

TEACHING TO THE TEST: HIGH CRIME, MISDEMEANOR, OR JUST GOOD INSTRUCTION*

W. James Popham
University of California, Los Angeles

American teachers are experiencing enormous pressure these days to raise their students' scores on the kinds of high-stakes tests being used for diploma-denial, grade-to-grade promotion, or to evaluate the effectiveness of a school's staff. Sometimes, as a consequence of this score-boosting pressure, teachers provide classroom instruction incorporating, as practice activities, the actual items on the high-stakes test. In other instances, teachers provide their students with practice exercises featuring "clone items," that is, items so remarkably similar to the test's actual items that it's tough to tell which is which. In either case, these teachers can surely be said to be "teaching to the test."

What is "Teaching to the Test?"

Although many people employ the phrase, "teaching to the test," in a fairly cavalier manner, it is important for educators to understand exactly what this often pejorative phrase really means, and what it doesn't. To start with, educational tests are typically intended to *represent* a particular set of skills and/or knowledge. For example, a teacher's 20-item spelling quiz can be employed to represent a much larger collection of, say, 200 spelling words. Therefore, it is possible to distinguish between (1) a test's items and (2) the knowledge and/or skills represented by those items.

If a teacher directs instruction toward the body of knowledge and/or skills a test is supposed to represent, we typically applaud that teacher's efforts. Good instruction ought to be aimed at knowledge and/or skills that are sufficiently important for us to test students' mastery. This kind of instruction can be called teaching toward the knowledge and/or skills *represented* by a test. Some people, not inclined to turn a phrase with all that much precision, refer to such instruction as "teaching to the test."

Yet, other people will describe "teaching to the test" as a situation arising when the teacher either uses the test's *actual* items in classroom instructional activities or uses items so similar to the test's actual items as to be almost indistinguishable. Clearly, this is a very different kind of "testing to the test" than when a teacher aims instruction at *test-represented* curricular targets.

If I could, I would immediately expunge the phrase "teaching to the test" from our educational lexicon, forcing folks to say *either* "teaching to the test's items" or "teaching to the knowledge/skills represented by the test." Because people really can't tell which of those two significantly different meanings is being employed when someone says "teaching to the test," a good deal of resultant confusion would be dodged if we never used that expression.

* A version of this essay appeared in the March 2001 issue of *Educational Leadership*, Vol. 58, No. 6, pp. 16-20.

For convenience, in the remainder of this essay, when I refer to teaching that's focused directly on items in a test or on items much like them, I shall describe such instruction as *item-teaching*. In contrast, when referring to teaching that's directed at the curricular content (knowledge and/or skills) *represented* by a test's items, I'll call that type of instruction *curriculum-teaching*.

When referring to *item-teaching*, I will be thinking specifically about teachers who organize their instruction, for instance, teacher-explained illustrative items or item-based practice activities—either around the actual items found in a test or around a set of look-alike items. For instance, imagine that in an actual high-stakes test there's a multiple-choice subtraction item in which Gloria has 14 pears, but ate three. The test-taker must choose, from four answer choices, the number of pears that Gloria now has left. Suppose the teacher revised this item slightly so that Joe has 14 bananas, but ate three. The test-taker is asked to choose from the same four answer-options whose order is slightly altered from the real test's item about Gloria and her pears. Only the kind of fruit being consumed and the gender of the fruit-eater have been altered in this clone item. The essence of the cognitive demand facing the test-taker in the Joe-and-the-bananas item is unchanged from that present in the actual Gloria-and-the-pears test item.

Curriculum-teaching, however, requires teachers to direct their instruction toward a specific body of content knowledge and/or a specific set of cognitive skills *represented* by a given test. I am not thinking of the fairly loose manner in which some teachers assert they are “teaching toward the curriculum” even though that curriculum consists of little more than a set of ill-defined objectives or, more recently, a collection of vague and often too-numerous content standards. Curriculum-teaching, in this context, refers to the aiming of a teacher's instruction at *test-represented* content rather than at test items.

Is Teaching to the Test's Items Wrong?

The purpose of most educational testing is to allow teachers, parents, and other interested parties to arrive at accurate inferences about the levels of mastery that students have 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 bodies of knowledge and the sets of skills that teachers must teach are typically too enormous to determine if students have mastered *everything*, educational tests typically *sample* those bodies of knowledge and/or skills.

So, based on a student's ability to write one or two assigned persuasive essays on given topics, we make an inference about the student's general ability to write persuasive essays. If our test-based interpretation about the student's skill in writing such 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 composed of pairs of triple-digit multiplication problems, we arrive at the inference that the student can satisfactorily do other problems of that ilk, hence appears to have mastered multiplying pairs of triple-digit numbers. If a test-based inference is valid, and teacher gets an accurate fix on students' current knowledge and/or skills, then the teacher can make appropriate instructional decisions about which students need additional help or whether, because all students are doing well, it's time to switch to new instructional targets.

To illustrate, suppose in a district-developed "reading vocabulary" test, 25 items have been based on randomly drawn words from a set of 500 words reflecting the target vocabulary words at a particular grade level. If the test yields valid interpretations, a student who answers 60 percent of the test's 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, then teachers can make suitable decisions about which students need to be pummeled with more vocabulary instruction. Similarly, district-level administrators can make appropriate resource-allocation decisions, for example, regarding how much district staff-development attention should be given to the enhancement of students' reading vocabularies.

Curriculum-teaching, if it is effective, will simultaneously elevate students' scores on a test's items and, more importantly, elevate students' mastery of the knowledge and/or skills on which the test items are based. However, if a teacher happened to capture a copy of the district test, photocopied its 25 reading vocabulary items, and drilled next year's students relentlessly on those 25 items before the test was administered, this would render valid test-based interpretations impossible. No longer would a student's score on the test indicate, even remotely, how many of the designated 500 vocabulary words the student really knew. A student, because of item-focused coaching, might answer *every one* of the test's 25 items correctly. And yet, that student might actually know only a small proportion of the 500 vocabulary words. Valid inferences disappear as a consequence of item-teaching.

Accordingly, because teaching either to a test's items or to clone-like replicas 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?

Detection of Inappropriate Test-Preparation

One way of deterring individuals from engaging in inappropriate conduct is to install detection schemes that will expose such misbehavior. For example, when professional athletes are informed that they will be subjected to unannounced, random urine-testing to determine if those athletes have been using prohibited substances, there is typically a dramatic reduction in the athletes' use of the banned substances. Even if people are inclined to engage in inappropriate conduct, they become reluctant to

do so if such misconduct is apt to be caught. The risk of penalties, at least to many people, clearly exceeds the rewards derivative from engaging in proscribed behavior.

Unfortunately, I have concluded that, as a practical matter, there are no procedures available to us that can be used to “threaten” teachers into the avoidance of inappropriate test-preparation. Let me illustrate the difficulties of doing so by describing an absolutely fictitious teacher along with some possible procedures that might be employed to determine if the teacher has improperly prepared students for a high-stakes test.

Suppose our fictitious teacher, pressured to raise scores on a high-stakes test, simply concludes there’s no instructionally defensible way to do so. For purposes of this illustration, let’s call our fictitious teacher Dee C. Ving, a fifth-grade instructor in a school mostly serving low-income youngsters. Dee has consulted the descriptive information accompanying the nationally standardized achievement test that her fifth-graders are to take in the spring, and she finds those descriptions altogether inadequate from an instructional perspective. The test publisher’s depictions of the content the fifth-grade version of the achievement test is supposed to represent are simultaneously terse and ambiguous. Dee simply can’t aim her instruction at the knowledge and skills represented by the test’s items because she has no clear idea about what those skills and knowledge are.

Frustrated by the overwhelming pressure she is experiencing to improve her students’ scores, Dee decides to engage in some full-scale item-teaching. One of her friends has access to a copy of the tests that Dee’s students will be taking, and loans it to Dee for a few days so, as her friend says, Dee can “understand what sorts of content your students will really need to know.”

Dee, having covertly made a copy of key sections of the test, decides to devote one or two days per week to what she rationalizes as “test-targeted” instruction. In her explanations and practice exercises for the class, she either uses actual items taken from the test or employs slightly modified versions of those items. Not surprisingly, when Dee’s fifth-graders take the standardized achievement test in the spring, most of them score very well. Using the test’s national norms, whereas her students last year, on average, scored at the 45th percentile, this year’s students’ earn a mean score equal to the 83rd percentile.

The scores, of course, provide invalid interpretations about students’ actual mastery of the content tested—such invalidity having been created because Dee taught directly to the test’s items. But let’s give Dee the benefit of the doubt by assuming she genuinely believed she was really helping her students get high scores and, at the same time, making her school look good when the district schools’ test performances were compared. Dee, we will assume, is not fundamentally evil. She just hasn’t devoted all that much careful thought to the appropriateness of her test-preparation practices.

Could we have detected what Dee was up to so she could have been stopped as she taught directly toward the test's items? Let's say that, at some murky level, she has recognized, at least in the view of several of her district's administrators, she has been doing something that wouldn't be completely approved. So, she is reluctant to reveal to colleagues or administrators that she is supplying instruction that relies on photocopied test items as well as slightly altered versions of those items. How could someone have determined that this year's students' high test scores were attributable to Dee's item-coaching rather than to good instruction?

Detection-Procedures Doomed to Fail

If we set out to apprehend Dee as she dished out item-teaching to her fifth-graders, let's consider several likely procedural options—and their merits. But each of these procedures, as you will see, has serious shortcomings.

Teacher self-reports. It is possible to survey a school's teaching staff, even devising the survey so that teachers' responses will be truly anonymous, to see if teachers will respond truthfully to questions about whether they had provided item-teaching. But teachers such as Dee did not tumble off the turnip truck yesterday, so would assuredly supply "socially desirable," even if inaccurate responses to such a self-report questionnaire. Few teachers gleefully let the world know that potentially unsavory teaching may be taking place in their classrooms. No, self-reports won't work.

Teacher-collected materials. It is also possible to require teachers to compile an ongoing set of all tests and practice exercises they have used in their classes. Theoretically, such materials could later be inspected to see if they contained any actual items from the high-stakes test or any mildly massaged versions of those items. Yet, Dee will surely be shrewd enough to sanitize the materials she puts in her required compilation of materials. She'll surely shred or otherwise destroy any incriminating stuff. She'll also probably rely on many more 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. No, the scrutiny of a teacher's collected assessment and practice exercises won't work either.

Pre-announced classroom observations. If Dee's principal lets her know that a classroom visit will take place on, say, Wednesday of this week, you can be assured that the principal will see no item-teaching taking place. Dee surely knows how to play the high-stakes score-boosting game by the rules. And allowing a principal to walk in on an item-focused teaching activity would violate one of the game's unspoken rules. The principal will see only good things going on. Pre-announced classroom observations, alas, also won't help detect inappropriate test-preparation.

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

First, it casts the unannounced visitor in a thoroughly negative “Gotcha!” role. Few school-site administrators will enjoy playing policeperson. Second, forcing a school principal, for instance, to undertake this surveillance duty will surely diminish the principal’s effectiveness as a teacher’s improvement-focused instructional ally. And reduced effectiveness of a principal’s improvement-focused activities, in the long run, is certain to harm the quality of instruction received by students. Finally, if school-site administrators are going to spend so much time visiting teachers’ classrooms to be sure no inappropriate test-preparation is underway, this will be enormously time-consuming activity. The administrator’s other responsibilities will surely suffer because of the many classroom visits required. No, although intuitively appealing as a winning detection-ploy, unannounced classroom observations won’t solve the problem.

Student self-reports. Besides teachers, there are other eye-witnesses to what goes on in a classroom, namely, the students themselves. Theoretically, then, it would be possible to have students periodically complete anonymous “instructional questionnaires,” containing actual or slightly altered versions of a high-stakes test’s items, then ask the students if the teacher had provided explanations or practice exercises focused on items very similar to the “instructional questionnaire’s” sample items.

Yet, most students would surely have difficulty in determining the required degree of similarity between a questionnaire’s sample items and the practice or explanatory items that had been used previously by the teacher. Besides, this sort of tattle-on-teacher activity could create an unsavory relationship between teachers and students. Indeed, as soon as students had figured out what the purpose of the questionnaire was (and this would take less than five minutes in most cases), unhappy students could readily gain a degree of “revenge” by falsely asserting they’d been given oodles of practice on items like those in their “instructional questionnaires.” No, using students as in-class monitors of a teacher’s action doesn’t seem to be a suitable detection model.

Score jumps. I have often advised parents to view with solid suspicion any really substantial year-to-year increases they see in students’ test scores at their child’s school. There’s far too much likelihood these days that, because of pressures to boost students’ tests scores, inappropriate test-preparation practices have taken place 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 scores can jump up, of course, because improved instruction actually took place. Suppose, for instance, that a school served a large number of children whose first language was not English. Well, students' poor test performance in the previous year may have been directly attributable to their inability to read the actual test items. Recognizing the problem, the school's staff may have directed all sorts of instructional energy toward the promotion of students' reading comprehension. And, as a consequence of many students' new-found ability to read the standardized test's items, students' scores could have improved dramatically.

A score jump, all by itself, may signal the presence of item-teaching or worse. On the other hand, a score jump can arise just because improved instruction actually took place. All by themselves, therefore, score jumps can't be used to detect improper instruction.

In sum, we must regrettably arrive at a negative judgment about each of the most likely detection procedures that might be employed as an up-front deterrent to teachers who might otherwise teach toward test items. Does this mean that we simply avert our eyes while inappropriate test-preparation becomes even more common in the nation's schools than it is today? Putting it in other words, can an inappropriate practice that is largely *undetectable* ever be effectively *deterred*? Surprisingly, I think the answer to this question is a decisive Yes.

A Dual-Direction Deterrence Strategy

Providing a hefty dose of assessment literacy. I have spoken to a good many teachers about their test-preparation practices, especially teachers who are being seriously pressured to raise their students' test scores. The vast majority of those teachers have never given any thoughtful consideration to the appropriateness of their test-preparation practices. Indeed, after having been informed that by teaching directly toward a test's actual items they were creating invalid inferences about their students, most teachers are both surprised and dismayed.

I am not suggesting that once teachers recognize instructional improprieties, such improprieties will instantly disappear. Some of the teachers I've discussed this issue with, unfortunately, already understand quite well what the effects their item-focused teaching will be. The score-boosting pressure those teachers are experiencing simply leads them toward practices that, absent such pressure, would be regarded as repugnant.

But I believe that the vast majority of teachers, if they really recognized the adverse effects item-teaching will have, will abandon such teaching. The first deterrence tactic to be followed, then, should be an aggressive attempt to enhance teachers' assessment literacy—especially as it relates to the impact on the validity of test interpretation that ensues from item-teaching. Teachers should clearly understand

not only the difference between item-teaching and curriculum-teaching, but also the impact on children of those two types of teaching.

Helping policymakers understand what kinds of high-stakes tests should/ shouldn't be used. The reason that some teachers succumb to item-teaching is 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 for a pressured teacher to provide students with curriculum-teaching if there is not a clear description available of the knowledge and/or skills represented by a test's items. Obviously, for a teacher to focus instruction on the curricular content a test represents, then that curricular content must be spelled out at a level of descriptive detail sufficient for purposes of a teacher's instructional planning. A teacher, upon looking over what curricular outcomes a high-stakes test represents, should understand those outcomes well enough so that the teacher can plan and deliver *curricularly* on-target lessons. Anything less in the way of descriptive clarity drives teachers down a no-win instructional trail leading to item-teaching.

Thus, the second tactic in my dual-direction solution strategy is to educate policymakers so they do not allow a test to be used for high-stakes decisions that is not accompanied by accurate, sufficiently detailed descriptions of the knowledge and/or skills being measured by that test. A high-stakes test unaccompanied by a clear depiction of the curricular content it assesses is a test destined to make teachers losers in a no-win score-boosting game. Moreover, because of the resultant item-teaching that's apt to occur, tests with inadequate content descriptors also will render invalid most test-based interpretations about students. Any test that is apt to induce instructional pressure on teachers must either provide clarified descriptions of the content it assesses or it should not be employed.

A recent example from Hawaii. In order for teachers to be able to direct their instruction toward tangible teaching targets, not only should there be clear descriptions available of the curricular content assessed by a test, but there should be some reasonable assurances that good teaching will pay off in improved students' test scores. In an effort to use such an approach, Hawaii educational authorities recently completed a thoroughgoing overhaul of the state's content standards, that is, the knowledge and skills the Hawaii Board of Education has directed the state's teachers to promote. One element of that revision process was to reduce the number of content standards to a smaller, more intellectually manageable number of curricular targets. A second purpose of the revision was to clarify more completely what it was that a content standard actually signified in terms of the knowledge or skill embodied in that standard.

State officials then enlisted the assistance of an established test-development contractor to develop a test suitable for ascertaining students' mastery of the revised content standards. Each item was designated by the contractor so as to primarily measure one of the state's content standards. Then, after all items had been developed, and the designated content standard for each item identified by the

contractor, committees of Hawaii educators reviewed each item's quality. One of the review questions asked, a question requiring a Yes/No answer from an item reviewer, was the following:

If a teacher has supplied effective instruction directed toward students' mastery of this item's designated content standard, is it likely that most of that teacher's students will answer the item correctly?

It is apparent Hawaii educational officials were attempting to create a test that would allow teachers to engage in curriculum-teaching (by aiming at the state's content standards) rather than item-teaching. If Hawaii's teachers can focus their instruction on curricular targets, yet be confident that students' test scores will rise if effective instruction is supplied, there will be no need for the state's teachers to engage in rampant item-teaching.

Deterrence and Detection

The core dilemma underlying this problem arena is easy to define. If students' scores jump up, is it because those students are really able to leap over higher hurdles, or have the students surreptitiously been given step-ladders? 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 legally ejected from office 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. Such instructionally criminal conduct is increasing in our nation. I want to see its frequency dramatically diminished.

I have suggested, however, that there is no realistic procedure available for identifying, hence for dissuading, those teachers who choose to engage in item-teaching. Accordingly, I believe our best approach to deterrence lies first in getting educators to understand the difference between, and the consequence of, item-teaching and curriculum-teaching.

Then, and most importantly, we must not permit high-stakes, pressure-inducing tests to be used that are not accompanied by content descriptions sufficiently clear for teachers' on-target instructional planning. If we prohibit instructionally opaque tests, then teachers will no longer be victims of a score-boosting game they cannot win. If, instead, we employ tests with clarified instructional targets, then teachers can focus their classroom efforts on getting students to master what they're supposed to learn.