Assessment and Grading in High School Mathematics Classrooms

Sharon L. Senk, Michigan State University
Charlene E. Beckmann, Grand Valley State University
Denisse R. Thompson, University of South Florida

The assessment and grading practices in 19 mathematics classes in 5 high schools in 3 states were studied. In each class the most frequently used assessment tools were tests and quizzes, with these determining about 77% of students' grades. In 12 classes other forms of assessment, such as written projects or interviews with students, were also used, with performance on such instruments counting for about 7% of students' grades averaged across all 19 classes. Test items generally were low level, were stated without reference to a realistic context, involved very little reasoning, and were almost never open-ended. Most test items were either neutral or inactive with respect to technology. Written projects usually involved more complex analyses or applications than tests did. The teachers' knowledge and beliefs, as well as the content and textbook of the course, influenced the characteristics of test items and other assessment instruments. Only in geometry classes did standardized tests appear to influence assessment.

The Curriculum and Evaluation Standards for School Mathematics of the National Council of Teachers of Mathematics (NCTM, 1989) both articulated a new vision of school mathematics and proposed changes in classroom assessment and program evaluation. More recently, the Assessment Standards for School Mathematics (NCTM, 1995) elaborated the vision that was described in the 1989 document and established additional criteria for student assessment. Largely because of these two documents, assessment has become a central issue in the discussion of ways to improve mathematics education in Grades K–12 in the United States. Some reports outline frameworks for changes in assessment (Mathematical Sciences Education Board, n.d., 1993; National Commission on Testing and Public Policy, 1990). Others provide guidelines to teachers on how newer forms of assessment might be incorporated into mathematics classrooms (Stenmark, 1989, 1991; Webb & Coxford, 1993). Still others examine barriers to implementing the assessment practices envisioned in the Curriculum and Evaluation Standards (Collis & Romberg, 1989; Hancock & Kilpatrick, 1993; Madaus, West, Harmon, Lomax, & Viator, 1992; Romberg & Wilson, 1992).

As suggested by Collis and Romberg (1989), Madaus, et al. (1992), and Romberg and Wilson (1992), one powerful barrier to implementing changes in mathematics
education is the standardized testing currently mandated by states and school districts. Standardized tests generally cover a very narrow range of subject matter and generally emphasize low-level thinking. Consequently, some scholars (e.g., Romberg, Zarinnia, & Williams, 1989) claim that teachers who spend the bulk of their effort teaching to current standardized tests forgo the intellectual engagement students need to develop the kind of mathematical power recommended by the National Council of Teachers of Mathematics. However, after reviewing dozens of studies on the effects of mandated testing, Hancock and Kilpatrick (1993) conclude that “the picture given by the available research is neither so bleak as the one advanced by foes of standardized, multiple-choice testing, nor so rosy as that offered by proponents of testing as the engine of school reform” (p. 169).

Even when district testing policies are relatively unobtrusive and when the secondary school mathematics curriculum is changed in the direction outlined by the *Curriculum and Evaluation Standards*, teachers’ knowledge and beliefs also affect their assessment practices. For example, in studies conducted by the University of Chicago School Mathematics Project (UCSMP), teachers who field-tested preliminary versions of UCSMP textbooks reported that, even when they felt confident teaching new content, they did not feel competent in writing exams about that new content themselves (Hedges, Stodolsky, Mathison, & Flores, 1986). Teachers also reported being troubled by issues relating assessment and grading. In particular, one teacher noted that, because the course she was teaching was much richer and more demanding than what her students had experienced earlier, her students’ grades were sometimes lower than they had been on tests covering only the simple skills on which they had been tested in the past (Usiskin et al., 1990/1993).

The discussion above suggests that changing assessment practices is a complex endeavor involving a variety of stakeholders. As the mathematics education community makes detailed recommendations for changes in assessment practices (Marshall & Thompson, 1994; NCTM, 1995), it seems prudent that current practices and the classroom teacher’s perspective on assessment be well documented. Surprisingly, when this study was begun there had been little reported research on how mathematics teachers assess what their students know and can do, what mathematics teachers think about current recommendations for changes in assessment, what factors they perceive to influence their assessment, or what problems they might encounter in trying to change current practices. Moreover, in the literature on assessment in mathematics education there has been virtually no discussion about the relation between assessment and grading.

The lack of discussion about the relation between assessment and grading in mathematics education is curious, because several studies have reported that among secondary school teachers in both the United States (Stiggins & Conklin, 1992) and Canada (Wilson, 1990), assessment is generally perceived to be summative and grading is the most frequently and consistently practiced assessment function. For instance, Stiggins and Conklin report that assessment for purposes of grading among teachers of mathematics, science, social studies, and language arts in Grades 7–12 consumes as much as a quarter of their instructional time in
some classes and almost always provides the most direct and important feedback to students on their academic progress. They also report that secondary school teachers are far less influenced by standardized tests than are elementary teachers working in self-contained classrooms.

Since the research reported here was begun, several other studies have reported on some aspects of high school teachers’ assessment practices. Cooney (1992) surveyed the assessment practices of secondary mathematics teachers in Georgia; Madaus and his colleagues (1992) surveyed teachers of mathematics in Grades 4–12 from across the United States; Taylor (1992) surveyed algebra and geometry teachers in Ohio; and Garet and Mills (1995) surveyed mathematics department heads in high schools in Illinois, Wisconsin, and Michigan about instruction in Algebra I classes. All four studies reported that teachers rely heavily on short-answer tests for assessment. Cooney, Taylor, and Madaus et al. each reported a strong influence of publishers’ assessment materials on classroom practices. According to both Cooney and Madaus et al., teachers regularly used the tests provided by their textbook’s publisher without making any modifications to them. In contrast, Taylor reported that teachers in her sample tended to use publishers’ tests more frequently as sources of questions for teacher-made tests. Both Madaus et al. and Taylor analyzed the textbook publishers’ tests and found that they reflected neither the content nor the processes recommended by the Curriculum and Evaluation Standards (NCTM, 1989). Analysis of sample test items submitted by the teachers in Cooney’s and Taylor’s samples showed similar results. Like Stiggins and Conklin (1992), Cooney reported a strong link between assessment and grading in the minds of high school teachers.

The goals of the research reported here are to document high school mathematics teachers’ practices and to understand teachers’ perspectives as they assess their students’ performance and translate the results of their assessment into grades. Thus, this study examines some of the same issues addressed in the recent surveys by Cooney (1992), Taylor (1992), Madaus et al. (1992), and Garet and Mills (1995). However, it looks at teachers’ assessment practices in much more detail and over a longer period than any previous study. It also examines relations between assessment and grading, and it analyzes teachers’ perspectives on their practices.

There are four major research questions:

1. What types of instruments do high school mathematics teachers use for assessment in their classrooms?

2. How do teachers use these instruments to determine students’ grades?

3. To what extent are teachers’ assessment and grading practices consistent with the recommendations about assessment in the Curriculum and Evaluation Standards (NCTM, 1989)? In particular, to what extent have teachers implemented recommendations regarding the use of multiple assessment techniques, the use of calculators and computers in assessment, and the assessment of communication and reasoning? To what extent do assessment instruments use problems set in realistic contexts? To what extent do
they emphasize exercises or word problems requiring only one or two steps?  

4. What factors account for teachers’ current assessment and grading practices? In particular, what influences do standardized tests, the subject matter, the textbook, or the available technology have on assessment and grading practices?

METHOD

Sample

_Schools._ Schools were recruited to represent a variety of socioeconomic and educational opportunities. Because the researchers wanted to probe for teachers’ use of alternative forms of assessment, schools in which teachers were using newer forms of assessment, such as oral or written reports, were actively sought. In order to examine the extent to which assessment instruments were aligned with recommendations regarding the use of technology, schools where there was some evidence of the use of calculators and computers in mathematics classes were also recruited. Ultimately, five schools in three states were chosen to participate in this project. In this study the schools are called Highpoint, Roan Valley, St. Bridgette’s, Broadview, and Jackson High Schools.

Highpoint High School is located in an upper-middle-class suburb near a very large midwestern city. About 90% of the students are White; the largest minority group is Asian American. Virtually all of Highpoint’s graduates attend college, with most attending 4-year colleges. Scores on college entrance exams are well above the national average, and many students earn Advanced Placement credit in calculus. At the time of the study the district had recently implemented a mandate to incorporate writing in all school subjects. Soon after data collection was completed, Highpoint’s mathematics department received a national award for its high-quality program.

Roan Valley is located in an affluent suburb near a large city in the midwest. All but a few students are White. Students are virtually all college-intending, and as is true at Highpoint, college entrance test scores are above the national average. Advanced Placement calculus is offered.

St. Bridgette’s is a small Catholic school in a medium-sized midwestern city. Families are generally less affluent than those of students at either Highpoint or Roan Valley, but virtually all students are college bound. Calculus is not currently offered.

Broadview and Jackson High Schools are located in the same large county in the South. The county includes urban and suburban areas. A busing plan ensures that no school population is more than 30% minority. The largest minority group in each school is African American. For several years prior to the study, the district had offered summer in-service courses on such topics as the use of applications in teaching algebra and the use of calculators and computers in teaching mathematics.

---

1This study was begun before the _Working Draft of the Assessment Standards_ (NCTM, 1993) had been released. Hence, these standards did not influence this study’s design.

2All names of schools and teachers in this article are fictitious.
Broadview draws students from middle-income families from mostly suburban neighborhoods. College entrance scores at Broadview are at about the national average and Advanced Placement calculus is offered. Jackson High School’s students are mostly from lower- to middle-income families. The school does not offer Advanced Placement calculus, but it has a magnet program for the arts. The average SAT-M (Scholastic Aptitude Test—Mathematics) score at Jackson is about 50 points below the national average.

**Courses and teachers.** In this study, assessment was considered only in courses in which the majority of the content comprised topics in the *Curriculum and Evaluation Standards* for Grades 9–12 (NCTM, 1989). That is, the assessment practices of teachers of courses called algebra, geometry, advanced algebra, trigonometry, functions, and precalculus were studied, whether they were considered remedial, average, or honors by the school. Assessment was not studied in general mathematics, prealgebra, or calculus classes.

From among those teachers within each school who volunteered to participate in this study, one teacher was chosen from each core course as defined above. Twenty-two teachers began the study, but 3 were dropped either because they failed to supply most of the data requested, or because the assessment instruments they used were identical to those used by another teacher in the study. This article reports on the 19 teachers who supplied most of the data requested. Eight teachers are from Highpoint, 4 are from Broadview, 3 are from Jackson, and 2 each are from Roan Valley and St. Bridgette’s.

The sample consists of 8 men and 11 women. Their number of years of experience teaching mathematics ranges from 4 to 27, with a mean of 18. All but 2 of the teachers have at least one degree in mathematics or mathematics education; 13 have masters’ degrees, and one of these also has a Ph.D. in mathematics education. In contrast, Weiss (1994) reports that only 63% of mathematics teachers in Grades 9–12 have either an undergraduate or graduate major in their field.

Many teachers in the sample participate in professional activities. One teacher at Highpoint had been an officer in local and state professional organizations and is a leader in curriculum and research projects. One teacher at Roan Valley often conducted workshops in the midwest on uses of technology. The mathematics teachers at St. Bridgette’s regularly attended local and regional mathematics conferences and workshops, as did several teachers at each of Broadview, Jackson, and Highpoint. Thus, with respect to level of professional activity, the teachers in the sample are also above the national average (Weiss, 1994).

---

3In three schools the Department Head or administrator agreed to have his or her school participate in the study, then either asked for volunteers or assigned teachers to participate. In two schools one of the researchers contacted teachers directly to solicit volunteers, then obtained permission from administrators.

4Each teacher from Roan Valley and St. Bridgette’s provided data for two classes. However, only one course for each teacher was analyzed for this study.
Textbooks. Each school in the study used some materials in the vein of curriculum reform. In particular, textbooks developed by the University of Chicago School Mathematics Project (UCSMP) were used in at least one class in each study school (McConnell et al., 1990; Peressini et al., 1992; Rubenstein et al., 1992; Senk et al., 1990; Usiskin, Coxford, & Hirschhorn, 1991). Among the classes in the study, 12 used UCSMP texts and 7 used other texts. One of the other texts (Foerster, 1986), like the UCSMP materials, is known for its applications of mathematics. The remaining texts covered fairly traditional content, even though some were recent editions.

The textbooks developed by UCSMP for Algebra I, Geometry, and Algebra II assume that scientific calculators are available to students at all times. The UCSMP post–Algebra II textbooks assume that graphics technology (either a graphing calculator or function-graphing computer software) is also available. Several other texts used by teachers in the sample recommend, but do not assume, the use of calculators.

Technology. During the course of the study, each teacher in the sample had some access to technology. However, the nature and extent of the access varied considerably across schools.

When selected for the study, Highpoint had several computer laboratories, but because of high demand, they were often not available to mathematics teachers. During the study a laboratory with Macintosh and NEXT computers was installed for the exclusive use of the mathematics department. The department also had several sets of graphing calculators for teachers to use. Mathematics classrooms generally did not have computers.

During the study each mathematics teacher in Roan Valley had a Macintosh computer and video monitor in the classroom for demonstration, and access to Macintosh computers in a laboratory. Classroom sets of graphing calculators were also available for instruction.

The teachers at St. Bridgette’s shared a classroom set of graphing calculators and had occasional access to a computer laboratory containing Apple IIe and Macintosh computers.

The mathematics departments of Broadview and Jackson owned some scientific calculators for use by teachers and students, and each school had a computer laboratory for instructional use. At Broadview at least one teacher in the study had a computer in her classroom for demonstration. Teachers at Broadview also had access to several sets of graphing calculators for classroom use.

Among the teachers in the sample, 11 required calculators, 6 recommended them, one permitted their use, and one teacher’s use of technology was not clear. No teacher prohibited the use of calculators. In four of the five post–Algebra II classes graphing calculators were required. In all other classes in which a calculator was required, recommended, or permitted, a scientific calculator was listed. Computers were required by only two teachers, both post–Algebra II teachers at Highpoint.

Data Sources

Data were collected from three sources: questionnaires completed by the teachers, the assessment instruments used in each target class, and an interview with each teacher.
Questionnaires. At the end of the first marking period each teacher was asked to complete a questionnaire eliciting information about: (a) her or his academic background and experience, (b) the target course (including content goals and access to technology), and (c) her or his assessment goals and grading practices. Questionnaires administered during the year, generally one at the end of each marking period, probed for further information about assessment and grading practices and potential influences on those practices. Descriptive statistics were calculated for quantitative data. Responses to other questions were transcribed and analyzed qualitatively.

Tests and quizzes. At the end of each marking period, each teacher in the study was asked to supply a copy of each written assessment instrument used in the target class. Upon receipt of the instruments they were sorted by type (e.g., test, quiz, written report). Teachers’ responses to the initial questionnaire indicated that written tests were the primary assessment tool in each class. Examination of the tests and quizzes submitted indicated that except for the number of items on each instrument, the tests and quizzes were quite similar in content and format. Hence, only the tests were analyzed. Reported here are the analyses of all tests given by the teachers in the sample during the first semester of the 1991–92 school year.

Categories and coding schemes were developed for individual items on the basis of the terminology used in the *Curriculum and Evaluation Standards* (NCTM, 1989) and other recent reports on assessment. Descriptions of the coding categories used in this study appear in Table 1.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item format</td>
<td>Answer given</td>
<td>True or false, yes or no, multiple choice, or matching</td>
</tr>
<tr>
<td></td>
<td>Answer not given</td>
<td>Fill in the blank or longer free response</td>
</tr>
<tr>
<td>Skill</td>
<td>Yes</td>
<td>Solution requires applying a well-known algorithm such as solving equations or inequalities or bisecting an angle. Item does not require translation between representations.</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>No algorithm is generally taught for answering such questions, or item requires translation across representations.</td>
</tr>
<tr>
<td>Level</td>
<td>Low</td>
<td>A typical student in that course would use 1 or 2 steps to solve.</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>A typical student in that course would use 3 or more steps to solve, and the content is new to the course.</td>
</tr>
<tr>
<td>Realistic context</td>
<td>Yes</td>
<td>The item is set in a context outside of mathematics (e.g. art, fantasy, science, sports).</td>
</tr>
</tbody>
</table>

(table continued)
The code for item format was included in order to examine the extent to which teachers used multiple-choice or other answer-given formats. Recommendations in the *Curriculum and Evaluation Standards* (NCTM, 1989) to decrease reliance on tests using exercises requiring only one or two skills suggested to us the inclusion of codes for level of complexity of the item and for whether or not only skill was assessed. The standard on connections led to codes for the presence or absence of a realistic context and for the role of figures or diagrams. The standard on reasoning led to a code for deductive reasoning. Arguments for decreased emphasis on questions that force students into a predetermined solution strategy suggested a code for whether or not an item was open-ended. Finally, the recommendation for an increased emphasis on using calculators and computers in assessment led to an examination of the impact of technology on potential solutions to each item.

Two codes for technology dependence, each using the options *technology active*, *technology neutral*, or *technology inactive* developed by Harvey (1992), were assigned to each item. The first resulted from an analysis of the item with respect to the numerical features found on a typical scientific calculator, including keys for trigonometric and logarithmic functions. The second resulted from an analysis of the item with respect to the potential impact of technology with graphics or drawing features. This might be a graphing calculator, function-graphing software, or software for drawing geometric figures. Each code was assigned without regard to the teacher’s assumptions about technology or to the students’ access to technology.
Most codes were independent of the course in which the item was given. However, coding for level was an exception. If an item required only one or two steps to solve, it was coded as low. If an item required more than two steps to solve and if it covered content new to the course, it was coded as other. However, if an item was considered to test a routine prerequisite for a particular course, it was coded as low, even though it might have required more than two steps to solve. For instance, a linear equation of the form $ax + b = cx + d$, requiring more than two steps to solve, would be coded other in Algebra I but would be coded low in courses beyond Algebra I. Level is the only code that takes into account the course in which the item is being used.

Six items from the tests used by the teachers in the sample are given in Figure 1. The codes assigned to each item are given in Table 2. Because none of these items was considered open, the category open is not listed in the table.

1. Gunther’s corporation has a charity matching program. For every $20 Gunther gives to charity, his corporation will give $25. If Gunther gives $400 to a charity for the homeless, how much will the corporation give?

2. If $r_m(A) = B$ and $r_m(C) = D$, then which of the following is not justifiable?
   a. $AC = BD$
   b. $m$ bisects $BC$
   c. $r_m(B) = A$
   d. $m \perp AB$
   e. All are justifiable

3. Given: $\overline{AD} \parallel \overline{BC}$
   E is the midpoint of $AC$
   Prove $\triangle AED \cong \triangle CEB$

4. The matrix below gives figures for paid circulation of several papers on 31 March 1986:

<table>
<thead>
<tr>
<th>Paper</th>
<th>Morning</th>
<th>Sunday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chicago Tribune</td>
<td>760,031</td>
<td>1,163,083</td>
</tr>
<tr>
<td>New York Times</td>
<td>1,035,426</td>
<td>1,625,649</td>
</tr>
<tr>
<td>Washington Post</td>
<td>781,371</td>
<td>1,091,307</td>
</tr>
<tr>
<td>Los Angeles Times</td>
<td>1,088,155</td>
<td>1,353,376</td>
</tr>
</tbody>
</table>

   If each paper wanted to triple its paid circulation, what matrix describes the new circulation figures?

5. Simplify $\frac{2x^{-1}}{y^2} \left(\frac{x^2}{y}\right)^{-1}$

6. Graph the function $y = 2 \cos\left(2x - \frac{\pi}{2}\right) + 1$

*Figure 1. Sample test items, each from a different teacher*
Table 2

*Coding for Sample Items in Figure 1*

<table>
<thead>
<tr>
<th>Item</th>
<th>Course</th>
<th>Format</th>
<th>Skill</th>
<th>Level</th>
<th>Realistic context</th>
<th>Reasoning required</th>
<th>Role of diagram</th>
<th>Technology</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Algebra</td>
<td>Answer not given</td>
<td>No</td>
<td>Other</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>Neutral</td>
<td>Inactive</td>
</tr>
<tr>
<td>2</td>
<td>Geometry</td>
<td>Answer given</td>
<td>No</td>
<td>Other</td>
<td>No</td>
<td>No</td>
<td>None</td>
<td>Inactive</td>
<td>Neutral</td>
</tr>
<tr>
<td>3</td>
<td>Geometry</td>
<td>Answer not given</td>
<td>No</td>
<td>Other</td>
<td>Yes</td>
<td>Yes</td>
<td>Interpret</td>
<td>Neutral</td>
<td>Neutral</td>
</tr>
<tr>
<td>4</td>
<td>Algebra II</td>
<td>Answer not given</td>
<td>No</td>
<td>Low</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>Active</td>
<td>Inactive</td>
</tr>
<tr>
<td>5</td>
<td>Algebra II</td>
<td>Answer not given</td>
<td>Yes</td>
<td>Low</td>
<td>No</td>
<td>No</td>
<td>None</td>
<td>Inactive</td>
<td>Inactive</td>
</tr>
<tr>
<td>6</td>
<td>Post-Algebra</td>
<td>Answer not given</td>
<td>No</td>
<td>Other</td>
<td>No</td>
<td>No</td>
<td>Make</td>
<td>Neutral</td>
<td>Active</td>
</tr>
</tbody>
</table>
Each item was coded independently and blindly by two researchers. The overall mean percentage agreement on coding was 89%, with the following percentage agreements for each code: item format (88%), skill (82%), level (75%), context (98%), role of diagram (91%), reasoning (97%), open-ended (99%), scientific calculator technology (79%), and graphing/drawing technology (85%). When two raters disagreed, the third rater resolved the coding.

After each item on a test was coded, the percentage of items assigned each code was calculated for the test. Then for all tests given by each teacher during the semester, the mean percentages of items receiving each code were calculated.

Other assessment instruments. Other assessment instruments were categorized by type—for instance, written report or computer laboratory experiment—on the basis of the teacher’s self-report. The content of each instrument and a brief analysis of the extent to which it reflected recommendations in the NCTM Standards (1989) were also prepared. Because these instruments varied widely in style and substance and because each had fewer questions than a typical test, the analysis was less detailed than the analysis of the classroom tests.

Interviews. At the end of the school year each teacher was interviewed by one of the researchers. Most interviews lasted from 30 to 60 minutes. Interviews were audiotaped and transcribed. Each researcher used the same script but was free to modify it if modification seemed reasonable. Thus, although not all interviews touched on exactly the same points, each provided additional information about the teacher’s assessment and grading goals and practices.

RESULTS

Instruments Used for Assessment and Grading

Although most recent reform documents separate assessment from grading, 11 teachers (58%) graded all assessment tasks. The 8 teachers who said that they assessed some work that was not graded gave as examples certain activities done in class using calculators or computers, or problems to be solved individually or in small groups. As Ms. K noted, “Kids don’t like to do things that aren’t graded.” Mr. Q explained why some students might not be willing to do assignments that aren’t graded:

Our students are good “crap detectors.” They have terribly tight schedules. The kids I had were the most active in the school in terms of taking courses …. They were taking physics or chemistry and so on. If I were a kid in their shoes, I would get a sense of what was really going to be graded, and what wasn’t, and put the thing that wasn’t graded down low on the priorities. So I tend to have things that I grade, and if it is a requirement, if it’s an assignment, it’s going to be graded.

Each teacher used a variety of assessment techniques. Table 3 lists the types of assessment instruments used by each teacher during the first semester and the average weight per semester given to each type of assessment.

---

6At Broadview High three teachers chose not to be taped, so only the notes taken by the researcher during the interview were transcribed.
Table 3
Percentage of Grade from Various Types of Assessment Instruments by Teacher and Course

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Algebra I</th>
<th>Geometry</th>
<th>Algebra II</th>
<th>Post–Algebra II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tests</td>
<td>45</td>
<td>60</td>
<td>72</td>
<td>50</td>
</tr>
<tr>
<td>Quizzes</td>
<td>35</td>
<td>25</td>
<td>17</td>
<td>25</td>
</tr>
<tr>
<td>Homework</td>
<td>20</td>
<td>15</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Written reports</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

a Ms. C, Mr. F, and Ms. L reported percents that did not sum to 100%. b Initial questionnaire data for Ms. D were lost. Assessment instruments and interview revealed tests, quizzes, homework, and written reports were used to determine students’ grades. c Portfolios and personal interviews. d Oral reports, notebooks and class participation. e Oral reports and interviews. f Group problem-solving activities. g Classwork problems.
Every teacher gave tests, generally one at the end of each chapter in the text. Ten teachers also gave cumulative tests either at the end of the marking period or at the end of the semester. The mean number of tests given per teacher per semester was 6.8.

In all classes, performance on written tests or quizzes was the primary determinant of students’ grades, with these two types of assessment accounting for between 50% and 100% of a student’s grade. The mean weight given to tests and quizzes was 77%, about 6% higher than the weight given by the mathematics teachers in Grades 7–12 surveyed by Cooney (1992). Although every teacher relied heavily on tests and quizzes, there was considerable variability in the relative weight given to each type of instrument. The weight given to tests ranged from 25% to 80%, with a mean of 55%. Similar variety in grading practices was also noted by Cooney (1992).

For this study, homework was the third most important contributor to students’ grades, with an average weight of 15%. Seven teachers relied exclusively on written tests, quizzes, and homework papers for determining grades. Twelve teachers from four of the five schools used some other form of assessment. Among the 12 were 7 from Highpoint, where all teachers are required to give a grade in writing.

The other types of assessment used were written reports \((n = 8)\), oral reports \((n = 2)\), interviews with students \((n = 2)\), portfolios \((n = 1)\), notebooks \((n = 1)\), group problem-solving \((n = 1)\), classwork problems \((n = 1)\), and class participation \((n = 1)\). For the entire sample the mean weight given to these other forms of assessment was 7%. Among the 12 teachers who used something other than tests, quizzes, and homework to determine grades, the mean weight of other forms of assessment was 12%.

During the interviews 15 teachers were asked, “Of all the instruments you used to assess students’ performance during the year, which were the most informative to you, and why?” They answered as follows: tests \((n = 5)\), writing assignments \((n = 4)\), quizzes \((n = 2)\), notebooks \((n = 1)\), personal interviews \((n = 1)\), and “it depends” \((n = 2)\). Eight of the respondents had used at least one of written or oral reports, interviews, or portfolios during the year. Five of these eight teachers listed one of the newer forms of assessment as most informative. Below are some rationales for the teachers’ responses.

I think quizzes told me the most about how things were going on a day-to-day basis in the class. I think that the tests told me a lot about the student’s ability to capture large groups of information and synthesize it …. But, I think daily homework tells me more about the kid’s willingness to participate. [Mr. M]

Quizzes were more limited because they focused on a smaller section of material …. I guess, the projects, I thought, were more interesting. They really helped me see a little bit where their interests lie. But I suppose the tests would tell me more about what they actually learned. [Ms. P]

The writing assignments are the ones that have given me a little bit different insight to what the student knows …. The writing assignments I’ve gotten have given me a much better indication of who really understands the processes, because they can explain them to me. [Mr. O]

I’d have to say those personal interviews, hands down. I guess if you’re writing a test, you can just put the answer down … but when you’re talking to them, one-to-one, you know whether they know what they are talking about or not. [Ms. G]
Characteristics of Assessment Instruments

Tests. Table 4 summarizes the results of the coding of the test items. It describes the extent to which the teachers’ test items reflect many of the changes recommended in the Curriculum and Evaluation Standards (NCTM, 1989) and other reform documents.

The use of answer-given formats (e.g., true/false, multiple-choice, or matching) by teachers in the sample ranged from 3% to 42%, with a mean of 19%. No consistent pattern was found in the use of an answer-given format in terms of course or text. However, the teachers who used answer-given formats most frequently on chapter tests also used such items on their end-of-quarter or semester exams. During the interviews the department head at Highpoint reported that often the end-of-term tests must be graded by the next day. This rapid turnaround time seemed to put pressure on teachers to use mostly multiple-choice items or items with short numerical responses that can be scored by people or machines in a very short time.

The emphasis on skill also varied considerably across teachers’ tests. The percentage of items coded as skill ranged from 4% to 80%, with a mean of 36%. In general, geometry teachers had fewer skill items on their tests than teachers of other subjects.

Most items on the tests were low level. That is, most questions either tested new content that could be answered in one or two steps or tested prerequisite knowledge. The percentage of low-level items on teachers’ tests ranged from 53% to 90%, with a mean of 68%. The percentage of low-level items was rather consistent across courses.

One of the most consistent recommendations for reform in mathematics education has been to put mathematics in context, that is, to relate mathematics to the world in which the student lives. The results in Table 4 show that there was a great deal of variety in the extent to which teachers’ tests reflected this recommendation. The percentage of test items with realistic contexts ranged from 0% to 41%, with a mean of 17%. In eight classes the percent of questions set in a context exceeded 20%. In contrast, in five classes no more than 5% of the test items throughout the semester were set in any context.

In general, the use of realistic contexts appears to be related to two factors: the text used and the course content. Those teachers who gave test questions set in realistic contexts most frequently were those whose textbooks supported this recommendation. Specifically, each of the teachers in the 10 classes with the highest percentage of test items set in realistic contexts used a textbook noted for emphasizing applications of mathematics. Overall, the percentage of items set in a realistic context was lower in geometry than in other mathematics classes.

There is very little evidence that the test items required students to use reasoning. The percentage of items involving reasoning ranged from 0% (four teachers) to 15%, with a mean of 5%. The three teachers with the highest percentage of items requiring explanations, justifications, or proofs taught geometry, the school mathematics course that traditionally has emphasized proof-writing. Some evidence of reasoning was also found in the post-Algebra II courses taught by Mr. O, Ms. P, and Mr. Q. The reasoning items on their tests involved such items as explaining why a function is or is not even, proving trigonometric identities, or justifying number-theoretic arguments. There was virtually no evidence of reasoning being assessed on tests in Algebra I or II classes.
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Algebra I</th>
<th></th>
<th>Geometry</th>
<th></th>
<th>Algebra II</th>
<th></th>
<th>Post–Algebra II</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ms.</td>
<td>Mr.</td>
<td>Ms.</td>
<td>Mr.</td>
<td>Ms.</td>
<td>Mr.</td>
<td>Ms.</td>
<td>Mr.</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>Item format</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Answer given</td>
<td>11</td>
<td>10</td>
<td>22</td>
<td>13</td>
<td>6</td>
<td>32</td>
<td>31</td>
<td>10</td>
</tr>
<tr>
<td>Skill</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>50</td>
<td>36</td>
<td>33</td>
<td>48</td>
<td>80</td>
<td>23</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>78</td>
<td>68</td>
<td>79</td>
<td>69</td>
<td>79</td>
<td>90</td>
<td>80</td>
<td>53</td>
</tr>
<tr>
<td>Realistic context</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>28</td>
<td>32</td>
<td>33</td>
<td>12</td>
<td>5</td>
<td>9</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Reasoning required</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>Open-ended</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Role of diagram</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpret</td>
<td>12</td>
<td>17</td>
<td>19</td>
<td>0</td>
<td>1</td>
<td>44</td>
<td>36</td>
<td>47</td>
</tr>
<tr>
<td>Superfluous</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Make</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td>Number of tests coded</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>
Virtually no teachers used open-ended items on tests. The percentage of open-ended items on tests ranged from 0% to 10%, with a mean of 2%. Occasionally, Ms. H, a geometry teacher, and Ms. P, a teacher of a post–Algebra II course, asked open-ended questions. Often they were relatively simple items asking students to draw examples of certain figures or to give examples of certain types of statements or data sets.

The standard on connections (NCTM, 1989) calls for emphasis on different but equivalent representations of concepts and procedures and for translation among different representations. Table 4 contains information about the extent to which teachers in the sample used graphical representations on their tests. Again, there was wide variation among the classes. The percentage of items in which students had to interpret a diagram given in the question ranged from 0% to 67%, with a mean of 22%. The percentage of items in which a superfluous diagram was given ranged from 0% to 15%, with a mean of 3%. The percentage of items in which students had to make a graph or figure ranged from 0% to 19%, with a mean of 7%.

Visual representation, particularly interpretation of diagrams, was more evident in geometry than in other courses. However, in all five geometry classes these were generally low-level questions involving the identification of parts of figures. The tests given by Ms. G and Ms. H required students to make graphs or draw figures more frequently than did the tests given by the other three geometry teachers. In general, these were items asking students to use a protractor to draw angles or to use a ruler and compass to construct certain geometric figures.

Among the 14 Algebra I, Algebra II, and post–Algebra II teachers, 4 (Ms. D, Ms. E, Mr. N, and Mr. S) seldom asked their students to interpret or to draw a graph or diagram. Again the influence of the textbook was strong. These teachers all used texts in which visual representation played a very small role, or in which most use of graphs occurred in the second half of the texts. In contrast, in those classes in which drawing or interpreting graphs was tested more frequently, the textbooks emphasized visual representation.

The extent to which each item depended on technology is given in Table 5. In general, the use of technology-active items on tests was low. With respect to the features of scientific calculators, the percentage of calculator-active items ranged from 0% (seven teachers) to 40%, with a mean of 8%. Only four teachers gave tests with more than 15% calculator-active items. All four taught at the post–Algebra II level. In the geometry classes most items were calculator-inactive. In the other classes most items were neutral as related to scientific calculators.

The demand for technology on test items was strongly influenced by content. Items involving trigonometry, logarithms, very large or very small exponents, and statistical calculations were often technology-active. Mr. R and Mr. S, the two teachers whose tests contained the most calculator-active items, taught courses that were almost entirely devoted to trigonometry during the first semester. The classes taught by Mr. O and Ms. P, who were the next most likely to use calculator-active items, included a substantial amount of statistics during the first semester.

Test items that actively encouraged the use of graphing or drawing technology were even less common than those that encouraged the use of scientific calculators.
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Algebra I</th>
<th>Geometry</th>
<th>Algebra II</th>
<th>Post–Algebra II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ms.</td>
<td>Mr.</td>
<td>Ms.</td>
<td>Ms.</td>
</tr>
<tr>
<td>Scientific calculator</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>Active</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Neutral</td>
<td>54</td>
<td>53</td>
<td>58</td>
<td>80</td>
</tr>
<tr>
<td>Inactive</td>
<td>39</td>
<td>40</td>
<td>34</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>Graphing technology</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Active</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Neutral</td>
<td>99</td>
<td>100</td>
<td>99</td>
<td>78</td>
</tr>
<tr>
<td>Inactive</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 5: Percentage of First Semester Test Items Dependent on Technology by Teacher
The percentage of graphing or drawing technology-active items ranged from 0% (nine teachers) to 16%, with a mean of 3%. Only one teacher (Ms. P) gave tests with more than 15% graphing technology-active items. Another five teachers gave graphing technology-active items either 7% or 8% of the time. They were the other four post–Algebra II teachers, and Ms. L, an Algebra II teacher. The rest of the sample virtually never used items to take advantage of graphing or drawing features, with the mean percentage of such technology-inactive items being 86%.

**Other assessment instruments.** Eight teachers described or submitted copies of a total of 30 assessment instruments other than tests or quizzes that were used to determine students’ grades during the semester. Five teachers supplied only 1 or 2 assignments each. Three teachers described or provided at least 3 such other assessment instruments.

Ms. E, an Algebra I teacher, gave six writing assignments during the first semester. Each included the objective, a description of the task to be accomplished, a sample solution to a similar assignment, and the criteria to be used for grading the work. Each writing assignment was to be one page and to be done by an individual. Assignments included comparing and contrasting an expression and an equation, making up and solving word problems, and explaining in detail how to solve a given linear inequality.

Mr. M, an Algebra II teacher, gave seven alternate assessment tasks: two writing assignments, two group quizzes, and three other group activities. These included determining the weight of a box by using a calibrated seesaw, finding equations to model given data sets, and searching for Fibonacci numbers on a pineapple.

Ms. P, a post–Algebra II teacher, assigned seven nontraditional tasks: one written project in each of the three marking periods, which were to be chosen from those in the textbook or from projects she created; a 5-minute oral presentation to the class about one of the projects; two group problem-solving assignments; and one other assignment involving analysis of tables or graphs from a newspaper. The projects available to students included simulating a capture-recapture experiment, determining a model for the relation between neck and wrist sizes of students in the class, and exploring the locus of vertices of parabolas with equations of the form $y = ax^2 + bx + c$, where $a$ and $b$ are constant and $c$ varies.

In general, each of the instruments identified by the eight teachers as other than a test, quiz, or homework paper covered a single concept or situation. Overall, these tasks asked higher-level questions than those asked on the tests. Several projects involved using physical materials. Although most tasks were highly structured, several provided opportunities for students to choose options, and several were somewhat open-ended.

**Factors Influencing Assessment**

The results reported above indicate numerous ways in which the subject matter and the textbook influenced the nature of the items on the teachers’ tests. A more detailed look at technology and some other factors influencing teachers’ assessment and grading practices is presented below.

**Technology.** Of the 18 teachers who described the frequency and nature of their use of calculators or computers in the target class, 12 (67%) said their students used
calculators extensively or virtually every day, and 5 (28%) said calculators were used whenever needed. Overall, the geometry teachers seemed to report using calculators less frequently than teachers of other courses. Seven teachers (39%), distributed among all four course levels, said they used computers during the first marking period. For three of the seven, computer use was limited to demonstrations. In every case teachers reported using computers far less frequently than calculators, with computer use ranging from once during the marking period to once every 2 weeks. These results are consistent with the use of technology in Algebra I classes reported by Garet and Mills (1995).

Every teacher permitted students to use calculators during tests. Generally, when graphing calculators were available for instruction, they were also available for tests. Only one teacher mentioned using computers for testing. Nine teachers gave specific examples of how technology was used in assessment. The comments below are representative.

We used calculators in class every day, for computation, for attempting to find solutions by trial and error, and for verifying solutions …. Students have not been graded specifically on their use of technology. [Ms. E]

Calculators are used daily for computations. Two computer activities using Geo-Explorer were done in Chapter 4 …. Students were permitted to use an automatic drawing program on homework assignments. A few students used the computer to answer test questions, but most did not, because of the time needed to get a hard copy. We have only one printer. [Ms. H]

Once or twice a week we used graphing calculators for linear relations and scatter plots—getting used to setting range, etc. (i.e., changing viewing window and zooming). We used the computer to illustrate any BASIC programs that were shown in the text, for example, for recursion and explicit formulas. [Ms. L]

During the interview at the end of the school year, teachers were asked, “How has the students’ access to technology influenced your assessment?” Of the 12 teachers who responded to this question, only Ms. K suggested that technology had not changed her assessment practices. By this she seemed to mean that she used the textbook’s published tests and hence was not creating any assessment instruments. She claimed that “all tests were geared for technology already.”

All other teachers stated that students’ access to technology had influenced assessment, and in many cases their responses to this question were very detailed. Teachers of all subjects pointed out that technology allowed them to deal with situations that would have involved tedious calculations if no technology had been available. They explained that “not-so-nice,” “nasty,” or “awkward” numbers arise from the need to find the slope of a line, the volume of a silo, the future value of an investment, or the 10th roots of a complex number. Additionally, some teachers of Algebra II or post–Algebra II classes noted how technology influenced them to ask new types of questions, how it influenced the production of assessment instruments, and how it raised questions about the accuracy of results.

I think you have to ask different kinds of things …. When we did trigonometry, you just can’t ask them to graph \( y = 2 \sin x \) or something like that. Because their calculator can do that for them … I do a lot of going the other way around. I do the graph, and
they write the equation … The thing I think of most that has changed is just the topic of trigonometry in general. It’s a lot more application type things … given some situation, an application that would be modeled by a trigonometric equation or something like that. [Ms. P]

I use it [the computer] to create the papers, and I can do more things with it … not just hand-sketched things. I can pull in a nice polynomial graph from Mathematica, put it on the page, and ask them questions about it. So, in that way, it’s had a dramatic effect on me personally …. We did talk about problems with technology. Sometimes it doesn’t tell you the whole story. And sometimes it fails to show you the right graph. If you do the tangent graph on the TI-81, you see the asymptotes first. You know, that’s really an error. It’s not the asymptote. [Mr. M]

Textbook publishers’ tests. Twice during the year teachers were asked to identify the sources they used to make up tests and other assessment instruments. Eleven teachers responded to these items. Six teachers said that their main source for the first test of the third marking period was a test provided as an ancillary by the publisher of their text. Five teachers listed themselves as the principal source of the test items, but they each consulted a published test in writing their questions. Eight of the 11 teachers relied only on the publisher’s or their own ideas for testing. The other 3 also consulted colleagues or other commercially available materials when creating tests.

Teachers reported that they planned to use more varied sources in making up the final exam than they had used in making up the first test of the third marking period. Nine teachers expected to use at least two sources. Six teachers expected the publisher’s ancillary materials to be their main source, one planned to use other commercially available materials, two expected to rely primarily on colleagues, one planned to write the test himself, and one teacher planned to use last year’s final exam.

Some tests had a copyright notice from the textbook publisher, indicating the source of that particular instrument. Seven teachers regularly used tests provided by the publishers without making any changes; three teachers volunteered that they regularly modified published tests, and nine generally made up their own tests. Teachers who regularly used published tests gave reasons such as the following:

Convenience. This is the first time through the book. [Ms. K]

The tests in the ancillaries give us an advantage of keeping us honest and matching the four dimensions of learning that are reflected in the students’ review and the students’ self-test …. I found the ancillaries a necessary resource because I could not have written a test from scratch. I could have in previous years in a more traditional course. [Mr. Q]

When asked if the tests provided by the publisher had been used in the past, most teachers said “no,” saying that in the past either they did not have copies of publisher’s test or that when they did, the tests weren’t nearly as well written.

Teachers who either modified the tests published for their textbook or who made up their own tests gave the following rationales and methods:

Basically I go through and use the homework problems the kids have and from them take problems similar to that [and] use them as the basis of the test …. If it’s a problem that tends to be one we’ve spent a lot of time with, then I’ll make sure that I place it on the test someplace, too. Give them new numbers; change them around. [Mr. F]

When I make up a test I mostly make it up from the problems that were on the assignments, and I try to put … the majority of the problems are problems that every student
should be able to answer. And then I try to always put maybe a small percentage of problems that would be for the better students. And I usually include bonus problems from the more difficult sections that we covered. [Ms. J]

I usually look through the sections we’ve done and they have done all the odd problems, which they have all the answers for. And I have done some of the even ones in class, but quite often I’ll take some of the even ones, either the same problem or a problem similar to those and just create my own …. And looking at each section, if there’s like, I think five or six sections in the chapter, I try to get an equal number from each section … being careful to word the directions almost exactly as the book did so they’ll be familiar with the terminology and not questioning what do I mean. [Mr. S]

**Standardized tests.** During the interview teachers were asked about the extent to which they take into account what students might be asked on standardized exams such as the SAT, the American College Testing (ACT) program, or state proficiency exams when they write their own test questions. Of the 15 teachers who responded to this question, 12 reported little or no influence of standardized tests on their teaching or testing.

The only teachers who claimed that standardized tests were a very important influence on their classes were three geometry teachers. However, when asked how these exams influenced either the content taught or the form of questions asked in their courses, these teachers generally cited relatively minor adjustments.

I don’t think I change my questions particularly. But I try to make sure I teach the things that, you know, the Pythagorean Theorem, the basic things. I always say, you know, “If you’ve taken high school geometry they’re expecting you to know this.” [Ms. G]

Basically, I would say … the questions are traditional, asked in a traditional way …. We’ve introduced some … comparison questions: “A, if it’s greater than …; B, if it’s less.” However those are. So, we’ve put some of those on our tests. [Ms. I]

Probably not a lot. I teach what’s in the book. It’s just that when I run across things that I know are typical SAT question types, I do concentrate on them or at least zero in on them. A lot of times I’ll say to the kids, “Now this is a typical SAT problem.” [Ms. J]

**Other factors.** During the interviews several additional factors that might influence teachers’ assessment and grading practices were explored. Two seemed to have strong impact: (a) teachers’ knowledge and beliefs about assessment, and (b) the time teachers need to create and grade various forms of assessment.

**Knowledge and beliefs about assessment.** Ten teachers had studied assessment techniques in a course, workshop, or conference. Seven reported having received training directly about assessment, either through college courses or in conferences or workshops organized by their school or district, the Woodrow Wilson Institute, UCSMP, or local, state, or national professional organizations. Two of these seven teachers and three others had received some information about assessment indirectly, for instance, in workshops about uses of graphing calculators or cooperative learning.

---

7 Each school was required to give a high school proficiency exam. In one state it determined whether a high school diploma could be awarded; in the other states it did not affect graduation.
Virtually all the teachers said they were familiar with the curriculum standards, but only three teachers said they were familiar with the assessment standards. Regardless of their familiarity with the evaluation standards, when shown a list of aspects of assessment that might receive either increased attention or decreased attention (NCTM, 1989, p. 191), every teacher said that he or she supported the recommendations.

Because of time constraints in the interviews, only seven teachers were asked to describe the extent to which their own assessment practices in the target class were consistent with these recommendations. All but one teacher expressed some progress toward implementing recommendations for changes in assessment. The following comments illustrate the range of responses:

Well, as opposed to what one would like to have, probably not all that good. Let’s see what they want increased [emphasis] on …. for example, “using multiple assessment techniques, including written, oral, and demonstration formats.” We don’t do much of that. We have a problem with both the Algebra II and Geometry. You’re fighting the clock every day of the year just to cover the amount of material that we’re required to, which means that a lot of these things you really have difficulty working in, unless they’re going to be outside of the classroom situations. [Mr. N]

I think overall, I try to do all these things. But I don’t think I do enough of them. Like, I agree that you should use multiple assessment techniques. I think I don’t do that enough. [Ms. A]

Very. We use calculators every day. The computer is used for demonstration purposes. The whole book is with problem situations. I use multiple assessment techniques. But I used standardized tests too much. Maybe I need more variety. [Ms. L]

Virtually every teacher said that he or she would be interested in attending a workshop or course on assessment if one were offered. Learning about different types of assessment beyond tests was the topic in which teachers most frequently (n = 9) expressed an interest. Those who seldom or never used anything but tests, quizzes, or homework were interested in learning how to use other forms of assessment, and those who had been using written projects wanted ideas about interpreting their students’ work. The topic with the second highest degree of interest was the relation between assessment and grading. One teacher wanted to know how to grade assignments more fairly; another wanted to learn how to weight tests, quizzes, and homework to assess students fairly. The use of calculators, computers, and manipulatives in assessment was also of interest to this sample. Many teachers requested specific examples of classroom-tested assignments and detailed instructions on how to implement the ideas in their classrooms.

When I see what other people do and the samples of other kids’ work,… it gives me different expectations. In my own limited way I have a vision of what I think writing is, but there may be all kinds of creative things out there that I don’t know about, and if I saw them, they’d be my ideas too then. [Mr. M]

Make it age-level appropriate, content appropriate, and look at alternatives, plus the psychology of how to get students to do the things that they’re not necessarily going to get a grade for. [Mr. B]

In addition to their own lack of knowledge of new assessment techniques, several teachers explained how students’ backgrounds and attitudes influenced the teacher’s ability to implement some assessment techniques.
If you have better students who are self-motivated, you can have them working on things outside of class, special research projects, and computer work, and so on, I think that would be great. But we don’t have that here. I’ll tell you frankly, we have difficulty getting our kids to do anything out of the classroom. [Mr. N]

This first one, “assessing what students know and how they think about mathematics.” It’s difficult for us in high school with juniors and seniors, and 15-, 16-, 17-year-olds. It’s hard for them to talk mathematics because it’s not cool …. You would like them to stand up and explain the problem or tell me everything you know about it, but it’s not, that’s not the socially acceptable thing to do, you know, you’re weird. It’s wonderful theory, but it’s just, it’s not the “in” thing to do, and I don’t think we’re overcoming that. [Mr. S]

Time. At the end of the year teachers were asked about the time required to make up and grade each of the various types of assessment instruments they used. Most of the 16 teachers responding to these questions answered with an interval. For instance, to make up the last test given before the interview, the minimum amount of time needed was 20–40 minutes [Mr. S], and the maximum amount of time was 3–4 hours [Mr. F]. The mean interval of time spent making up the last test was 1 hour and 20 minutes to 1 hour and 43 minutes, for an average of 1 hour and 31 minutes. The time estimated to grade these tests for one section of the course ranged from 40 minutes [Mr. R] to 6–8 hours [Ms. P]. The mean time spent grading the last test was 1 hour and 28 minutes per class. These results were almost identical to the data for a “typical test,” where the mean time reported to make it up was 1 hour and 19 minutes and to grade it was 1 hour and 25 minutes. The mean time spent making up a typical quiz was 36 minutes, and the mean time spent grading such a quiz was 28 minutes per class.

For the five teachers who reported specific estimates of the time needed to make up assessment tasks such as oral or written projects or computer laboratory assignments, the mean was 2 hours and 9 minutes. The mean time for grading such assignments was 3 hours and 50 minutes per class. Two other teachers reported spending “several hours” or “a couple of hours” per class grading such assignments. Thus, newer forms of assessment generally took, on average, twice as much time to prepare and twice as much time to grade as chapter tests.

DISCUSSION

Because the schools selected for this study had access to textbooks consistent with many recent recommendations for reform, to state-of-the-art calculator and computer technology, and to above-average opportunities for professional development, the sample has a positive bias. That is, the picture painted by this article is probably more optimistic than that of the entire set of high school classes in the United States. Even though the sample is somewhat selective, we believe the data suggest trends and issues that may apply to many other high schools. We discuss these trends and issues here to stimulate future research and development and to offer guidance to reform efforts currently underway.

Tests Versus Other Forms of Assessment

New forms of classroom assessment. The Assessment Standards for School Mathematics calls for “a shift toward using multiple and complex assessment
tasks, projects, writing assignments, oral demonstrations, and portfolios, and away from sole reliance on answers to brief questions on quizzes and tests” (NCTM, 1995, p. 29). Among our relatively reform-minded sample, this shift had begun to occur in about two-thirds of the classes.

Why are newer forms of assessment used in some classes and not in others? Two factors appear to be critical: (a) teachers’ knowledge and beliefs, and (b) the instructional materials available to teachers. Thus we suggest that future research investigate effective models for preservice and in-service education to help teachers address issues in assessment and that future curriculum development efforts include the creation of assessment tasks. In particular, teachers need examples of worthwhile assessment tasks geared to specific courses they are teaching, rather than examples that are meant to assess more general levels of mathematical performance.

Teachers are also eager for training in reliable and easy-to-use methods for interpreting and reporting students’ work on complex tasks and for ideas on how to implement new assessment techniques in their classrooms. Because some teachers lack confidence in their own writing ability or in their ability to comment fairly and constructively on student writing, we suggest that preservice and in-service work in assessment deal with affective, as well as cognitive, issues related to assessment techniques. Researchers can help identify the major benefits—be they cognitive, affective, or practical—of various alternative assessments.

Classroom tests. The generally low level of test questions, the paucity of questions set in realistic contexts or of those asking students to justify a conclusion, and the apparent lack of impact of technology on the types of test items teachers use suggest that assessment on chapter tests often does not “reflect the mathematics that all students need to know and be able to do” (NCTM, 1995, p. 11).

Initially, we were dismayed by the large number of test items that emphasized low-level thinking and fairly simple responses. However, given that chapter tests must often address a dozen or more objectives and that they must often be administered in 40- to 50-minute periods, teachers’ methods of selecting and creating test items are both rational and consistent with traditional methods of test design.

Furthermore, one can argue that there should be some questions that can be answered with one or two steps. Such items give students a chance to be successful and to demonstrate some knowledge, even if they might run into roadblocks when doing a more complex task. Short-answer questions also potentially make it easier for a teacher to decide which objectives have yet to be mastered, so that appropriate remediation can be determined. On complex tasks, when students have difficulty, it is sometimes difficult to decide where the student has gone wrong.

Additionally, our coding of test items does not describe the content or the variety of the mathematics that was being tested. In some cases, teachers whose tests consisted of mostly low-level items almost exclusively tested skills, such as factoring polynomials, that are considered to be of limited value (Usiskin, 1980/1995). In other cases, even though teachers consistently gave tests with mostly low-level items, there was considerable variety in the questions. For instance, some items were set in realistic contexts; some required graphical representation; some were technology-active;
and in some the mathematics being tested involved new content, such as interpreting a boxplot or describing the effects of a scale change on a figure. We argue that the two types of tests are different in ways not detected by our analysis and that the latter tests do show a shift in the directions outlined in NCTM’s *Standards* (1989, 1995). Thus, in this respect also, we believe that the state of assessment, at least in those high schools using more innovative curricula, is a little rosier than our research might indicate.

Initially, we were also surprised by how teachers’ perceptions of the impact of technology on assessment were much greater than their use of technology on classroom tests would indicate. There are several possible explanations for this phenomenon. First, having made major changes in their instructional practice, teachers may have simply assumed, incorrectly, that these changes resulted in major changes in their assessment practices. Second, it is possible that a few years before we conducted our research, no items on tests were technology-active and that the change from 0% to 5%, 10%, or 20% technology-active was considered a big change. Third, the teachers may have thought that the technological demands of some test items were higher than we thought. Clearly, classroom teachers, researchers, and test developers could all benefit from further discussion of how to recognize and write worthwhile technology-active test items or other tasks.

In recent years many curriculum projects have incorporated new content into the secondary school mathematics curriculum, put mathematics in context, and worked to connect graphical and symbolic representation in instructional materials. There is some evidence of these features being incorporated into published tests and into teacher-made tests. However, greater efforts need to be made to incorporate more reasoning and multistep problem-solving, as well as more substantive use of both numerical and graphical technology on tests.

Recent work (de Lange, 1993; Thompson, Beckmann, & Senk, 1997; van den Heuvel-Panhuizen & Gravemeijer, 1993) provides some guidance on ways to think about improving classroom tests. Perhaps a checklist based on the coding scheme reported in this paper could also help both teachers and professional test developers analyze the characteristics of tests and make them aware of the tests’ strengths or weaknesses.

Research reported by Garet and Mills (1995) suggests that teachers’ emphasis on short-answer classroom tests will continue to be strong in the immediate future. Hence, improving classroom tests may be the surest way to improve the quality of assessment in high school mathematics.

**Balancing tests and other forms of assessment.** Even when newer forms of assessment are used, why are they used so seldom? Why is so little weight given to forms of assessment that many teachers report to be quite valuable? We suggest several explanations.

First, as mentioned above, teachers’ limited knowledge of alternate assessment techniques, and their lack of confidence in their ability to apply those techniques, is a clear deterrent to their using newer forms of assessment as often as more familiar chapter tests.

Second, time is a factor. Newer forms of assessment take about twice as much time to create and twice as long to score as textbook-supplied classroom tests. When
time is limited, people will use ready-made instruments or instruments that are quick and easy to create.

Third, although there has been much written about the general characteristics of good assessment practices or examples of new forms of assessment (e.g., NCTM, 1995; Stenmark, 1989, 1991), there has been little professional guidance about how to balance older and newer forms of assessment. Thus, teachers are left to decide for themselves how to weight tests, quizzes, homework, and newer forms of assessment. Without the support of a professional dialogue to respond to concerns of students, parents, and administrators about new kinds of assessment, teachers are likely to be conservative in their first attempts.

Fourth, even with support, very few people can make sweeping changes in their behavior in one year. As Leinwand suggests, a change of about 10% per year in a teacher’s classroom practices probably is “large enough to represent real and significant change, but small enough to be manageable” (1994, p. 393). In many cases our research involved teachers who were using new forms of assessment for the first or second time. When viewed from Leinwand’s perspective, our findings are actually not that discouraging.

Last, even among teachers who have used written projects, oral demonstrations, or other newer forms of assessment, there appears to be no clear consensus as to which is the most useful kind of assessment tool.

Clearly, no one can or should prescribe an exact mix of how to balance older and newer forms of assessment. We contend that, when aligned with a rich curriculum and when sensitive to the other general characteristics outlined in the Assessment Standards for School Mathematics (NCTM, 1995), tests and quizzes, like writing assignments, oral demonstrations, journals, and interviews, can each provide valuable information to teachers. It is time for public discussion and systematic research on possible mixes of assessment strategies that might be implemented in typical classrooms. Case studies describing the circumstances under which newer forms of assessment can play a larger role than they currently do would be particularly welcome.

Assessment versus Grading

Despite attempts in recent literature to distinguish between assessment and grading, many teachers tend to equate the two. That is, many teachers assign grades to virtually every assessment task except for daily homework assignments, and when asked about their assessment practices, they refer to tests, quizzes, and other graded assignments. One explanation for this phenomenon is that teachers do not understand the distinction between assessment and grading. To the extent that this is true, greater attention to defining terms such as assessment, evaluation, and grading, as is done in the Assessment Standards for School Mathematics (NCTM, 1995) will help clarify communication.

A second explanation, and one that we expect is true of most teachers, is that although teachers understand the distinction between assessment and grading, the system in which they work does not easily allow them to ignore grading. Virtually every high school teacher in the country has to give grades, whether letters or numbers, and
colleges have come to expect them. As Wilson points out, “what gets graded is what gets valued” (1994, p. 412).

Teachers who are in the process of making the types of shifts in assessment practices advocated by the profession need guidance on how to aggregate results of students’ performance on these new tasks and on how to report summary comments to students, parents, employers, and colleges. Because teachers need such guidance as soon as they begin to try new forms of assessment, research on issues of aggregation and reporting of high school students’ performance on complex mathematical tasks is a critical and immediate need. In the meantime, grading should come out of the shadows and onto the center stage of the discussion on how to improve the quality of assessment in schools.

Some Other Issues

The fact that the high school mathematics teachers in this sample rarely considered standardized tests when developing assessment instruments confirms the findings of Stiggins and Conklin (1992). Apparently, except for geometry teachers, high school teachers do not feel pressured by external tests in the same way that elementary and middle school teachers often do. Thus, although we encourage further research and policy statements regarding the improvement of standardized tests, we believe that the bulk of the effort to improve assessment in high school should be focused at the assessment taking place in classrooms.

We hypothesize that geometry teachers may be more concerned than other mathematics teachers about standardized tests because, until recently, geometry often has been taught in only one course. As a result, geometry teachers often cannot expect students to know much geometry when they begin the course, and they cannot count on much follow-up after students leave their course. Hence, they are unique in feeling solely responsible for a body of content.

The enormous variability among courses suggests that much information about teachers’ practices is masked by collapsing and aggregating survey results. In fact, different strategies for changing assessment practices may be necessary at different grade levels or in different courses. For instance, whereas geometry teachers appear to need to consider ways to assess reasoning via something other than formal proofs, Algebra II teachers need to consider how to assess reasoning in any form. Whereas Algebra I, Algebra II, and post–Algebra II teachers all need to consider the impact of graphing technology on their courses, the appropriate balance between paper-and-pencil graphing and graphing with technology is likely to be different for beginning algebra students than for more advanced students. Thus, we suggest that future research on assessment not only be sensitive to potential differences in issues at various school levels and in various courses, but that the manner of reporting research results portray as much as possible the diversity among teachers, courses, and schools.

The teachers in this sample were more professionally active than average and spoke thoughtfully on many subjects. They realized that creating worthwhile assessment tasks is a time-consuming and difficult enterprise. When they are dealing with changes
in curriculum and instruction, it is often difficult for them to find time to make changes in assessment. Our research suggests that teachers use published assessment materials either because they think the quality of those materials is higher than those they can create themselves, or because they think they can get the same quality in less time. Thus, teachers’ use of published materials is not simply a matter of blind trust.

The *Assessment Standards for School Mathematics* (NCTM, 1995) provides general criteria for judging the quality of an assessment program, instrument, or task. Such standards are helpful in thinking about assessment programs at the national, state, and district levels. However, the assessment that affects the daily instruction of millions of students occurs in the classroom and is directed by the classroom teacher. The research reported here suggests that assessment issues in high school are different from those in elementary or middle school, and that assessment issues are different across high school courses. We suggest that further research, teacher education, and materials development related to assessment be directed more toward the classroom than they have been thus far.

**REFERENCES**


**Authors**

S. L. Senk, Associate Professor, Department of Mathematics, Michigan State University, East Lansing, MI 48824, senk@math.msu.edu.

C. E. Beckmann, Associate Professor, Department of Mathematics, Grand Valley State University, Allendale, MI 49401, beckmanc@gvsu.edu.

D. R. Thompson, Assistant Professor, College of Education, University of South Florida, EDU 208B, Tampa, FL 33620, thompson@tempest.coedu.usf.edu.