

THE SCHOOL-REFORM EVALUATOR: APPRAISER OR IMPROVER?*

W. James Popham
University of California, Los Angeles

Preamble: The following remarks are focused on an important issue currently facing school-reform evaluators. Although I shall be addressing a group of international colleagues, my understanding of this issue is based on almost 40 years of my personal experiences as an educational evaluator in the United States. It would be presumptuous of me to suggest that any evaluation-related conclusions that I have based on America's school-reform activities have relevance for evaluators elsewhere such as those in the Gulf Region.

Thus, I hope you will regard my comments not as advocacy of the way that school-reform evaluators ought to function irrespective of the settings in which they operate. Rather, I hope you will consider my recommendations according to your own nation-specific context. Some of my suggestions will surely not be applicable. I hope that a few will.

School reform is supposed to make education better. Indeed, the goal of enhancing educational quality undergirds all school-reform initiatives, whether those endeavors are undertaken nationally or locally. It would seem, therefore, that the success of any school-reform program could be straightforwardly determined according to whether the program did, in fact, improve education. And that's the point at which educational evaluators typically enter the fray. Although the actual conduct of an evaluation of any school-reform program is carried out in the midst of all sorts of program-specific particulars, the overriding mission of a school-reform evaluator must be to find out whether, as a consequence of school reform, education has been improved.

Yet, while the final aim of a school-reform evaluator must be to discover if a reform program has led to improved educational quality, an important along-the-way choice faces today's school-reform evaluators. Surprisingly, it is a choice-point that, until recently, has not been widely recognized.

As implied by the title of this address, the evaluative choice that I intend to consider revolves around the degree to which evaluators carry out a hands-off *appraisal*

* A keynote address presented at the *International Conference on the Reform of Secondary Education* organized by the Sultanate of Oman's Ministry of Education in cooperation with UNESCO, Muscat, Oman, December 21-23, 2002.

of school-reform programs or, instead, take a more proactive role in nurturing the *improvement* of such programs.

To get underway, I wish to briefly examine four considerations that, in concert, have led me to adopt a preference regarding an evaluator's *appraisal-versus-improvement* choice. Having registered that preference, and after attempting to defend it, I will then illustrate how an evaluator of school-reform programs might implement such an approach. Let's begin by taking a brief look at the precursors of today's educational evaluation activities.

Antecedents of Current School-Reform Evaluation

Formal educational evaluation's brief existence. People have been evaluating education for as long as education has existed. It seems probable that most Neanderthals who taught their young how to forage for food probably wondered, at least occasionally, whether their food-foraging instruction had actually worked. (Such wonderment probably took place most frequently just before meal-times.)

Teachers, throughout time, have surely speculated about the success of their efforts. But almost all such speculations were decisively *informal*. At least in America, it was really not until the middle of the twentieth century that a *formal* specialization in educational evaluation emerged. Although there may have been similar developments elsewhere in the world, and possibly earlier than those that occurred in the United States, *formal* educational evaluation in America was born with the enactment in 1965 of a federal law that, for the first time, dispensed substantial federal dollars to U.S. public schools. America's public schools, prior to that time, had been almost exclusively supported by local tax dollars. This historic federal law, the Elementary and Secondary Education Act (ESEA), not only changed the way U.S. public education was financed, but it also significantly altered the way U.S. educators evaluated their own effectiveness. Let's see why.

During the Congressional deliberations that preceded ESEA's passage, there was substantial concern voiced by legislators that the proposed new federal monies might be misspent by local educators. Thus, led by then U.S. Senator Robert F. Kennedy, provisions were installed in the law so that each year's ESEA-supported projects must be *evaluated*. Moreover, the subsequent allocation of federal funds for any ESEA project would depend on the *evaluated* quality of that ESEA-funded project. In other words, in order for local educators to obtain *next year's* federal subsidy, they were compelled to evaluate *this year's* federally financed program. And the more positive that the evaluation was, of course, the more likely it was that future funding would be provided. Overnight, at least in the U.S., formal educational evaluation was born.

The formative/summative evaluation. Given the almost total absence of any serious thinking in the U.S. about educational evaluation prior to ESEA's passage, and the substantial funds newly linked to the conduct of project-evaluations, American

educators immediately sought guidance about how to carry out the new federally imposed evaluations of ESEA-supported projects. Fortunately, several remarkably insightful academics turned their attention to this topic, notably, Robert E. Stake (a psychometrician by training) and Michael Scriven (a philosopher by training). Individuals such as these authored a series of seminal essays regarding the procedures that ought to be employed when one carries out an evaluation of any educational program.

One early distinction, drawn by Scriven in an influential 1967 essay*, contrasted two functions of an evaluator's work, namely, *formative evaluation* and *summative evaluation*. This potent conceptual contrast helped fledgling evaluators recognize that there were two quite distinctive roles in which they could function when they attempted to evaluate an educational program.

Formative evaluation, according to Scriven, refers to the appraisal of a yet-malleable program's quality—an appraisal aimed clearly at improving that program's effectiveness. Formative evaluators, therefore, strive to strengthen programs that are still capable of being modified. In contrast, Scriven contended that *summative* evaluation appraises the quality of an already completed program. Whereas the formative evaluator supplies information to enhance the caliber of a still-alterable program, the summative evaluator supplies information contributing to a go/no-go decision regarding a mature, non-malleable program.

Since the mid-sixties, there has been substantial energy expended by the world's educational evaluators in refining the nuances of the evaluator's summative and formative functions. But Scriven's original 1967 distinction has proven to be both remarkably resilient, and also particularly helpful to those educational evaluators who seriously contemplate the nature of their endeavors.

School-reform evaluation. During the last two decades, especially in the U.S., public school educators have been subjected to ever-increasing scrutiny regarding the quality of their efforts. To illustrate, America's educational accountability movement of the past 20 years arose chiefly because many high-level policymakers became dismayed with what they regarded as an ineffectual U.S. public education system. The rallying cry of accountability programs' proponents ran along these lines: "Okay, educators, your time's up! Now come up with hard, convincing *evidence* that you are really doing a good instructional job!"

And the same skeptical policymakers who helped install our increasingly prevalent educational accountability programs have, not surprisingly, also played a prominent role in stimulating the emergence of educational reform itself. Let's face it, if any system is thought to be working satisfactorily, you really don't need to *reform* it. Thus, school-reform programs have emerged from the same cynical Zeitgeist that led to

* Michael Scriven, "The Methodology of Evaluation," in *Perspectives of Curriculum Evaluation*, R.E. Stake (Ed.), American Educational Research Association Monograph Series on Evaluation, no. 1 (Chicago: Rand McNally, 1967)

the genesis of educational accountability—a fundamental dissatisfaction with the caliber of schooling.

This historical background has inclined most school-reform evaluators to adopt a decisively *summative* orientation to their work. After all, most school-reform programs were installed as a direct consequence of public displeasure with existing school programs. It is not surprising, then, that school-reform evaluators would regard themselves as the public's designees whose role it was to assure the world that educational improvement has, in fact, occurred. And this sort of proof-focused evaluative stance, of course, constitutes a distinctively *summative* approach to educational evaluation.

I regard such a methodological orientation for school-reform evaluators as thoroughly defensible. School-reform evaluators should, indeed, function *summatively*. Concerned stakeholders ought to be supplied with nonpartisan evidence regarding the success of any school-reform program. And we have ample experience, from almost four decades of formal educational evaluation, that it is essentially impossible for any educational evaluator to *simultaneously* carry out both a formative and a summative evaluative mission, at least to carry both missions out with any sort of success. Clearly, therefore, a summative orientation for school-reform evaluators makes substantial sense.

It would, however, be folly for any school-reform program not to employ on-staff *formative* evaluators to assist in the ongoing improvement of such a during-development program. Skilled formative evaluators can play a powerful role in strengthening the efficacy of any educational intervention, hence formative evaluators are certainly needed as soon as any substantial school-reform activity is launched.

But the appraisal-versus-improvement choice I shall soon address is one that must be faced by *summative* evaluators of school reform. Particularly in the evaluation of major school-reform efforts, such as those carried out at a regional or national level, relevant decision-makers deserve to know whether a school-reform program worked. And a well-concerned summative educational evaluator will always do the best job in supplying the kind of nonpartisan evidence necessary for those decision-makers to arrive at a termination/continuation decision about a school-reform program.

At first glance, it may seem contradictory to be dealing with a summative evaluator's *improvement* activities (because such activities are typically carried out by formative evaluators). However, an improvement-oriented approach is precisely one option that now presents itself to today's summative evaluators of school reform.

So, to recapitulate this first point, although there is an important and continuing role for formative evaluation during the time that school-reform programs are being born, refined, and installed, the evaluation of any mature school-reform program should always be *summative* in nature.

Impact of Summative Evaluation's Evidence-Gathering Instruments

The move toward increasingly more stringent educational accountability programs has introduced a significant new wrinkle into the summative evaluation enterprise. Put simply, the more important that a summative evaluation is, the more *influence on educational practice* will be attributed to the data-gathering instruments employed during that evaluation.

To illustrate, during recent years in the U.S. we have seen the installation of rigorous educational accountability programs in which “high-performing” schools are given meaningful financial rewards, but “low-performing” schools are singled out for a series of increasingly substantial penalties. These penalties typically end up with a low-performing school’s being taken over by the government or simply shut down. Is it any wonder that the educators who staff the schools that operate in such an accountability environment strive to be evaluated positively? We have seen, in short, that those educators who operate a school due to be summatively evaluated, particularly when the consequences of the evaluation are significant, will be greatly influenced by whatever evaluative evidence-gathering instruments are employed.

The most important evidence in any school-reform evaluation these days, of course, almost always consists of *students’ performances on tests*, that is, the tests being used by the evaluators. In any U.S. setting where tests play a prominent role in the conduct of an important educational evaluation, whatever is emphasized on the tests will most assuredly be emphasized by classroom teachers.

Although this impact of an evaluation’s data-gathering devices is understandable, and even anticipatable, what has proved so surprising in the U.S. is the enormity of classroom impact wielded by the tests employed in any sort of “high-stakes” evaluative effort. High-stakes tests, in the U.S., are those assessments whose results either (1) have a direct impact on an individual student, such as the award/denial of a high school diploma, or (2) play a prominent evaluative role in determining the degree to which educators are seen to be successful.

Any major school-reform initiative is almost certain to be scrutinized carefully. And it is equally certain that the dominant determiner of the initiative’s evaluation-based success (or failure) will be the quality of students’ test performances. Accordingly, we can be confident that, in most instances, the tests being used by school-reform evaluators will not only be regarded as “high-stakes” assessments but, in addition, will also play a meaningful role in shaping the instructional events that transpire in classrooms.

In short, the instructional impact of the tests employed by today’s school-reform evaluators has turned out to be far greater than had been anticipated during the early days of educational evaluation. And that instructional impact, as I shall attempt to demonstrate, can either be educationally helpful or educationally harmful.

The Harm Caused by Inappropriate Evaluative Assessments

If the wrong kinds of tests are employed by school-reform evaluators, then not only is a significant improvement-opportunity missed, but those tests can also lead to a decisive erosion of educational quality. I can best illustrate this unfortunate consequence by recounting what has often happened in the U.S. when unsound data-gathering instruments have been used by educational evaluators.

American educators, wishing to be evaluated positively, will typically try to have their students perform successfully on the chief tests used in any educational evaluation. The higher the stakes associated with the evaluation, the greater will be the effort that educators give to promoting their students' higher test scores. For instance, if individual schools in a state are to be evaluated according to their students' scores on a statewide achievement test, then it is almost certain that the administrators in those schools will exert considerable pressure on a school's teachers to elevate their students' test scores.

Instructional insensitivity. But this pressure often turns out to have a negative impact on classroom instruction because most statewide tests are not instructionally supportive in at least two important ways. First, many of these tests *attempt to measure too many curricular aims*. A state's curriculum specialists often have, with well-intentioned zeal, identified many more skills and bodies of knowledge for students to master than can possibly be *taught* during the available instructional time or *tested* during the available assessment time. As a result, teachers are forced to *guess* about which of their state's myriad curricular targets will actually be assessed by a given year's tests. And they often guess incorrectly.

Second, many of the standardized achievement tests being used for statewide evaluative purposes, because they were originally constructed to provide norm-referenced comparative interpretations of students' performances, turn out to be *instructionally insensitive*. This is because the test items that maximize the score-spread of students' test performances—score-spread so necessary for fine-grained normative comparisons—are often the very items that are least sensitive to instruction. For example, test items linked to students' (1) socioeconomic status or (2) inherited academic aptitudes tend to do a great job in spreading out students' scores because both of those variables are, themselves, nicely distributed. But to the extent that an achievement test is composed of many items linked to socioeconomic variables or many items linked to students' inherited academic aptitudes, that test will be instructionally insensitive. If teachers are trying to teach their students well, so that their students' test scores will improve, then those teachers are destined to experience a major disappointment. Even first-rate instruction will not raise students' scores on instructionally insensitive tests.

Adverse classroom consequences. Pressured to raise their students' scores on high-stakes evaluative assessments, far too many American teachers have recently transformed their classrooms into test-worship cathedrals wherein students are

relentlessly drilled in ways thought to contribute to higher test scores. More often than not, for example, students will be obliged to complete hours of practice tests that, insofar as possible, resemble the actual evaluative assessments to be used. Thus, rather than a classroom being the locus of students' *learning*, it becomes a drudgery-dominated test-focused drill factory. As a consequence, any genuine joy that students might have experienced in school soon evaporates. Such *jettisoned joy* constitutes a serious adverse consequence of evaluations based on the use of inappropriate assessment instruments.

What is especially troubling about this increasingly prevalent preoccupation with score-boosting in America's classrooms is that U.S. teachers rarely know *where to aim* their instructional efforts. Teachers are typically forced to estimate what's to be assessed by a high-stakes evaluative test—and they frequently are mistaken. Yet, even though teachers are often wrong about an upcoming test's actual emphases, many of those teachers have still chosen to discard curricular content that, in their estimate, does not coincide with what they believe is likely to be tested. As a consequence, we see rampant *curricular reductionism* in the U.S. Content that it is not thought to be assessed turns out, increasingly, to be content that is simply not taught.

Finally, increasing numbers of American teachers—pressured to raise their students' scores on typically opaque and instructionally insensitive tests—have begun to display outright *teacher dishonesty*. For instance, students are often prepared for tests with exercises in which a test's *actual* items are used. And, of course, when students learn what the correct answer is to an item during in-class practice sessions, those students are apt to answer that item correctly during the test administration itself. And when students take the actual test, they discover that they have been made unwitting co-conspirators in a teacher-engineered assessment fraud.

In addition to unethical *test-preparation* practices, we now find more and more U.S. teachers violating *test-administration* protocols in an attempt to get their students to score higher. For instance, numerous teachers have been caught (1) allowing students to have more than the stipulated test-taking time or (2) supplying lavish hints to students about which answers are correct during the test-administration session.

Although we can readily understand why it is that some teachers, in frustration, have engaged in such deplorable practices, those practices are nonetheless deplorable. They must be eliminated. We clearly do not want teachers to be modeling dishonesty for their students.

But these three negative consequences of educational evaluation I have just cited, namely, jettisoned joy, curricular reductionism, and teacher dishonesty, are by no means the *necessary* side effects of educational evaluation. Rather, they are the consequences of educational evaluations based on instructionally dysfunctional evaluative instruments.

School Reform's Raison D'être

There is one more point I need to make before confronting the appraisal-versus-improvement choice for school-reform evaluators and that point deals with the fundamental reason school reformers and, indeed, educational evaluators exist in the first place.

Based on ample experience, I know all too well how enticing it is for educational evaluators to get caught up in the procedural and theoretical nuances of their specialization. For instance, evaluators love to debate the answers to such methodological questions as “How many evaluative resources should be devoted to quantitative versus qualitative evaluation procedures?” or “What are the most effective tactics for ascertaining the unanticipated side effects of an educational program?” Specialists, as is well known, often revel in stirring the viscera of their specialties. Educational evaluators are no different.

But when we step back a pace or two from the evaluation of school reform, we should recognize that the underlying reason school reformers or school-reform evaluators exist at all is *to make education better for children*. To illustrate, how deplorable it would be if an indefensible evaluation of a school-reform program were to provide decision-makers with the wrong evidence—so that those decision-makers either (1) shut down a truly effective program or (2) extended the life of a truly ineffective program. The losers in such a situation are the very children we are attempting to educate.

Thus, I contend that the ultimate criterion by which school-reform evaluators should determine what they do is *not* whether they are properly employing the most up-to-date, professionally sanctioned evaluative methods. Rather, the criterion by which school-reform evaluators should govern their own conduct is *whether their actions are apt to have a beneficial impact on children's learning*.

This point may appear to be little more than a reminder of the obvious, but specialists (in this instance, school-reform evaluators) sometimes need to be reminded of the obvious. Later in this analysis, I shall point out *why* it is imperative that school-reform evaluators continually regard children's well-being as their specialization's, and their personal, *raison d'être*.

Appraisal Versus Improvement

Let me now provide a brief en-route review. I have tried so far to isolate four considerations that bear on the way today's school-reform evaluators should conduct themselves. First, I suggested that school-reform evaluators must function in a summative mode. Second, I contended that the nature of the tests employed by school-reform evaluators will play a remarkably influential role in determining what happens in those classrooms involved in the school reform. Third, I claimed that inappropriate evaluative assessments tend to have an adverse impact on the quality of schooling.

Finally, I argued that the overriding concern of any school-reform evaluator should be to increase the likelihood that children will receive a better education. With these four points in mind, I now turn to an important procedural choice-point for anyone who sets out to evaluate a school-reform program.

Solomon's seductive allure. There is something terrifically appealing about functioning as a dispassionate judge who, having weighed the positives and negatives of an issue, renders a thoughtful judicial conclusion. Judges who dispense their decisions with Solomonic detachment often serve as attractive role models for school-reform evaluators. Thus, many current evaluators of school-reform initiatives conceive of their mission as one of aloof, data-based judgment-making. Their adoption of a Solomonic stance is altogether understandable. After all, summative school-reform evaluators have sometimes been heard to imperiously say such things as: "Formative evaluators, though useful, have typically been totally co-opted by the program's staff. What educational policymakers need is a completely non-partisan, hands-off appraisal of any school-reform program. Formative evaluators are too partisan."

There's no doubt about it—people like to be regarded as nonpartisans. In most circles, even-handedness is applauded while partisanship is deplored. We can, therefore, readily understand why it is that many school-reform evaluators succumb to the obvious appeal of Solomonic appraisal. Yet, I believe that they are wrong. In a choice between dispassionate appraisal and proactive partisanship, I opt with enthusiasm for the latter.

Proactive partnership—its limits. I am *not* suggesting that the evaluation of a school-reform initiative should be carried out by individuals who have already pre-judged the program to be a winner. Nor am I recommending that the caliber of a school-reform initiative be evaluated exclusively by a flock of formative evaluators. No, I want school-reform programs to be evaluated with as much dispassionate rigor as possible *in a consummately summative manner.*

But I want those summative evaluators, *up front*, to help the school reformers be influenced by the most *instructionally supportive* tests possible. Early on, I want school-reform evaluators to work collaboratively with school-reform personnel to devise assessment instruments (for the upcoming evaluation) that will not only produce genuinely credible evidence of the school reform's success (or lack of it), but will also have a decisively positive impact on the instruction taking place as part of the reform itself.

In essence, I want the summative school-reform evaluator to work proactively in constructing suitable evaluative measures of a school reform's effectiveness. Because of the certain impact of those assessments on classroom instruction, such assessments will increase the likelihood that more effective teaching takes place. I will attempt to suggest how this might be accomplished in the remainder of this analysis. However, as soon as suitable data-gathering tests have been installed, the school-reform evaluator

should immediately return to the role of a stringent, hands-off appraiser of whether the school-reform program has, in fact, improved education.

By working collaboratively with the reform-staff's curriculum and instruction personnel, and devising suitable assessments for the evaluation, then a school-reform evaluator will have given the school reformers a fair chance to become successful and, even more importantly, will have increased the odds that children will be better taught. Those are two powerful payoffs for school-reform evaluators, payoffs that I hope will persuade today's school-reform evaluators to opt for the improvement side of the appraisal-versus-improvement dichotomy.

Recapping, then, I believe evaluators of current school-reform initiatives must, though functioning summatively, enter into early-on collaborative efforts to fashion assessment devices that will, hopefully, have a beneficial impact on classroom instruction. In the remainder of this presentation, I will attempt to illustrate how this might be done.

Instructionally Supportive Assessments

Briefly, I intend to describe the nature of the kinds of *instructionally supportive* tests I would like to see school-reform evaluators install as the key evidence-gathering instruments for their evaluations. For those who wish to read further about such tests, there are several recent reports available from the Commission on Instructionally Supportive Assessment that will provide additional insights regarding the composition of these sorts of assessments.*

Four characteristics of appropriate evaluative assessments. What should be the nature of the tests that a *proactively partisan, but summatively stringent* school-reform evaluator should attempt to install? Well, in my opinion, instructionally supportive tests need to possess the four characteristics cited below:

1. *Importance:* The tests must measure unarguably significant skills and/or knowledge so that a school reform initiative can be evaluated according to whether students have mastered genuinely worthwhile curricular aims. School-reform initiatives, more often than not, constitute substantial, publicly funded undertakings. Such undertakings must lead to the attainment of important, not trivial educational outcomes.

* The Commission on Instructionally Supportive Assessment. (1) *Building Tests That Support Instruction and Accountability: A Guide for Policymakers*, (2) *Illustrative Language for an RFP to Build Tests That Support Instruction and Accountability*. Washington, DC: Author, 2001. Available online at www.aasa.org, www.naesp.org, www.principals.org, www.nea.org, www.nmsa.org; Popham, W. James, *Implementing ESEA's Testing Provisions: Guidance from an Independent Commission's Requirements*, March 2002. Available online at www.aasa.org, www.naesp.org, www.principals.org, www.nea.org, www.nmsa.org.

2. *Describability*: The skills and/or knowledge assessed by the tests must be described well enough so that teachers will understand the nature of the skills and/or knowledge represented by those tests. That is, teachers will be able to understand what's being measured at a level of clarity sufficient for purposes of their day-to-day instructional planning.
3. *Reportability*: The skills and/or knowledge being measured must be assessed so that an individual student's degree of mastery can be determined, *and reported*, for each skill or body of knowledge being tested. This means, typically, that such tests must measure fewer skills and/or bodies of knowledge than is typically the case. Thus, there is an even greater need for instructionally supportive tests to assess indisputably important skills and/or knowledge.
4. *Teachability*: The important skills and/or knowledge measured by the tests must be instructionally addressable by typical teachers so that students' mastery of those skills and/or knowledge can realistically be promoted. It is patent folly to try to appraise a school-reform program via instructionally insensitive tests or tests designed to measure essentially unalterable student attributes.

It should be the task of all school-reform evaluators to employ instructionally supportive tests in their work, that is, incorporating these four qualities: importance, describability, reportability, and teachability. Typically, to do so will require a meaningful degree of early-on collaboration between school-reform evaluators and the school reformers whose programs are to be evaluated. Then, once the *nature* of the tests to be used in the evaluation has been determined, the evaluators should make clear to the school reformers what those tests will be like.

I am *not* suggesting that the tests themselves be given to school-reform personnel. Rather, a set of clear descriptions of the skills and/or knowledge to be measured by the tests should be made known to the school reformers. School reformers ought to know how their work is to be evaluated, and providing the school-reform staff with succinct but lucid descriptions of what the tests will assess—along with sample items/tasks akin to those to be used in the actual tests—should be sufficient for instructional planning by the architects of any school-reform initiative (and by the classroom teachers involved in the school reform). What school-reformers should aim their instruction at is the skills and/or knowledge *represented* by an evaluation study's tests—not at the tests themselves.

Happily, to the extent that the evaluation's tests assess significant skills and/or knowledge that are truly teachable, any instructional attention given to the curricular aims represented by those tests will be apt to benefit students.

An illustration of an instructionally supportive assessment. The best example of instructionally supportive assessments in the U.S. are our nation's writing-sample tests. To determine how well a student can compose an essay (for example, a persuasive essay), we employ performance tests in which the student must generate an original essay of a designated type dealing with a previously unencountered topic.

Students' essays are then scored using a rubric (scoring guide) based on a modest number of evaluative criteria (for instance, the essay's "organization" or its "use of proper mechanics"). All of these evaluative criteria are instructionally addressable, that is, they can be taught directly so that students develop an evaluative framework suitable for judging their current and future compositions. As a consequence of the widespread use of writing samples throughout the U.S. during the past 20 years, American students can now write better than they could a few decades ago when a student's ability to write was typically measured using multiple-choice items.

America's writing-sample performance tests are instructionally supportive because they satisfy the four characteristics of an appropriate evaluative assessment. That is, such tests (1) focus on a patently important instructional outcome; (2) are accompanied by a clear description of how a student's performance will be evaluated; (3) can be reported in a straightforward fashion to teachers, students, and students' parents; and (4) can be effectively taught to students by their teachers.

These are the sorts of data-gathering devices that school-reform evaluators must try to employ in their work. Instructionally supportive tests give school reformers a reasonable chance to be successful. Instructionally supportive tests increase the likelihood that students will be properly taught.

America's sorry story of "standards-based reform." To illustrate how a well-intentioned educational reform strategy can collapse if the wrong kinds of assessments are employed to evaluate it, we need look no further than to the "standards-based reform" movement that has dominated U.S. educational reform for the past decade. Although the essential idea underlying this attempt to improve schools was eminently sensible, the efforts to evaluate its success have led to a clear erosion of educational quality, not to the anticipated improvements in children's schooling.

The heart of standards-based reform hinges on the idea that a state's educators will first spell out their state's authorized curricular aims in the form of *content standards*, that is, the skills and knowledge the state's students are supposed to learn. Then teachers attempt to instructionally promote their students' mastery of those content standards. Finally, "standards-based tests" are thereafter employed to ascertain whether students have attained the skills and knowledge reflected in the state's officially approved content standards. Ostensibly, these standards-based tests have been "aligned" to the content standards. Further, it is thought, the actual instruction provided in classrooms will also be aligned with those content standards. This all sounds quite reasonable. Unfortunately, standards-based reforms in America

have proved almost totally ineffectual chiefly because the tests that have been used to evaluate these reforms have been anything but instructionally supportive.*

Frequently, evaluators of standards-based reform have simply selected off-the-shelf standardized achievement tests to use as their evaluative assessment tools. Such tests are rarely “aligned” (in any honest sense) with a state’s content standards. Moreover, because those tests were constructed chiefly to yield norm-referenced comparative information, they are usually instructionally insensitive. In-the-schools educators, of course, invariably fear the adverse consequences of negative evaluations. Accordingly, many educators have attempted to “appear” effective (based on students’ test scores) by devoting inordinate amounts of classroom time to outright test-preparation drills—sometimes (as I indicated earlier) using practice items remarkably akin to the standardized test’s actual items. The quality of classroom instruction has usually plummeted in all those locales where improper evaluative assessments have rendered standards-based reform little more than an exercise in thoughtless score-boosting.

In other instances, evaluators of standards-based reform have pinned their hopes on customized tests that, supposedly, have been custom-made to more accurately represent a state’s content standards. But, often because the content standards themselves are too vague and too numerous, the resultant customized tests fail to do a satisfactory job either in (1) assessing students’ mastery of those standards or (2) clarifying what curricular aims are actually to be emphasized in the evaluation’s assessment instruments.

Summing up, standards-based reform in America has been a much touted effort to improve a public-school system that most observers believe needs improvement. Yet, because the evaluators of this reform initiative have not been proactively partisan, standards-based educational reform in the United States has turned out to be a well-intentioned washout.

With more suitable evaluative instruments—devised in advance of the actual implementation of this very reasonable improvement strategy, standards-based reform might well have benefited American children. If the evaluative tests had been focused on a modest number of truly significant content standards that were (1) accurately measured, (2) well described, (3) reported on a standard-by-standard basis, and (4) addressed instructionally by teachers, then standards-based reform might have improved U.S. education. As it is, however, the impact of this promising improvement strategy has been dissipated by the use of inadequate evaluative assessments.

Only One Major Point

As indicated at the outset of this analysis, I have based my views exclusively on evaluation developments in the U.S. Hence, my comments may not have relevance in

* I have addressed this problem in greater detail elsewhere, Popham, W. James, “Combating the Fraudulence of Standards-Based Assessment,” *American School Board Journal*, *in press*.

other settings where education, accountability, or the roles of test-based evidence are viewed differently. Listeners must reach their own judgments about the pertinence of my remarks.

However, in the U.S. we have learned one enormously important lesson related to the evaluation of school-reform initiatives. The lesson is that the school-reform evaluator's choice of evidence-gathering instruments, specifically the tests of students' performance, is immensely influential on the instruction that transpires as part of the school-reform initiative. It was suggested that if a school-reform evaluator installs instructionally supportive tests as the evaluation's dominant data-gathering devices, a number of positive educational consequences will ensue.

For some who serve as educational evaluators, it is tempting to sit back and— from a dispassionate distance—render nonpartisan appraisals regarding the merits of school-reform programs. I think that's fine, but only *after* installing the kinds of instructionally supportive tests that can help school reformers create better instructional activities for students.

I am confident that, if given this appraisal-versus-improvement choice, Solomon himself would have made the right selection.