

SCRUTINIZING HIGH-STAKES TEST ITEMS FROM AN INSTRUCTIONAL PERSPECTIVE*

W. James Popham
University of California, Los Angeles

As America's educational accountability juggernaut gains strength, the nation's teachers find themselves increasingly pressured to boost students' scores on assessments that, without question, are properly described as *high-stakes tests*. Moreover, the consequences now linked to students' performances on these tests have become greater and greater. Not only are students being denied diplomas or grade-to-grade promotions, but significant cash rewards for educators (as much as \$25,000 per teacher) are now riding on how students' test scores turn out.

Harmful Classroom Consequences

Unfortunately, the enormous score-boosting pressures that teachers currently experience have led to a host of classroom practices that are unarguably harmful to children. We see teachers reducing otherwise rich curricula to a paltry collection of only those things measured by a high-stakes test. We see teachers transforming their otherwise interesting classrooms into drudgery-laden drill factories that soon stamp out any joy students might derive from their schooling. And, finally, we see teachers who engage in test-preparation tricks that even naïve students can recognize as blatant or borderline cheating. To deny the harmful impact of such practices on children these days is to deny reality.

But, in many instances, teachers adopt instructionally unsound classroom practices because the high-stakes tests currently being used do not give teachers a legitimate chance to be successful. Most of today's high-stakes tests, unfortunately, have not been built so that teachers can organize and deliver genuinely effective instruction aimed at the domain of knowledge and skills represented by these significant tests. The trick, of course, is to create high-stakes tests that can simultaneously serve their accountability functions while, at the same time, help teachers make more defensible instructional decisions. Such instructionally facilitative tests *can* be built. In this paper I will describe one state's efforts to do so. More specifically, I will recount an attempt by officials in Hawaii to create large-scale assessments to measure students' mastery of that state's content standards. Even more specifically, I will describe activities carried out by the Hawaii Department of Education (DOE) in an attempt to make certain the items for their new tests would not only assess the state's content standards accurately, but would do so in a manner that could make a meaningful contribution to teachers' instructional success.

Stage Setting

First off, because the context for any such an assessment undertaking is always important, it should be noted that Hawaii, our nation's only single-district state, had in late 1998

* Presented at the annual meeting of the National Council on Measurement in Education, Seattle, Washington, April 11-13, 2001.

hired a new state superintendent, Dr. Paul LeMahieu, who was committed to a standards-based reform strategy. One of the new superintendent's first acts was to support an effort to refine the state's existing content standards (that is, the knowledge and skills Hawaii's public schools were supposed to promote). The original standards were published in June 1994 as the *Hawaii Content and Performance Standards*.

Those original content standards were not only numerous (1,544 in all), but quite varied in the levels of specificity explicated in different subjects. For instance, there were 495 and 418 content standards, respectively, in language arts and science. Yet, in mathematics and social studies there were, respectively, only 119 and 133 content standards. Moreover, a good many of the original content standards were—because of their general phrasing—somewhat difficult to interpret.

As a consequence of a substantial refinement effort, in late 1999 a revised set of content standards was published, namely, the *Hawaii Content and Performance Standards II* (HCPS II). The knowledge and skills embodied in HCPS II were the ones that a new, to-be-built statewide standards-based test would attempt to assess. Architects of Hawaii's standards-based reform initiative recognized that this new standards-based test would play a pivotal role in their state's reform efforts.

Even though the revised content standards presented in HCPS II reflected a substantial effort to clarify and reduce the number of such standards, there remained a considerable number of curricular outcomes represented in the revised standards. In mathematics, for instance, the new conceptualization called for a total of five "strands" (such as *measurement*) to embrace 14 content standards. These 14 standards, however, were often broad, e.g., "Understand various types of patterns and functional relationships," or multifaceted, e.g., "Pose questions and collect, organize, and represent data to answer those questions." As a consequence, the 14 mathematics content standards are subdivided by grade clusters (K-1, 2-3, 4-5, 6-8, 9-12) into "benchmarks," many of which are also multifaceted in nature, e.g., "Represent and use whole numbers, fractions, decimals, whole number percents, and ratios."

Because, even with all these subdivisions, there is still less than satisfactory clarity regarding the meaning of certain of the standards and benchmarks, DOE officials subsequently intend for a series of "performance indicators" to be created as a further attempt to operationalize the outcomes contained in HCPS II.

The situation in language arts is similar with respect to standards, grade-cluster benchmarks, and to-be-developed performance indicators. In language arts there are six strands covering three components, namely, (1) reading and literature, (2) writing, and (3) oral communication. The grade-cluster benchmarks are, as was seen in mathematics, often multifaceted in nature, e.g., "Use strategies for constructing meaning that include annotating, interpreting, connecting, and analyzing." As in mathematics, performance indicators are to be constructed to exemplify students' mastery of particular standards/benchmarks.

Prior to the use of a new standards-based test, Hawaii's children had most recently been tested annually using the *Stanford Achievement Tests*. Although it was foreseen that some items

from the ninth edition of those tests (the SAT-9) would be used for purposes of deriving national norm-based comparative data, the new test was to consist largely of newly constructed items specifically designed to measure HCPS II skills and knowledge. The test-development contractor designated to create the new test was Harcourt Educational Measurement (HEM), also the publisher of the SAT-9.

A Pivotal Item-Review Activity

Recognizing that an educational test is no better than its items, DOE officials determined that a significant step in the test-development process would occur during a January 10-12, 2000 review of HEM-developed items by committees of Hawaii educators. Based on the results of that January review, items would then be field-tested in Hawaii by HEM during the spring of 2000.

Because this item-review activity would significantly determine the nature of the new test, DOE assessment personnel devoted considerable thought to the procedures and the evaluative dimensions that would be employed to govern the review process. Four evaluative dimensions were finally chosen, each of which will be described below. Once these evaluative factors had been identified, they were immediately transmitted to HEM so that their firm's item-developers would be aware of the factors that would be used to scrutinize items during the upcoming item-review activity.

The Four Item-Review Evaluative Dimensions

Each item in the pool of to-be-reviewed items prepared by HEM was to be judged on four separate bases. Every reviewer participating in the three-day January 2000 review activity was to render a per-item judgment according to each of the four evaluative criteria to be described below. The review criteria were formulated as questions that were to be responded to, positively or negatively, by the item reviewers.

Standards Congruence. The first question a reviewer was to answer regarding each test item dealt with the degree to which the item coincided with the content standard in the HCPS II that it was chiefly intended to assess. HEM personnel, acting upon a request by DOE officials, had identified the primary content standard and grade-range benchmark being assessed by each item. This designated content standard, as you will see from the initial item-review question presented below, was central in an item-reviewer's judgment.

Standards Congruence: *Will students' responses to this item help educators accurately determine whether students have mastered the knowledge and/or skill embodied in the designated content standard the item is intended to assess?*

When reviewers answered this question, Yes or No, they were to remember that a test item—all by itself—can rarely provide a definitive answer to a student's mastery of a content standard. Thus, reviewers were reminded that the question asked whether students' responses to the under-review item will *help* educators make a determination about students' mastery of a

particular content standard. In other words, reviewers were to consider whether the item was sufficiently congruent with the designated content standard so that, in concert with other standard-congruent items, a valid inference about students' mastery of that standard was likely to be made.

To answer the *standards-congruence* question accurately, reviewers were told to think about just what it was that the architects of the content standards had in mind when they crafted the specific standard designated for the under-review item. In written orientation materials, reviewers were told, "This is not an occasion when any old sort of content relevance should lead you to a Yes answer. For instance, suppose a social studies content standard called for students to draw inferences about the meaning of certain events in the Civil War, but an item merely asked students to display memorized information about Civil War facts. Even though the item is, in some loose way, relevant to the content standard, the item ought to be judged negatively with respect to the *standards-congruence* question. Even in concert with similar items, it would be impossible to make valid judgments about a student's inference-making skills."

Instructional Sensitivity. The second review question for each item dealt with the ability of the new standards-based tests to detect successful instruction if Hawaii teachers did a good job in promoting their students' mastery of the HCPS II standards.

Instructional Sensitivity: *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?*

As you can see from the question, a Yes or No answer was once more required of reviewers. Notice that the *instructional sensitivity* question asks reviewers to judge whether, if a teacher is doing a solid instructional job in promoting students' mastery of a given standard, *most* of the teacher's students will be able to answer the item correctly. Obviously, not every student will always answer an item correctly, even if the teacher has done a superb instructional job. But what this question asked item-reviewers to judge is this: If the teacher has taught reasonably well toward the content standard the item is supposed to help measure, will the bulk of the teacher's students come up with a correct answer?

In the orientation materials, reviewers were told that, "As part of the overall standards-based reform strategy now being implemented in our state, it is imperative that if the state's teachers provide increasingly effective instruction focused on the HCPS II content standards, then the results of such improved instruction will clearly be mirrored by students' improved test scores. If the new standards-based state tests are instructionally sensitive, then improved instruction will be seen when it takes place. If the tests are instructionally *insensitive*, then such improvements will be masked. The new tests must be instructionally sensitive. But this will only happen if the test's items are, themselves, instructionally sensitive."

Out of School Factors. The third evaluative dimension dealt with two out-of-school variables over which teachers have no influence. If either of those two factors were present, it was argued, then the item under review should not be included in the new standard-based tests.

Out-of-School Factors: *Is this item essentially free of content that would make a student's likelihood of answering it correctly be dominantly influenced either by the student's socioeconomic status or the student's inherited academic aptitudes?*

One of the factors this question asked reviewers to focus on was the extent to which children from advantaged backgrounds were likely to do better on an item than children from less advantaged backgrounds. For example, suppose a test item in science revolves around an exotic fruit that might be purchased by an affluent family, but would never be purchased by a low-income family. It is clear that children from affluent homes would have an advantage on that exotic-fruit item (over their less affluent counterparts) simply because of *socioeconomic status* (SES). If a child's socioeconomic status would meaningfully influence the child's likelihood of getting an item correct, reviewers were told to answer this third item-review question with a No.

A second element of this third item-review question dealt with *inherited aptitudes*. Item-reviewers were reminded of the increasingly popular view of intelligence among today's educators and psychologists that there are a number of intelligences, not just one. For example, children possess differing degrees of *interpersonal* intelligence (regarding other people) as well as varying amounts of *intrapersonal* intelligence (regarding oneself). The kinds of academic aptitudes referred to in this third item-review question, however, were the traditional quantitative, verbal, and spatial abilities that, at least to some degree, are inherited.

Reviewers were told that, "Clearly, a child's inherited intellectual aptitudes, especially the academic ones (such as a child's verbal, spatial, or quantitative capabilities) play *some* role in how a child answers *any* test item. What this out-of-school factors question is asking you to focus on, however, is whether the item is really getting at things taught in school or, instead, is aimed at intellectual aptitudes the child brings to school. Is inherited aptitude the *dominant* factor that the item is assessing? If you find yourself reviewing a test item that is really intended to see how academically quick-witted a child is, then you should answer No for this item-review question. Because there are *two* factors involved in this third item-review question, if you think that *either* SES *or* inherited academic aptitude is dominantly influencing in a student's likelihood of answering this item correctly, then mark No on your response form. To mark a Yes, the item should be free of both factors."

Absence of Bias. The final item-review evaluative dimension dealt with the important issue of assessment bias.

Absence of Bias: *Is this item essentially free of content that might offend or unfairly penalize students because of personal characteristics such as race, gender, or religion?*

This fourth item-review question, as was the case with the previous item-review question, was based on two factors that could incline reviewers to judge an under-review item negatively. Item-reviewers were told to judge an item negatively if it contained content that might *offend or unfairly penalize* students because of their personal characteristics such as gender or ethnicity.

Reviewers were informed that a test item might *offend* certain students if, in the item, members of a group were portrayed in a stereotyped manner. To illustrate, if minority youngsters were depicted in an item as members of rowdy gangs while majority youngsters were not, then the item should be judged to be biased because it might offend minority students. Another example of an offensive test item would be one that portrayed members of a minority group as dull-witted or an item that implied females could not succeed in “the hard world of business.” Test items that offend students are apt to have an adverse effect on those students’ performances because an offended student will often be distracted when completing subsequent items and will, therefore, perform more poorly on those later items than would otherwise be the case. Item-reviewers were urged to think of how they would be apt to perform on an examination if some of the test’s items disparaged their own race, religion, or background.

Item-reviewers were told that a test item might *unfairly penalize* a particular group of students if those students performed less well than another group of students, even though both groups were at the same achievement level with respect to the knowledge or skill being tested. This difference could be caused, for example, by dissimilar interests of the two groups. It could also be caused by differences in the two groups’ mastery of a skill (or knowledge) that was irrelevant to what was being tested. To illustrate, suppose a test item required students to draw a conclusion from a reading selection about cosmetics. It is possible that males would perform less well on such an item than females, *not* because males are less able to draw conclusions (the skill being tested), but because males may be less interested in and have less knowledge about cosmetics than females.

Reviewers were asked in the orientation materials to recognize that a biased test item is one that *unfairly* penalizes a student. “If test items are well constructed, they may very properly penalize students who *ought* to be penalized. If students haven’t studied certain content, they may answer an item incorrectly. But, of course, there’s nothing *unfair* about that sort of penalty.” Reviewers were also told that if they found that an item *might* be biased, on either of the two bases identified in the *absence of bias* question, they were to indicate briefly what the nature of the bias was. The orientation materials pointed out, “For instance, if you think an item might offend physically disabled people, you could simply write (in the space provided) ‘offensive to the handicapped.’ Note that you do not have to be *certain* an item contains biased content. If you think there’s a meaningful possibility that the item *might* offend or unfairly penalize a student because of personal characteristics, answer No to this fourth item-review question.”

Procedural Fundamentals

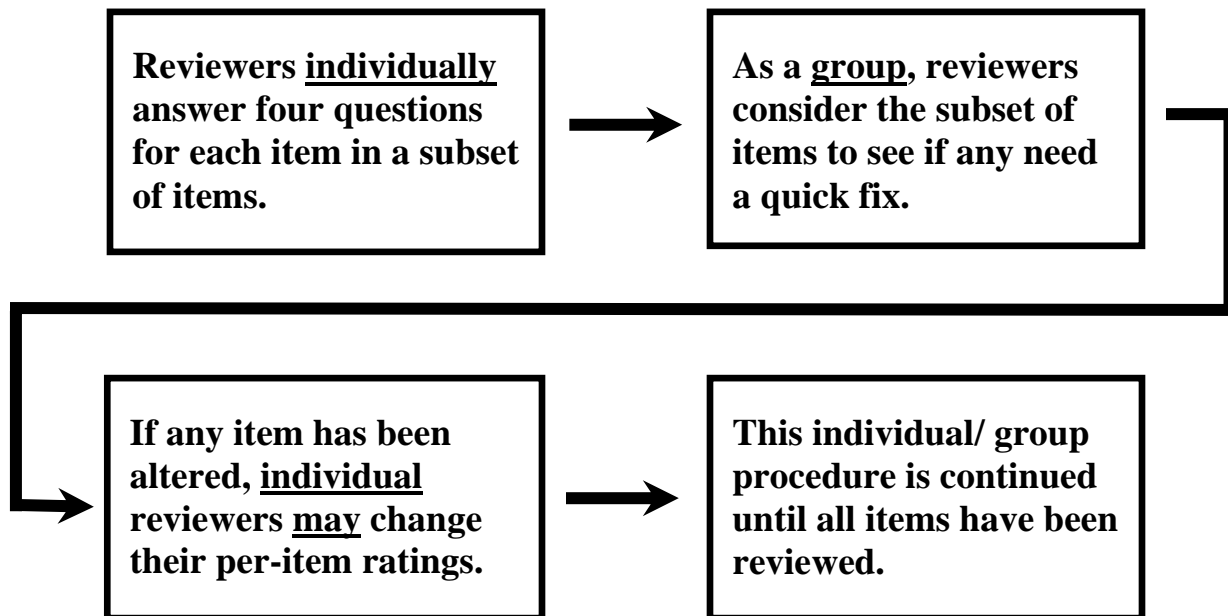
Item reviewers, prior to the January meeting, had been sent by mail a brief description of the upcoming three-day review activity and have been urged to familiarize themselves with the HCPS II outcomes in the subject area for which they were to review items.

At the beginning of the January item-review session, an orientation session of approximately 90 minutes was conducted to clarify the nature of the item-review activity. Approximately 100 reviewers were told that the purpose of the item-review activity was “to

determine whether items in the pool are appropriate for assessing students' mastery of the HCPS II content standards.”

An overview of the item-review process was presented. Figure 1 contains a graphic representation of the process.

Figure 1. The Item-Review Process



Reviewers were also alerted to factors that might cause them to judge items too stringently or too leniently. These factors are set forth in Figure 2.

Figure 2. Factors that may cause you to judge items

too stringently:

- Believing that one item, all by itself, should measure a student's mastery of a particular standard.
- Recognizing that, at this point, many of the state's teachers may not have zealously promoted HCPS II content standards.

too leniently:

- Wanting our state's students to be able to answer a test item correctly.
- Being deferential to the professional test-developers who constructed the items.

After explanations had been given of each of the four item-review questions, the reviewers received practice in responding to each of those questions. Reviewers were also given ample opportunities to raise questions, make comments, or seek clarification during the orientation session. At the end of the orientation session, all reviewers went to their assigned review rooms where they received review forms along with the items they were to review. Review committees, for a given subject area and grade range, typically consisted of 10-15 members. HEM had, as requested, designated the HCPS II content standard to which a particular item was dominantly addressed.

Security-control procedures were employed throughout the three-day review session. At the conclusion of the review, HEM and DOE staff summarized and analyzed the data as needed for the next step in the test-development process, namely, the isolation of items to be included in the Spring 2000 field test.

Looking Back

In retrospect, this was the not the first time Hawaii's DOE staff had reviewed items prepared by an external contractor. It was, however, the first time that DOE had built in an item-review process around four evaluative factors clearly intended to produce "a different kind of test." It was hoped that the item-review operation described herein would lead to statewide standards-based tests that not only would yield useful accountability evidence, but would also support Hawaii's teachers as they attempted to promote the state's content standards.

Although those involved in the January 2000 item-review operation surely have personal opinions about the fruitfulness of this sort of item-review process, the ultimate proof of this assessment pudding will be the nature of the tests it produces and the impact of those tests on classroom practices in Hawaii.