

LESSONS FOR LARGE-SCALE ASSESSORS*

W. James Popham

University of California, Los Angeles

Large-scale educational assessments can either enhance or erode the quality of instruction given to students. Whether large-scale tests help or harm children's education, however, depends dominantly on the nature of the tests themselves. The following analysis, therefore, will focus on how to create large-scale educational assessments so they are likely to improve instructional quality, not diminish it.

More specifically, I will describe a set of lessons learned about large-scale assessment in the United States. These are high-cost lessons, for in order to have arrived at these test-related insights, American educators have made a number of profound assessment and instructional mistakes—mistakes that have done serious damage to the quality of education received by thousands of US children. More troubling, however, the assessment lessons I shall identify in this paper are far from universally accepted by American educational policymakers. Even worse, the lessons are far from universally accepted by American educators themselves.

An astonishing amount of assessment illiteracy exists in my country among members of the educational community. Although there is ample evidence that large-

* Presented at a conference, "Future Trends for the Assessment of Students' Achievement," Abu-Dhabi, United Arab Emirates, November 11-12, 2001.

scale “assessment as usual” is harming many of our nation’s children, the bulk of America’s educational policymakers, whether at the national or state level, persist in endorsing *traditional* large-scale educational assessment. It is my belief that, for the most part, these policymakers are well-intentioned, but appallingly ignorant of the negative impact their assessment-related policies are currently having on children.

Fortunately, there is some activity underway in the US intended to overcome such assessment illiteracy among educational policymakers. To illustrate, only a few months ago, five national associations of US teachers and administrators convened an independent group of assessment specialists to offer recommendations about how to develop large-scale accountability tests that could help American educators make better instructional decisions. That group, the Commission on Instructionally Supportive Assessment, was convened in July 2001 by the American Association of School Administrators, the National Association of Elementary School Principals, the National Association of Secondary School Principals, the National Middle School Association, and the National Education Association. Based on the Commission’s deliberations, a report* was distributed nationally in October listing the requirements that a large-scale test must satisfy if that test were to (1) supply accurate evaluative evidence of school quality and (2) support teachers’ instructional decision-making. It is the Commission’s hope that, as its requirements for instructionally supportive large-scale assessment are disseminated among US educators and, particularly, among US educational

* The Commission on Instructionally Supportive Assessment, *Building Tests To Support Instruction and Accountability: A Guide for Policymakers*. Washington, DC: Author, 2001.

policymakers, there may be a meaningful shift in perceptions regarding the kinds of large-scale tests that should be administered in American schools.

Potential Applicability

The eight lessons I shall describe in the following analysis will be in complete accord with the recent recommendations of the Commission on Instructionally Supportive Assessment. Nonetheless, these lessons are offered with a clear recognition that the United Arab Emirates and the United States of America are different nations whose educational needs and whose histories of educational testing, especially large-scale educational testing, are quite different.

It would be presumptuous of me to suggest, therefore, that the assessment-related lessons to be identified in this paper are appropriate for the UAE. This is not my intention. Rather, I intend to lay out a series of lessons for those large-scale assessors in the US who wish to make assessment an instrument for instructional improvement in America's schools. Educational policymakers and educational assessors in the UAE will themselves need to determine which, if any, of these lessons have any relevance for large-scale assessment in the UAE.

Given my lack of conversance with the particulars of instructional practices and educational assessment in the UAE, I am personally unable to determine the UAE-relevance of this assessment-related advice. However, I look forward to subsequent

discussions with colleagues in the UAE regarding the applicability of these US-derived assessment lessons.

A Capsule History of Large-Scale Assessment in the US

So that readers of this analysis will understand why those in my nation have made major mistakes regarding the uses of large-scale educational assessment, I need to provide a brief history of large-scale educational testing in the US. The need for brevity will force me to overlook events of lesser import, but with that warning having been given, I hope the following synopsis will set my subsequent lessons for large-scale assessors in a more meaningful context.

The Army-Alpha. The large-scale test that has been most influential in shaping America's large-scale educational testing was developed during World War I to help the US Army identify, from the thousands of recruits inducted into the military, the most suitable candidates for officer training programs. At the behest of the US Army, a test was developed to accomplish this sorting-out function. It was known as the *Army Alpha* and it was administered to approximately 1,750,000 men during World War I. The success of the *Alpha* made it a model for all subsequent large-scale tests in the US. That *Alpha*-influence still is substantial in all US large-scale assessments.

The *Army Alpha* was a *predictor* test intended to identify those recruits who would be most likely to succeed in the Army's officer training programs. The *Alpha* has

often been described as an *aptitude* test but it was, in truth, a *group-administered intelligence test* that attempted to sort out examinees on the basis of their intellectual abilities to deal successfully with diverse sorts of problems such as those involving verbal analogies and quantitative puzzles of one sort or another.

Examinees' *Alpha* scores were compared with the scores of a representative group of previous test-takers known as the *norm group*. Using this comparative scheme, when "George scored at the 92nd percentile," George was enrolled in an officer training program. On the other hand, when "Bill scored at the 43rd percentile," then Bill was first sent to basic training, then whisked off to the trenches. In order for the *Alpha* to provide sufficiently fine-grained *comparative* interpretations (that is, *norm-referenced* interpretations), substantial *score-spread* had to be present in test-takers' scores. Insufficient score-spread precluded the exact comparisons that were needed. And, to accomplish the *Alpha's* unabashed *sorting* mission, the necessity for score-spread was altogether sensible. The *Alpha*, as noted earlier, was a universally acclaimed success. That success, as you'll see, has unfortunately spawned a series of subsequent mistakes in America's large-scale assessment activities.

Here's how those assessment mistakes took place. Shortly after World War I's conclusion, a number of educational tests were developed in the US. For example, the widely used *Standard Achievement Tests*, now available in their ninth edition, were first published in 1923. Such tests were described as *achievement* tests because, unlike the *Alpha's* predictive mission, the emerging large-scale tests (soon to be known as

standardized achievement tests) were supposed to measure students' *knowledge and skills*. It was these new tests' attempts to assess students' knowledge and skills (rather than assessing their intelligence) that allowed American educators to distinguish between so-called *aptitude* tests and so-called *achievement* tests.

Later, during the 1960s, when America's federal policymakers began, for the first time, to distribute substantial funds to US schools, many elected officials (notably Robert Kennedy, then a US senator from New York) demanded *test-based evidence* that these federal tax dollars were being well spent. To satisfy a set of new, federally imposed requirements for evaluation of federally funded educational programs, most American educators selected "off-the-shelf" nationally standardized achievement tests such as the *Iowa Tests of Basic Skills*. In American educators' assessment naiveté, it was assumed that such *achievement* tests would measure what students had learned in school. That assumption, however, turned out to be mistaken.

You see, in the years following World War I, the developers of the large-scale assessment instruments known as "achievement" tests had subscribed totally to the *Alpha's* sorting-out mission. Thus, developers of the nation's "achievement" tests were preoccupied with producing sufficient score-spread so that precise comparative interpretations among students would be possible. This unrelenting quest for score-spread, unfortunately, has rendered today's US "achievement" tests largely unsuitable for measuring what students have *achieved* in school. Let's see why.

Instructionally insensitive tests. In order to construct standardized achievement tests that do a good job of creating score spread, US developers of those tests usually do three things that are *Army-Alpha appropriate*, but reduce the suitability of the tests for evaluating teachers' instructional effectiveness.

For one thing, because *mid-difficulty items* optimize a test's score-spread, the creators of traditional standardized achievement tests prefer to use items answered correctly by between 40 percent and 60 percent of the students who take those tests. Items answered correctly by, say, 90 percent of the students fail to contribute adequately to score-spread, hence are usually jettisoned from a standardized test as soon as the test has been revised. Unfortunately, items on which students perform well are often based directly on content which, because of that content's importance, will be stressed by teachers. The more important the content, the more that teachers will emphasize it instructionally. Yet, the better the job that teachers do in teaching high-import content, the less likely it is that items assessing this teacher-stressed content will remain in an oft-revised standardized achievement test. Too many students will have answered those items correctly, so the items are not contributing sufficiently to the production of score-spread. Thus, there is a systematic tendency to exclude from US large-scale assessments those items measuring the most important things that teachers teach.

Second, there are a good many items in America's standardized achievement tests that are more likely to be answered correctly by children whose family is fairly

affluent, that is, possesses a higher, rather than lower, socioeconomic status (SES). *SES-linked items* are included in these achievement tests because such items do an excellent job in producing score-spread. SES is a nicely distributed variable and is not readily altered. But, of course, SES-linked items measure what students bring to school, not what they learn there.

Finally, many items will be found in America's traditionally constructed standardized achievement tests that are clearly dependent on children's *inherited* academic aptitudes, that is, children's innate verbal, quantitative, and spatial potentials. These *inheritance-linked items* also do a wonderful job in producing score-spread because children's innate academic potentials are well distributed. But, as was true with SES-linked items, inheritance-linked items measure what students bring to school, not what they learn there.

Taken together, the heavy reliance on mid-difficulty items, SES-linked items, and inheritance-linked items renders traditionally constructed US standardized achievement tests remarkably insensitive to detecting the impact of even first-rate instruction.

Educational accountability. During the last two decades of the twentieth century, the effectiveness of American educators came under increasing scrutiny from many quarters. As a consequence of genuine doubts regarding the quality of American schools, *educational accountability programs* were installed in almost every US state. The cornerstone of these accountability programs was always a statewide test that

attempted to inform the state's citizens about how much the state's children had learned. Typically, these statewide tests were either off-the-shelf nationally standardized achievement tests or, in some instances, were state-customized versions of those national tests. (It is unfortunate that, because the state-customized accountability tests were usually built by the same firms that created America's nationally standardized achievement tests, the state-customized tests often did not differ significantly from the national tests.)

America's educators, as a consequence of state-level accountability programs, often found their schools ranked in local newspapers based exclusively on students' standardized achievement test scores. Students who scored low on these tests were often denied diplomas or grade-level advancement. In some states, financial incentives were given to teachers from high-scoring schools. In contrast, low-scoring schools were sometimes taken over by a state or, in a few instances, those low-performing were closed down altogether. Standardized achievement tests, either national or state-specific, truly became *high-stakes* tests in every sense of that term. Students' scores on standardized achievement tests, therefore, became the paramount indicator of American educators' success. And yet, as noted above, those tests are remarkably insensitive to detecting instructional impact. Those tests, in short, should not be used to evaluate educators. They are the *wrong* tests for what is an altogether *right* evaluative task.

Negative consequences. Because of the prevalent use in my nation of the wrong kinds of accountability tests, deplorable things are currently happening in America's classrooms. For one thing, because of so much pressure on teachers to raise their students' scores on large-scale tests, there have been widespread examples of *curricular reductionism*, that is, teachers' instructional avoidance of *any* content not assessed by a statewide accountability test.

A second negative consequence of using the wrong kinds of accountability tests is that many teachers are forcing their students to take part in *drudgery drilling*, that is, seemingly interminable practice sessions devoted exclusively to students' practicing items similar to (or, in some instances, identical to) the actual items used in a state-level accountability test. Such practice sessions soon stamp out any joy that students might otherwise derive from the act of learning.

Finally, we now hear weekly, if not daily, reports of American educators who have engaged in *unethical test-preparation* or *unethical test-administration* related to statewide accountability tests. Surely, such improper conduct sends an unacceptable message to those students who become aware of educators' unethical behaviors. Although such conduct by test-pressured teachers and administrators is understandable, it is nonetheless unacceptable.

Summing up, then, this brief walk-through of America's large-scale assessment history indicates that, because an improper *Army-Alpha* assessment strategy was

adopted to satisfy a nation's evaluative needs regarding its schools, a series of adverse consequences has led to a deterioration of instructional quality in many parts of the US. The task before us in America is simple. We must correct this situation by employing appropriate large-scale assessments. Here, then, are eight lessons that I believe must be followed in our country. Some may have applicability in the UAE.

Large-Scale Assessment Lessons

In the remainder of this analysis, I will present, in turn, eight lessons for large-scale assessors. Each bold-faced lesson will be followed by a brief rationale, explanation, or clarification. If large-scale US educational assessment programs were created in accord with these eight lessons, the resultant assessment programs would not only provide credible evaluative evidence for purposes of educational accountability, but would also help stimulate improved instructional quality.

I wish to remind readers that these eight lessons are derived from US-based experiences in educational assessment. The applicability of the lessons to other settings must be determined by individuals who are conversant with educational contexts in those other settings. UAE educational policymakers and assessment specialists, therefore, must clearly be the judges of whether any of these eight lessons are germane to their nation's large-scale assessment needs.

Lesson 1: Content standards must be prioritized.

American educators use the expression *content standards* to describe the knowledge and skills that we want children to master. In earlier times, US educators often described such curricular aims as “educational goals” or “instructional objectives.” Now, however, in every one of our 50 states, a *content standard* is the preferred label that is used to signify the skills and knowledge educators are supposed to promote for that state’s students. (A *performance standard*, in contrast, refers to the desired level of proficiency students must display to demonstrate mastery of a given content standard. My eight lessons are focused on content standards, not performance standards.) In many US states, we have seen a clear commitment to “standards-based educational reform” as a strategy for improving educational quality. Such reform efforts, as might be guessed, revolve around a set of state-approved content standards.

Typically, a state’s content standards are determined by a group of curricular specialists. (We have, in the US, no nationally sanctioned content standards, only state-by-state content standards—many of which are quite similar from one state to another.) Well, when a state’s curriculum specialists convene to decide what skills and what knowledge children should be taught at different grade levels or, possibly, at different grade ranges, those curriculum specialists typically want students to learn just about *everything* that can be learned. Indeed, the curricular aspirations of these determiners of content standards are often grandiose beyond belief. As a consequence, many of our states now possess officially approved lists of content

standards that are staggeringly long—far too numerous either to be taught or to be tested in the time available. Elsewhere*, I have referred to these near-endless litanies of curricular aims as “wish-list content standards” because they appear to represent all the good things that educators *wish* their students would be able to learn.

In some states, for example, there are more than 500 content standards approved for a single subject field such as mathematics. It is ludicrous to contend that such interminable sets of content standards can be *sensibly taught* or *accurately tested* in the time we have available for either of those purposes.

Therefore, the first thing that must be done to make a large-scale educational assessment program *instructionally* viable is to *prioritize* the curricular aims that are to be assessed. When doing so, it is insufficient to merely *rate* the importance of a set of content standards. Rather, it is necessary to *rank* the educational significance of those standards, one-by-one. This import-ranking must be carried out at least for the most importantly rated standards. This import-ranking must be done separately for each subject area, for instance, in language arts, mathematics, science, and so on.

What the designated curriculum personnel must turn over to the assessment personnel is a set of prioritized content standards. What the assessment personnel must do is determine how many of the highest ranked content standards can be

*Popham, W. James, “Assessing Mastery of Wish-List Content Standards,” *NASSP Bulletin*, December 2000, 84(620).

appropriately assessed in the time that has been made available for large-scale assessment. As you'll see in a moment (Lesson 3), in order for a large-scale assessment to be instructionally supportive, it must provide standard-by-standard results. Clearly, the provision of standard-by-standard results will necessitate the assessment of fewer content standards. These fewer, yet *most important* content standards will become the foci of an instructionally supportive large-scale assessment program.

It is blatantly misleading to pretend that a state can accurately assess its students' mastery, during a one-hour or two-hour testing session, of literally hundreds of content standards. In truth, such claims by measurement specialists represent a clear form of *assessment hypocrisy*. However, a comparable form of *instructional hypocrisy* is also encountered when a curricular staff contends that many hundreds of content standards can actually be taught during a typical school year. It is time for us to be truthful with respect to both assessment and instruction.

This first lesson's chief function, then, is to call for a prioritization of content standards so that honest claims can be made about what should be taught and what should be tested. By an unequivocal isolation of our most important content standards, we set the stage for curricular and assessment integrity.

Lesson 2: High-priority content standards must be clearly described so that the knowledge and skills sought from students are evident.

Historically, educators' curricular aims have often been stated at a level of descriptive rigor insufficient for teachers' day-and-day instructional planning. If Lesson 1 has been properly followed, and the mission of a large-scale assessment program is to measure students' attainment of a modest number of *high-import* content standards, then teachers must understand, unequivocally, what each high-import content standard entails.

One way of promoting teachers' understanding of the nature of the skills and/or knowledge represented in a particular content standard is to create an *assessment description* to accompany each high-import content standard. Each such assessment description must isolate the essential *cognitive demands* placed on students by any test items intended to measure students' mastery of a particular content standard.

Because the most significant dividend of a properly constructed assessment description will be the clarification it supplies to teachers, all assessment descriptions must not only be brief (one-to-three paragraphs), but must be written in teacher-palatable language. To derive optimal *instructional* benefits from these assessment descriptions, teachers must not regard those descriptions as dense, lengthy, or otherwise off-putting.

Assessment descriptions are typically constructed collaboratively by those who possess assessment expertise and those who possess instructional expertise. Each

assessment description should identify any key enabling subskills or bodies of knowledge that students must acquire in order to master the content standard for which the assessment description has been supplied. Teachers need to recognize that any important enabling subskills and knowledge must either already be possessed by students or, if not, must be taught.

Finally, each assessment description should be accompanied by several *illustrative but nonexhaustive* test items to help teachers understand how a student's mastery of a content standard's cognitive demands might be assessed in different ways. Ideally, therefore, these illustrative items should reflect a variety of tasks that might be used to assess students' content-standard mastery. Because the illustrative items are nonexhaustive, teachers will recognize that their challenge is to promote students' *generalizable* mastery of a content standard, that is, a mastery so powerful that it could be demonstrated in myriad ways.

After a teacher has considered a high-priority content standard, and its accompanying assessment description, the teacher should have a clear idea about the knowledge and/or skills represented in a content standard. The teacher can then design instruction that is aimed at the content standard itself, not at a particular set of test items intended to measure students' mastery of that standard.

Lesson 3: Assessment of high-priority content standards must be reported standard-by-standard for each district, school, and student.

In the US at present, a state's officially approved content standards are supposedly measured by some kind of achievement test, typically a test constructed in the traditional *Army Alpha* fashion. These tests almost always fail to yield *per-standard evidence of students' mastery*. Rather, the tests typically provide a single, overall score that purports to represent students' mastery of the entire slate of content standards. From an instructional perspective, this kind of reporting-practice is patently dysfunctional.

How are teachers to know which segments of their instruction are working and which segments should be refurbished if those teachers are not given *per-standard* feedback about their students' performance? Standard-by-standard reporting must be present if any version of standards-based reform has a chance of benefiting children. If teachers find out that their instructional programs have been successful in promoting students' mastery of particular content standards, then the instruction related to those standards need not be altered. If, however, teachers discover that students have failed to master the content standards that teachers have addressed instructionally, then improvements in that instruction must be made.

And per-standard reporting must be provided at the *student* level. How else will teachers discover *which* students need further instruction? How else will parents, or the students themselves, find out how a particular child is truly faring with respect to the mastery of important content standards?

It is apparent that per-standard reporting, especially for each student, will require more items per content standard than are typically encountered in large-scale assessments. And this means, in turn, that fewer content standards can be accurately assessed in an instructionally supportive large-scale test. Clearly, this is the reason that Lesson 1 and Lesson 3 are closely linked. It is *because* per-standard reporting requires more items for each content standard that content standards must be prioritized. It is preferable to accurately assess a modest number of high-import content standards (in a manner that benefits instruction) than it is to inaccurately assess a large number of content standards (in a manner that doesn't benefit instruction).

Standards-based assessment without standard-by-standard reporting is like playing soccer without being able to find out whether goals were scored. Without standard-by-standard reporting, standards-based testing typically turns out to be a fine-sounding but ineffectual assessment charade.

Lesson 4: Educators must be provided with optional classroom assessments that can measure students' progress in mastering content standards unassessed by large-scale tests.

Although Lesson 1 and Lesson 3, if followed, can lead to instructionally beneficial assessment of the most important content standards, the remaining (unassessed) content standards are also important. We want our students to master as many of those unassessed, albeit important, content standards as possible. One way of

assisting teachers in their instructional promotion of those unassessed content standards is to provide teachers with a menu of easy-to-use and easy-to-score classroom assessments dealing with many of these content standards.

Teachers can then choose which of these *optional* classroom assessments can be used to support a given teacher's instructional emphases. (It is recommended that these classroom assessments be optional, rather than obligatory, so as to reduce the amount of *mandated* testing of children.) The availability of a menu of readily usable classroom assessments will not only improve the effectiveness of a teacher's instruction (directed toward unassessed content standards), but will also increase the efficiency of that instruction. Those effectiveness and efficiency benefits will occur because of the teacher's being able to arrive at a more accurate picture of each child's standards-mastery status for those content standards the teacher has chosen to address instructionally.

Lesson 5: The breadth of the enacted curriculum must be monitored to ensure that instructional attention is given to all content standards and subject areas, including those not measured by a large-scale assessment.

The role of large-scale assessments as curricular magnets has been well documented in the US for many years. What is tested by an important large-scale test tends to be taught by teachers. This is particularly true for high-stakes tests, that is,

those tests whose results have important consequences for students or for the teachers who taught those students. Typically, such tests are associated with state-level accountability programs.

The first three lessons cited here are, in a nutshell, intended to create instructionally supportive large-scale assessments that facilitate teachers' effective promotion of the most important curricular aims sought. But, as is true with most such improvement strategies, the chief potential strength of this kind of assessment system is its chief potential weakness. If teachers realize that only a modest number of high-import content standards will be assessed (Lesson 1), if they have a clear idea about what those content standards mean (Lesson 2), and if they receive per-standard feedback on students' standards-mastery (Lesson 3), then many teachers are apt to focus their instruction *too heavily* on the content standards measured by a large-scale, high-stakes test.

Lesson 4 and Lesson 5 are intended to counteract the quite natural inclination of teachers to devote excessive instructional attention to those content standards measured by a large-scale, high-stakes test. Lesson 4 indicates that educators should be given optional classroom assessment instruments to measure otherwise unassessed content standards. This fifth lesson calls for the establishment of some kind of monitoring mechanisms, that is, quantitative, qualitative, or a combination of both, to help educators attend to the importance of providing a broad and rich curriculum to students, not merely a narrow, test-constrained curriculum.

This fifth lesson refers to the importance of monitoring the *enacted* curriculum. This means that if a suitable system for breadth-monitoring is devised, that system must focus on what actually goes on in classrooms (the *enacted* curriculum) rather than what is represented “on paper” as a set of curricular goals. Lesson 5’s focus is on what transpires in classrooms, not what is intended to transpire by curriculum planners.

Lesson 6: Adequate time must be allowed for the development of instructionally supportive large-scale assessments.

Those who are unfamiliar with the procedural particulars of large-scale test-development are often astonished by the amount of time it takes to develop a first-rate large-scale test. All too frequently, US educational policymakers demand the installation of accountability tests using timelines that, because of the inherent brevity of those timelines, preclude the development of defensible large-scale tests. Naturally, once a policymaker has decided that a large-scale accountability test should be created, that policymaker often wants to get the test “up-and-running” as soon as possible. But the time allowed for the development of many large-scale tests in America makes it, quite literally, *impossible* to develop tests that both supply accurate accountability evidence and also support teachers’ instructional decision-making. Short test-development timelines lead to shoddy large-scale assessments.

The most influential guidelines governing test development in the US are the *Standards for Educational and Psychological Testing**. Any American test-development organization that seriously tries to comply with the recommendations contained in the *Standards* will find that, at minimum, two or three years will be required to create an instructionally supportive large-scale test.

Lesson 7: Educators must be provided with professional development focused on how to optimize children’s learning via instructionally supportive large-scale assessments.

Most American educators would not know how to get the most instructional mileage from instructionally supportive assessments even if such assessments were suddenly made available. Accordingly, the installation of the sorts of large-scale assessments implied by the foregoing lessons creates a concomitant need for powerful professional development to be supplied to both teachers and administrators.

Teachers, for example, must learn how to design instruction aimed at promoting students’ cognitive mastery of the skills and knowledge identified in a content standard and elaborated on in that content standard’s assessment description. Teachers must also be aware of how much confidence can be placed in any test-based inference about a student’s mastery of a specific content standard when only a modest number of test items explicitly assess mastery of that content standard. Teachers must also learn how

* American Educational Research Association (1999), *Standards for Educational and Psychological Testing*, Washington, DC: Author.

to employ standard-by-standard feedback to enhance their instructional planning and delivery.

Administrators must learn how to guide classroom teachers in the most efficient promotion of students' standards-mastery, not only for those content standards measured by large-scale tests, but also for those content standards not assessed by such tests. Teachers will need meaningful administrative support so that curricular reductionism does not occur. Teachers must be given administrative support in order to derive optimal instructional benefits from well-formed large-scale assessments. Instructionally supportive assessments make it possible for teachers to be successful in their classrooms. Administrators must help teachers learn how to translate those possibilities into realities.

In short, the potential dividends of instructionally supportive large-scale assessments are enormous. But those dividends will be realized only if educators know how to utilize instructionally supportive tests. Powerful professional development can help make certain that the inherent instructional payoffs of properly fashioned large-scale assessments will be attained.

Lesson 8: Evidence should be continually collected to determine whether large-scale assessments are appropriate for accountability and instructional enhancement, yet have no negative consequences.

Even the best-conceived plans of well-intentioned assessment professionals, when implemented, sometimes fail or, at least, sometimes require repairs. This final lesson reminds us that a large-scale assessment program's potentials, both positive and negative, should force us to evaluate the assessment program's worth on a continuing basis. Does the assessment program yield the kind of evaluative evidence so necessary for a sound educational accountability program? Are teachers really benefited by what is believed to be an *instructionally supportive* assessment program? Are there unforeseen negative side effects such as dramatically increased student-dropout rates or the emergence of teachers who care more about raising test scores than they do about teaching children?

A variety of evaluative evidence bearing on the assessment program's merits should be collected in a regular, ongoing manner. Ideally, these evaluative studies should be carried out by nonpartisan agencies or individuals, that is, those who have no vested interest in the assessment program's success or its failure.

In the early years of any large-scale assessment program's existence, it is quite likely that a number of adjustments and corrections will need to be made so that the program becomes more effective. The rigorous collection of evaluative evidence by nonpartisans can provide the information so that such improvements can be made in a large-scale assessment program.

Detail-Free Lessons

In conclusion, let me acknowledge that the foregoing eight lessons were provided without an enormous amount of accompanying detail. It was my view that this analysis was not the appropriate locus to elaborate extensively on each lesson's implementation. Those wishing to secure additional details regarding the creation of instructionally supportive tests can contact me for such information (wpopham@ucla.edu). However, please be assured that each one of the eight lessons set forth here can definitely be accomplished, and can be accomplished without Herculean effort.

Instructionally supportive large-scale assessments can be built in the US. And I believe that instructionally supportive large-scale assessments can also be built in the UAE. Instructionally supportive large-scale assessments should be built in any nation, because those sorts of assessments will benefit children.