

ASSESSMENTS THAT ILLUMINATE INSTRUCTIONAL DECISIONS*

by

W. James Popham
University of California, Los Angeles

How does someone create a test that helps teachers make better decisions about how to instruct their students? That's a question I've been trying to answer for over 30 years. It's a difficult question. It's an important one.

A Bit of Stage-Setting

Two years ago, I had the opportunity to open this conference by challenging members of the measurement community to create large-scale assessments that not only satisfied today's accountability requirements, but also addressed the instructional needs of classroom teachers. If sufficient ingenuity were employed, I argued, tests could be built to simultaneously serve both an accountability and instructional function.

After that 1998 presentation (see Popham, 1999), more than a few of my colleagues put to me, in one form or another, the following question: "All right, Jim, what would these dual-purpose tests look like that could, at the same time, satisfy an accountability and an instructional mission?" In the remainder of this paper, I'm going to try to supply an answer.

Fortunately, during this session two colleagues whom I respect are also going to take a shot at answering that same question. Perhaps, by pooling our three answers, a serviceable set of insights can be derived about how to build large-scale assessments that can do a solid accountability job, yet still offer teachers guidance in how to promote students' mastery of the knowledge and skills those assessments are supposed to measure.

As I began to organize my thinking about the essay, that is, about the creation of instructionally illuminating educational tests (English teachers who subscribe to the *writing process* would refer to this as "pre-writing."), I paused to read Bob Linn's recent analysis dealing with assessment and accountability (Linn, 2000). In that insightful essay, Linn somberly concludes that, during his measurement career, "in most cases

* A presentation at the 30th Annual National Conference on Large-Scale Assessment, Council of Chief State School Officers, Snowbird, Utah, June 25-28, 2000.

the instruments and the technology have not been up to the demands that have been placed on them by high-stakes accountability.” Regretfully, he asserts that the use of large-scale, accountability oriented assessments has not “improved education and student learning in dramatic ways.” Linn suggests that, at least in part, teachers can be discouraged from “narrow teaching to the test” if the officials of high-stakes accountability programs annually install test-forms that have been equated to previously used test-forms. He recognizes, of course, that the costs of annually birthing creating brand new forms of high-stakes tests may prove prohibitive.

I concur with Linn’s pessimistic appraisal of the impact of high-stakes accountability tests, and I’m sure his new-forms-each-year approach, albeit costly, would dissuade many teachers from teaching toward specific test items. Nevertheless, I think that there is another way, a better way, to make high-stakes tests positively influence instruction. And that way is to deliberately build tests with teachers’ instructional decisions in mind. I’ll turn now to how such tests could be created.

Given space and time constraints, I am only going to provide four how-to-do-it rules that should be followed by those who wish to create instructionally illuminating assessments. Although I’ll not be able to elaborate all that much on each of these rules. I hope they convey a reasonable idea of how I think instructionally illuminating tests can be built. The rules have been painfully derived. That is, they are based on three decades worth of pain-inducing test-construction errors that I’ve personally made.*

I’ll conclude the paper with an example to illustrate the kinds of assessments I have in mind. Along the way, I’ll assume that the rules I cite are being written for the test-development staff of an agency that has been commissioned to create a test for a high-stakes accountability system. I doubt if many busy classroom teachers will have the time to grind out the sorts of tests I’ll be identifying. Yet, I think there are some elements of the rules that, at least in part, are relevant to the creation of teachers’ classroom assessments.

* In passing, I’d like to note that for about 25 years I headed an R&D assessment group that specialized, at least for a decade or so, in the creation of large-scale assessment programs. We decided to develop high-stakes tests for about a dozen states because we were eager to create criterion-referenced assessment approaches that would serve as an alternative to the norm-referenced assessment strategies so widely used in the late seventies and early eighties.

Because, when you are in that sort of business, your test-development group is competing for contracts against numerous other test-developers, you are usually reluctant to share any insights about test-development with your competitors. After all, it is precisely those kinds of insights that may win the next contract. The kinds of rules I’m about to discuss, therefore, would have been a closely guarded secret when we were actively bidding on new test-development projects. However, about a decade ago, our group downsized dramatically and exited from the test-development business. I am, therefore, a “recovering test developer.” If the rules I set forth here are of any use to *anyone*, please follow them with my blessing.

Combating Wish-List Curricula

My first rule is intended to counter the predilections of subject-matter curriculum specialists to want to teach children *everything*, that is, everything that exists in a specialist's content area. This tendency is commonly seen these days whenever we find subject-matter specialists generating a set of content standards (the knowledge and/or skills sought for students) in particular content areas such as reading, science, or mathematics. Such content-standards deliberations take place nationally (sponsored by major subject-specialty associations) as well as at the state or district level.

What typically happens is that a collection of teachers and curriculum experts carefully decide just what skills and knowledge children should master, at given age or grade levels, for the subject area involved. Characteristically, because these folks not only know their subject field inside out, but also *love* that subject, the group engages in rampant "wish-listing" by citing as quintessential content standards just about everything extant in their subject field. These subject-matter mavens really do want students to master all the stuff in their field.

In recognition that the resultant litany of knowledge and skills embodied in these wish-lists often is so voluminous as to be off-putting to ordinary teachers, some of the content-standards connoisseurs have re-packaged their desired sets of knowledge and skills into a smaller number of broader and, supposedly, coalesced constructs. But, if you look closer, you'll usually see that underneath each broad construct there will still be listed the numerous skills and bodies of knowledge that were there in the first place. I regard this somewhat specious, but often well-intentioned practice as *counterfeit coalescence*.

I am not knocking the virtues of these curriculum zealots. Naturally, they want children to master all the nifty things in their fields. And they'd obviously like to make sure students really do master it by having such mastery assessed. But it won't work! There are simply too many skills and too much knowledge represented in most of today's content standards that serve as the springboard for high-stakes accountability programs. I am certain that all of the content standards in most of these wish lists will rarely be taught, much less measured.

It is perfectly acceptable if curriculum folks want to wish-list up a storm by identifying a galaxy of skills and knowledge they'd like to see children master. Maybe a good many of those content standards will actually be mastered as children wind their way through school. But, *from an assessment and an instructional perspective, too many curricular targets turn out to be no targets at all*. And that leads to my first rule.

Rule 1. Require curricular personnel to prioritize the most important outcomes they want children to achieve, then develop tests to assess only the highest priority outcomes that can be both accurately assessed and instructionally accomplished.

Accurate assessment. We must oblige the individuals who determine content standards, that is, the curricular targets on which the test is to be based, to first identify the *most* important knowledge and/or skills to be taught, then have these content standards ranked from most to least important. To illustrate how the prioritizing process might work, suppose a state department of education has assembled a Content Standards Committee of 35 curriculum specialists and classroom teachers, to help state officials identify the important knowledge and skills the state's children should master in Subject Area X. Let's assume that the Content Standards Committee has, after much deliberation, identified 65 content standards.

Now, as a requisite step before any assessments are to be built, state officials should require the committee to split the 65 content standards into three groups of, say, *essential*, *highly desirable*, and *desirable*. Let's assume that, after extended deliberation, the committee designates 22 content standards as *essential*. At that point, as painful as it will most certainly be—using consensus-seeking procedures or some form of committee-member balloting—the committee must rank the *essential* content standards from “most important” to “least important.” Starting at the top of the importance list, the test constructors should then attempt to develop a test to measure as many of the highest ranked content standards as, in the assessment time actually available for administration of the test, can yield reasonably valid inferences about a child's mastery of those content standards being assessed.

In some instances, this process may lead to the assessment of only a small proportion of the content standards originally thought praiseworthy by curricular personnel. It is true, of course, that the *entire* original array of content standards are still eligible for *instruction*. Ideally, students will learn all that they really need to learn in a subject field, and this may clearly exceed what is assessed by any accountability system. Subject matter specialists must understand that educators are not precluded from teaching boys and girls *all* the wonderful things that ought to be taught simply because only the highest priority content standards are being assessed by an accountability-focused test. But curricular folks need to come face-to-face with assessment reality. It is simply *impossible* to assess properly all the good things that kids need to learn. So, rather than squandering a modest number of measurement arrows on targets far too broad to assess with any accuracy, the required prioritizing of the curricular outcomes will at least make sure that the very most important content standards are going to be satisfactorily assessed. Teachers can still, and *should* still, pursue the non-assessed outcomes for their students. Many of those other important content standards can, of course, be straightforwardly measured via teachers' in-class observations or classroom tests.

Remember, for measurement specialists to pretend that a huge array of curricular outcomes can actually be well assessed is hypocrisy. Moreover, if assessors claim they can (and will) measure a total set of multitudinous content standards via a high-stakes test, many teachers will be tempted to dilute their instructional activities to try reaching *all* of the assessed curricular outcomes. As a result, *none* of those curricular outcomes may be reached well. Teachers are still free to seek students'

mastery of a full set of curricular outcomes, not only those deemed sufficiently important to be assessed by a high-stakes test.

Instructional accomplishability. Only those content standards should be assessed that, as Rule 1 asserts, *can be instructionally accomplished*. An *instructionally accomplishable* content standard is one that can be taught to students with reasonable success, by typical run-of-the-classroom teachers. To teach such a content standard successfully, teachers need not be siblings of Socrates. But to determine which of the high priority content standards can be taught with reasonable success, the test-developer must think *instructionally*, not only psychometrically. Two of the remaining three rules deal specifically with the instructability of potentially assessable content standards.

Fortunately, in most instances it is possible to conceptualize many content standards (some of which are often stated too broadly or too vaguely) so that typical teachers can successfully promote their mastery. The probability that successful instruction will ensue, however, must be a concern of the test-developer from the very outset of the test-building process. A focus on the likelihood of successful instruction certainly should guide the test-developer with respect to decisions about the actual number of content standards to assess. Some content standards, after being subjected to careful scrutiny, actually turn out to be tapping a student's inherited intellectual aptitude rather than anything that can be promoted instructionally. Test developers must be constantly asking themselves, "Can this potentially assessable content standard actually be taught by most teachers, using non-Herculean instructional approaches, in the time that's available?"

It will be useful to remember an important instructional lesson drawn from the widespread advocacy of behavioral objectives in the sixties and seventies. By framing instructional objectives into hundreds of small-scope, although well explicated, instructional aims, the advocates of behavioral objectives unfortunately ended up overwhelmingly teachers with *too many* instructional targets. As a result, most teachers paid little attention to the endless lists of crisply stated but often trifling behavioral objectives. In a similar fashion, if a high-stakes test's assessment targets are going to function as an instructional focus for teachers, then it is imperative that only an intellectually manageable number of a field's content standards be designated for high-stakes assessment. In my experience, a half-dozen or so assessment (instructional) targets make much more sense than a dozen or two. As always, *less is more*; and *more* is definitely *not* magic.

Because high-stakes accountability tests have often been installed to let policymakers and the public know if our schools are being successful, it is desirable to assess content standards that are of such obvious worth (and, sometimes, difficulty) that—once they are mastered by students—onlookers will regard that accomplishment as clearly commendable.

So, after Rule 1 has been followed, the test-developer will have in hand a series of reasonably well identified bodies of knowledge and/or cognitive skill in a particular subject area. Let's assume the test-developer has been told that 90 minutes of student time, on one test-administration occasion in the spring, will be made available for assessment. The test-developer's task, then, is how to devise assessments for as many of the highly ranked content standards as possible. However, because any assessed content standard must pass muster with respect to both assessment validity *and* instructional accomplishability, it is still too early to tell which particular content standards will be measured. As the test is being developed, and as the test-development staff gets into the innards of a given content standard, instructional accomplishability questions will be more readily resolvable. So, for the time being, it is probably sensible to try to tackle a few more content standards than, given the test-administration time, are feasible for assessment.

I've treated Rule 1 in a fair amount of detail because if it isn't followed, the rest of the rules don't make any difference. The curricular springboard for any instructionally illuminating test must be defensible. Too many high-stakes tests are developed with no chance of success because they started with a lengthy laundry list of content standards. Content standards for instructionally illuminating tests should be few in number, but as significant as they can be. Let's turn, then, to Rule 2.

Conceptualizing Assessment Tasks from an Instructional Perspective

To display their ability to master any assessed content standard, students make responses to the tasks that constitute a test. Based on those responses, educators make an inference about the degree to which a student possesses mastery of the assessed content standard. My second rule focuses on the nature of the tasks themselves.

And as we consider Rule 2, it will become clear that to build an instructionally illuminating test, the developer of that test must constantly be employing both an assessment *and* an instructional point of view. The test developer must continually be raising questions such as, "Can this skill be taught to students?" or "How might teachers promote a student's mastery of this body of knowledge?" The successful developer of instructionally illuminating tests will typically be shifting back and forth between attempting to answer the two questions listed below:

1. *How can I build a test that will yield valid score-based inferences about the content standard to be measured?*
2. *How can I build a test that will help teachers promote the content standard being measured by the test?*

In short, an effective creator of instructionally illuminating tests will need to simultaneously "think teaching" and "think testing." Rule 2 will make this dual-focused thinking somewhat more operational.

Rule 2. Construct all assessment tasks so an appropriate response will typically require the student to employ (1) key enabling knowledge and/or subskills, (2) the evaluative criteria to be used in judging a response's quality, or (3) both of these.

As you can see, this second rule deals with the *tasks* to which students must respond. Those tasks can either elicit some sort of *selected response*, such as when students must answer multiple-choice items, or some sort of *constructed response*, such as when students are asked to write an essay or to make an oral report. Irrespective of whether selected-response items or constructed-response items are used, Rule 2 requires the test-developer, whenever feasible, to deliberately incorporate in an assessment task any precursive knowledge and subskills that will need to be taught if students are to master the content standard being assessed.

To comply with Rule 1, it is apparent the test-developer needs to understand thoroughly the nature of the cognitive demands that any task requires. For example, if a fairly straightforward vocabulary test is being built, then when students are given a vocabulary term, the student must cognitively be able to generate (or recognize) a suitable definition of the term. To illustrate this point, if multiple-choice items (tasks) are employed, then the student merely needs to choose the best definition from a set of optional definitions. If short-answer items (tasks) are employed, then the student needs to create suitable definitions from scratch. In either event, the cognitive demand of the task centers on the ability of the student *to recall, from memory, definitions for terms*. That, then, is the key cognitive demand to be kept constantly in mind by the test-developer.

Clearly, higher-level intellectual skills will require greater cognitive demands than would be necessary in the foregoing example of a knowledge-only vocabulary test. For instance, if a student were required to write a persuasive essay about a current controversial political issue, the student would need to rely on knowledge about the issue itself as well as a considerable ensemble of subskills that are required to create a well-written persuasive essay. The essay-writing task (and the important content standard it represents) is obviously much more complex than the tasks required by a knowledge-based vocabulary test.

The tasks created by a test-developer should typically call for students to use the complete array of necessary knowledge and subskills required by the task. For ease of communication, I shall refer to such knowledge and subskills *enabling K/S* because their mastery is deemed to be precursive to the mastery of the content standard assessed by a given test.

By building tasks of this sort, the test-developer will be trying to send a message to a teacher that, in order for students to perform well on such a task, the students will

need to master all important enabling *K/S*. And that means, of course, such enabling *K/S* must be *taught* to students.

The test-developer, therefore, will need to carry out a careful *task analysis* for every content standard to be assessed. In other words, for each cognitive skill or body of knowledge that is to be assessed, a careful analysis of all key enabling *K/S* must be completed. This means, of course, that the test-developer is smack in the middle of figuring out how a given content standard might be taught.

In passing, let me make clear that the kind of instructionally illuminating assessments I am talking about will definitely not require that a *particular* form of instruction be employed to promote students' learning. I believe, for most curricular outcomes, there are many roads to an instructional Mecca. Different teachers will be able to successfully employ quite different approaches to the promotion of students' mastery. Yet, irrespective of *how* a teacher provides instruction, the test-developer will invariably find that a careful analysis of an assessment *task* will reveal there are certain things that, in order for students to respond successfully to the task, they will need to learn. These necessary subskills and bodies of knowledge are precisely the things that, insofar as is sensible, should always be incorporated in most assessment tasks.

What the test developer is trying to do, unabashedly, is call teachers' attention to the importance of any key enabling *K/S* or any criteria that are to be used when judging the quality of students' responses. For instance, if a well-written persuasive essay should incorporate *appropriate mechanics* (spelling, punctuation, grammar), *suitable content*, *skillful organization*, as well as one or more effective *persuasion strategies*, then it is likely that a scoring rubric for judging students' responses will incorporate those four evaluative criteria. And, assuming such evaluative criteria can be taught to students (as, in this instance, they certainly can), then it is clear that a teacher who attends to the cognitive demands of the task will realize students must learn how to employ these four evaluative criteria as they create their persuasive essays.

Creating a Lucid Assessment Description

However, many teachers will need assistance, even with an instructionally illuminating test, in determining just what the critical instructional elements of the content standard are. That's where Rule 3 comes in.

Rule 3. Concurrent with the construction of a test, create a sufficiently clear description of the knowledge and/or skills represented by the test so that teachers will have an understanding of the cognitive demands required for students' successful performance.

Briefly, this third rule calls for test-developers to be simultaneously attentive not only to the test that they're building, but also to the creation of a companion *assessment description* that spells out for teachers the essence of what's being measured by the test's items or tasks. The rationale for Rule 3 is simply that teachers who adequately understand the nature of the content standard to be measured will, as a consequence of that enhanced understanding, be better able to promote students' mastery of the content standard. Simply put, teachers can teach better toward clear targets than toward murky ones.

This is the point at which a reasonable amount of *instructional acumen* is needed. A test-developer either possesses it, acquires it, or makes sure that there are members on the test-development team who have it. In the test-developer's consideration of what needs to go into a test (in order to accurately assess students' mastery of whatever content standard is to be measured), the test-developer must continuously be *thinking instructionally*. That is, at every step in the test-construction process, the persons creating the test must be continually asking themselves, "Can students really be taught to master this content standard and, if so, how?"

Remember that in Rule 2 assessment tasks were to be constructed so they typically required students to employ (1) key enabling *K/S*, (2) the evaluative criteria used to judge a student's response or (3) both. A test-developer who "thinks instructionally" will try to make sure that such enabling *K/S* or evaluative criteria can definitely be taught by teachers. So, for example, let's say that an instructionally astute test-developer is trying to build a test to measure a student's mastery of a chosen content standard. Let's call it *Intellectual Skill X*. After carefully task-analyzing the cognitive demands of that skill, the test-developer identifies the following set of enabling subskills and knowledge a student must achieve in order to master *Intellectual Skill X*:

- Body of Knowledge Q
- Body of Knowledge Z
- The subskill to apply Knowledge Q and Z so that previously unencountered phenomena can be classified as either Q or Z.
- The subskill to use a set of three stipulated evaluative criteria to determine if a given Q is better than a given Z.

The test-developer then must decide whether, using any conventional teaching approach, a typical teacher could instruct students so they master this set of enabling *K/S*. I say "conventional" teaching approaches because I want to re-emphasize that, in the main, instructionally illuminating tests should not require some sort of atypically splendid teachers. Nor, for that matter, should any particular instructional method be required in order to get students to master the assessed content standard. All the test-developer needs to do is make reasonably sure that at least a few accepted instructional approaches will do the teaching job satisfactorily.

The assessment description. In crafting the assessment description that will accompany the under-development test, those who must do so will need to employ a

fair amount of *descriptive artistry* in order to generate a succinct, yet communicative description of the content standard as it will be assessed. The description must isolate the key enabling *K/S* and any relevant evaluative criteria for the content standard. This must be done at a level of detail suitable for a teacher to make on-target instructional plans, but not so detailed as to become too lengthy. If the assessment description is so dense and interminable that most teachers will find it aversive, then its instructional impact will be negligible. The assessment description must not only be sufficiently comprehensive to capture the important features of the assessment, but must do so in a way that the resulting description will be *teacher-palatable*. Such an undertaking will take more than a little descriptive suave on the part of the individuals creating an assessment description.*

Given the considerable variety in the nature of content standards to be assessed and their related enabling *K/S*, it is impossible to provide an assessment-description template that will always work satisfactorily in diverse settings. But the two chief attributes of a well-written assessment description are that it (1) accurately isolates the enabling *K/S* and evaluation criteria, but (2) does so concisely. That's a nontrivial challenge for test developers.

Focusing on skills. If the high-stakes test attempts to assess students' mastery of an important cognitive skill (as will often be the case), the assessment description must spell out, in understandable language, the essence of the intellectual operation(s) required if the student is to respond satisfactorily to the test's tasks. Typically, this intellectual essence is characterized as the *cognitive demand* called for by the test's tasks. In addition, an assessment description for a high level cognitive skill, as indicated previously, must identify any important enabling knowledge and/or skills for the skill as well as any key evaluative criteria by which the adequacy of a student's response will be judged.

Focusing on knowledge. Let's deal for a moment with *knowledge*. In order for a teacher to promote students' mastery of a body of factual knowledge (whether the knowledge is an enabler for a content standard or is the content standard itself), the teacher needs to know what the entire set of knowledge is. Thus, even if the actual test only requires a student's familiarity with a *sample* of that knowledge, it is necessary to lay out for teachers the *complete* set of information, facts, rules, etc. that students should be taught to master. This can usually be done most efficiently by providing some sort of separate supplement to the assessment description.

So, for example, if an important cognitive skill in science required a student to have mastered, as enabling knowledge, vocabulary from a set of 90 scientific terms,

* A test developer can, of course, create *separately* very detailed specifications for the generation of test items. Such test-item specifications typically deal with such important variables as readability-levels, length of questions, and so on. But these specifications exist *in addition* to the assessment description that focuses on the instructionally relevant aspects of the content standard being assessed. Experience makes it abundantly clear that most teachers will rarely read a detailed set of test-item-specifications. If the assessment description is reasonably concise, and written in simple and straightforward expository prose, it will be read (and, hopefully, used) by teachers.

then all 90 terms should be identified in a *description-supplement*. If this is truly important enabling knowledge for the student to master, then it should be laid out so the teacher can plan instructional activities to promote students' mastery of the 90 vocabulary terms. If the content standard being assessed deals exclusively with a set of to-be-memorized information, then such information must be explicitly provided to teachers—so they can teach it to their students. Teachers should not have to guess about what knowledge they need to transmit to students.

Sometimes, of course, there are fundamental sorts of basic enabling *K/S* that students should already have been taught in earlier grades. For instance, if a high school student needs to be able to read reasonably well in order to pass a printed social studies test, it is not necessary to isolate “reading” as an enabling subskill precursive to the assessed content standard in social studies. The enabling *K/S* to be identified in the description should be what is *distinctively* necessary for a student's mastery of the content standard being assessed.

Clearly, if a social studies teacher discovers that some students are non-readers, then the teacher will need to tackle this important deficit instructionally. But such basic enabling subskills as reading, writing, and so on, need not be spelled out in an assessment description.

Nonexhaustive exemplar items. Illustrative items are almost always helpful in communicating to teachers how a particular content standard will really be assessed. So, it is useful to accompany an assessment description with a modest number of *illustrative but non-exhaustive* items that could be employed to measure students' mastery of the assessed content standard.*

It is especially important to stress that these sample items are non-exhaustive. We want teachers to be promoting students' *generalized* mastery of a designated set of knowledge or skills, not merely how to respond to a particular type of test item. Thus, a teacher who understands it is not the specific form of a test item but, rather, the *cognitive demands* of *any* appropriate test item, will recognize that successful instruction must so thoroughly promote students' mastery of the assessed content standard that such mastery could be displayed by students in a wide variety of ways. Accordingly, a teacher dare not engage in hyperfocused teaching toward one or two item-types. Instead, an effective teacher will nurture students' generalized mastery of the content standard being measured.

The set of illustrative but non-exhaustive sample items accompanying any assessment description should, so as to promote a teacher's push toward students' generalized mastery, be reasonably varied. Thus, if possible, the sample items should include not only selected-response items but also constructed-response items. Ideally, different sorts of response-modes should also be employed, that is, both written and

* I have previously recommended this approach to illustrating descriptions of measured knowledge or skills. See Popham, W. James, “The Instructional Consequences of Criterion-Referenced Clarity,” *Educational Measurement: Issues and Practice*, Winter 1994, 13(4), 15-18, 30.

oral. What we want teachers to do is promote students' generalized mastery of the assessed content standard. If it's clear that other types of test items (beyond those illustrations accompanying the assessment description) might be employed, then sensible teachers will instruct students so these students can satisfy the cognitive demands represented by a considerable variety of test items.

Every item in the set of exemplars accompanying the assessment description, of course, should be capable of contributing (along with other items) to a valid inference about a student's mastery of the content standard being assessed. Great care should be taken to make sure that every exemplar item could, in fact, make a contribution to an accurate performance-based interpretation regarding the student's mastery or nonmastery of the content standard being measured.

If a test is intended to measure four content standards, then that test should be accompanied by four assessment descriptions (all the more reason for those descriptions to be as concise as possible). If, on the other hand, only one skill is being assessed, as is often the case when we measure students' composition skills, then only a solitary assessment description is needed.

Making Sure

And now for one final rule. Actually, it's not really a rule governing how to *construct* instructionally illuminating tests. Instead, it's a rule regarding how to determine whether those tests have, in fact, been properly constructed. And, as you'll see, Rule 4 depends on the uses to which test results will be put, that is, it depends on the consequences of test usage.

Rule 4. For any high-stakes test, its items and assessment description(s) should be judgmentally reviewed at a level of effort and rigor commensurate with the intended consequences of the test's use.

This rule is intended to deal with important high-stakes tests such as those that might be used for grade-to-grade promotion or diploma denial. It's not that classroom teachers will never create their own instructionally illuminating tests. Some might. But it's a load of work for busy teachers to do. And I honestly can't think of too many sensible teachers who would wish to spend their discretionary hours churning out oodles of these instructionally illuminating tests. Moreover, if many teachers ever did, they'd rarely have enough energy left over to secure judgmental reviews of their efforts.

No, I'm thinking about high-stakes test that, because of the significance of their assessment mission, can command the meaningful financial resources required to construct such important tests—and to make certain that the tests have been well-constructed. The more important the test, the more exacting should be the judgmental review of that test's items and its assessment descriptions.

The judges. Who is to do all this judging? Well, because we are attempting to create tests that illuminate teachers' instructional decision-making, I suggest that most of the judges should be classroom teachers who have had experience in trying to teach the assessed content standard. Thus, for a significant high-stakes test we might assemble a review panel of, say, 15-25 individuals. To the panel I'd probably appoint a few instructional specialists (for example, instructional psychology professors) and curriculum experts (for example, district-office curriculum supervisors). But, in the main, I'd want my review panels to consist of firing-line teachers.

The judgments. What would we ask the members of such a review panel to do? Well, of course, they'll need to recognize that they will be functioning as *external quality monitors* and, as such, it will be their job to determine if the test is good enough to use in the high-stakes setting for which it is intended. The panelists' mission is to make a careful *post facto* analysis of the test's tasks and its assessment description(s) so that it can be determined whether the test is good enough to be used.

This essay is not the appropriate place to describe a detailed, step-by-step description of how a review panel should judge a high-stakes test, but I suggest there are several salient foci around what any decent review should be organized. Attention should be given to the test's items with respect to *curricular congruence*, *instructional sensitivity*, and *bias*. Attention should be given to an assessment description(s) with respect to *instructional illumination* and *palatability*. If multiple assessment descriptions are being provided for a test that measures multiple content standards, is the collection of assessment descriptions *in aggregate* too overwhelming for teachers?

One or more questions should be posed to review panelists with respect to each of these evaluative dimensions. Panelists can respond in a binary fashion, for example, Yes or No, or perhaps on some sort of multi-point scale. If you wish, an "Uncertain" response-option might be added to the Yes or No choices.

To illustrate the kinds of questions that could be given to panelists, I have listed below one example for each of the evaluative dimensions I cited above. These sorts of questions would typically need to be carefully crafted for the specific high-stakes test under review, hence the following illustrations are intended *only* to be illustrative.

For Each of a Test's Items (or Tasks)

- *Curricular Congruence: Would a student's response to this item, along with others, contribute to a valid determination of whether the student has mastered the specific content standard the test's developers indicate the item is supposed to be measuring?*
- *Instructional Sensitivity: If a teacher is, with reasonable effectiveness, attempting to promote students' mastery of the content standard that this item is supposed to measure, is it likely that most of the teacher's students will be able to answer the item correctly?*

- *Bias: Is this item free from content that might offend or unfairly penalize students because of personal characteristics such as race, gender, ethnicity, or socioeconomic status?*

For Each of a Test's Assessment Descriptions

- *Instructional Illumination: After having read this assessment description, would a teacher have a sufficiently clear idea about the cognitive demands required of students so that the teacher could plan effective instruction to have students master the content standard being measured?*
- *Palatability: Is this assessment description sufficiently understandable and concise so that a teacher would be willing to read it carefully?*

If review panels have been judiciously selected, and panelists have been properly oriented to their judgmental tasks, the resultant data regarding the test's items and its assessment description(s) should prove adequate to determine whether the test is, indeed, an instructionally illuminating test suitable for use in the specific high-stakes context for which it was created.

Two In-Passing Observations

With the four rules now having been treated, I'd like to register two points that are potentially pertinent if one were to follow the suggestions I've set forth for creating instructionally illuminating tests. First, it should be recognized that few, if any, teachers have been given direct training in how to read a succinct description of an assessed content standard, then plan instruction so that such instruction has a high likelihood of promoting students' mastery of what's being assessed. Accordingly, it is naïve to think that the mere generation of instructionally illuminating tests will automatically lead to improved instruction. It won't. Most teachers will not know how to make use of such tests and the clarity they provide.

That leads me to my second point. Accompanying each instructionally illuminating high-stakes test should be a menu of instructional suggestions from which teachers can choose—if they wish to do so. In other words, because most teachers will not know how to take advantage of instructionally illuminating tests, such teachers must be helped. That help could come in the form of ideas about how to instruct students so they master the assessed content standard. Teachers can consider these diverse suggestions, adopt some, or adopt none. It must be the teacher's choice. Yet, a meaningful attempt will have been made to help the teacher do the kind of improved instructional job that will be possible because of the availability of instructionally illuminating tests.

An Illustrative Assessment Description

Let's turn, now to an example of an assessment description for an instructionally illuminating test. I've been claiming that such assessment descriptions will provide teachers with a sufficiently clear idea about the content standard being measured so that a teacher could provide instruction targeted accurately at the knowledge and/or skills being described. These assessment descriptions, built to reflect what's being measured, but spelled out in teacher-palatable language focusing on the test's cognitive demands, are the chief factor that can make assessments actually contribute to more successful instructional decision-making.

Incidentally, if Rule 1 were followed, the resultant assessments would deal with only a small number of genuinely significant outcomes. That being so, the assessments will rarely assess knowledge alone. Genuinely significant outcomes will almost certainly deal with high-level cognitive skills—skills that may, indeed, call for students to first master substantial bodies of enabling knowledge. But, to be candid, a truly significant *knowledge-only* assessment is difficult for me to imagine.

What follows now will be my attempt to carve out an illustrative assessment description along the lines of those I've been advocating. I don't pretend that this description constitutes the "final word," and that it cannot be improved by teachers who have more experience than I in teaching the content involved. But I hope this assessment description will give you an idea of the sort of descriptive scheme I have in mind.

A Sample Assessment Description: Using History's Lessons

Introduction. This content standard was deliberately chosen because it represents a particularly challenging cognitive skill in social studies. If it can be demonstrated that students have, over time, increased their mastery of this high-powered skill, both parents and policymakers should readily concede that some first-rate learning has taken place. And first-rate learning usually occurs because of first-rate teaching.

This assessment description, the illustrative tasks that follow it, and the set of instructional suggestions provided thereafter are intended for use by eleventh- and twelfth-graders who are taking a U.S. history course. If the skill were to be sought at earlier grade levels in which U.S. history is taught, the illustrative items would be simplified with respect to both language as well as the cognitive complexity of the tasks. For any attempt to measure this skill, it will be necessary for appropriate state or district curriculum authorities to identify (and promulgate to educators) the eligible historical events for students to consider. As an example, the following historical events were identified as suitable by one large school district for eleventh- and twelfth-grade U.S. history courses:

<p style="text-align: center;">AMERICAN HISTORY TIMELINE MAJOR EVENTS For Eleventh and Twelfth Grade U.S. History</p>	
Constitution Territorial Expansion Civil War Reconstruction Industrial Revolution Imperialism World War	Depression New Deal World War II Cold War Civil Rights Viet Nam Communication Revolution

The Assessment Description. Given a prose account of a real or fictitious current problem, as well as a proposed solution to that problem, students will be able to respond appropriately to any on or any combination of the following subtasks:

- Subtask 1. Identify at least one significant historical events (such as the industrial revolution) that is, at least in part, germane to the problem and its proposed solution
- Subtask 2. Justify the relevance of the identified historical event(s) to the problem and its proposed solution.
- Subtask 3. Make a defensible history-based prediction regarding the proposed solution's likely consequences.
- Subtask 4. Support the prediction on the basis of parallels between the identified historical event(s) and the proposed problem-solution.

Students will be presented, orally or in writing, with a real or fictitious current day problem-situation along with a proposed solution to that problem. (See illustrative items below.) After students have had an opportunity to consider the problem and the proposed solution, they will be asked to supply written or oral responses to one or more of the four subtasks listed above.

Although students may be asked to supply responses to individual subtasks so that their mastery of those types of subtasks can be determined, all students will ultimately be required to respond to a comprehensive task such as the one illustrated in Item No. 1 below. When students respond to a comprehensive task, of course, they will be given a *different* problem and proposed solution than the ones used for individual subtasks.

Illustrative, Nonexhaustive Tasks (Items)

A Sample Problem Situation and Proposed Solution

Directions: Read the fictitious problem described in the box as well as the proposed solution to that problem, then respond to the numbered tasks presented below the box.

WAR OR PEACE

Nation A is a large, industrialized country whose population is almost 100,000,000. Nation A has ample resources, and is democratically governed. Nation A also owns two groups of islands that, although distant, are rich in iron ore and petroleum.

Nation B is a country with far fewer natural resources and a population of only 40,000,000. Nation B is about one-third as large as Nation A. Although much less industrialized than Nation A, Nation B is as technologically advanced as Nation A. Nation B is governed by a three-member council of generals.

Recently, without any advance warning, Nation A was ruthlessly attacked by Nation B. As a consequence of this attack, more than half of Nation A's military equipment was destroyed. After its highly successful surprise attack, Nation B's rulers have proposed a "peace agreement" calling for Nation A to turn over its two groups of islands to Nation B. If the islands are not conceded by Nation A, Nation B's rulers have threatened all-out war.

Nation A's elected leaders are fearful of the consequences of the threatened war because their military equipment is now much weaker than that of Nation B. Nation A's leaders are faced with a choice between (1) peace obtained by giving up the islands or (2) war with a militarily stronger nation.

Nation A's leaders decide to declare that a state of war exists with Nation B. They believe that even though Nation B is now stronger, in the long term Nation A will prevail because of its greater industrial capability and richer natural resources.

- Task 1. In an essay, drawing on your knowledge of American history, select one or more important historical events that are especially relevant to the fictitious situation described above. Then justify the relevance of your selection(s). Next, make a reasonable history-based prediction about the likely consequences of the decision by Nation A's leaders to go to war. Finally, defend your prediction on the basis of the historical event(s) you identified.

Note: The evaluation of a student’s response to this four-step, comprehensive task will be based on the quality with which each of the following have been carried out: [1] event(s) selection, [2] event(s) justification, [3] history-based prediction, and [4] defense of prediction.

(Individual subtasks would, apart from this illustration, require a different problem and proposed solution.)

Task 2. In the spaces below, name one or more important events in American history that are particularly relevant to the problem and proposed solution described in the box.

Task 3. In an oral presentation of one-to-two minutes, identify at least one important historical event that is especially pertinent to the situation described in the box, then justify why you believe this to be so.

Task 4. In the situation given in the above box, two fictitious nations are described. From the four choices below, choose the two nations and the armed conflict *most* comparable to those described in the box.

- A. Conflict: World War I
Nations: U.S. and Italy
- B. Conflict: Korean “Police Action”
Nations: U.S. and North Korea
- C. Conflict: Spanish American War
Nations: U.S. and Spain
- *D. Conflict: World War II
Nations: U.S. and Japan

Instructional Suggestions

One immediate benefit of teachers’ understanding a to-be-taught skill is that they can then carry out more effective instructional task analyses. Simply put, you need to decide if there are any enabling knowledge or subskills that must be mastered by your students before they attain mastery of the *Using History’s Lessons* skill. If you identify such knowledge or subskills, then you’ll need to promote them during instruction.

During the past few decades, American educators have identified an impressive collection of effective, research-based instructional procedures. You should employ as

many of these proven instructional procedures as you believe appropriate. For instance, after describing the nature of the skill to your students and getting them to understand its importance (to them), it is usually helpful for you or selected students to model the skill. And after students understand what is being sought of them, it is often helpful to give them opportunities to discriminate between weak and strong illustrative responses. As soon as your students are ready, the provision of guided practice followed by independent practice will usually move students toward skill mastery. Such practice activities, guided or independent, will be more effective if a rubric (scoring guide) is consistently used by you and your students. Generation of a rubric suitable for the *Using History's Lessons* skill is discussed below.

In addition to conventional teacher-directed activities, you will also find such instructional variations as cooperative learning and peer critiquing can prove useful in promoting this skill. What you need to recognize is that the vast majority of what you know about instructional procedures will be highly relevant in promoting your students' mastery of this really challenging, but eminently teachable social studies skill.

The following specific suggestions, related to your promotion of the *Using History's Lessons* skill, are not offered in order of importance. Please review them to see if any suggestions appear to be appropriate for your own instruction. Clearly, these are suggestions, not mandates. You are free to use any or none of the specific suggestions provided here.

1. Although the most comprehensive display of this skill calls for students to carry out four separate subtasks in the proper order, it may be instructionally useful to break these subtasks into separate steps so that, by pursuing them one at a time, students can gain mastery of those four subskills. For example, you could describe a current-day problem to your students such as the possibility of emergent monopolies in the computer and communications fields, then ask how today's citizens might solve that problem. After a few problem-solutions had been offered, the class could select one, then try to follow the four-step sequence of subtasks in order to use history's lessons to deal with current problems. This activity could be carried out in subgroups or as a whole-class analysis so that you can supply necessary en route guidance.
2. One powerful way to get students to understand what's involved in this skill is to develop a *rubric* for evaluating students' responses to each of the individual subtasks as well as the comprehensive task in which the four subtasks are carried out in a specific sequence. The caliber of a student's response to each subtask (and to the synthesized sum of those subtasks) is dependent on the quality of each response. Thus, the rubric would need to isolate what it is that makes a response to each subtask either strong or weak. For instance, if students are asked to justify their selection of an historical event, just what must a student's justification contain in order to be of high or low quality?

3. Once a rubric has been developed, either by you or collaboratively with your students, that rubric's set of evaluative criteria can prove immensely helpful to you and your students as they attempt to generate suitable responses to the individual subtasks or to the comprehensive set of subtasks. The more familiar students are with the rubric's evaluative criteria, and the more they are called on to use them, the stronger their responses are apt to be.
4. Make sure that the students (and you) understand that this skill requires a distinctive way of thinking about history. To display mastery of the skill, your students must be able to (a) analyze a current problem-situation and proposed solution so that they can (b) *select* and *justify* relevant historical events, then (c) *predict* the proposed solution's consequences and *defend* that prediction. This three-step sequence, namely, (1) analyze, (2) select/justify, and (3) predict/defend can offer a framework for the instruction you provide for this skill. Alternately, you might wish to emphasize the four subtask components of the skill, that is, (1) identify an event, 2) justify the event's relevance, (3) make a history-based prediction and (4) defend that prediction.
5. Because this skill is intended to help students apply history's lessons to their own post-school lives, you'll need to get students to start viewing historical events not merely from the perspective of "what happened," but from the perspective of "what lessons, if any, can we draw from this historical event." Not all historical events, of course, yield lessons for the future. Much of history is so dependent on the particular actions of unique individuals, or is so complex, that it is difficult to identify any cause-effect generalizations that would be applicable to the future. But predictive cause-effect generalizations based on one or more historical events are often present if we search for them. Encourage your students to be on the lookout for such history-based lessons routinely. Help them see the difference between historical events that yield lessons for the future and historical events that don't.
6. Frequently have your students consider a current problem-situation and see if any events in American history are sufficiently relevant to the situation so that reasonable predictions about current-day consequences might be made. It is especially valuable if the current problems are seen by students as relevant to their own lives. You are encouraging students to routinely regard today's problems in light of relevant historical occurrences.
7. Teach students any special terminology associated with this skill, for example, "predict," "justify," "defend," etc.
8. Expand students' knowledge of history so they have a meaningful basis for drawing history-based lessons. Ideally, as you cover the necessary major events of U.S. history, you should stress the importance of drawing cause-effect lessons that can guide today's citizens. You might wish to analyze separately each of the

14 major events in American history to discern if there are lessons to be drawn from an event such as the Reconstruction Period.

9. Familiarize students with time-lines and how to use them. Make sure your students know that, for purposes of measuring this skill, the illustrative 14-event timeline provided earlier contains all “eligible” major events in U.S. history. Each major event, of course, embraces a number of specific historical occurrences from which historical lessons might be drawn.
10. Organize small group of students whose mission is to look back over the eligible events drawn from American history, then attempt to isolate one or more events from which important lessons are clearly derivative. Ask the groups to report to the entire class regarding the lessons they have isolated. The rest of the class, then, can judge the quality and the current-day relevance of the groups’ inferred history-based lessons.

References

- Linn, Robert L. “Assessments and Accountability,” *Educational Researcher*, March 2000, 29(2), 4-16.
- Popham, W. James. “Where Large-Scale Assessment is Heading and Why It Shouldn’t,” *Educational Measurement: Issues and Practice*, Fall 1999, 18(3), 13-17.
- Popham, W. James. “The Instructional Consequences of Criterion-Referenced Clarity,” *Educational Measurement: Issues and Practice*, Winter 1994, 13(4), 15-18, 30.