Methods for Improving Standardized Test Scores: Fruitful, Fruitless, or Fraudulent?

William A. Mehrens and John Kaminski
Michigan State University

You may be able to identify several ways students can be prepared to take an upcoming test, but which ways are ethical to use? There are available materials that you may purchase and use to prepare students for a test: Are they effective and is it ethical to use them? What constitutes an ethically acceptable test preparation procedure?

Can fantasies become facts? In Lake Wobegon, all the children were above average. Cynics with superior intellects (i.e., those who know the definition of average) smiled or perhaps even chuckled at such an idea. But it turns out the nation is making progress toward that goal. According to at least one report, the average in every state is above the national average (Cannell, 1987). Cannell suggests that inaccurate norms and teaching to the test may be the reasons for high scores (Cannell, 1987).

In this article, we consider the teaching to the test explanation. Although not discounting the many other explanations offered (see Volume 7, Number 2 of this journal), we strongly suspect a major reason for the high scores is teaching to the test—and that this practice has increased because of the increasing efforts to use scores as measures of school and teacher quality. Other experts seem to agree (Rothman, 1988).

Teaching to the test is not a practice that, in the abstract, is good or bad. Both the content of the instruction and the content of the test should be related to educational objectives. One hopes that quality instruction will result in higher scores on relevant achievement tests. However, if one wishes to infer to a broader domain from a sample of items (or objectives), then teaching too directly to the items or the specific sample of objectives lowers the validity of the inference.

In this paper we (a) discuss the increased use of test scores as measures of school quality and the concomitant increased pressure to teach to the test, (b) consider legitimate and illegitimate practices with respect to test preparation, (c) describe some commercial test preparation programs, and (d) draw some conclusions about the effectiveness and desirability of using such programs.

Testing and Accountability

The public is for educational accountability and believes that holding teachers responsible for the achievement of their students will result in better education. Many believe that the best data regarding students' levels of achievement come from standardized achievement tests. The states have responded to the public demands. Thirty-two states test elementary children on a statewide basis (Cannell, 1987). Results can even affect how far a teacher will advance up a "career ladder program" (see Georgia Department of Education).

William A. Mehrens is Professor of Educational Measurement, Michigan State University, 462 Erickson Hall, East Lansing, MI 48824. He specializes in educational measurement.

John Kaminski is a graduate student, Michigan State University and a public school teacher, Traverse Area Public Schools, Traverse City, MI 49684. He specializes in education.
1985, & South Carolina Department of Education [Ryan & Rowzi, 1987], for examples). Such test data may well represent the "best data around," for accountability purposes, but educators recognize the many limitations of such data, and some fear the public will misuse the results and blame the educators for the "inadequate" performance of their children. Whereas many educators correctly understand test data limitations, they must tread carefully in their criticisms. Negative statements are viewed as defensive, and arguments against gathering or using such test data are viewed as attempts to cover up inadequate teaching (see Citizens Research Council of Michigan, 1979, p. 17).

When tests are used for important decisions, teachers will teach to the test (Haney & Madaus, 1986). Although teaching to the test is not a new concern, there is no doubt that increased accountability could lead to increased teaching toward the test. Such teaching could be productive or counterproductive. Under what circumstances teaching for the test is a harmful educational practice is a question that deserves careful consideration. Indeed, a National Council on Measurement in Education (NCME) Ad Hoc Committee of Past Presidents has listed this issue (among 7 others) as being one about which "NCME should take some kind of action" (Gardner, 1987).

Legitimate Versus Illegitimate Test Preparation: An Overview

The issue of what constitutes legitimate preparation of students for tests is the major focus of this paper. There are two somewhat extreme points of view on the issue, and many positions somewhere in between. One might (somewhat inaccurately) characterize the two extremes as the old and new points of view. But many current measurement experts would prefer the old viewpoint. An example of the old position is Wrightstone’s 1967 suggestions about the preparation of pupils for the Metropolitan Reading Tests. He states that one should never use any form of a nationally standardized test for preparation or practice and that, "It is unethical to administer or discuss with pupils a parallel form or the specific form of the test to be used" (Wrightstone, 1967). An example of the new position is Cohen who advocates Criterion Referenced Instruction (CRI).

CRI presents the identical task to be learned in both the instructional process as well as in the final assessment, an ideal way to insure the precise match among what is taught, what is measured, and what is intended to be learned. The effect is near perfect learning. . . . (Cohen, 1987, p. 16)

Popham’s notion of measurement-driven instruction, although somewhat vague on the specific instructional processes, seems to have more in common with Cohen's position. "The competencies that are covered by the test will become curricular magnets that draw instruction toward themselves" (Popham, Cruse, Rankin, Sandifer, & Williams, 1985, p. 629).

The inferences made from a score on a standardized test should be to a different domain than those inferences drawn from a criterion-referenced test. (The terminology here, although widely used, is not what we would prefer—see Mehrens & Lehmann, 1987, p. 18.) But many practitioners apparently do not understand the distinction. The following section provides a closer look at the various inferences we may wish to draw from test scores and their relationship to what may be appropriate and inappropriate teaching to the test.

Inferences From Test Scores

As Mehrens (1984) has stated:

The only reasonable direct inference we can make from a test score is the degree to which a student knows the domain of material that the test samples. Any inference about why the student knows the domain to that degree . . . is clearly a weaker inference . . . (p. 10)

Nevertheless, the public may still wish to make indirect, weaker, and possibly inaccurate causal inferences. What everybody should agree on, however, is that causal inferences about student achievement cannot possibly be accurate unless the direct inference about student achievement is made to the correct domain.

Only on rare occasions does one wish to limit the inference about knowledge to the specific questions asked in a specific format. Even for making decisions about the adequacy of instruction, it is common for a test to sample the information taught. The inference is from this sample to the total domain of information taught. Of course, if one wished to show instruction to be maximally effective one could, as Cohen (1987) advocates, present an identical task in instruction and in final assessment. But then one could not generalize to slight variations in the task or be sure there is understanding rather than rote regurgitation of memorized information.

For making inferences only about the degree to which students learned the objectives (skills) taught, one should probably devise what is popularly called a criterion-referenced test that samples only those objectives (skills). Such a test will influence instruction. If the total set of objectives is important, it may be quite appropriate, as Popham suggests, to have measurement-driven instruction. However, if the test sampled from the total set of objectives, it would be inappropriate to focus instruction on only those objectives that happened to be tested. Such a practice would make invalid any inference from the test score to a student’s knowledge of the total set of objectives.

Many times educators and the public do not wish to make inferences just about the specific content that has been taught. For example, if parents wish to infer how well their children will do in another school next year, they need to infer to a general domain, not to the narrow and perhaps idiosyncratic domain of a single teacher’s immediate objectives. Lindquist (1951) suggested that a skillful teacher should develop broad and meaningful generalizations and it is the generalized outcomes that are the important ultimate objectives. (See Mehrens, 1984, for further discussion.)

Because the domain sampled by a standardized test is frequently considerably broader than the domain sampled by a criterion-referenced test, it is considered less appro-
prietate to teach to the specific objectives that have been sampled on such a test. Few, if any, standardized tests would cover all the objectives. There is a sampling of objectives as well as a sampling of items within objectives.

Although all the above should, and probably does, seem obvious to most measurement specialists, it seems wise to "state the obvious" as relevant background for the major portion of the paper—a discussion of what does and does not represent appropriate test preparation for standardized tests.

In the remainder of this paper, we will (a) present a continuum of types of test preparation activities from providing general instructions to teaching the test, (b) explore in some depth four sets of commercial materials designed to increase test scores, (c) discuss the attitudes and practices of educators toward test preparation activities, (d) review briefly the literature on the effectiveness of various test preparation programs, and (e) draw some conclusions and implications regarding the effectiveness and legitimacy of using commercial materials.

Types of Test Preparation Activities

In order to discuss whether test preparation methods and materials are fruitful, fruitless, or fraudulent it is necessary to define some terms and discuss some distinctions. As Shepard and Kreitzer (1987) make clear, at some point legitimate teaching to the test can cross over an ill-defined line and become inappropriate. Educators may not always agree on where to place that point on the continuum, but it seems useful to describe the continuum using some descriptive headings to identify points located on a reasonably straight line. We suggest the following descriptive headings:

1. general instruction on objectives not determined by looking at the objectives measured on standardized tests;
2. teaching test taking skills;
3. instruction on objectives generated by a commercial organization where the objectives may have been determined by looking at objectives measured by a variety of standardized tests. (The objectives taught may, or may not, contain objectives on teaching test taking skills);
4. instruction based on objectives (skills, subskills) that specifically match those on the standardized test to be administered;
5. instruction on specifically matched objectives (skills, subskills) where the practice (instruction) follows the same format as the test questions;
6. practice (instruction) on a published parallel form of the same test; and
7. practice (instruction) on the same test.

We submit that (1) above is always ethical and (6) and (7) are never ethical (although practices (6) and (7) would make standardized tests more "sensitive to instruction," which some regard as a desirable virtue); (2) would typically be considered ethical. Most measurement specialists believe that making students equally testwise will increase validity. Thus the point where one crosses over from a legitimate to an illegitimate practice on the ill-defined line must be somewhere between (3) and (5). We submit that the location of this point changes depending upon the inferences one desires to make from the test scores.

In the next section, we provide examples of commercial programs that fit under headings (2) through (5), and give examples of how closely the preparation materials match the objectives or questions of the tests.

Commercial Test Preparation Materials

We identified 12 commercially prepared achievement test preparation programs currently being sold. We selected four commercial test preparation programs for analysis as being representative of descriptive categories (2) through (5) on the continuum of test preparation activities listed above.


This program consists of 15 booklets organized into five categories that are identified as test-taking skills: following test directions, using time wisely, and analytical skills in reading, language, and mathematics.

The publisher did not attempt to match standardized achievement test questions or objectives with the instruction or practice activities in this program. The Riverside Publishing Company makes this point clear in their Teacher's Guide Manual (Pritchard, 1983, p. 5).

We compared the skills covered in the Riverside program with the skills covered in the California Achievement Test (CAT) Level 15, Form E; word analysis, vocabulary, comprehension, math computation, math concepts, and math application test sections. We found only 4 of the 69 skills on the CAT were covered in Improving Test-Taking Skills (see Table 1).

(9): Scoring High—Subject Centered: Random House

The second program we reviewed was the Scoring High in Math and Reading by Random House. (Random House also publishes test-specific workbooks to be reviewed later). The publisher states:

Each workbook in Scoring High's Subject Centered series covers a single content or process area—Reading, Language, Math, or Cloze—providing students with an intensive review and practice in the subject. All exercises are based on an in-depth analysis of the most frequently used standardized achievement tests...and cover the skills most often required on these examinations. Youngsters using the Scoring High will become familiar with—and prepared for—the varying kinds of formats found on the major standardized achievement tests. (Random House, 1987, p. 2)

There are also 37 test-taking skills taught in the math and reading books.

We compared the subskills in the Scoring High—Subject Centered Program with the 69 subskills tested on the CAT Level 15, Form E. One match point was given for those items that were tested on the CAT and were also included in the Scoring High—Subject Centered Programs. An example of a full 1 point match would be subtraction of a five-digit number from a five-digit number in a vertical format. If, for example, addition was tested on the CAT with vertical and horizontal formats, but only vertical formats were practiced in the Scoring High
—Subject Centered program, only 1/2 point match would be awarded.

Although the program contains more pages than the test-specific program to be described later, we found approximately 14% fewer matched items than in the test-specific program. Obviously the publisher either found it difficult or did not wish to include all the tested skills found in the five leading achievement tests into just one preparation program. Although not as many test items were matched by this program as those described in (4) and (5) to follow, the program still matched 55 out of 69 CAT subskills. Additional skills not included in the CAT Level 15, Form E, were found in this preparation program (see Table 1 for a summary of the comparisons).


CTB/McGraw-Hill (1987) advertises that, "use of the material . . . will increase levels of achievement.

. . . The materials are entitled CAT Learning Materials because they are based on objectives tested by the California Achievement Tests, Forms E and F" (p. 18).

The program includes excellent teaching strategies, suggested student activities, and application and practice sheets matched to each objective tested on the CAT. The instructional activity cards have the corresponding CAT test objectives printed on them. The program does not concentrate on test-taking skills or use the same test format found in the CAT.

We compared the CAT Learning Materials with the actual subskills tested on the CAT using the approach described earlier. The CAT Learning Materials' program was awarded 66 1/2 points out of a possible 69 (see Table 1).

An example of how close some practice items are to the actual test questions is the CAT item that displays a picture of a thermometer with each degree line also representing 2 degrees, the student is required to choose the letters (degree lines are labeled with letters) that indicate what the temperature would be if it was 10 degrees above and 10 degrees below the pictured temperature. Another example of a very close match is the test items and learning materials on consonant digraphs. The test has questions using charcoal, shingle, theme, bashful, and attach. The learning materials include chart, shower, those, washer and pouch.

The program contains additional objective activity cards and practice for subskills not tested on the CAT Level 15, Form E, because the CAT Learning Material program we reviewed is titled Level 14 and 15, for Forms E and F, which includes instruction and practice for tested objectives on Level 15, Form F, and Level 14, Forms E and F.

The program is too comprehensive to use all the materials as a brief test preparation program just prior to taking the achievement test. However, one could give students a pretest such as CTB/McGraw-Hill's California Diagnostic Mathematics and Reading Tests or use last year's CAT test results to identify the specific skills a student needs to practice. Following that diagnosis by selecting the skill lessons and the related practice items in the CAT Learning Materials' lessons would probably result in a significant increase on the subsequent resulting test scores. Whether one could still infer to a broad domain from the test scores is very problematic.

(5): Scoring High—Test Specific: Random House

The Test Specific Scoring High program published by Random House includes materials for various levels of the following tests: Comprehensive Tests of Basic Skills (CTBS), CAT, Iowa Tests of Basic Skills (ITBS), and the Metropolitan Achievement Tests (MAT).

---

**Table 1**

**Comparison of Several Test Preparation Programs to the California Achievement Test**

<table>
<thead>
<tr>
<th>Program</th>
<th>(2) Improving Test-Taking Skills</th>
<th>(3) Scoring Subject Centered</th>
<th>(4) CAT Learning Materials</th>
<th>(5) Scoring Test Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-taking skills</td>
<td>Yes (15)*</td>
<td>Yes (37)</td>
<td>No</td>
<td>Yes (33)</td>
</tr>
<tr>
<td>Matched subskills with the 69 on the CAT Level 15, Form E</td>
<td>4</td>
<td>55</td>
<td>66.5</td>
<td>64.5</td>
</tr>
<tr>
<td>Additional subskills covered in the Program not found on the CAT</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>CAT format followed</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Note.** Table 1 is based on comparisons of the California Achievement Test Level 15, Form E: Test A, Word Analysis; Test 1, Vocabulary; Test 2, Comprehension; Test 6, Mathematics Computation; and Test 7, Mathematics Concepts and Applications with the following test preparation programs: Improving Test-Taking Skills for Grades 4-6; the Subject Centered Scoring High in Math book B; and Scoring High in Reading book B; CAT Learning Materials Mathematics Level 14-15; and CAT Learning Materials Reading, Study Skills Level 14-15; and the Test-Specific Scoring High on the California Achievement Test, Book B. *The number in parentheses indicates the number of test-taking skills.
The publisher states that:
Scoring High helps raise student performance on today's most widely administered achievement tests. Each Scoring High work- book in the test-specific series covers the multiple subject areas which are addressed in a single, specific achievement test. The exer cises in these workbooks provide your students with the concentrated practice and review of the very skills necessary to score high on the latest edition of each test. (Random House, 1987, p. 2-3)

We compared the subskills in the Scoring High on the CAT prepara tion program with the actual subskills tested in the CAT using the same procedure previously described. We found a match of 64.5 points out of a possible 69 between the actual tested subskills and those practiced in the preparation program (see Table 1).

The program consists of a student workbook about 70 pages long containing practice with skills tested on the CAT, Forms E or F. The list of the reading skills and math skills tested in the CAT and the lessons in which they appear in the program, are listed in the Scope and Sequence chart located at the beginning of the Teacher's Edition book. We did not find any discussion or emphasis on subskills not tested on the CAT. Each practice page has the same format as the CAT. The program also includes instruction on 33 test-taking skills and practice on timed tests. The program's design requires students to do a page and then the teacher is instructed to review the answers and work through the answer choices with the students. Future researchers need to pay close attention to this aspect of the program. The amount of practice and instruction students receive on each item they had difficulty with, would seem to determine how successful the program will be. It is our belief that this program serves as a pretest for the CAT in the same manner as if one actually used the CAT as a pretest prior to giving the same CAT at a later time.

The publisher seems to have made an effort to change slightly the way some questions are worded while being sure not to omit skills tested on the CAT. For example, the CAT asks students to identify the figure that does not show a line of symmetry. The Scoring High program asks the students to find the figure in which the dotted line is not a line of symmetry. Another example is the thermometer pictured in Scoring High; it has each degree line representing 1 degree compared to the CAT picture representing 2 degrees. As with the CAT Learning Materials, the consonant digraphs in the test are all covered with very similar words in the Scoring High materials.

### Attitudes and Practices Towards Test Preparation

As we, and others, have suggested, there is an increasing tendency to judge both the quality of schools and individual teachers based upon students' scores on standardized tests. The result is to increase the pressures to teach inappropriately to the test (see Lewis, 1986; Pechman, 1985; Prell & Prell, 1986). Indeed, as we have suggested earlier, there is a philosophical base for teaching to the test (Cohen, 1987; Popham et al., 1985), but the wisdom of that philosophy is debatable.

### Views on Cheating

What do teachers and other educators consider inappropriate test preparation—cheating? To what extent do educators engage in inappropriate practices? An interesting survey was conducted in a Dallas Independent School District on this issue (Gonzalez, 1985). A sample of 147 teachers and test administrators were asked seven questions about what constitutes cheating and asked to respond on a 4 point scale from "not at all" to "definitely."

Four of the questions and their responses are given in Table 2 (in a different order than in the original survey). It is comforting to see that 89% believe teaching actual test items is definitely cheating—but discounting that 11% do not realize this. We are more dismayed by the other results. We would consider all four practices as inappropriate (cheating may be too strong a word) if one wishes to infer from the test results to a broader domain of content.

What is and is not inappropriate test preparation (cheating) will likely remain a subject of debate for some time. As Kurtines and Gewirtz (1984) point out, however, the likelihood of cheating is a function of the likelihood of detection, the magnitude and importance of the gains that cheating will bring and the anticipated consequences if caught. Gonzalez (1985) wonders whether currently the severity of the consequences are sufficient to deter cheating. She points out that in ancient China "the death penalty was in effect for examinees as well as examiners if anyone was found guilty of cheating during civil service examinations." She indeed does add that "No one suggests such penalties today" (p. 5).

### Proper Preparation Defined

Ligon and Jones (1982) give an excellent definition of an appropriate activity for preparing students for standardized testing: "One which contributes to student's performing on the test near their true achievement levels, and one which contributes more to their scores than would an equal amount of regular classroom instruction" (p. 1). This is admittedly a bit abstract, because the "true score" changes with the domain to which one is inferring. Nevertheless, it is a good conceptual definition. Ligon and Jones, as well as other professionals, go far beyond this conceptual definition and provide fairly specific guidelines regarding what is considered appropriate test preparation activities.

Matter (1986), for example, discusses legitimate ways to prepare students for testing. A partial quote follows: "Ideally, test preparation activities should not be additional activities imposed upon teachers. Rather, they should be incorporated into the regular, ongoing instructional activities whenever possible" (p. 10).

Other measurement experts employed by school districts have made similar, very sensible statements. There is reason to believe, however, that such well thought out approaches/guidelines for preparation activities are still not as widespread among school districts as we would ideally hope. Covert (1986) sent all members of the National...
Association of Test Directors a survey on practices and policies regarding test preparation. He found that although most (86.5%) districts provide help to local schools in identifying acceptable standardized test preparation activities, 71.2% do use practice tests (which could be inappropriate). Only 15.4% have a formal test Code of Ethics.

Various states do have guidelines regarding test preparation materials. For example, South Carolina requests that districts not use the Scoring High materials. They have an agreement with the California Test Bureau that the form of the CAT used by the state will not be sold to individual districts or teachers in the state. Of course, one could use alternate forms as test preparation materials or indeed, one could probably obtain a copy of the form from someone residing in another state. Michigan has published what we consider an excellent set of guidelines for test preparation for their state administered Michigan Educational Assessment Program (MEAP) (Michigan State Board of Education, 1987).

Publishers have also written guidelines for appropriate test preparation activities. An example is CTB/McGraw-Hill's (1985) guidelines. They encourage the use of practice tests “to give students experience in taking standardized tests and to identify students who may need help with the mechanics of test taking” (p. 1). They are opposed to coaching using materials that “are essentially forms of the test” because this “may increase test scores without any real increase in the skills being measured.” They go on to correctly point out that:

A standardized achievement test is meant to sample the broad domains of skills represented in the test. Coaching or instruction that focuses only on content as defined by sample test items mean that the test may no longer be a valid sample and that generalizations that may be made about a student’s acquisition of the broad domain of skills based on performance on the test are severely limited. CTB believes that the Scoring High materials published by Random House fall into this category and we do not support their use . . . (p. 1)

Interestingly, but not surprisingly, CTB/McGraw-Hill supports what they call targeted instruction:

Instruction focused on the skills measured by the test as defined in terms of broad category objectives—not test items—can facilitate the learning process and lead to increased levels of achievement in the curriculum topics sampled by the test. Test results can be used to identify from among these curriculum topics, the greatest needs of individual students, as well as groups, and to focus instruction on those needs. CTB believes that objective-based instructional aids such as its BASIS Kits, CAT Learning Materials, and Classroom Kits used with its criterion-referenced tests can provide legitimate support for instruction in the curriculum topics sampled by the test. Because the Learning Materials, BASIS, and the Classroom Kits cover the topics thoroughly, this instruction need not constitute ‘teaching to the test.’ (p. 2)

Dale Foreman (1987, personal communication), who is a coauthor of the subject matter (not test specific) Scoring High materials states that his original intent was:

To provide students with some type of learning-oriented practice in the test taking skills over an extended period of time. Short periods of concentrated study would not be of much use to the student if he/she was to improve. (p. 1)

---

### TABLE 2

**Educators’ Views Regarding What Is Cheating in Preparation for Standardized Tests**

<table>
<thead>
<tr>
<th>A form of cheating?</th>
<th>Not at all</th>
<th>Somewhat</th>
<th>Mostly</th>
<th>Definitely</th>
</tr>
</thead>
<tbody>
<tr>
<td>If copies of the actual test are made available by colleagues and actual test items are taught in the classroom, does this constitute a form of cheating?</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>89</td>
</tr>
<tr>
<td>If a teacher is aware of vocabulary words tested on a standardized test and teaches these words, does this constitute a form of cheating?</td>
<td>21</td>
<td>21</td>
<td>12</td>
<td>46</td>
</tr>
<tr>
<td>If a teacher remembers specific questions and then proceeds to teach next year’s students the same questions does this constitute a form of cheating?</td>
<td>24</td>
<td>21</td>
<td>17</td>
<td>37</td>
</tr>
<tr>
<td>Does the use of former versions of standardized tests for test-taking practice in a district using the current form of the test constitute a form of cheating?</td>
<td>54</td>
<td>30</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

*Adapted from Gonzalez, 1985.*

---

Spring 1989
He therefore wished to design a program which would:

(1) provide meaningful instruction on important objectives of the subject area of concern, which would also reflect the most commonly occurring objectives in the frequently used norm-referenced testing programs, (2) provide practice in taking a test using a variety of testing formats that existed in the testing programs, and (3) provide the student with some test-taking skills which would make him/her a more efficient and accurate test taker. It was important that the program not be an attempt to teach the test, but to teach content-based test taking skills which would improve the student's ability to obtain test scores that would more honestly reflect his or her classroom performance in the skill area of concern. (1987, p. 1)

**Effectiveness of Test Preparation Programs**

There are few systematic evaluations of commercial test preparation programs because they are reasonably new on the market. Most of the studies relate to noncommercial programs that attempt to teach test wisdom (test taking skills). Studies of the impact of training programs have produced somewhat mixed results.

Most of the research and controversy about the effectiveness of training programs has centered around those programs designed to impact scores on aptitude tests. However, there is some research on the coachability of achievement tests (see Bangert-Drowns, Kulik, & Kulik, 1983).

**Research on Teaching Test-Taking Skills**

Much of the research has been done on the effects of teaching test-taking skills. We will not review all that literature. Indeed it is somewhat conflicting in results. Scruggs, White, and Bennon (1986) reviewed the literature and computed 65 mean effect sizes based on 24 investigations of programs that specifically taught test-taking skills. They computed separate effect sizes for methodologically adequate and inadequate studies and for criteria they categorized as achievement tests, tests of test wisdom, self-esteem, and anxiety. The average of 44 effect sizes for achievement test scores from methodologically adequate studies was only .10. The authors concluded that it was not clear cut whether training in test-taking skills should be provided instead of spending the same amount of time teaching reading or math.

Samson (1985) conducted a meta-analysis of 24 studies and found a mean effect size of .33. Many writers believe that both evidence and logic suggest that test wisdom can be taught and that it will increase scores on both measures of test wisdom and standardized achievement tests.

**General Research**

In 1983, Bangert-Drowns, Kulik, and Kulik conducted a meta-analysis on 30 controlled studies focusing on the effectiveness of coaching (defined quite generally) for achievement tests. They found that 25 of the 30 studies showed the coaching program had a positive effect on test performance. The average effect size was .25. The average effect size varied with the duration of the training intervention. For short test-taking orientation and practice sessions the average effect size was .17; it was .43 for those programs that included cramming on sample test questions; and it was .66 for a single study that provided general instruction in "broad cognitive skills" (Bangert-Drowns et al., p. 578).

**Research Using Materials Produced By Commercial Publishers**

The Chicago Public Schools (Byrd, 1987) conducted a study comparing the effectiveness of four test preparation programs: Random House's Scoring High materials, Riverside's Improving Test-Taking Skills, Continental Press' On Target for Tests, and Hammond's Reading Skills for Standardized Testing. The criterion measure was the Reading Comprehension Subtest of the Iowa Test of Basic Skills. Although the authors admit to some problems in the design of the study, they "failed to find any decisive differences between students who received commercially prepared instruction in test-taking strategies and those who did not..." (Byrd, 1987, pp. viii, ix).

Deaton, Halpin, and Alford (1987), compared students who received instruction from Scoring High on the California Achievement Test (Random House) with a control group. Although multivariate analyses revealed some significant differences between the groups, results "suggested that the Scoring High program did not produce consistent increases in student scores" (p. 149).

**Research Using Matched Skills and Matched Format**

Data on the effectiveness of matching both objectives (skills) and format in instruction come from those who argue for measurement-driven instruction. If one has a reasonably small set of objectives, builds items measuring these objectives based on specific item specifications, and then publishes these objectives and item specifications as an aid to test preparation, it could reasonably be expected that instruction using the study guide would result in higher scores. Although we are not aware of any experimental research on this question, anecdotal evidence suggests such study guides are effective in raising scores. Shepard (1987) (see also Shepard & Kreitzer, 1987), describes the monumental effort that went into preparing for the Texas Examination of Current Administrators and Teachers (TECAT). In the section on Basic Skills and 'Teaching to the Test' she clearly explains why the research team members "were not entirely convinced that all of the gains on the test reflected real increases in teachers' skills" (p. 69). The report gives examples of what Shepard terms the "exploitation of the test specifications" (p. 71). This exploitation alters the validity of the broader inference we really wish to make.

**Research on Practice**

In 1984, Kulik, Kulik, and Bangert conducted a meta-analytic synthesis of 40 studies that used practice forms of the tests as the treatment variable. They found an average effect size of .42 in 19 studies in which subjects had one practice trial on a test identical to the criterion, an effect size of 1.89 for seven practice trials, and an effect size of .23 for
a single practice session on a test parallel to the criterion. In concluding their article, they suggest that programs that do not produce as large a gain as the ones they studied may be “inefficient in design. The programs might produce better results if practice trials on strictly parallel tests were made an important instructional element in the programs” (p. 444). This suggestion should give one pause if the goal is to make an inference to a domain of objectives which the test only samples.

**Summary, Conclusions, and Implications**

Standardized tests are used increasingly in evaluating the quality of the local schools. This places pressure on the administrators and teachers to engage in activities that are intended to increase students’ scores. Commercial materials are widely available and heavily advertised as serving this purpose. Their effectiveness is not widely researched but appears to be limited at best. Such materials may be inappropriate depending upon both how closely they match the tests and the inference one wishes to make from the test scores. The literature suggests that many of the leading school test directors understand the issue of legitimate ways to prepare students for standardized tests but that many districts do not have any formal policy regarding test preparation. Anecdotal information suggests that some actual cheating goes on with respect to teaching the test.

A summary of our comparison of the four selected test preparation programs to the CAT is presented in Table 1. In our analyses of these commercial materials, we found a very close match between the CAT and both Scoring High on the CAT and the CAT Learning Materials. We judge the match in both cases to be too high to recommend using either set of materials prior to testing if one wishes to make an inference from the test score to a broader domain than the test samples. It is interesting, and a bit ironic, that the CTB/McGraw-Hill guidelines quoted earlier find the Scoring High materials objectionable because they match too closely the CAT test items but support the CTB Learning Materials. Although the CTB/McGraw-Hill materials do cover objectives tested in both CAT levels 14 and 15, the match to items is close indeed.

It should be noted, however, that the CAT Learning Materials do not use the format of the CAT. This decreases the legitimacy of the charge that they “teach to the test.” Further, we agree with the notion that, “Test results can be used to identify from among these curriculum topics, the greatest needs of individual students, as well as groups, and to focus instruction on these needs” (CTB/McGraw-Hill, 1985, p. 2).

Thus, because the CAT Learning Materials are well designed curricular materials we would support their judicious use following testing. Whether these materials are more effective in teaching the “broad category objectives” than other available commercial materials is, as far as we know, a matter of conjecture.

The match between the Scoring High—Subject Centered materials and the CAT is less close (55 out of 69 subskills with additional skills not tested on the CAT included in the program). Thus, using the subject centered materials would not be as apt to invalidate an inference from a score on the CAT to a more general domain. The match between Riverside’s *Improving Test Taking Skills* and the CAT is even less close and using those materials would not invalidate a broader inference. (Although we did not do a detailed analysis of CTB/McGraw-Hill Skills Bank materials they seem to cover the more general domain and using them would not invalidate the broader inference.)

Although one cannot generalize to other standardized tests and other commercial programs from our analysis, it seems reasonable to “draw the line” between legitimate and illegitimate practices between points (3) and (4) of our descriptive list in *those cases where we wish to infer to objectives, skills, or subskills not sampled by the test*. The same would be true for criterion referenced tests if the objectives actually on the test were only a sample from a broader domain. For example, for teacher licensure, the objectives considered important for a teacher to be minimally competent frequently exceed the number that can be tested on one test. In such a case the total set of objectives should be communicated. It would be appropriate to publish a test preparation guide that prepared a student for the total set of objectives. It would be inappropriate to publish a guide where the objectives were isomorphic with those on the test.

For those criterion-referenced tests where the objectives comprise (do not sample from) the entire domain of interest, the line should be drawn between (4) and (5) [or perhaps between (5) and (6)]. Although this paper did not focus on such tests, we submit that it is possible for the format to be so close that the inference must be limited to the particular objectives when the items are written in a specific manner. One is hardly ever interested in limiting the inference of, say, differentiating fact from opinion in a reading passage to making a choice among two fact and two opinion options. Very little of our everyday reading presents us with two fact and two opinion options (one of each in and one of each not in the passage) for the paragraphs we read. If we really wish to know whether individuals can make the differentiation in real life we should not put in our test preparation guides the specific details of the items because that results in “exploitation of the test specifications” (Shepard, 1987, p. 71). As states increasingly publish study guides for state built and mandated tests, this is becoming an increasingly important issue.

We hope this paper has served to increase the understanding of the old, but increasingly relevant, issue of teaching to the test. What is and is not appropriate must ultimately be based on the closeness of the match of the preparation materials to the tests and the inference one wishes to make from the test scores. From an examination of several commercial test preparation programs and a review of the literature on the effectiveness of such programs, we conclude that it is likely that at least some of them are both fruitless and fraudulent. *Caveat emptor!*
References


Cannell, J. J. (1987). Nationally normed elementary achievement testing in America’s public schools: How all fifty states are above the national average. Daniels, West Virginia: Friends for Education.


Call for Manuscripts

Educational Measurement: Issues and Practice seeks quality manuscripts from NCME members and others who are concerned with issues in educational measurement and with the practice of educational measurement. For information, contact Anthony J. Nitko, Editor, 5812 Forbes Quadrangle, Department of Psychology in Education, University of Pittsburgh, Pittsburgh, PA 15260 or Suzanne Lane, Assistant Editor, 5812 Forbes Quadrangle, Department of Psychology in Education, University of Pittsburgh, Pittsburgh, PA 15260.

Letters

Readers are encouraged to write letters to the editor commenting on material appearing in this and the previous issues of EmP. Short letters are given preference over longer letters. Letters may be edited to increase brevity. Authors will be offered the opportunity to respond to comments made by letter writers. Letters not published will not be returned to the writer.