown Jewish Academy, Chicago

Ke Wu Norman
University of Minnesota

Approximately 1400 middle-grades students who had used either the Connected Mathematics Project (CMP) or the MATHThematics (STEM or MT) program for at least 3 years were assessed on two widely used tests, the Stanford Achievement Test, Ninth Edition (Stanford 9) and the New Standards Reference Exam in Mathematics (NSRE). Hierarchical Linear Modeling (HLM) was used to analyze subtest results following methods described by Raudenbush and Bryk (2002). When Standards-based students’ achievement patterns are analyzed, traditional topics were learned. Students’ achievement levels on the Open Ended and Problem Solving subtests were greater than those on the Procedures subtest. This finding is consistent with results documented in many of the studies reported in Senk and Thompson (2003), and other sources.

Key words: Achievement; Assessment; Curriculum; Middle grades, 5–8; Multivariate techniques; Reform in mathematics education

This research was supported by the National Science Foundation (NSF) under Grant ESI-9618741 (1996-2004). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.
This study examined achievement patterns of middle school students enrolled in Standards-based curricula, in particular those curricula that were funded from a solicitation of proposals through the National Science Foundation (NSF) in the early 1990s (NSF RFP 91-100). The focus was on traditional topics in mathematics as measured by two nationally normed achievement tests.

This study builds on and extends our existing understanding of student achievement in Standards-based programs in three 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, we focus on Standards-based curricula used 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 as many earlier field-test teachers were volunteers and therefore may not be typical of all teachers who 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. Thus, HLM allows us to consider student and classroom results simultaneously.

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). An earlier study (Harwell et al., 2007) discussed similar issues for secondary students. Schoenfeld (2002) reviewed this emerging body of work and concluded 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 Connected Mathematics Project (CMP) (Reys, Reys, Lapan, Holliday, & Wasman, 2003; Ridgway, Zawojewski, Hoover, & Lambdin, 2003; Riordan & Noyce, 2001) and MATHThematics (MT) (Billstein, 1998; Reys et al., 2003; Kilpatrick, 2003), when referring to the 13 chapters in Senk and Thompson (2003), concluded that “[t]he studies reported in this volume offer the best evidence we have that Standards-based reform works” (p. 487).

The research on 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 et al. (2003) found that sixth-grade CMP students started 1 year behind non-CMP students on the Iowa Test of Basic Skills (ITBS), and at the end of Grade 6 were 1.5 years behind the other group. The CMP Students who started .52 standard deviation (SD) behind non-CMP students were .61 SD behind after 1 year. At the end of eighth grade, however (3 years later), CMP students were .32 SD ahead. In related investigations, the authors concluded that there is no immediate short-term advantage to CMP, 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.
There is research suggesting that the benefits from a Standards-based curriculum extend beyond increases in mathematics achievement on Open Ended and Problem Solving subtests. Billstein and Williamson (2003) found that students who used MT improved in their attitudes toward mathematics and had higher scores on the language achievement subtest of the ITBS 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 et al., 1999, p. 309; cf. Ball & Cohen, 1996). Briars and Resnick (2000) looked at fidelity of implementation in Everyday Mathematics, a K–6 program used 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 Reference Examination (NSRE). 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, and 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, although a Standards-based curriculum alone can positively influence teacher pedagogy, the results are especially promising if combined with high fidelity of implementation and effective instruction of these new materials.

Another variable in this complex interaction between curriculum and achievement is the students themselves. The ways that students react to and interface with the curriculum in the classroom can affect implementation of the curriculum (Cooney, 1985; Henningsen & Stein, 1997). In addition, it has long been known that characteristics that students bring with them to the classroom help to shape their achievement. For instance, with respect to the School Mathematics Study Group (SMSG), 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. . . . 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. (pp. 213–214)

Student SES has also been shown to play a role in how students interact with Standards-based curricula (Lubienki, 2000). Past research on student achievement in Standards-based classrooms has used SES either as a variable in matching groups for comparison purposes (Reys et al., 2003; Riordan & Noyce, 2001) or as a predictor of student achievement in regression analyses (Schoen et al., 2003). School environment also affects successful implementation of any curriculum (cf.,
Post, Harwell, Davis, Maeda, Cutler, Andersen, Kahan, and Norman (2001; Eisenhart et al., 1993). Schoen et al. (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 1990s, the NSF, 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” (NSF, 1997, p. 5).

The “Minneapolis and St. Paul Merging to Achieve Standards Project” (MASP)², one of these 47 projects, provided professional development 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 each year thereafter.

Of these 21 school districts, 5 were invited to participate in the study of student achievement reported here. The districts were selected by (MASP)² project managers to provide a range of district types while remaining within the budget that was available for the testing phase of the NSF grant. The districts contained different combinations of middle school and senior high NSF-funded curricula and represented all types of districts in the project—urban, suburban, and boundary districts. Purposive sampling of districts served two purposes. First, it provided information to key groups of constituents connected to the project. Second, following the arguments of Cochran (1983), it allowed generalizations to be made to a target population of similar school districts.

The 5 districts included in our sample used either CMP (Lappan, Fey, Phillips, & Anderson, 1998) or MT (Billstein & Williamson, 1998) at the middle school level. 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 (1989), they also share many similarities: recurring integration of topics within grade levels, extended explorations, and a decreased emphasis on paper-and-pencil computation. Although we recognize the danger of combining similar curricula (Davis, 1990), our research questions referred to students in broadly defined Standards-based mathematics classrooms, not those studying from a specific Standards-based curriculum. Therefore, after initially finding no significant differences in our descriptive results, we pooled student results from these two curricula.
The location, curriculum, rationale for choice, and assessment for each district are shown in Table 1. There were sharp differences among some of the districts in geographic location, student enrollment, and student characteristics. The purely urban school district had substantially greater enrollment than the others and showed the greatest diversity in student ethnicity, eligibility for free or reduced 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. We note that the criteria for a student to be classified as eligible for free or reduced lunch, as a nonnative speaker, or as a special education student are state-mandated and so are the same across the five school districts.

<table>
<thead>
<tr>
<th>District</th>
<th>Geographic location</th>
<th>Middle-grades curriculum assessed in this study</th>
<th>Rationale for choice of district</th>
<th>Assessment used</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Urban-suburban</td>
<td>MT</td>
<td>High fidelity of implementation with much parental support</td>
<td>Stanford 9</td>
</tr>
<tr>
<td>B</td>
<td>Urban</td>
<td>CMP</td>
<td>Large urban population—variable implementation</td>
<td>Stanford 9</td>
</tr>
<tr>
<td>C</td>
<td>Suburban</td>
<td>CMP</td>
<td>Wholesale adoption sabotaged by a few faculty dissenter and parents</td>
<td>Stanford 9</td>
</tr>
<tr>
<td>D</td>
<td>Suburban</td>
<td>MT</td>
<td>Wholesale adoption with district authorized supplements</td>
<td>Stanford 9</td>
</tr>
<tr>
<td>E</td>
<td>Suburban</td>
<td>CMP</td>
<td>Enthusiastic adoption</td>
<td>Stanford 9 &amp; NSRE</td>
</tr>
</tbody>
</table>

**Data Collection**

Cross-sectional data for two groups of eighth-grade middle school students, one tested in the Spring of 2001, and the other in the Spring of 2002, were collected. These students had been studying from either CMP or MT for a total of 3 years. 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 consisted of approximately 1600 Standards-based middle school students, most of whom (85%) took the Stanford Achievement Test, Ninth Edition (Stanford 9), which consisted of Open Ended, Problem Solving, and
Procedures subtests. Approximately 25% of the students took both the Stanford 9 and the New Standards Reference Exam in Mathematics (NSRE). However, the statistical analyses were based on fewer than 1600 students, primarily because of missing data on the Stanford 9 subtests. Overall, the percentages of missing data on the Open Ended, Problem Solving, and Procedures subtests were 8.9%, 10.3%, and 11.2%, respectively. Also, 9% of the students who failed to provide data for the Open Ended subtest also failed to provide data for Problem Solving. 10.9% failed to provide data for the Open Ended and Problem Solving subtests, 3.5% failed to provide data for the Problem Solving and Procedures subtests, and 5% failed to provide data for any of the Stanford 9 subtests.

Proceeding with statistical analyses in the presence of missing data would typically invoke the listwise deletion option popular in data analysis software. Listwise deletion requires all subjects with missing data be eliminated from the analysis. Listwise deletion also requires that the missing data be missing completely at random. Under this condition, the missing data (if obtained) would convey the same information as the available data, or, more formally, the distributions of the missing and available data would be identical. If the assumption that the missing data are missing in a completely random fashion is not satisfied, the statistical results will be biased by an amount depending on the extent to which this assumption is not true. If a biasing effect is present, it will be exaggerated for groups with disproportionately greater amounts of missing data through its impact on statistics computed for those groups. Accordingly, we examined the data for evidence that particular demographic groups had substantially different amounts of missing data.

In general, the amount of missing data on the Stanford 9 subtests appeared to be similar across the student demographic variables. The median percentages of missing data across the Stanford 9 subtests for nonnative and native-English speakers were 10% and 9%, respectively; for students who were or were not eligible for free or reduced lunch, these percentages were 12% and 8%, respectively; for Black, Asian, Hispanic, and White students, the median percentages of missing data across the Stanford 9 subtests were 16%, 5%, 10%, and 8%, respectively.

The largest difference in the amount of missing data was for Black and Asian students (16% and 5%, respectively). However, given the numbers of Black and Asian students in our sample (248 and 124, respectively), and the average Stanford 9 scores for these groups, it is unlikely, at least in terms of mean differences, that the missing data introduced serious bias. For example, the Open Ended mean for Black students was 42.6 (N = 204), whereas for Asian students this mean was 50.9 (N = 118). The 44 missing Black students would need to score on average 89 for the entire sample to produce an Open Ended mean equal to that of the Asian students (i.e., 50.9), and an average of 61 to produce an Open Ended mean halfway between 42.6 and 50.9 (i.e., 46.7). Similarly, if the missing Asian students each scored 1 on the Open Ended subtest, the mean for the entire sample of Asian students would still be much larger than 42.6. Similar results appear for the Problem Solving and Procedures subtest. Thus, the largest observed difference in missing Stanford 9 data among demographic groups is unlikely to change the basic findings.
On the whole, these results provide evidence that missing Stanford 9 data were not disproportionately located in a particular Stanford 9 subtest or student demographic group, which suggests that the reasons data were missing cannot be explained by which subtest a student may have failed to take or by a student’s demographic profile. This in turn makes biased results due to missing data less likely. However, we cannot be certain that the results reported in this article would be similar to those obtained if the missing data had been available, and it would be prudent to interpret our findings in light of this potential bias.

**Design**

A nonexperimental 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. No control group of students who had experienced a traditional curriculum existed since we were testing students in districts that had adopted, in wholesale fashion, a reform curriculum in each of their middle schools. 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.

We determined that an appropriate comparison would be the standards established by the publisher of the standardized test instrument used in the assessment—the Stanford 9 or the NSRE. The Normal Curve Equivalent (NCE) mean of 50 was selected as our benchmark because it reflected average performance on the Stanford 9 based on national norms. 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” (Harcourt Assessment, Inc., 2005, p. 4).

**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 (MASP)² provided professional development. We had a responsibility to document achievement patterns of students related to national norms on traditionally oriented standardized tests. The school districts wanted testing instruments with national norms, and after reviewing several instruments from national publishers and consulting with our districts, we narrowed the list to the Stanford 9 and the NSRE. Both sets of tests were published and scored by Harcourt Brace.

The mathematics portion of the Stanford 9 has three subtests. The Problem Solving subtest contains 30 multiple-choice problems that require students to solve problems set within real-world and mathematical contexts. The Procedures subtest has 20 multiple-choice questions that require students to perform one of the four basic arithmetic operations with whole numbers, integers, and fractions. The Open Ended subtest uses realistic problems to evaluate students’ concepts and skills.
Each of the Stanford 9 subtests tests 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 takes two 50-minute school periods to administer. Calculators are allowed on all subtests except for Procedures. The Stanford 9 reports both scale scores and NCEs, and 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 mathematics portion of the NSRE consists of three parts. The first consists of 20 multiple-choice questions that are a subset of Stanford 9 multiple-choice test items. In addition, this first section contains short tasks. The Stanford 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; medium and long tasks are extended-response items that require detailed answers.

The NSRE is a criterion-referenced test that sets levels of constructed-response performance in three areas: Skills, Concepts, and Problem Solving. The performance levels (achieved with honors, achieved the standard, nearly achieved, below the standard, and little achievement) are derived from national Standards developed by a conglomerate of assessment-related organizations (Wiley & 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.

Other studies have noted (Begle, 1973; Reys et al., 2003; Riordan & Noyce, 2001) that prior achievement in mathematics is an important predictor of student achievement. As is often the case, different districts administered different mathematics tests to students. At the middle school level, the ITBS, Northwest Achievement Level Test (NALT), Minnesota Comprehensive Assessment (MCA), Metropolitan Achievement Test (MAT7), and Terra Nova were used. In three of the districts, a subsample of students had scores on two of these mathematics tests. We only had access to total scores for these tests.

Because we wished to have a single (common) prior mathematics achievement score for each student, we began by examining the characteristics of these tests. We felt there was considerable overlap in the objectives, content, and format of the various tests, providing strong evidence for treating these tests as reflecting a single construct of mathematics proficiency.

Next we examined available correlation evidence of student performance on these tests. A subset of students had scores on two prior mathematics tests, producing a Pearson correlation of .79 between the MCA and ITBS ($N = 172$), .50 for the MAT7 and MCA ($N = 108$), and .69 for the MCA and NALT ($N = 302$). These correla-
tions provided empirical support for the conclusion that these tests assess a common construct of mathematics proficiency.

We also fitted multiple regression models to the Stanford 9 student data within each district using the available prior mathematics measures 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 explained variance attributable to prior mathematics achievement when the effects of the other predictors were held constant.

The logical and empirical analyses above 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. NCEs 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 40%. 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 sensitive to the metric of the combined prior mathematics achievement variable (i.e., our findings are similar regardless of whether the original [combined] prior mathematics scores or their deciles are used).

**Student and Classroom Samples**

Students in 43 Standards-based classrooms were tested. The teachers in these classrooms had participated in three types of professional development provided through the (MASP)² LSC. (NSF mandated 130 hours of targeted professional development experience for all LSC teachers.) First, teachers participated in 2 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. In-depth consideration of some of the mathematics underlying these activities was also provided by (MASP)² staff. 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, and meetings with teachers and administrators to discuss administrative issues and with their counselors relating to the scheduling
of students. Last, (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 by one-on-one debriefings and, in some cases, demonstration lessons. Teachers requested professional development beyond year 1 that was designed for the next level of the curriculum to be used. Higher Education Eisenhower funds provided for additional professional development in years 2 and 3. Middle-grades teachers in this study had completed an average of 162 professional development hours over a 3-year period.

The number of students that could be tested within a district was limited by the cost of test administration and grading and was allocated in proportion to the number of mathematics teachers in the various schools. District administrators, in collaboration with project personnel, selected a number of classrooms whose totals contained the requested number of students. Administrators were provided direction to purposively select classes containing students who were representative of the entire spectrum of the student body in that school. Administrators were also asked to select classes that were perceived to have a high fidelity of curriculum implementation; this judgment of fidelity was confirmed by the assigned mentor teacher. Assessing the fidelity of implementation can be difficult for a variety of reasons. Reys and colleagues (2003) found that teaching strategies consistent with the reform text author’s suggestions were used from 10% to 90% of the time in the middle-grades Standards-based classrooms they observed.

As part of our effort to determine implementation levels, the assigned (MASP)\(^2\) mentors for teachers provided their informal assessment of their teacher’s implementation level. This assessment was done for all teachers. Another effort was to send selected mentors into a sample of classrooms using an instrument, “Core Evaluation Classroom Observation Protocol” (Lawrenz, Huffeman, & Appeldoorn, 2002), developed under another NSF grant and modeled after “Inside the Classroom: Observation and Analytic Protocol” (Weiss, Pasley, Smith, Banilower, & Heck, 2003). The authors of the protocol trained the observers in its use. These observers then practiced using videos of various classrooms. Eleven of the 43 classrooms tested in our sample were observed using this protocol. On a Likert scale of 1 to 5, with 5 representing exemplary implementation, all observed classrooms fell in the 3 to 5 range. Finally, we relied on knowledge of individual teacher’s classrooms possessed by district evaluation personnel and district curriculum directors. While any set of efforts cannot guarantee full fidelity, these efforts and associated evidence of their success indicate that the curricula were implemented in our test classrooms at a satisfactory level.

RESULTS

**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, but for the Procedures subtest four of
the five districts scored below the national mean. Recall that this subtest of the
Stanford 9 covers the four basic operations with whole numbers, integers, and frac-
tions within purely symbolic settings and sometimes with one-step word problems
solved without calculators.

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

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

The results for District E, the district that also administered the Mathematics
portion of the NSRE, a subsample of the Stanford 9, are in Table 3. The results
show that 86% (57 + 29) of the students tested in District E achieved or exceeded
the mathematical skills standard, whereas 33% (11 + 22) of the students at the
national level performed at this 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, whereas
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 NSRE instrument comes from the
Northeastern part of the United States, an area that typically has higher scores than
other geographical areas (National Center for Education Statistics, 2003).
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, District B had approximately 21% of its middle school students classified as nonnative English speakers, whereas the remaining districts had values ranging from 0% to 6%. The percentage of special education students varied from 2% to 12%, with the highest value attached to a suburban district. SES showed comparatively more variability across all districts. One district had more than 60% of its middle school students eligible for free or reduced lunch (low SES on Figure 1), two districts had 18% to 20% eligible, another about 15%, and in one district 3% of the middle school students were eligible.

Average performances for the Stanford 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 average mathematics proficiency and prior mathematics knowledge across the districts. However, as shown in the analysis to follow, these differences shrink when various demographic variables are statistically partialled out.

There was also evidence of substantial variability in average mathematics performance at the classroom level. Figure 3 shows the classroom means for the Stanford
Figure 1. District demographic data.

Figure 2. District achievement data
There was variation in average mathematics performance across SES and English language (native vs. nonnative speaker). Students not eligible for free or reduced 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 Figure 4 and shows that their combination had 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).

Additional descriptive statistics, including effect sizes, appear in Table 4. The effect sizes help to quantify differences apparent in Figure 4. 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 size were gender (not reported) and Asians versus 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 the Open Ended subtest across all subgroups; the Procedures subtest produced the lowest NCEs for all subgroups.

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.

Figure 4. Achievement by ethnicity.
Table 4
Descriptive Data and Effect Sizes for Various Subgroups

<table>
<thead>
<tr>
<th></th>
<th>Open Ended</th>
<th>Problem Solving</th>
<th>Procedures</th>
<th>Prior Achievement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group</td>
<td>N</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>SES</td>
<td>High</td>
<td>982</td>
<td>62.9</td>
<td>18.5</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>444</td>
<td>45.3</td>
<td>23.0</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>White</td>
<td>1004</td>
<td>63.3</td>
<td>18.2</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>204</td>
<td>42.6</td>
<td>24.1</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>118</td>
<td>50.9</td>
<td>18.5</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>111</td>
<td>39.7</td>
<td>23.0</td>
</tr>
<tr>
<td>Location</td>
<td>Suburban</td>
<td>981</td>
<td>61.6</td>
<td>18.7</td>
</tr>
<tr>
<td></td>
<td>Urban</td>
<td>456</td>
<td>48.8</td>
<td>24.6</td>
</tr>
<tr>
<td>Language status</td>
<td>Native</td>
<td>1315</td>
<td>59.8</td>
<td>20.4</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-native</td>
<td>122</td>
<td>33.4</td>
<td>20.1</td>
</tr>
<tr>
<td>Special Education</td>
<td>Non-special</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ed.</td>
<td>1356</td>
<td>58.5</td>
<td>21.1</td>
</tr>
<tr>
<td></td>
<td>Special</td>
<td>81</td>
<td>41.3</td>
<td>23.4</td>
</tr>
</tbody>
</table>

*p < 0.05
HLM Analyses of Student and Classroom Data

The Stanford 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 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), three models were fitted. First an unconditional model of the form

\[ Y_{ij} = \beta_{0j} + r_{ij} \]  
\[ \beta_{0j} = \gamma_{00} + u_{0j}, \]

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

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

in which \( \beta_{1j} \) 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 \( j \)th 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 Stanford 9 subtest varied across classrooms, we constructed a predictive model to try to explain variation in these slopes with classroom-level predictors. Key classroom predictors included Class SES (percentage of students eligible for free or reduced lunch in a classroom) and average prior mathematics knowledge in a classroom. Other classroom predictors that 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 class-
room attendance and the percentage of female students in a classroom could be removed because they did not contribute 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

\[
\beta_{0j} = \gamma_{00} + \gamma_{01} (\text{Class SES}_1 - \bar{W}_1) + \gamma_{02} (\text{Class Black}_2 - \bar{W}_2) \\
+ \gamma_{03} (\text{Class Asian}_3 - \bar{W}_3) + \gamma_{04} (\text{Class Hispanic}_4 - \bar{W}_4) \\
+ \gamma_{05} (\text{Class nonnative English speakers}_5 - \bar{W}_5) \\
+ \gamma_{06} (\text{Class special education}_6 - \bar{W}_6) \\
+ \gamma_{07} (\text{district}_7 - \bar{W}_7) + \gamma_{08} (\text{professional development}_8 - \bar{W}_8) \\
+ \gamma_{09} (\text{prior}_9 - \bar{W}_9) + u_{0j},
\]

in which where \(\gamma_{01}\) is the classroom-level slope capturing the effect of class SES (percentage of low SES students) on average mathematics performance, \(\bar{W}_1\) is class SES averaged across classrooms, Class Black is the percentage of Black 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) were 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% to 7%, and, at the other end of the distribution, 10 of the classrooms had values ranging between 14% to 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 (see Table 5). Thus, each of the classroom predictors above was transformed to a scale in which each was represented with four values corresponding to the Table 5 quartiles.

The HLM5 software (Raudenbush, Bryk, Cheong, & Congdon, 2001) does not permit missing data when fitting the student level model in equation (3) to each classroom’s data, employing listwise deletion to ensure that only complete student datasets are analyzed. This raises the possibility of biased statistical findings due
to omitting cases showing missing data systematically differing from cases that provided data. Earlier results suggested that the percentages of missing data were generally similar across demographic groups and that average Stanford 9 scores were not seriously affected by the presence of missing data. We continued to explore possible effects of missing data by focusing on HLM5’s use of listwise deletion. We also fitted the model in equation (3) to the student Stanford 9 data using the AMOS5 (SmallWaters Corporation, 2003) structural equation modeling software (AMOS5 does not perform hierarchical linear modeling). Rather than employing listwise deletion, AMOS5 uses whatever data students provide in estimating regression parameters for a classroom. Although HLM5 and AMOS5 use different methods to estimate parameters (Ordinary Least Squares and Full Information Maximum Likelihood, respectively), they should produce similar estimated parameters except when class sample sizes are quite small, which we did not think was a serious problem given the median classroom size of 26 in our data.

For the Open Ended subtest, we then compared the attendance, SES, gender, and prior slopes estimated using HLM5 and AMOS5 for each classroom. Similar slopes

<table>
<thead>
<tr>
<th>Variable</th>
<th>Quartile</th>
<th>N</th>
<th>Range (R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class SES</td>
<td>1</td>
<td>10</td>
<td>R \leq 15.38%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>12</td>
<td>15.38% &lt; R \leq 23.53%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11</td>
<td>23.53% &lt; R \leq 69.57%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>10</td>
<td>R &gt; 69.57%</td>
</tr>
<tr>
<td>Class English language status</td>
<td>1</td>
<td>15</td>
<td>R = 0%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7</td>
<td>0% &lt; R \leq 5.36%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11</td>
<td>5.36% &lt; R \leq 14.29%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>10</td>
<td>R &gt; 14.29%</td>
</tr>
<tr>
<td>Class special education</td>
<td>1</td>
<td>15</td>
<td>R = 0%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7</td>
<td>0% &lt; R \leq 4.17%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11</td>
<td>4.17% &lt; R \leq 10.26%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>10</td>
<td>R &gt; 10.26%</td>
</tr>
<tr>
<td>Class Black</td>
<td>1</td>
<td>19</td>
<td>R \leq 5.27%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>9</td>
<td>5.27% &lt; R \leq 10.71%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11</td>
<td>10.71% &lt; R \leq 33.3%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>10</td>
<td>R &gt; 33.3%</td>
</tr>
<tr>
<td>Class Asian</td>
<td>1</td>
<td>15</td>
<td>R = 0%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7</td>
<td>0% &lt; R \leq 3.85%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11</td>
<td>3.85% &lt; R \leq 8.7%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>10</td>
<td>R &gt; 8.7%</td>
</tr>
<tr>
<td>Class Hispanic</td>
<td>1</td>
<td>17</td>
<td>R = 0%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>0 &lt; R \leq 3.33%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>12</td>
<td>3.33% &lt; R \leq 9.52%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>11</td>
<td>R &gt; 9.52%</td>
</tr>
<tr>
<td>Professional development hours</td>
<td>1</td>
<td>6</td>
<td>R = 65</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>8</td>
<td>65 &lt; R \leq 130</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>15</td>
<td>130 &lt; R \leq 158</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>14</td>
<td>R &gt; 158</td>
</tr>
</tbody>
</table>
suggest that HLM5’s use of listwise deletion did not have a large effect on the analysis. We summarized the differences by subtracting the HLM5 estimated slope for each student predictor from that produced by AMOS5, and then computed the median of these differences. For example, the median difference between the HLM5 and AMOS5 estimated slopes for attendance was .16, meaning that, on average, there was a very small difference among the estimated attendance slopes (< 1 NCE point).

The median difference between the HLM5 and AMOS5 estimated slopes for the gender and prior predictors was –.004 and –.04, respectively, again suggesting little impact of HLM5’s use of listwise deletion. For SES slopes, the median difference was noticeably larger at –3.6, meaning that, on average, the HLM5 estimated SES slopes under listwise deletion were smaller than those estimated by AMOS5 by 3.6 NCE points. Put another way, the SES slopes produced by HLM5 on average underestimated the effect of this predictor on Open Ended scores. It is hard to pinpoint the source of this dampening effect in HLM5 other than its use of listwise deletion, but the consequence is that the impact of SES on Open Ended scores is probably greater than suggested by the findings reported below.

We performed the same analyses with the Problem Solving and Procedures scores, and found little difference between the HLM5 and AMOS5 estimated slopes. For example, the median difference for SES slopes for Problem Solving was –.84, indicating that HLM5 on average underestimated SES slopes by less than 1 NCE point; for Procedures the median difference for SES slopes was –.07. On the whole, these results suggest that the only bias among the estimated slopes as a result of HLM5’s use of listwise deletion is for SES for the Open Ended subtest. This in turn suggests that HLM5’s use of listwise deletion did not have much effect when estimating the impact of attendance on subtest scores. We interpret the HLM results presented below accordingly.

The HLM cross-sectional results are summarized in Table 6 using a Type I error rate of \( a = .05 \) for each statistical test. Several general findings emerged across the Stanford 9 subtests. First, there was substantial between-classroom variation in the Open Ended, Problem Solving, and Procedures subtest scores with variation between 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 a statistically significant predictor of Open Ended scores, and its effect is probably underestimated. Student SES was also a significant predictor of Problem Solving scores, producing a moderate effect. 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 a difference in average classroom performance (level 2) between the large urban district and the remaining districts even when demographic variables (e.g., SES and
Table 6
Results of HLM

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>HLM results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Ended</td>
<td>There was significant between-classroom variation in means (34%).</td>
</tr>
<tr>
<td></td>
<td>1. There were significant within-classroom student effects for SES, which produced an average slope of (-3.38) and is likely an underestimate, prior mathematics knowledge (.72), and attendance (.27). Classroom Open Ended means and prior knowledge slopes varied significantly across classrooms.</td>
</tr>
<tr>
<td></td>
<td>2. 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.</td>
</tr>
<tr>
<td></td>
<td>3. 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% &lt; 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).</td>
</tr>
<tr>
<td></td>
<td>4. Significant variation in classroom Open Ended means remained unexplained.</td>
</tr>
<tr>
<td>Problem Solving</td>
<td>1. There was significant between-classroom variation in means (38%).</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>3. There was a strong district effect ((-11.4)) for classroom means, accounting for approximately 19% of the variance in classroom means.</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td>Procedures</td>
<td>1. There was significant between-classroom variation in means (34%).</td>
</tr>
<tr>
<td></td>
<td>2. The only significant within-classroom predictor was prior mathematics knowledge (.46). Procedures means showed significant variation within classrooms.</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
</tbody>
</table>
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 free or reduced 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 ($.72$), and the concentration of special education ($-1.57$) and Asian (1.28) students. The latter finding means that, with other predictors held constant, increasing the concentration of Asian students by one quartile (e.g., using Table 5, from 0% to 3.85%) 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 ($-5.1$), whereas at the classroom level there was an even larger difference between urban and 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 free or reduced 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 the concentration 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 ($.46$), whereas at the classroom level the concentrations of special education students ($-3.9$) and prior mathematics knowledge ($.53$) were significant predictors of classroom means. There were no between-classroom effects for Procedures.

In sum, the HLM cross-sectional results for the Open Ended and Problem Solving subtests were 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, and results presented above relating to the effect of
HLM5’s use of listwise deletion suggests its impact on Open Ended scores may be even more substantial. 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 free or reduced lunch. For the Procedures subtest, only prior mathematics knowledge and the concentration of special education students in a classroom were significant predictors.

SUMMARY

Descriptive Data

The descriptive results reported in Table 2 show that on the Stanford 9 subtests, which are 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; Senk & Thompson, 2003).

Although students were, on the whole, performing at or above expectations on two of the three Stanford 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 Black and Hispanic students scoring significantly lower. There were no differences among male and female students on any of the Stanford 9 subtests. There was substantial variation in average Stanford 9 performance among classrooms, however.

One set of district results is 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 NSRE, 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 NSRE and was the highest achieving and most affluent of our districts.

HLM Across-District

The HLM results for the Stanford 9 Open Ended and Problem Solving subtests indicated that prior mathematics knowledge was a consistent predictor of mathematics performance at the student and classroom levels, although the effect was modest in size. SES was a relatively strong 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.

There is long-standing documentation of the effect of prior achievement on future performance (e.g., Begle, 1973). Our analysis again suggests that prior mathematics achievement is a significant predictor (though small, less than 1 NCE point on average) of student achievement on all three Stanford 9 subtests. This finding underscores the position that achievement gaps need to be seriously addressed early in students’ academic careers.

Boaler (2003) suggested 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 Stanford 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), whereas the Procedure subtest relates only to paper-and-pencil algorithmic calculations.

High percentages of nonnative speakers, low SES students, and high percentages of minority students are commonly associated with urban schools (Grant & 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).

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 percentage 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 percentage of special education students in the classroom increases, prior knowledge plays a bigger role in predicting student achievement on Stanford 9 subtest scores.

**BROADER IMPLICATIONS**

These results indicate that those middle-grades students in the five districts discussed, each of whom have been involved with either the MathThematics or the Connected Mathematics Project (CMP) Standards-based mathematics curricula for 3 consecutive years, demonstrated achievement patterns on the Open-Ended and
Problem-Solving subtests of the Stanford 9 that exceeded the means of the national samples upon which the Stanford 9 was normed. Results on the Stanford 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 results are very promising on the NSRE, with roughly 2.5 to 4 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 means of the national norming group. It must be noted that these results were from our highest achieving district, the only district in our sample that used the NSRE.

Students in this study were not mathematically disadvantaged in the topics contained in the Open Ended and Problem Solving Stanford 9 subtests nor on the topical areas represented in the NSRE. Student performance on computational items has always been a problem area for students, teachers, parents, and educators. It seems that no one is ever fully satisfied with student performance in the procedural area. The evidence presented here does not suggest a resolution to this concern.

The Stanford 9 and the NSRE, although not the most “conservative” measures available, are primarily attuned to traditionally oriented content. There are topics that Standards-based students have studied that are either inadequately assessed or are absent altogether from the two assessments used here. There is published evidence from other studies suggesting that when a Standards-based curriculum is fully implemented with fidelity, students achieve at a rate that is significantly higher than in classrooms in which teachers simply sample, supplement, and significantly modify the Standards-based mathematics curriculum (Briars & Resnick, 2000; Senk & Thompson, 2003).

The HLM results document that when prior mathematics knowledge and several demographic variables were taken into account there continued to be significant achievement differences between the urban and suburban classrooms. 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.

**FUTURE DIRECTIONS**

*Consensus on Curriculum Content*

There presently is no agreed upon core of essential mathematics for school mathematics programs. Indeed one of the primary reasons for the “math wars” is disagreement about the role of speed and accuracy with the manipulation of standard arithmetic and algebraic algorithms. The NCTM’s *Standards* documents
Post, Harwell, Davis, Maeda, Cutler, Andersen, Kahan, and Norman (1989, 2000) have provided such a roadmap, but its implementation as embodied by the various NSF supported Standards-based curricula is a divisive issue. It is the relative balance between algorithmic mastery, applications, and understanding the more structural aspects of the subject (place value, rational number, proportionality, Cartesian space, etc.) that are the bones of contention.

Our experience has also suggested that there are operational vocabulary differences among the groups. For example, some colleagues in the mathematics department were inclined to call every type of mathematics activity “problem solving,” whereas mathematics educators tend to differentiate problem solving from algorithmic activity. Semantic clarification would serve all parties well.

Education is a cooperative venture that requires support from a broad range of school and extra-school constituencies, including of course teachers, students, parents, school administrators, and counselors. Individuals who are external to the school, such as legislators, business representatives, curriculum developers, and others involved in teacher professional development including mathematics educators and mathematicians, also must be invited to the negotiating table. This was essentially what NCTM did over the past 20 or so years as preparation for the development of the two Standards documents. It is no small task to get everyone on the same page, given our traditional interest in maintaining local control over our schools and their curricula. It is not essential that there be unanimity of agreement, but it will be important that there be a substantial core of confluence. It is our belief that additional and more concerted efforts must be made to inform the uninformed and to negotiate the parameters of desired mathematical outcomes with the various parties, as mentioned above. This seeming impasse is ironic in that all parties ostensibly want the best possible mathematics education for all students.

These research results provide yet another platform for teachers, researchers, and the public to discuss what exactly is valued in a middle school mathematics curriculum. Can schools do it all? For all persons? What computational skills are basic? Who decides what mathematics society should value? What mathematical abilities does the average citizen need to function well in our society? Is there a separate set of computational skills that future mathematics-oriented students need for success? If so, what are they?

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. (Mathematical Association of America, 2000, p. 3)

Can society afford to implement a curriculum whose primary purpose is to prepare the 4% to 5% who pursue mathematics-related careers?

Assessment Tools and Needs

Valid and reliable instruments that adequately measure the new content and processes inherent in Standards-based curricula are challenging to develop and may
be cumbersome to administer and time-consuming to score. The variety of instruments to be used in future No Child Left Behind assessments hold little 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 we value, and what are the dimensions of programs most likely to produce them?”

REFERENCES


**Authors**

Thomas R. Post, Department of Curriculum and Instruction, 175 Peik Hall, University of Minnesota, Minneapolis, MN 55455; postx001@umn.edu

Michael Harwell, Department of Educational Psychology, 323 Burton Hall, University of Minnesota, Minneapolis, MN 55455; harwe001@umn.edu

Jon D. Davis, Department of Mathematics, 4422 Everett Tower, Western Michigan University, Kalamazoo, MI 49008; jon.davis@wmich.edu

Yukiko Maeda, Department of Teacher Education, 313 Erickson Hall, Michigan State University, East Lansing, MI 48824; maed0009@msu.edu

Arnie Cutler, deceased; formerly Department of Curriculum and Instruction, 175 Peik Hall, University of Minnesota, Minneapolis, MN 55455

Edwin Andersen, Department of Curriculum and Instruction, 175 Peik Hall, University of Minnesota, Minneapolis, MN 55455; ander630@umn.edu

Jeremy A. Kahan, Ida Crown Jewish Academy, 2828 West Pratt Boulevard, Chicago, IL 60645; jkahan@icja.org

Ke Wu Norman, Department of Curriculum and Instruction, 175 Peik Hall, University of Minnesota, Minneapolis, MN 55455; kwu@umn.edu