March 2001 | Volume **58** | Number **6** Helping All Students Achieve Pages 16-20

Teaching to the Test?

W. James Popham

In an era of high-stakes and high-stress testing, how do we ensure that classroom instruction does not give way to inappropriate teaching?

American teachers are feeling enormous pressure these days to raise their students' scores on high-stakes tests. As a consequence, some teachers are providing classroom instruction that incorporates, as practice activities, the actual items on the high-stakes tests. Other teachers are giving practice exercises featuring "clone items"—items so similar to the test's actual items that it's tough to tell which is which. In either case, these teachers are teaching to the test.

What is Teaching to the Test?

Although many use the phrase, educators need to understand exactly what *teaching to the test* means. Educational tests typically represent a particular set of knowledge or skills. For example, a teacher's 20-item spelling quiz might represent a much larger collection of 200 spelling words. Therefore, the teacher can distinguish between test items and the knowledge or skills represented by those items.

If a teacher directs instruction toward the body of knowledge or skills that a test represents, we applaud that teacher's efforts. This kind of instruction teaches to the knowledge or skills represented by a test. But if a teacher uses the actual test items in classroom activities or uses items similar to the test items, the teacher is engaging in a very different kind of teaching. For clarity, I will refer to teaching that is focused directly on test items or on items much like them as *item-teaching*. I will refer to teaching that is directed at the curricular content (knowledge or skills) represented by test items as *curriculum-teaching*.

In *item-teaching*, teachers organize their instruction either around the actual items found on a test or around a set of look-alike items. For instance, imagine that a high-stakes test includes the multiple-choice subtraction item "Gloria has 14 pears but ate 3." The test-taker must choose from four choices the number of pears that Gloria has now. Suppose the teacher revised this item slightly: "Joe has 14 bananas but ate 3." The test-taker chooses from the same four answers, ordered slightly differently. Only the kind of fruit and the gender of the fruit-eater have been altered in this clone item; the cognitive demand is unchanged.

Curriculum-teaching, however, requires teachers to direct their instruction toward a specific body of content knowledge or a specific set of cognitive skills represented by a given test. I am not thinking of the loose manner in which some teachers assert that they are "teaching toward the curriculum" even though that curriculum consists of little more than a set of ill-defined objectives or a collection of vague and numerous content standards. In curriculum-teaching, a teacher targets instruction at *test-represented* content rather than at test items.

Is Teaching to Test Items Wrong?

The purpose of most educational testing is to allow teachers, parents, and others to make accurate inferences about the levels of mastery that students have achieved with respect to a body of knowledge (such as a series of historical facts) or a set of skills (such as the ability to write particular kinds of essays). Because the amount of knowledge and skills that teachers teach is typically too great to test everything, tests sample those bodies of knowledge or skills. For example, on the basis of a student's ability to write one or two persuasive essays on a given topic, we can infer the student's general ability to write persuasive essays. If our interpretation of the student's skill in writing essays is accurate, we have arrived at a valid performance-based inference about the student's mastery of the skill represented by the test.

Similarly, when a student scores well on a 10-item test consisting of multiplication problems with pairs of triple-digit numerals, we infer that the student can satisfactorily do other problems of that ilk; hence, he or she appears to have mastered multiplying pairs of tripledigit numbers. If a test-based inference is valid and the teacher gets an accurate fix on students' current knowledge or skills, then the teacher can make appropriate instructional decisions about which students need additional help, or, if all the students do well, whether it's time to switch to new instructional targets.

To illustrate, suppose a district-developed reading vocabulary test includes 25 items from a set of 500 words that reflect the target vocabulary words at a particular grade level. If the test yields valid interpretations, a student who answers 60 percent of the items correctly will, in fact, possess mastery of roughly 60 percent of the 500 words that the 25-item vocabulary test represents. If the test yields valid inferences, of course, teachers can make suitable decisions about which students need to be pummeled with more vocabulary instruction. Similarly, district-level administrators can allocate appropriate resources—for example, staff-development focused on enhancing students' reading vocabularies.

Curriculum-teaching, if it is effective, will elevate students' scores on high-stakes tests and, more important, will elevate students' mastery of the knowledge or skills on which the test items are based. If a teacher, however, gets a copy of the district test, photocopies its 25 vocabulary items, and drills next year's students on those 25 items, valid test-based interpretations become impossible. A student's score on the test would no longer indicate, even remotely, how many of the designated 500 vocabulary words the student really knows. Valid inferences disappear as a consequence of item-teaching.

Because teaching either to test items or to clones of those items eviscerates the validity of score-based inferences—whether those inferences are made by teachers, parents, or policymakers—item-teaching is reprehensible. It should be stopped. But can it be?

Detecting Inappropriate Test Preparation

One way of deterring inappropriate conduct is to install detection schemes that expose misbehavior. For example, when professional athletes are informed that they will be subjected to unannounced, random urine testing to determine whether they have been using prohibited substances, there is typically a dramatic reduction in the athletes' use of banned substances. The risk of penalties, at least to many people, clearly exceeds the rewards from engaging in proscribed behavior.

Unfortunately, I have found no practical procedures to detect teachers who are using inappropriate test preparation. Let me illustrate the difficulties by describing a fictitious teacher, Dee C. Ving. A 5th grade instructor in a school mostly serving low-income youngsters, Dee has consulted the descriptive information accompanying the national standardized achievement test that her 5th graders will take in the spring. She finds those descriptions inadequate from an instructional perspective: They are both terse and ambiguous. Dee simply can't aim her instruction at the knowledge or skills represented by the test items because she has no clear idea about what knowledge or skills are represented.

Frustrated by the overwhelming pressure to improve her students' scores, Dee engages in some full-scale item-teaching. One of her friends has access to a copy of the test that Dee's

Popham, W. J. (2001). Teaching to the test. Educational Leadership, 58(6), 16-20.

students will take and loans it to Dee for a few days so that Dee can "understand what content your students will really need to know."

Dee, having covertly made a copy of key sections of the test, devotes one or two days each week to what she rationalizes as test-targeted instruction. In her explanations and practice exercises, she uses either actual items taken from the test or slightly modified versions of those items. Not surprisingly, when Dee's 5th graders take the standardized achievement test in the spring, most of them score very well. Her students last year scored on average in the 45th percentile, but her students this year earn a mean score equal to the 83rd percentile.

The scores, of course, provide invalid interpretations about the students' actual mastery of the content. But let's give Dee the benefit of the doubt by assuming that she genuinely believed she was helping her students get high scores and, at the same time, was making her school look good when the district compared schools' test performances. Dee, we assume, is not fundamentally evil. She just hasn't devoted much careful thought to the appropriateness of her test-preparation practices.

Could we have detected what Dee was up to? Let's say that, at some level, she recognizes that she has done something inappropriate. She is reluctant to reveal to colleagues or administrators that she relied on photocopied test items and slightly altered versions of those items. How could we determine that this year's high test scores were attributable to Dee's item-coaching rather than to good instruction?

Detection Procedures Doomed to Fail

What options might we have to apprehend Dee as she dished out item-teaching to her 5th graders?

Teacher self-reports. We might survey a school's teaching staff, and even devise the survey so that teachers' responses will be anonymous, to see whether teachers respond truthfully to questions about item-teaching. But teachers like Dee did not tumble off the turnip truck yesterday and would undoubtedly supply socially desirable, if inaccurate, responses to a self-report. Few teachers gleefully let the world know that they engage in questionable teaching practices.

Teacher-collected materials. We might also require teachers to compile a set of tests and practice exercises that they have used in their classes. Theoretically, we could inspect such materials to see whether they contained any actual items from the high-stakes test or any massaged versions of those items. But Dee will surely be shrewd enough to sanitize the materials that she puts in her required compilation. She'll destroy any incriminating papers and probably rely on chalkboard explanations and practice exercises. Chalkboards can be erased ever so completely.

Oral exercises also are difficult to monitor. Once uttered, they evaporate. Moreover, it is both naive and professionally demeaning to ask teachers to assemble a portfolio of potentially self-incriminating evidence. In most schools, such a requirement would be a genuine morale-breaker.

Pre-announced classroom observations. If Dee's principal lets her know that he or she will visit the classroom on Wednesday of that week, that principal will see no item-teaching. Dee knows how to play the high-stakes score-boosting game. And allowing a principal to walk in on an item-focused teaching activity would violate the rules of the game. The principal will see only good teaching.

Unannounced classroom observations. Whereas pre-announced classroom observations by a school administrator give teachers ample time to display appropriate lessons, unannounced observations do not. Unannounced visits, therefore, ought to work better than pre-announced ones. But this detection ploy is not promising on three counts.

First, it casts the unannounced visitor in a negative "Gotcha!" role. Few school-site administrators enjoy playing police officer. Second, forcing a school principal or other administrator to undertake this surveillance duty will diminish that person's effectiveness as an ally for a teacher's improvement. And reduced effectiveness, in the long run, is certain to harm the quality of instruction for students. Third, visiting teachers' classrooms to ensure that no inappropriate test preparation is underway is enormously time consuming. The administrator's other responsibilities may suffer.

Student self-reports. There are other eyewitnesses to what goes on in a classroom—the students. Theoretically, students could periodically complete anonymous instructional questionnaires, containing actual or slightly altered versions of high-stakes test items. We could then ask them whether the teacher provided explanations or practice exercises focused on items similar to those on the instructional questionnaire.

Yet most students would have difficulty determining the degree of similarity between a questionnaire's sample items and the practice or explanatory items that they had already seen. Besides, this tattle-on-teacher activity could create an unsavory relationship between teachers and students. Indeed, as soon as they figured out the purpose of the questionnaire, unhappy students could readily get revenge by falsely asserting that they had been given oodles of practice items.

Score jumps. I often advise parents to view with suspicion any substantial year-to-year increases that they see in their children's test scores. There is far too much likelihood that because of pressures to boost students' test scores, teachers have engaged in inappropriate test preparation—or, worse, violations of the prescribed test-administration procedures. When student scores jump dramatically from one year to the next, I urge parents to look into what's going on instructionally at the school. Standardized achievement tests are notoriously insensitive to instruction. That is, such tests typically fail to detect the impact of even first-rate instructional improvements.

But, of course, scores can jump because of improved instruction. Suppose, for instance, a school served a large number of students whose first language was not English. Students' poor test performances in the previous year may be directly attributable to their inability to read the actual test items. Recognizing the problem, the school's staff may have directed instructional energy toward students' reading comprehension. And, as a result, students' scores could have improved dramatically.

On the one hand, a score jump may signal the presence of item-teaching or worse. On the other hand, a score jump may arise because of improved instruction. By themselves, score jumps can't detect improper instruction.

Does all this mean that we simply avert our eyes while inappropriate test preparation becomes even more common in U.S. schools? Can this inappropriate practice ever be effectively deterred? Surprisingly, the answer is a decisive *yes*.

Deterrence Strategies

Provide a hefty dose of assessment literacy. I have spoken to many teachers about their testpreparation practices, especially teachers who are seriously pressured to raise their students' test scores. The vast majority of them have never considered the appropriateness of their

Popham, W. J. (2001). Teaching to the test. Educational Leadership, 58(6), 16-20.

test-preparation practices. Indeed, after learning that teaching directly toward test items created invalid inferences about their students, most teachers are both surprised and dismayed.

I am not suggesting that once teachers recognize instructional improprieties, such improprieties instantly disappear. Some teachers, unfortunately, already understand quite well the effects of their item-focused teaching. The score-boosting pressures that those teachers experience lead them toward practices that, absent such pressure, they would regard as repugnant.

But I believe that the vast majority of teachers, if they recognize the adverse effects of itemteaching, will abandon such teaching. The first deterrence should be an aggressive attempt to enhance teachers' assessment literacy—especially as it relates to the impact on the validity of test interpretation. Teachers should understand not only the difference between item-teaching and curriculum-teaching, but also the impact that those types of teaching have on their students.

Help policymakers understand what kinds of high-stakes tests they should use. Some teachers succumb to item-teaching because, if they truly believe they are obliged to raise test scores, they think they have no alternative. More often than not, those teachers are correct.

There's no way a pressured teacher can provide students with curriculum-teaching if he or she doesn't have a clear description of the knowledge and skills represented by the test items. Obviously, for a teacher to focus instruction on the curricular content that a test represents, that content must be spelled out sufficiently for the teacher's instructional planning. A teacher, looking over what curricular outcomes a high-stakes test represents, should understand those outcomes well enough to plan and deliver targeted lessons. Anything less descriptive drives teachers down a no-win instructional trail leading to item-teaching.

Thus, the second tactic is to educate policymakers to support only high-stakes tests that are accompanied by accurate, sufficiently detailed descriptions of the knowledge or skills measured. A high-stakes test unaccompanied by a clear description of the curricular content is a test destined to make teachers losers. Moreover, because of the item-teaching that's apt to occur, tests with inadequate content descriptors also will render invalid most test-based interpretations about students.

For teachers to direct their instruction toward tangible teaching targets, not only should they have clear descriptions of the curricular content assessed by a test, but they should also have some reasonable assurances that good teaching will pay off in improved student test scores. In an effort to use such an approach, Hawaii education authorities recently overhauled the state's content standards—the knowledge and skills that the Hawaii Board of Education has directed the state's teachers to promote. One element of the process was to reduce the number of content standards to a smaller, more intellectually manageable number of curricular targets. A second element of the revision was to clarify what a content standard actually signified in terms of the knowledge or skill embodied in that standard.

State officials then enlisted an established test-development contractor to develop a test suitable for ascertaining students' mastery of the revised content standards. Each item measured one of the state's content standards. After the contractor developed the test items and identified the designated content standard for each item, committees of Hawaii educators reviewed each item's quality. One of the review questions was "If a teacher has supplied effective instruction directed toward students' mastery of this item's designated content standard, is it likely that most students will answer the item correctly?"

Hawaii education officials attempted to create a test that would allow teachers to engage in curriculum-teaching, rather than item-teaching, by targeting the state's content standards. If Hawaii's teachers can focus their instruction on curricular targets yet feel confident that student test scores will rise with effective instruction, they will have no need to engage in rampant item-teaching.

Deterrence and Detection

The core issue underlying this problem is easy to define. If students' scores jump, is it because those students are really able to leap over higher hurdles, or have the students been surreptitiously given stepladders? We surely do not wish to penalize a teacher who delivers instruction so stellar that students' performances go into orbit. But we don't want that orbit to be illusory.

In 1999, we learned that a United States president can be impeached for high crimes and misdemeanors. I'm not sure whether item-teaching is, technically, a high crime or a misdemeanor. But because it can harm children, I lean toward the high crimes label—and such instructionally criminal conduct is increasing.

No realistic procedure identifies and, hence, dissuades those teachers who choose to engage in item-teaching. Our best approach to deterrence lies first in getting educators to understand the difference between, and the consequences of, item-teaching and curriculum-teaching. Then, we must not use high-stakes, pressure-inducing tests that are not accompanied by content descriptions sufficiently clear for teachers' on-target instructional planning. If we prohibit instructionally opaque tests, teachers will no longer be victims of a score-boosting game that they cannot win. If we use tests with clarified instructional targets, teachers can focus their classroom efforts on getting students to master what they're supposed to learn.

Author's note: This article was adapted from a presentation at the annual meeting of the American Educational Research Association, New Orleans, Louisiana, April 24–28, 2000.

W. James Popham is UCLA Emeritus Professor, Graduate School of Education and Information Studies, University of California, Los Angeles, CA 90095; <u>wpopham@ucla.edu</u>.