High-Stakes Testing and Mathematics Performance of Fourth Graders in North Cyprus

OSMAN CANKOY
Atatürk Teacher Training Academy, North Cyprus

MEHMET ALI TUT
Eastern Mediterranean University, North Cyprus

ABSTRACT **The authors attempted to determine the effects of a high-stakes standardized testing-driven instructional approach on mathematical performance. The authors developed a multiple-choice mathematics performance test for 1,006 Grade 4 students in 28 North Cyprus schools. Analysis revealed that students who spent more time on test-taking skills performed better, especially in routine mathematics items, than did students who spent less time on test-taking skills. There was no difference observed in test results from nonroutine story problems. However, analysis did indicate that spending too much time on test-taking skills led to memorizing procedures and cuing on surface attributes of a problem.**

Key words: high-stakes standardized tests, routine and nonroutine mathematics items, test-taking skills

With the progress of globalization and the growth of competition between communities, the education of future generations is attracting critical attention. That scrutiny has led to the rethinking of important aspects of the education systems. One of the most important elements of any education system or teaching and learning process is assessment, or testing. Researchers have found that testing, when poorly prepared, is not objective and has negative as well as detrimental effects on students and the education system as a whole (Amrein & Berliner, 2002; Paris, 1995; Popham, 1999, 2001). Although educators routinely give testing prime importance and concentrate on producing reliable test items, they sometimes overlook unexpected influences and the consequences that test items can have on teachers and learners. It is the unexpected influences that ultimately affect the quality of testing.

A crucial sort of testing in education is high-stakes standardized testing, which is standard in the sense that all students answer the same questions under the same conditions and are scored in the same manner. The tests are high stakes because test answers are used for making major decisions about the students (Marcus, 1994; Popham, 1999). Students usually complete high-stakes standardized tests once or twice a year, in comparison with classroom quizzes or end-of-unit tests that usually are taken weekly or monthly.

We attempted to determine how a high-stakes, standardized test-driven instructional approach affected the mathematics performance of fourth graders. We considered that time spent on test-taking skills was an important linkage to the high-stakes, standardized test-driven instructional approach that consisted of several classroom practices such as (a) working on test questions from a current or prior test, (b) giving students test questions for drill, and (c) teaching students standard algorithms and procedures (especially ways of seeking for cue words and surface characteristics of a problem) for answering multiple-choice test questions and rule memorizing.

Many cited publications emphasize solving nonroutine mathematics problems as good indicators of mathematics performance (e.g., National Council of Teachers of Mathematics [NCTM], 1989, 1991, 2000), so we analyzed the ability of fourth graders to solve routine and nonroutine mathematics problems. In this study, a routine mathematics problem represented a textbook-like problem that could be solved or answered with a standard algorithm or procedure. For routine mathematics problems, the student had to implement only a limited number of steps. However, for nonroutine mathematics items, the students did not have to apply any formal algorithms. If there were an algorithm that the student could follow, the student had to examine the mathematics problem and apply the algorithm flexibly. Contrary to the nonroutine mathematics items, the contexts that the students used for the routine mathematics problems often were used in the classroom practices and textbooks.

In this study we sought answers to the following questions.

1. Does the performance of the students in solving and answering routine and nonroutine mathematics problems differ?
2. Is there a gender effect on the mathematics performance of students when they solve routine and nonroutine mathematics problems?

Address correspondence to Osman Cankoy, 12. Gelibolu Sok. No:24, Lefkosa, North Cyprus, via Mersin 10, Turkey. (E-mail: cankoy@kktc.net)

3. Is there any difference among the groups of students (those who spent more and those who spent less time on test-taking skills) in terms of mean scores on routine and nonroutine number problems?
4. Is there any difference among the groups of students (those who spent more and those who spent less time on test-taking skills) in terms of mean scores on routine and nonroutine operation problems?
5. Is there any difference among the groups of students (those who spent more and those who spent less time on test-taking skills) in terms of mean scores on routine and nonroutine story problems?
6. Is the choice of distractors, which could reveal rote memorization, surface understanding, or searches for only cue words in mathematical items, group dependent?

Related Literature

Problems With High-Stakes Standardized Testing

Education systems that are oriented toward high-stakes standardized testing are commonplace around the world. In such systems, especially those in which the tests are prepared badly, educators might have difficulty producing critical and reflective students. Those students who spend too much time on high-stakes standardized tests might have problems integrating what they are learning or applying their knowledge and skills to real situations.

The most fundamental problem with an examination-oriented education system is that examinations might distort students' motivation and learning by overemphasizing the importance of scores as outcomes and measures of student abilities. Overstressing on examination results also might subvert students' learning strategies because examination-taking strategies usually are inconsistent with learning strategies taught every day in the classroom (Paris, 1995).

Educators primarily use high-stakes standardized testing to sort large numbers of students in as efficient a manner as possible; however, this narrow objective usually results in short-answer or multiple-choice questions. Also with that type of test construction, important skills such as writing, acting, speaking, and creating, which can and should be taught in schools, are relegated to second-class status (Bowers, 1989).

Teaching to the test, a heavy reliance on high-stakes standardized testing or on an examination-oriented education system, also might produce important negative effects on teaching activities. In general, any form of teaching to the test raises scores without increasing students' knowledge and skills in the subject being tested (Kober, 2002). In one of the nationally representative surveys conducted in the United States, 79% of teachers said that they spent "a great deal" of their time instructing students in test-taking skills (Quality Counts, 2001). Teaching to the test not only produces unproductive and uncritical students but also can be misleading. When teachers teach directly to a specific question on a test, the resulting scores likely give an inflated picture of students' understanding of the broader domain (Kober). A teacher who is familiar with a state English test could prepare students by drilling them in a few dozen vocabulary words that have often appeared on earlier tests, out of the hundreds of vocabulary words that students are expected to learn (Popham, 2001).

High-stakes standardized testing is not only a barrier for good teaching practices and the resulting higher order student skills but also might be a waste of money. Haney, Madaus, and Lyons (1993) estimated that "taxpayers in the USA are devoting as much as $20 billion annually in direct payments to testing companies" (p. 95). In such a testing system, students' test performances should be a valid and reliable measure of their knowledge and skills (Thurlow, Quenemoen, Thompson, & Lehr, 2001). Even if one can guarantee the validity of high-stakes standardized tests, other problems might exist. Minorities, and students with disabilities in particular, may suffer as a result of traditional assessment practices that are inaccurate and inconsistent yet continue to be used for prediction, decision making, and inferences about student performance and lifelong success (Dais, 1993).

Good Testing Practices

Positive features of high-stakes standardized testing might be attained through replacement by performance testing or portfolios of work samples in which assessment is linked to the classroom curriculum and is part of an ongoing process in which students monitor their personal progress (Corono, 1992). Although replacement of high-stakes standardized tests by other more positive tests may be difficult, educators can at least change some aspects of teaching to the test. For example, teachers could change the curriculum and manner of teaching by (a) teaching the most important knowledge, skills, and concepts contained in the standards for a particular subject; (b) addressing standards for basic and higher order skills; (c) using test data to diagnose areas in which students are weak and focusing in those areas; and (d) giving students diverse opportunities to apply and connect what they learn (Kober, 2002). In a study of New Jersey teachers, Rutgers University researchers Firestone, Monfil, Mayrowetz, and Camilli (2001) found that the state's elementary school assessments in mathematics and science, which included a mix of test-item formats, encouraged teachers to place greater emphasis than did other states on writing, problem solving, use of hands-on materials, and student discussion and explanation of their thinking.

High-Stakes Standardized Testing-Oriented Education in North Cyprus

In North Cyprus, education at the elementary and secondary school level is highly centralized and under the control of the Ministry of Education. Students enter elementary school at age 6 and leave at age 11 (from first grade to

fifth grade). Most of the people in North Cyprus value education and educated people. As a result of that attitude, people tend to overemphasize testing and preparation consonant with entrance examinations, which also can be called high-stakes standardized tests. In addition, North Cypriots value knowing mathematics and obtaining good grades in mathematics. At the elementary school level, the general instructional approaches and techniques used in the classrooms are aligned with behaviorist learning theories; in the first 3 years, one can observe thought-provoking mathematics activities congruent with constructivist learning perspectives. At the end of each school year, most fifth graders (nearly one third of all elementary school graduates) in North Cyprus take the Entrance Examination for the Middle Schools (EEMS), for which the general medium of instruction is English. The examination is considered by the majority of families in North Cyprus as the most important key in the future academic life of students. The EEMS is prepared and administered once a year by the Ministry of Education. Because of this high-stakes standardized testing, which usually begins at the fourth through the fifth grade, instructional approaches in elementary schools of North Cyprus are geared mostly to teaching to the test. Each year, many families spend a large amount of their family budget on private lessons to prepare their children for the test. Although EEMS is considered the most important high-stakes standardized test in North Cyprus, the Ministry of Education has not gathered empirical evidence about its reliability and validity for the last 25 years. That lack of evidence could be problematic and misleading for the overall education system in North Cyprus.

In the academic year 2000–2001, data collected through interviews and structured questionnaires during the inservice training activities with fourth- and fifth-grade elementary school teachers showed that (a) 65% of the teachers spent at least 70% of their class time working on actual test questions from a current test, (b) 85% of the teachers gave their students actual test questions for drill, and (c) 75% of the teachers taught their students test-taking skills and had them practice with tests from prior years. The data also showed that fourth- and fifth-grade teachers spent nearly 70% of their semester time on language and mathematics because the EEMS consisted of two batteries of tests for mathematics and language skills.

Method

Participants

We randomly selected 28 schools out of 83 schools (n = 1,006) in North Cyprus. Then, from each selected school, we chose a number (ranging from 1 to 4) of fourth-grade classes for observation to determine the percentage of class time that was spent on test-taking skills in mathematics. We trained 28 volunteer preservice elementary teachers to perform the observations. Using a structured time unit observation sheet, each preservice teacher coded teachers' instructional activities in the classrooms. The test-taking skills categories listed on the observation sheet were (a) working on test questions from a current or prior test, (b) giving students actual test questions for drill, (c) teaching students standard algorithms and procedures for answering especially multiple-choice questions, and (d) memorizing rules.

The preservice teachers tried to observe which one of the listed categories had occurred in each 1-min period. Each class was observed for at least 6 class hr (each class hour was nearly 40 min). The students from schools in which nearly 70% of class time was spent on test-taking skills formed the *high-emphasis group* (HEG; n = 351). That group spent the rest of its time on noninstructional activities without a textbook. Teachers generally used current and prior test items and worksheets as the main source of instruction. The students from schools in which nearly 50% of the class time was spent on test-taking skills formed the *moderate-emphasis group* (MEG; n = 207). That group spent most of its time on noninstructional activities and instructional activities guided by the textbook recommended by the Ministry of Education for regular classroom practices. The students who attended schools in which nearly 30% of the class time was spent on test-taking skills formed the *low-emphasis group* (LEG; n = 448). That group spent the remainder of its time on noninstructional activities, along with instructional activities guided by the textbook recommended by the Ministry of Education. That group spent more time on those activities compared with the MEG. According to the observations for all groups, the category "teaching students standard algorithms and procedures to be applied in answering specifically multiple-choice-type test questions" was used most, and the category "rule memorizing" was used least. The book recommended by the Ministry of Education was based primarily on traditional instructional approaches and had few indirect relations with conceptual understanding and nonroutine problem solving.

Instrument

We developed a 36-item, multiple-choice Mathematical Performance Test (MPT) and adapted it for this study to measure the mathematical performance of fourth graders. The test included three subtests (Number Skills, Operations With Numbers, and Story Problems) and six dimensions: (a) Routine Number Items (RNI), (b) Nonroutine Number Items (NRNI), (c) Routine Operation Items (ROI), (d) Nonroutine Operation Items (NROI), (e) Routine Story Problems (RSP), and (f) Nonroutine Story Problems in Nonroutine Contexts (NRSP). There were seven, three, five, seven, eight, and six items, respectively, in each dimension (see Table 1). We previously gave the items to 13 experienced elementary school teachers and 5 inspectors from the Ministry of Education and asked them to categorize the

TABLE 1. Selected Mathematics Problems

Category	Test question: Routine	Test question: Nonroutine
Number skills	Item 1: How is "three hundred thousand and fifteen" represented by numerals? (A) 315 (B) 3,015 (C) 30,015 (D) 300,015	Item 9: Which one of the following numbers cannot be written in the place of "a" if "13a" is a three-digit even number? (A) 4 (B) 3 (C) 2 (D) 0
Operations with numbers	Item 15: What is the result of "2,500 × 26" (A) 65,000 (B) 62,000 (C) 55,500 (D) 6,500	Item 14: Which operation can be used in order to find the value of "a" with the shortest way in "524 + a = 816?" (A) Addition (B) Subtraction (C) Multiplication (D) Division
Story problems	Item 24: A factory can produce 2,007 toys daily. How many toys can be produced in 2 weeks? (A) 4,004 (B) 4,014 (C) 10,035 (D) 28,098	Item 36: Ali, Mehmet, Oya, and Cem completed a race in 10, 8, 9, and 7 seconds respectively. Who is the first? (A) Ali (B) Mehmet (C) Oya (D) Cem

items as routine and nonroutine by considering the teaching and learning practices in North Cyprus elementary schools. The fourth graders had 1 hr to answer the items on the MPT. Alpha reliability of the subtests—Number Skills, Operations With Numbers, and Story Problems—were .75, .71, and .61, respectively. Alpha reliability for the test, including the 36 items, was .86. We asked a measurement and evaluation expert and two mathematics educators from Middle East Technical University in Turkey to judge the content validity of the test. We provided each of the experts with information that included the definition of what we intended to be measured, the instrument, and the description of the intended sample. We also asked the experts to judge whether the distractors of the items could enlighten rote memorization, surface understanding, or a search for only cue words. All the experts concluded that the content and format of the items were consistent with the definition of the variable and the study participants.

Procedures

Before we conducted this study, we sought permission from the Elementary School Education Division of the Ministry of Education to develop and administer the EEMS within a high-stakes context for the 2000–2001 academic year. The Elementary School Education Division noted that permission could be given for only the data that would be collected from the EEMS. Unfortunately, because of the lack of evidence for the reliability and validity of the EEMS, conducting the research with that requirement became impossible. However, after we agreed to provide feedback to the Ministry of Education, the Elementary School Education Division offered permission and assistance for us to conduct the present study and administer the MPT in the 28 schools simultaneously. We trained 28 preservice elementary teachers for a week so they could conduct the live classroom observations. We divided the participants into three groups (HEG, MEG, LEG) and, according to the observation results, the observers administered the MPT.

Analysis of Data

We based the data analysis mainly on quantitative methods. We used analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA) procedures to observe the differences among the three groups in terms of performance in answering and solving routine and nonroutine mathematics problems that were related to number skills, operations with numbers, and story problems. We used chi-square procedures to conduct a detailed analysis and to determine whether the responses given to the choice of items were dependent on the group variable (HEG, MEG, LEG). We based the interpreta-

tions of the effect sizes observed on the suggestions of Cohen (1988); the level of significance used throughout the study was .05.

Results

Table 2 shows that five of the six dimensions of the MPT were intercorrelated. The one exception was the dimension related to nonroutine story problems. Because of the difference, we used MANOVA procedures to analyze the first five dimensions (RNI, NRNI, ROI, NROI, RSP). When we analyzed the performances of the three groups (HEG, MEG, LEG) in solving nonroutine story problems, we used one-way ANOVA procedures (see Table 3).

TABLE 2. Intercorrelations Between Subdimensions of the MPT for Fourth Graders (N = 1,006)

Subdimension	1	2	3	4	5	6
1. RNI	—	.41*	.68*	.60*	.62*	–.02
2. NRNI		—	.45*	.48*	.46*	–.05
3. ROI			—	.64*	.61*	–.02
4. NROI				—	.66*	–.02
5. RSP					—	–.03
6. NRSP						—

Note. MPT = Mathematical Performance Test; RNI = routine number items; NRNI = nonroutine number items; ROI = routine operation items; NROI = nonroutine operation items; RSP = routine story problems; NRSP = nonroutine story problems in nonroutine contexts.
*$p < .05$.

A paired samples t test showed that mean score on items in routine mathematics items was significantly higher than was the mean score on nonroutine mathematics items, $t(1{,}005) = 43.285$, $p < .05$. That finding also was confirmed by a large effect size ($d = 1.36$). Later analysis showed significant but small group-dependent main effects (considering small effect sizes) on mean scores of routine, $F(2, 1{,}006) = 35.239$, $p < .01$, $\eta_p^2 = .07$, and on nonroutine mathematics problems, $F(2, 1006) = 24.644$, $p < .05$, $\eta_p^2 = 0.05$, favoring the HEG and MEG compared with the LEG. However, gender had no significant main effect on mean scores for routine and nonroutine mathematics problems. MANOVA results in Table 4 revealed that group and gender and their interaction had significant effects on the linear combination of mean scores that resulted from RNI, NRNI, NROI, and RSP with small effect sizes.

Number Skills

Table 5 shows that group had significant effects on mean scores of RNI and NRNI with small effect sizes. However, group was more effective on RNI compared with NRNI (see effect sizes in Table 5). Bonferroni post-hoc tests ($p < .05$) revealed no significant difference between the HEG and the MEG in terms of mean scores on RNI and NRNI. However, both groups had significantly higher mean scores than did the LEG on RNI and NRNI. Results also revealed that gender had no significant effect on mean scores on RNI and NRNI (see Table 5).

TABLE 3. Descriptive Statistics for Routine and Nonroutine Number Problems

	Number problem				
	Routine		Nonroutine		
Group	*M*	*SD*	*M*	*SD*	*n*
High emphasis					
Girls	4.56	1.61	1.39	0.85	172
Boys	4.27	1.66	1.23	0.90	179
Total	4.41	1.63	1.30	0.88	351
Moderate emphasis					
Girls	4.32	1.77	1.21	0.83	108
Boys	4.45	1.59	1.36	0.83	99
Total	4.38	1.68	1.29	0.83	207
Low emphasis					
Girls	3.69	1.73	1.06	0.79	233
Boys	3.56	1.85	0.94	0.89	215
Total	3.63	1.79	1.00	0.84	448
Total					
Girls	4.11	1.74	1.20	0.83	513
Boys	4.00	1.77	1.13	0.89	493
Girls and boys	4.06	1.76	1.17	0.86	1,006

Operations With Numbers

Results showed that group had significant effects on mean scores of ROI and NROI with small effect sizes (see Table 5). Bonferroni post-hoc tests ($p < .05$) revealed no significant difference between the HEG and the MEG in terms of mean scores on ROI and NROI. However, both of those groups had significantly higher mean scores than did the LEG on ROI and NROI. Results also revealed that gender had no significant effect on mean scores on ROI (see Table 6). How-

TABLE 4. Wilks's Lambda Test of Significance for Effects of Level of Preparing for High-Stakes Standardized Mathematics Test on Fourth Graders' Performances

Effect	Λ	*F*	Hypothetical *df*	Error *df*	*p*	η^2_p
Group A	.93	7.53	10	1,992.00	.000*	.036
Group B	.98	3.43	5	996.00	.004*	.017
A × B	.02	1.99	10	1,994.00	.031*	.010

Note. Fourth graders' performances involved answering routine and nonroutine number and operation problems and solving routine story problems.
*$p < .05$.

TABLE 5. Tests of Between-Subjects Effects for Performance in Answering Routine and Nonroutine Number and Operation Items and Solving Routine Story Problems

Variable	*SS*	*df*	*MS*	*F*	*p*	η^2_p
			Group A			
RNI	147.11	2	73.56	25.02	.000*	.048
NRNI	21.77	2	10.88	15.10	.000*	.029
ROI	141.94	2	70.97	31.15	.000*	.059
NROI	143.04	2	71.52	20.21	.000*	.039
RSP	151.06	2	75.53	21.47	.000*	.041
			Group B			
RNI	2.13	1	2.13	0.72	.395	.001
NRNI	0.46	1	0.46	0.63	.427	.001
ROI	5.82	1	5.82	2.56	.110	.003
NROI	27.92	1	27.92	7.89	.005*	.008
RSP	0.49	1	0.49	0.14	.709	.000
			Group A × Group B			
RNI	5.17	2	2.58	0.88	.416	.002
NRNI	3.54	2	1.77	2.45	.086	.005
ROI	4.62	2	2.31	1.01	.363	.002
NROI	10.41	2	5.20	1.47	.230	.003
RSP	30.82	2	15.41	4.38	.013*	.009
			Error			
RNI	2,940.13	1,000	2.94			
NRNI	721.00	1,000	0.72			
ROI	2,278.38	1,000	2.28			
NROI	3,538.54	1,000	3.54			
RSP	3,517.86	1,000	3.52			

Note. RNI = routine number items; NRNI = nonroutine number items; ROI = routine operation items; NROI = nonroutine operation items; RSP = routine story problems.
*$p < .05$.

ever, results showed that girls had a significantly higher mean score than did boys on NROI, $t(1,004) = 3.33$, $p < .05$, with an effect size near to a medium degree ($d = 0.3$).

Story Problems

Table 5 shows that mean scores of RSP groups were significantly different with a small effect size. Bonferroni post-hoc tests ($p < .05$) revealed that no significant difference existed between the HEG and the MEG in terms of mean scores on RSP. Both groups had significantly higher mean scores on RSP than did the LEG (see Table 7). Results also revealed that gender and group had significant interaction effects on mean scores on RSP with a small effect size. In

TABLE 6. Descriptive Statistics for Routine and Nonroutine Operation Problems

	Operation problem				
	Routine		Nonroutine		
Group	*M*	*SD*	*M*	*SD*	*n*
High emphasis					
Girls	3.50	1.46	4.37	1.79	172
Boys	3.35	1.48	3.97	1.81	179
Total	3.42	1.47	4.16	1.81	351
Moderate emphasis					
Girls	3.45	1.46	4.17	1.96	108
Boys	3.46	1.59	4.11	1.98	99
Total	3.46	1.52	4.14	1.97	207
Low emphasis					
Girls	2.85	1.59	3.70	1.87	233
Boys	2.51	1.47	3.10	1.94	215
Total	2.68	1.54	3.41	1.92	448
Total					
Girls	3.19	1.55	4.02	1.88	513
Boys	3.00	1.56	3.62	1.95	493
Girls and boys	3.10	1.56	3.83	1.93	1,006

TABLE 7. Descriptive Statistics for Routine and Nonroutine Story Problems

	Story problem				
	Routine		Nonroutine		
Group	*M*	*SD*	*M*	*SD*	*n*
High emphasis					
Girls	3.99	1.90	0.65	0.70	172
Boys	3.56	2.02	0.75	0.71	179
Total	3.77	1.97	0.70	0.71	351
Moderate emphasis					
Girls	3.30	1.86	0.85	0.71	108
Boys	3.84	1.91	0.79	0.80	99
Total	3.56	1.90	0.82	0.75	207
Low emphasis					
Girls	2.92	1.81	0.70	0.69	233
Boys	2.94	1.79	0.74	0.70	215
Total	2.93	1.80	0.72	0.70	448
Total					
Girls	3.36	1.90	0.71	0.70	513
Boys	3.35	1.93	0.75	0.72	493
Girls and boys	3.35	1.92	0.73	0.71	1,006

the LEG, there was no significant difference between girls and boys in terms of mean scores on RSP. In the HEG, girls had significantly higher mean scores than did boys on RSP, $t(349) = 2.02$, $p < .05$, but the difference was insignificant because of the small effect size ($d = .01$). In the MEG, however, boys had a significantly higher mean score than did girls on RSP, $t(205) = -2.067$, $p < .05$. The effect size for the MEG ($d = .2$) was not as small as for the HEG. Results also showed that group had no significant effect on mean scores on NRSP (see Table 8). In other words, mean scores of the three groups were not significantly different on NRSP.

We also wanted to determine whether the choice of the distractors of the items, which could enhance rote memorization, surface understanding, or search for only cue words, were group dependent. Chi-square statistics for each item (see Table 1) showed that in 25 items (16 of 25 were in nonroutine contexts) out of 36 items, the choice of the distractors reported in the Instruments section was dependent on the group (see Table 2). For example, in Item 14, the typical distractor "A" was selected by fewer students from LEG (44.2%) compared with students from HEG (62.8%) and MEG (57.3%), $\chi^2(8, N = 1{,}006) = 32.114$, $p < .01$. There was a similar finding in Item 36 (see Table 1) in which the typical distractor was A. For Item 36, fewer students from the LEG (60%) compared with students from HEG (67%) and MEG (64%), $\chi^2(8, N = 1{,}006) = 16.183$, $p < .01$, selected the typical distractor, A.

Discussion

The overall results show that the fourth graders' performances in routine mathematics items were better than were their performances in nonroutine mathematics items. That finding is reasonable when one considers the overall examination-oriented education system in North Cyprus in which routine procedures and surface characteristics, especially those of problems, are generally emphasized (Davidson, Deuser, & Sternberg, 1994). Although the amount of time that students spent preparing for a high-stakes standardized test was an important factor in their mathematics performances, the differences favoring the students who were instructed with a test-driven approach resulted from their answering mostly routine number, operation, and story problems. The students who were instructed with a more test-driven instructional approach also performed better on nonroutine number and operation items. However, relatively small effect sizes indicates that a test-driven instructional approach has little to do with improving higher order skills like solving nonroutine story problems (e.g., Kenney & Silver, 1997; Lindquist, 1989). There were no differences among the three groups in terms of performances in nonroutine story problems. Therefore, teaching students standard procedures to solve different types of problems is obviously not the way to teach them to problem solve. Although learning algorithms might work on some occasions in the classroom, it does not guarantee that students will be able to determine when to use the procedures and how to use them correctly. Kantowski (1981) stated that problem solving means different things to different people; the majority of people believe that it is solving word problems, which includes nonroutine problems.

A further reason that we observed no differences in nonroutine problems was that untested concepts are less likely to be emphasized (Quality Counts, 2001). Before we conducted this research, we analyzed the high-stakes standardized tests used at the elementary school level in North Cyprus for the last 10 years; we found that only 5% of the test questions emphasized nonroutine story problems. The results of this study show that nonroutine problem solving is an important dimension of mathematics activities that should be emphasized during mathematics instruction. For example, in 2000, the NCTM released Principles and Standards for School Mathematics, a revised version of the 1989 standards. Unlike previous standards, that version defines problem solving as engagement in a task for which "the solution method was not known in advance and in order to find a solution, students must draw on their knowledge, and through this process, they will often develop new mathematical understandings" (NCTM, 2000, p. 51).

Many researchers notice gender differences in mathematics performances (e.g., Fennema, Carpenter, Jacobs, Franke, & Levi, 1998; Gallagher & De Lisi, 1994). In this study, although gender was not as effective as group on students' mathematics performance, girls responded better than did boys to nonroutine operation items. When solving routine story problems, in the group in which a large emphasis on high-stakes standardized tests was placed, girls performed better than did boys. On the contrary, in the moderate-emphasis group, boys outperformed girls. In North Cyprus,

TABLE 8. Tests of Between-Subjects Effects for Nonroutine Story Problems

Source	*SS*	*df*	*MS*	*F*	*p*	η^2_p
Group (A)	2.093	2	1.046	2.070	.127	.004
Group (B)	0.196	1	0.196	0.387	.534	.000
A × B	0.913	2	0.456	0.903	.406	.002
Error	505.524	1,000	0.506			

families have stricter rules for girls than for boys because of the country's social and ethical perspectives. That rigidity might cause girls to follow certain rules and prescriptions more precisely than boys do. Therefore, it is likely that in a more examination-oriented environment, girls benefit more than boys. For example, Cai (1995) and Hyde, Fennema, and Lamon (1990) stated that girls performed better than boys at the elementary school level.

We found that spending more class time on test-taking skills did not influence nonroutine story problem solving. We considered conceptual and qualitative reasoning aspects of nonroutine story problems and found that educating students in a nonconstructivist perspective may not guarantee improvement in higher order mathematics skills that are important for scientific thinking and human life. Kenney and Silver (1997) and Lindquist (1989) have shown that high-stakes standardized, test-driven education does not lead to higher educational quality. Teaching to the test or spending too much time on test-taking skills or test-driven instruction may tend to produce only some inflated scores.

The groups that used more test-driven instruction performed better on the problems that emphasized algorithms, but one should not conclude that there was no problem with the group in which test-driven instruction was least emphasized solely because we found no differences among the groups in higher order mathematics skills. Classroom observations showed that there was limited use of constructivist, nontraditional teaching practices. Therefore, careful selection of instructional approaches that foster higher order mathematics skills likely is essential.

Moreover, groups that used a test-driven approach largely performed better on many mathematics problems, but small effect sizes throughout the study showed that these groups were not so different from those that spent less time on test-taking skills. Small effect sizes also implied that test-driven instructional approaches did not contribute much to better understanding. One of the most interesting findings of this study was the dependency of the distractors of items that could enlighten rote memorization, surface understanding, or search for only cue words in group. Surprisingly, the HEG and MEG were distracted more than were low-emphasis groups. That finding implies that more time spent focusing on procedural skills such as drills, test taking, or practice with tests from prior years, with little connection with conceptual understanding and qualitative reasoning, can distract students and encourage them to memorize procedures and to search for a single path to a single answer (Kenney & Silver, 1997; Lindquist, 1989; Mestre, 1991; Sacks, 2000).

Unfortunately, this study, as well as current practice, indicates that the majority of mathematics instruction in elementary schools (estimates range up to 80%) is spent introducing, developing, practicing, and establishing proficiency with written algorithms and solving computations without total conceptual understanding (e.g., America 2000, 1991; R. E. Reys & Nohda, 1992; Pape & Tchoshanov, 2001; Sowder, 1992). If the instructional emphasis is mainly on procedural skills (rote memorization) and not conceptually based, then the students not only suffer from (meta)cognitive disadvantages but also might view mathematics as only a discipline of basic rules, formulas, and algorithms that do not require understanding and reasoning (B. J. Reys & R. E. Reys, 1998). Therefore, Reys and Nohda suggested that the narrow view of mathematics in elementary school as computation and routine problem solving must be changed. Researchers need to emphasize conceptually based activities especially at the elementary school level and prepare teachers to maintain a strong content and pedagogical knowledge of mathematics with a constructivist perspective. Also, there is no doubt that educators should attempt to change the views and beliefs (in line with changing positive approaches in the education field) of society as a whole, beginning with the experts who make decisions about overall education.

To summarize, we offer the following recommendations in light of our findings and current practice.

1. If it is not possible to eliminate high-stakes standardized tests, all aspects of mathematics knowledge and its connections might be assessed.
2. Multiple sources of assessment information should be used in making high-stakes decisions.
3. Tests and classroom instruction should emphasize and foster problem-solving skills, especially those that are nonroutine.
4. Students should be trained and encouraged to become skilled problem solvers with the ability to conduct qualitative analyses of problems before they perform quantitative solutions.
5. Preservice and inservice teachers should have the opportunity to view and teach mathematics in a more constructivist way.
6. Teachers should use open-ended problems that encourage the use and integration of conceptual knowledge in assessments and classroom practices.

REFERENCES

America 2000. (1991). *An educational strategy to move the American educational system ahead to meet the needs of the 21st century*. Washington, DC: U.S. Department of Education.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Archives, 10*(18), 1–70.

Bowers, B. C. (1989). *Alternatives to standardized educational assessment*. Eugene, OR: ERIC Clearinghouse on Educational Management. (ERIC Document Reproduction Service No. ED312773)

Cai, J. (1995). A cognitive analysis of U.S. and Chinese students' mathematical performance on tasks involving computation, simple problem solving, and complex problem solving. *Journal for Research in Mathematics Education* (Monograph Series 7). Reston, VA: National Council of Teachers of Mathematics.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Corono, L. (1992). Encouraging students to take responsibility for learning and performance. *Elementary School Journal, 93*, 69–83.

Dais, T. A. (1993). *An analysis of transition assessment practices: Do they recognize cultural differences?* Washington, DC: Vol. 2, pp. 1–19. (ERIC Document Reproduction Service No. ED372519)

Davidson, E., Deuser, R., & Sternberg, R. J. (1994). The role of metacognition in problem solving. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition* (pp. 207–226). Cambridge, MA: MIT Press.
Fennema, F., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher, 27*, 6–11.
Firestone, W. A., Monfil, L., Mayrowetz, D., & Camilli, G. (2001, April). *The ambiguity of teaching to the test: A multiple method analysis*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
Gallagher, A. M., & De Lisi, R. (1994). Gender differences in scholastic aptitude test—mathematics problem solving among high-ability students. *Journal of Educational Psychology, 86*, 204–211.
Haney, W., Madaus, G., & Lyons, R. (1993). *The fractured marketplace for standardized testing*. Boston: Kluwer.
Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*(2), 139–155.
Kantowski, M. G. (1981). Problem solving. In E. Fennema (Ed.), *Mathematics education research* (pp. 111–126). Reston, VA: National Council of Teachers of Mathematics.
Kenney, P. A., & Silver, E. A. (1997). *Results from the seventh mathematics assessment of the NAEP*. Reston, VA: National Council of Teachers of Mathematics.
Kober, N. (2002). *Teaching to the test: The good, the bad, and who's responsible. Test talk for leaders, no. 1*. Center on Education Policy. Retrieved January 20, 2003, from http://www.ctredpol.org/pubs/testtalkjune2002.html
Lindquist, M. M. (1989). *Results from the fourth mathematics assessment of the NAEP*. Reston, VA: National Council of Teachers of Mathematics.
Marcus, E. (1994). Rite of passage: French teachers demand reasoning. *Educational Vision, 2*(2), 18–19.
Mestre, J. P. (1991). Learning and instruction in pre-college physical science. *Physics Today, 44*(9), 56–62.
National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.
National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston, VA: Author.
National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
Pape, S. T., & Tchoshanov, M. A. (2001). The role of representation(s) in developing mathematical understanding. *Theory Into Practice, 40*(2), 118–127.
Paris, S. G. (1995). Why learner-centered assessment is better than high-stakes testing. In N. M. Lambert & B. L. McComb (Eds.), *How students learn: Reforming schools through learner-centered education* (pp. 189–210). Washington, DC: American Psychological Association.
Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*(6), 8–15.
Popham, W. J. (2001). *The truth about testing*. Alexandria, VA: Association for Supervision and Curriculum Development.
Quality Counts. (2001). A better balance. *Education Week, 20*(17), 36.
Reys, B. J., & Reys, R. E. (1998). Computation in the elementary curriculum: Shifting the emphasis. *Teaching Children Mathematics, 5*, 236–241.
Reys, R. E., & Nohda, N. (1992). *Computational alternatives for the twenty-first century*. Reston, VA: National Council of Teachers of Mathematics.
Sacks, P. (2000). *Standardized minds: The high price of America's testing culture and what we can do to enhance it*. Cambridge, MA: Perseus Books.
Sowder, J. T. (1992). Estimation and number sense. In D. C. Grouws (Eds.), *Handbook of research on mathematics teaching and learning* (pp. 371–389). New York: Macmillan.
Thurlow, M., Quenemoen, R., Thompson, S., & Lehr, C. (2001). *Principles and characteristics of inclusive assessment and accountability systems* (Synthesis Rep. No. 40). Minneapolis, MN: National Center on Educational Outcomes.