EDUCATIONAL MISMEASUREMENT:
HOW HIGH-STAKES TESTING CAN HARM OUR CHILDREN
(And What We Might Do About It) *

W. James Popham
University of California, Los Angeles

In too many of our nation's classrooms these days, instructional quality is dropping as a direct result of the pressure for teachers to boost students' scores on poorly conceived tests. The consequences for kids are devastating. The impact on teachers' morale and sense of professional efficacy are also devastating. And, finally, the effects of such score-boosting pressures on the overall quality of public education, and on citizens' regard for public education, are equally devastating. When it comes to high-stakes testing programs these days, devastation abounds.

The vast majority of our teachers are doing what they've always done, namely, trying to provide children with a decent education. But in more and more localities, the relentless pressure to produce higher test scores has driven some teachers to (1) exclude any curricular content not covered by the applicable high-stakes test; (2) drill students so heavily on items akin to those on the test that some students' love of learning is extinguished; and (3) engage in questionable or downright dishonest test-preparation and test-administration practices. Classroom teachers are *not* the villains in this devastation-laden drama. On the contrary, teachers are the *victims* of unsound accountability systems designed in such a way that they foster unsound instructional practices.

In most parts of the nation today, it's not too difficult to collect a medley of horror stories about what happens when test-pressured teachers focus only on score-boosting. Whatever the actual number of teachers who have, though understandably, engaged in tawdry test-preparation practices, any number larger than zero is clearly a number too large. The students of such teachers are being educationally short-changed.

---

* Presented at a Joint NSAII and NALPSE Conference, National Education Association, Sanibel Island, Florida, November 19, 2000.

A Preview

In the following essay I will describe (1) what sorts of measurement misconceptions allow unsound high-stakes assessment programs to flourish and (2) what actions might be taken if the National Education Association wishes to impede today's test-driven erosion of educational quality. Before getting underway, however, please attend to one key word in the title of this essay. That one word is *can*.

Here's why such a petite, three-letter word is so pivotal. Most people believe in the instructional dividends of sensible educational testing. It is widely held that by teachers' using the proper tests for the proper purposes, those teachers will teach better, so that students will learn better. And even *high-stakes tests* can help teachers and students—if those tests are properly constructed. I am *not*, therefore, opposed to all high-stakes testing programs. I am, however, fiercely opposed to flawed high-stakes testing programs—the kind that end up harming children. Unfortunately, the high-stakes testing programs we now find throughout our nation are so flawed that, without delay, they should either be dumped or dramatically overhauled.

Questing for Culpability

Howard Cosell, a prominent TV sportscaster of yesteryear, loved to identify the chief villain whenever a sports blunder had taken place. Howard would say, "Who goofed?" He would then quickly answer his own question by singling out the person he regarded as the offending party. Cosell recognized that many people derive some serious, if voyeuristic, kicks from finding out "Who goofed?" Well, given the increasingly harmful educational impact of current high-stakes testing programs, many educators find themselves wondering who is at fault. How did this sorry state of affairs come to exist? Who fell down on the job?

The answer to the "Who goofed" query is quite simple: *We goofed*. That is, American educators have allowed unsound high-stakes testing programs to be born. And American educators have allowed unsound high-stakes testing to prosper. It is our fault.

Let me clarify who these "American educators" are that Howard Cosell and I would both label as the culprits. By "American educators" I refer to the policymaking leaders of the professional

organizations that represent teachers and administrators. By "American educators" I also refer to the university academics whose writings about key educational issues are often influential. And among the most guilty "American educators" are the members of the educational measurement community—myself included—who should have foreseen and forestalled the damaging consequences of poorly conceived high-stakes testing programs.[*] Collectively, then, I believe the entire array of our nation's educational professionals to be at fault. As a profession, we have allowed something insidious to flourish while we watched it happening.

Administrators' and teachers' failure to speak up about assessment issues is eminently understandable, of course, because in only a few states are prospective teachers or administrators required to complete any sort of measurement course as part of their preparation programs. And, until recently, those few required courses were altogether abstruse–dealing more directly with the nuances of exotic reliability coefficients than with what tests might actually have to do with *teaching*. One reason that many educational leaders have countenanced the creation and expansion of unsound high-stakes testing programs is because those individuals are, for the most part, not really knowledgeable regarding the essentials of educational measurement.

As a profession, we have permitted the emergence of an evaluative zeitgeist wherein the quality of schooling is being determined in a decisively dumb way. We have allowed this dumbness to continue because, for the most part, many educational professionals didn't really know how dumb it was. And the people who did know, or at least should have known, did not alert us to the emerging calamity. I refer specifically to the members of the educational measurement community who ought to have been creating a full-blown ruckus over what is a classic case of *educational mismeasurement at its worst.*

This nearly universal assessment acquiescence on the part of our profession, unfortunately, has permitted the creation of important assessment programs that *appear* to be doing the job for which they were created, but actually are not. Think of the numerous high-stakes testing programs that now purport to reveal how successfully our schools are. In many instances, the achievement tests used for these *school-success* assessment programs are off-the-shelf nationally standardized tests such as *The*

[*] Although I am a former high school teacher whose only graduate training in educational measurement consisted of one fairly vapid master's degree course, I have ended up working in the field of educational measurement because of assessment's potential impact, either positive or negative, on *instruction*.

Stanford Achievement Tests or *The Iowa Tests of Basic Skills*. All of these nationally standardized achievement tests are built and distributed by the three U.S. companies that sell such tests. In other instances, a customized achievement test may have been created especially for a state (often built by one of the same three test-publishing companies). Although these state-specific tests are typically intended to better reflect a state's curricular preferences, in many instances such customized tests actually function no differently than a nationally standardized achievement test.

Putting this point another way, even though a customized state-level accountability test may be referred to as a "standards-based" assessment or by some other positively-spun title, if the test was created by a company whose main stock-in-trade is the building of traditional standardized achievement tests, chances are that the customized test will end up working in a fairly traditional manner.

A Half-Dozen Pithies

The following analysis will focus exclusively on one prominent instance of educational mismeasurement, namely, *the reliance on students' standardized achievement test scores to evaluate the quality of instruction*. In order for you to see how wrong-headed such a use of test results is, a few crucial concepts about educational testing need to be considered. To keep this exposition suitably terse, I will cast it in the form of six pithy propositions—each proposition to be followed by a brief explanatory comment.[*]

Proposition 1: Educators assess students so that children's *overt* responses (to tests) will allow educators to draw inferences about children's *covert* knowledge, skills, and affect.

Comment: You can't tell how well a child can spell by observing the child, even with a magnifying glass or through a one-way mirror. Similarly, children's knowledge of U.S. history, their ability to write narrative essays, or their attitudes toward mathematics are *covert*. Educators assess children to secure a child's *overt* responses (to a test) that can indicate how much knowledge or skill a

[*] *Pithy*, according to the dictionary, refers to something that is "brief, forceful, and meaningful in expression." It can also signify something that is "of, like, or abounding in pith." In the current context, you must decide whether it is the first or the second meaning that is being employed. Think of this as a pith-quiz.

child possesses. For instance, when a student performs well on a test of reading, we use that (overt) test performance to help us figure out how much reading ability (covert) this student actually possesses.

Students' affect, of course, is assessed far less frequently than their skills or knowledge. But, as is true with the measurement of knowledge or skills, educators still end up using students' overt responses (for example, to an attitude inventory) as a way of determining students' covert affective dispositions.

Proposition 2: Tests evoke students' responses that are only samples of how students would respond to the full domain of knowledge, skill, or affect being assessed.

Comment: Most domains of knowledge, skill, or affect are far too large to assess in their entirety. To assess *completely* children's mastery of certain skills or bodies of knowledge, we would probably need to assess children continuously until they were drawing social security checks. AARP, fortunately, does not have an admissions examination. Educational tests, then, are intended to *represent* a given body of knowledge, skills, or affect. The sampling-based *representational* mission of educational assessment contributes significantly to the next proposition about assessment accuracy.

Proposition 3: **Educational tests are far less precise than is generally believed.**

Comment: Today's educators live in an era in which evidence, especially quantitative evidence, rules the roost. And because educational tests yield numbers, sometimes numbers even containing decimals, we often ascribe more accuracy to those numbers than is warranted. Every educational test has a "standard error of measurement" that, just as is seen with the media's sample-based opinion surveys, represents the test's plus-or-minus error margins. For an educator to reach a rock-solid conclusion about a student's covert capabilities on the basis of a single, sample-based test is naive.

Proposition 4: The nature of the inference that is based on students' test results plays a pivotal role in subsequent decisions or conclusions linked to that inference.

Comment: If a teacher administers a 20-item test assessing students' skill in solving double-digit multiplication problems, the teacher can draw reasonably accurate inferences about students' multiplication skills when dealing with these sorts of problems. Although the teacher's test-based inference might be somewhat off the mark because of the sampling nature of the test, the teacher can defensibly make instructional decisions based on students' test performances. It would be patently absurd, however, for the teacher to reach conclusions based on the multiplication test about whether students enjoyed mathematics or, indeed, planned to major in mathematics while in college. Those subsequent inferences, of course, are decisively different from the first inference. Using a multiplication test to make inferences about students' affective dispositions would result in invalid inferences. Lousy inferences lead to lousy inference-based decisions.

Proposition 5: It is *appropriate* to reach conclusions about the quality of education based on students' test scores *if* the items on a test measure what is supposed to be taught in school.

Comment: Recalling that the nature of sample-based testing always can reduce the accuracy of a score-based inference, if the items on a test do, indeed, measure what teachers ought to be teaching, then students' performances on that test should help us get a reasonable picture of what the students have been taught. Given the enormous amount of content to be assessed, there is a high likelihood that the *sample* of knowledge and skills measured by a standardized achievement test will not be well aligned with the curricular emphases in a given locality. And even if all the items on a test do, in fact, measure content that should be taught, there is still the possibility that such teaching-testing mismatches will be present. Hence, any test-based conclusion about educational quality should be quite guarded.

Care must be taken, of course, in how the test scores are collected. To compare the end-of-school performances of *this year's* sixth-graders with those of *last year's* sixth-graders is contrasting the test scores of two different groups of children. Given the imprecision of all educational

measurement, any resultant differences between the two groups of different children may be attributable to myriad factors, only one of which is instructional effectiveness.

Proposition 6: It is *inappropriate* to reach conclusions about the quality of education based on students' test scores *if* many of the items on a test do not measure what is supposed to be taught in school.

Comment: Test *scores* exist because of the way students respond to the *items* on a test. Yet, a rigorous review of the items on nationally standardized achievement tests will reveal that many of these items are not dominantly focused on what teachers are supposed to teach. A student's probability of supplying correct answers to a substantial numbers of items on these tests will be heavily influenced by a student's socioeconomic status or by that student's inherited academic aptitudes. In short, *many items on standardized achievement tests measure what students bring to school, not what they learn there*.

And, as noted earlier, most state-level educational accountability systems are either based on off-the-shelf standardized achievement tests or on customized achievement tests that function much the same way. Thus, most state accountability programs set out to measure educational quality using the wrong assessment tools. Trying to measure educational quality with a standardized achievement test is like trying to measure temperature with a tablespoon. It just won't work.

This concludes my proposition-based argument leading up to the final and most important one, namely, that if many of the items on an achievement test measure things *other* than what teachers are supposed to teach, then it is wrong to employ this test as a way of telling how effectively those teachers are teaching. The validity of Proposition 6 needs to be shored up, however, because it is this key proposition that should lead sensible folks to reject the use of standardized achievement tests as instruments to evaluate schooling. Let's turn, then, to a quick look at the innards of standardized achievement tests.

What Makes Standardized Achievement Tests Tick?

Standardized achievement tests are assessment instruments that are administered, scored, and interpreted in a standard and predetermined manner. Because the items in these tests typically deal

with students' skills and knowledge, and because the tests are called "achievement" tests, most people assume that these tests measure what students have *achieved* in school. That assumption is not warranted.

Standardized achievement tests are really quite marvelous measurement tools. If used properly, they can yield information that is useful to both parents and to teachers. If it is learned that Megan scored at the 95th percentile in reading, but only at the 37th percentile in mathematics, then that sort of information can be profitably employed both by Megan's teachers and by her parents. Educators should not be opposed to standardized achievement tests. But educators should insist that such tests be used appropriately. And judging the quality of schooling is not an appropriate use of these tests.

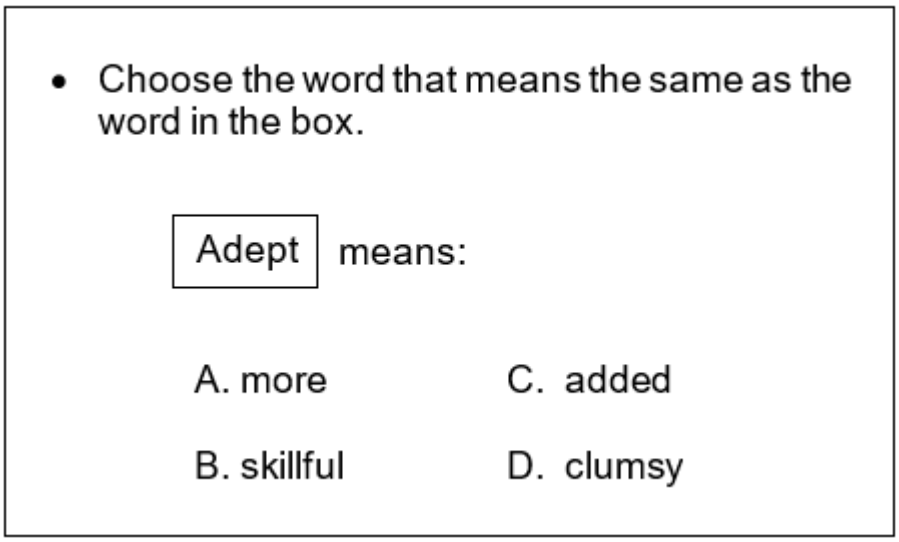
Standardized achievement tests trace their ancestry back to World War I when the *Army Alpha* was developed to help identify candidates for the U.S. Army's officer training programs. The *Alpha* was an *aptitude* test in the sense that it was designed to predict how well Army recruits would fare if they ended up as officer-trainees. The essence of the *Alpha's* measurement approach was comparative. It allowed Army officials to see who scored at the 96th percentile (in comparison to a norm group of earlier test-takers) and who scored at the 23rd percentile. In order for these comparative interpretations to be made, it was imperative for the *Alpha* to produce a reasonable degree of *score-spread*, that is, a range of performances so that different examinees would get different scores, thus making possible the fine-grained comparisons needed for the *Alpha* to work properly as a predictor test.

Today's standardized *achievement* tests employ a measurement strategy astonishingly similar to the one embodied in the Alpha, an admitted *aptitude* test. For today's standardized achievement tests to permit the kinds of fine-grained comparisons needed by educators, again we find a relentless quest for *score-spread*. The problem is that *the way such score-spread is attained* by standardized achievement tests renders those tests altogether unsuitable for judging the quality of schooling.

Three types of items. Three types of items will be found on nationally standardized achievement tests. These items measure (1) what students learn in school; (2) depending on their socioeconomic status, what students learn outside of schools and (3) what verbal, quantitative, or spatial aptitudes students have inherited. Remembering that students' test scores come from students' responses to a test's items, let's look briefly at each of these three item-types.

What's taught in school. A good many items on standardized achievement tests assess the sorts of knowledge and skills that are typically taught in school. Consider, for example, the item presented in Figure 1. It is a near-replica of an actual item currently found in one of today's nationally standardized achievement tests. The item has been altered slightly to preserve the test's security, but the cognitive demands imposed on the student who wants to answer this item correctly, and the other two items you'll soon see, are identical to those found in the actual items.

Figure 1. A sixth-grade reading vocabulary item based on a similar one from a nationally standardized achievement test.



Please look at the sixth-grade vocabulary item in Figure 1 calling for the student to select an appropriate synonym for the word "adept." Surely educators want their students to possess adequate reading vocabularies. But what if, in a particular school district, the word "adept" is not identified as a word that ought to be taught by the time the district's students complete the sixth grade? How fair is it to evaluate that district's instructional success on the basis of content that wasn't supposed to be taught?

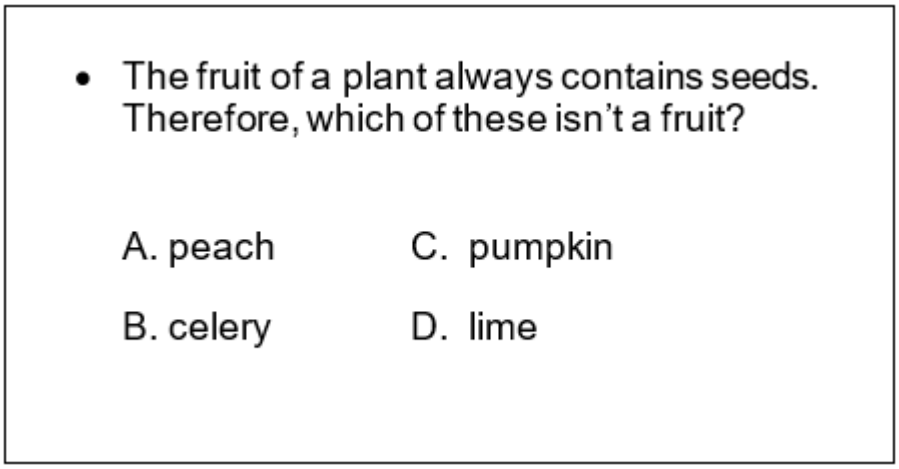
The mismatch between what's tested and what's taught will arise in part because a test publisher must, as noted in Proposition 2, *sample* content in order to complete a test's administration in a reasonable time period. Beyond that, a national test publisher needs to create a test that, from a curricular standpoint, meshes best with the diverse preferences of educators all across the land.

Although publishers try to base their items on the knowledge and skills that are most often pursued in our nation's schools, sometimes a test that's constructed according to a one-size-fits-all conception of content will, in a given community, fail to fit well what's taught in that community. One group of researchers at Michigan State University has concluded that between 50 and 80 percent of the content on certain standardized achievement tests is not apt to be covered meaningfully in some localities.

Thus, even for the items on standardized achievement tests that attempt to access what's taught in school, there will be some tests that don't work well in a given community because, for that particular community, a good many of the test's items will cover things that weren't even supposed to be taught.

What's learned outside of school. A second kind of item found in today's standardized achievement tests may look "educational" but, after closer scrutiny, turns out to measure the sorts of things that kids learn at home. And those sorts of things, as is well known, depend dramatically on the socioeconomic status (SES) of the child's family. Think about an affluent family in which both parents are well educated and all sorts of diverse learning opportunities exist. Consider, for instance, the learning opportunities that are present when children accompany their parents to the opera or hear conversations during dinner about the stock market's fluctuations. An SES-linked item is one that is more likely to be answered correctly by children from higher than lower SES backgrounds. For example, please look at the sixth-grade science item presented in Figure 2.

Figure 2. A sixth-grade science item based on a similar one from a nationally standardized achievement test.



It should be apparent that children will be advantaged on this item if their parents not only can afford to buy fresh celery and limes at the grocery store but also have the cash needed to transform a pumpkin into a jack-o-lantern each October. Such advantaged children will surely, on average, have an easier time with this item than will children whose parents are barely getting by on food stamps.

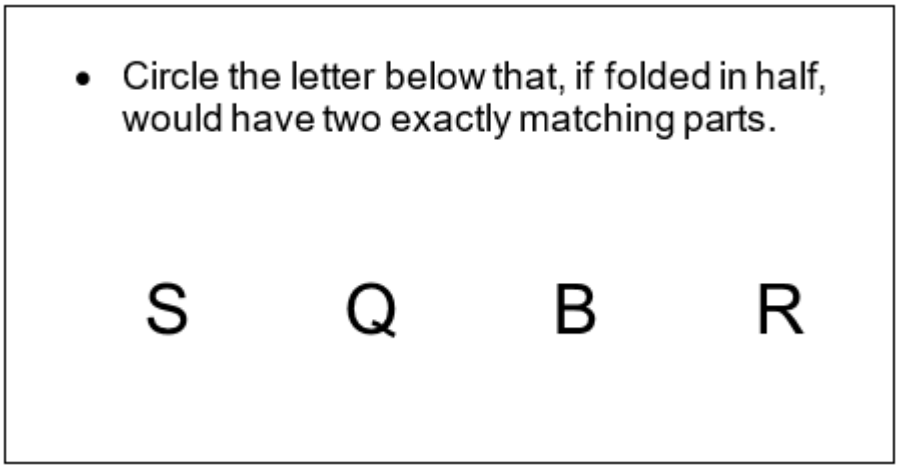
There are substantial numbers of these SES-linked items on standardized achievement tests. And why, one might ask, are such items used on an "achievement" test? The answer is all too simple. These sorts of SES-linked items do a terrific job in producing the score-spread so central to the assessment strategy underlying such tests. From the perspective of a test developer who wants a test that produces a meaningful spread of students' scores, SES is a delightfully spread-out variable, and a variable not readily altered! SES-linked items do, indeed, spread out students' test scores. But SES-linked items don't measure what is supposed to be taught in schools. To judge educators' effectiveness using a test containing many SES-linked items is wrong.

Aptitudes that children inherit. Some children are born smarter than others. Increasingly, of course, educators are accepting Howard Gardner's idea that there are multiple intelligences, so that a child can be *word-smart* without necessarily being *number-smart,* and so on. To a very substantial extent, children *inherit* key academic aptitudes such as their capacities to engage in accurate spatial visualization. The three kinds of inherited academic aptitudes measured by items on standardized achievement tests are children's *verbal*, *quantitative,* and *spatial* aptitudes.

Take a look at Figure 3's fourth-grade mathematics item patterned closely after an item in an existing standardized achievement test. This is an example of a kind of item commonly found in the mathematics section of standardized achievement tests. It is an item, of course, that depends heavily on students' abilities to visualize spatially. And some children are simply born with more of that aptitude than are other children.

Items such as the one in Figure 3 address content that may appear "mathematical," but what sensible fourth-grade teacher spends any instructional time having students practicing their "mental letter-bending skills?" And, once more, we might ask why these sorts of items (and there are plenty)

Figure 3. A fourth-grade mathematics item based on a similar one from a nationally standardized achievement test.



are used by the builders of standardized achievement tests? Same question; same answer. These inherited academic-aptitude items do a great job in producing that much cherished score-spread. Inherited academic aptitudes are not only nicely spread out but, by definition, they are not modifiable.

Yet, is it fair to judge a school staff's instructional success using items most likely to be answered correctly by students who got lucky in the genetic lottery? Today's achievement tests contain far too many items that really are IQ-items camouflaged in achievement-test costumes. It is fundamentally misguided to evaluate schools on the basis of tests containing many SES-linked or many inherited-aptitude items. It is this kind of educational mismeasurement that has led some teachers to engage in the sorts of educationally unsound practices described earlier. Those teachers realize they're not being properly evaluated, so they are driven toward score-boosting tactics that will work.

This is the reason that Proposition 6 needs not only to be understood, but also to be widely promulgated. To repeat, it asserts that: *It is* inappropriate *to reach conclusions about the quality of education based on students' test scores* if *many of the items on a test do not measure what is supposed to be taught in school*. Reasonable questions that a careful reader might raise are (1) "How many SES-linked or inherited-aptitude items are there in today's standardized achievement tests?" and (2) "Are we talking about a teaspoonful or a ton of such items?"

Well, I could hardly be classified as a nonpartisan with respect to this issue, but I did carefully go through the core batteries, item-by-item, of two currently used standardized achievement tests and found about 20 percent of their math items, 40-50 percent of their reading items, and 70-85 percent of their language arts items were unsuitable for purposes of judging school success. Even if you were to reduce my estimates by half, this would still leave way too many items on standardized achievement tests that confound any evaluations about the quality of schooling.

Unfixed Problems Fester

To recap, briefly, as long as the wrong sorts of high-stakes tests are used, it is certain that resultant pressure to raise students' test scores will drive some members of our profession, as a last resort, toward unsound instructional practices. I believe that the architects of today's educational accountability programs, with few exceptions, really thought that if educators were forced to display test-based evidence of their effectiveness, the quality of instruction for students would, over time, improve. Sadly, many accountability systems that were explicitly installed to improve the quality of education have actually degraded the educational programs they were intended to enhance.

A choice facing the National Education Association (NEA) is whether to allow the current status of educational assessment affairs to continue or, instead, to do something to change it. Wishing won't make it so. This is a clear choice-point for NEA and its state affiliates. It is reasonably easy to prophesy what will happen if organizations such as NEA adopt less-than-militant stance with respect to the use of unsound high-stakes tests. Things will get worse. The public will continue to clamor for the wrong kind of evidence. And increasing numbers of beleaguered teachers and administrators will succumb to the use of educationally harmful strategies that focus too heavily on score-boosting.

But what would happen if groups such as NEA and its state affiliates embarked on a serious campaign to halt the misuse of unsound high-stakes tests? What would happen if the Association, instead, actively supported the installation of accountability systems that simultaneously (1) provide the public with credible evidence of educational quality and (2) nurture instructional improvements leading to enhanced learning for our students? If educators can convince the public and pertinent policymakers that the current high-stakes testing situation is harming children, but that appropriate ways of evaluating our schools can be established, then more defensible kinds of educational accountability systems might just emerge.

I'd like to wrap up this analysis by describing a set of possible action options that might be considered by NEA leaders. Some of the activities are implementable at the state level with only modest support from NEA headquarters. Some options, on the other hand, could not realistically be carried out satisfactorily by a state affiliate operating solo. Those action options requiring national leadership will be identified. To be sure, not all of the following action options need be followed even if NEA policymakers decide to aggressively combat today's use of inappropriate high-stakes tests. Yet, a strategy incorporating several of these tactical options just might work!

A Menu of Action Options

Several of the potential activities to be described below would be precursive to any meaningful modification in a state's accountability program. Other activities would lead to substantive changes in the programs themselves. Certain of the activities to be described, then, should be seen as ancillary, but supportive, of revised state-level accountability programs. Other activities would fundamentally alter the nature of those accountability programs.

- ***Option 1: Initiate assessment literacy programs for teachers and administrators***. If a state's educators do not understand why certain high-stakes tests not only yield invalid estimates of instructional quality, but also are likely to lower educational quality, those educators cannot inform parents or policymakers about such problems. Moreover, assessment-*illiterate* educators will be unable to describe more appropriate evidence by which citizens and state policymakers can evaluate instructional quality. If an existing or proposed educational accountability program is likely to cause educational harm to the state's children, then that state's educators need to understand *thoroughly* why this is so.

  Although some of NEA's state affiliates could establish a variety of assessment-literacy promotion activities by themselves, NEA's leadership could certainly make this task less aversive by supplying affiliates with both print and non-print materials for use in state-level campaigns to promote teachers' and administrators' assessment literacy.

- **Option 2: Offer carefully structured briefing sessions to educational policymakers regarding appropriate/inappropriate ways of evaluating schooling**. At the state level,

concise explanatory sessions for state board members, district board members, and state legislators can be planned to clarify why an existing (or proposed) standardized achievement test is likely to provide a misleading picture of school and district instructional quality. The use of actual (or slightly altered) items from that achievement test can be especially helpful in allowing these individuals to understand why certain kinds of test items, items found in profusion on such tests, provide invalid estimates of educational quality. At the national level, of course, NEA would need to work with executive and legislative branches of the federal government.

The importance of this second action-option cannot be underestimated. Most of today's ill-conceived high-stakes testing programs were created because assessment-illiterate policymakers believed those programs would benefit children. Although there may be numbers of pro-voucher policymakers who would actually prefer to do away with our public schools, most educational policymakers simply didn't know any better when they supported the establishment of an educational accountability system based on unsound high-stakes tests. Such policymakers thought that standardized achievement tests were the proper measuring stick by which to judge a school's success. They need to learn why this is not so.

- ***Option 3. Provide briefing sessions for the media***. As the intensity of public interest in students' test scores increases, we can be certain that members of the media will attend to important events in this arena. Fortunately, an increasing number of education writers for the nation's newspapers are becoming knowledgeable about the nuts and bolts of educational testing. For example, Richard Lee Colvin and Martha Groves of *The Los Angeles Times* can now hold their own with educators in any discussion of test-based accountability systems. Members of the media who possess the sort of assessment acumen displayed by Groves and Colvin can make a real contribution to the public's understanding of assessment-related issues. Both nationally and locally, Association leaders and members can bring succinct, readily understandable explanations to media folks about what sorts of tests should/shouldn't be used to judge schools. And, as is true when providing explanations to policymakers (See action Option 2.), the use of actual or slightly altered items from real tests is an effective way for media representatives to grasp the true reality of how readily the implications of students' test scores can be misconstrued.

- ***Option 4: Implement meaningful assessment literacy programs for parents***. As soon as educator-focused assessment literacy programs have been concluded, NEA's state affiliates can provide outreach programs tailored to the interests of parents. Parents will almost always be supporters of assessment programs that are good for children. But parents need to truly understand the key measurement concepts involved. An ill-conceived educational accountability program will, therefore, be accurately seen by *assessment-literate* parents as an activity that will diminish the quality of schooling. Parents will recognize that the education being provided for their own children will almost certainly deteriorate as a result of an unsound accountability program. Informed parents can play a powerful role in combating assessment-sired silliness.

- ***Option 5: Foster establishment of* autonomous *parent-action groups***. Unfortunately, if *educators* protest the misuse of even a seriously flawed statewide accountability program, they will be regarded as self-serving, hence thoroughly unbelievable. However, if *nonpartisan* parent groups protest a poorly conceived accountability program, the views of those parents will be given more serious consideration by policymakers. NEA and its affiliates can have great confidence in the actions of *autonomous* parent groups, but only if the members of those groups are assessment-literate. A group of *assessment-literate* parents who study an unsound accountability program will almost always conclude that the program should be jettisoned.

- ***Option 6: Undertake a public-information campaign organized around educator-written letters and ed/op essays for local newspapers***. The nation's citizens need to understand that America's educators are *not* fleeing from evaluative scrutiny. Newspaper readers must learn that alternative accountability programs can be employed—programs even more rigorous than those that now exist. Such programs can monitor the success of a state's educators while stimulating even more effective instruction by its teachers and, as a consequence, more meaningful learning by students. The educators who write these letters or essays, of course, must thoroughly understand the difference between appropriate and inappropriate test-based accountability programs. The writers should include NEA and affiliate leaders and members.

- ***Option 7: Conduct security-monitored reviews of the items in the high-stakes test being used in the state's accountability program***. There are enormous insights to be gained if educators and noneducators carefully analyze, *one item at a time,* the actual items in a locally adopted high-stakes standardized achievement test. The protocol for such item-reviews must be carefully designed, of course, but the results of such rigorous item reviews can be remarkably illuminating. Appropriate authorization would typically need to be secured from a state department of education to have standardized achievement test items reviewed under strict, security-preserving conditions. If *nonpartisan* reviewers were employed in such item reviews, their conclusions would usually be regarded as credible. A state's citizens and its educators need to understand the nature of the test items that, when added together, produce students' scores. NEA could provide its state affiliates with possible protocols wherein an affiliate sponsored and arranged such item-reviews, but in which the moderators and item judges were parents or members of the business community.

- ***Option 8: Devise and implement valid, credible evaluative schemes suitable for school-level and district-level accountability***. There is nothing wrong with accountability-oriented programs for the evaluation of schooling *if* the appropriate kinds of evaluative evidence are incorporated in those evaluations. To reject an evaluative program based on the wrong data (e.g., standardized test scores), yet not replace the program with an evaluative system based on defensible data would be both professionally and politically unwise. However, ways *do* exist for collecting solid preinstruction-to-postinstruction evidence of a school staff's instructional effectiveness. School staffs must be assisted in learning how to collect and succinctly report credible evidence about their success in promoting students' mastery of their state's key content standards. Reports of such evidence, initially provided at the school level, then summarized at the district level, can meaningfully buttress the kinds of accountability evidence secured from other sources.

To illustrate, data-gathering designs can be employed by teachers so that students' post-instruction status can be contrasted with their pre-instruction status using blind-scoring models in which not only educators, but also parents or other members of the community participate in the evaluation of students' test responses. And not only can evidence of instructional effectiveness be based on students' growth in the mastery of significant skills and knowledge,

but also on students' affective growth. If appropriate data-gathering models are used, the resultant evidence of instructional success can not only be useful to teachers themselves, but will also be regarded by parents and educational policymakers as *credible*. Alternative data-gathering schemes must be valid—and they must be believable, even to skeptics.[*]

For this action option, it would be necessary for NEA headquarters to take the lead in supplying guidelines to state affiliates. These guidelines might take the form of relatively brief pamphlets containing rationales and step-by-step procedures to be followed by teachers who wish to assemble evidence regarding their own instruction's effectiveness. NEA state affiliates, then, could assist members in learning how to collect (and report) valid and credible evidence of instructional effectiveness.

- ***Option 9: Lobby for the use of custom-built statewide standardized tests that (1) accurately reflect mastery of a state's most important content standards, (2) provide appropriate instructional targets for the state's teachers, and (3) yield evidence from which valid inferences can be drawn about the instructional effectiveness of a state's educators***. As suggested in a related analysis,[**] it is possible to build large-scale assessments (such as the tests needed for a defensible state-level accountability program) that can measure instructional quality while, at the same time, providing appropriate clarification of the assessment program's targets. Such clarification is needed so that teachers can direct their instruction toward the important bodies of knowledge and skills being measured rather than toward the specific items on a test. The creation of a customized test for a state would, to be sure, cost more than simply using an off-the-shelf test. However, that test-development cost will be trivial when contrasted with the educational calamities certain to grip any state because of an accountability program that relies on the wrong kind of high-stakes tests.

  A customized test could be built in response to a state-issued request for proposals (RFP). This new test would need to satisfy the *instructionally oriented requirements* of that RFP. The new statewide tests might still be constructed by one of the major U.S. test-

[*] A description of such a data-gathering design, a split-and-switch version of the classic pretest-posttest model, is described in Popham, W. James, *Modern Educational Measurement: Practical Guidelines for Educational Leaders* (2000). Boston: Allyn and Bacon.

[**] Popham, W. James, *Assessments that Illuminate Instructional Decisions*, a presentation at the 30th Annual National Conference on Large-Scale Assessment, Council of Chief State School Officers, Snowbird, Utah, June 25-28, 2000.

development firms. Yet, because of the RFP's explicit instruction-related stipulations, the resulting tests would be dramatically different than customary off-the-shelf standardized achievement tests. However, having been created by an *established* national test-development firm, the customized high-stakes test would be seen as a reputable rather than home-grown.

For implementation of this action option, the leadership role of NEA headquarters would be indispensable. This is the kind of activity that few, if any, state NEA organizations could carry out by themselves. However, NEA headquarters could underwrite the development of a suitable *RFP template*, a document that could be readily modified to mesh with state-level particulars in different states. Then, using a state-specific plan, state affiliates could lobby legislators and/or other state policymakers to support the installation of statewide tests that still yield district-by-district and school-by-school accountability evidence, but do so while upgrading rather than downgrading the quality of education in that state.

Pick and Choose

The foregoing nine action options surely do not exhaust what NEA and its state affiliates might do if they wish to counter today's misuse of high-stakes tests. However, it would seem that a reasonable strategy for improving our current high-stakes testing environment could be fashioned from the use of a combination of several of these activities. I believe, however, that unless Option 8 (other credible evidence) or Option 9 (better state-level tests) are part of that array of action options, little support will be secured from our citizenry. The public has a right to know how its schools are doing. Options 8 and 9 supply evidence to help satisfy this need, but do so in a manner that enhances instructional quality, not degrades it.

It is with respect to Option 8 and 9 that NEA will find certain members of the educational measurement community who are eager to assist in the reversal of a phenomenon that, in no small way, members of that community have allowed to prosper. Not all specialists in educational measurement, however, would be useful colleagues in such an undertaking. Some of today's psychometricians are, candidly, downright elated with today's high-stakes testing world.

Going It Alone or With Allies

There is, in our land, a growing recognition among educators that something is terribly wrong in the way we are allowing our instructional activities to be influenced by students' performances on high-stakes tests. The Association's leaders and members surely must realize that this situation, if not tackled head-on, is only likely to worsen.

But in the leadership of other national organizations, similar recognitions are taking hold. If NEA decides to become a serious player in a major effort to halt the type of educational mismeasurement described herein, there would seem to be merit in collaborating with other professional organizations to create a powerful coalition that, because of its combined force, has a better chance of deterring the kind of folly arising from today's use of unsound high-stakes tests.

Something surely needs to be done. Who better to take the lead in dealing with this national educational problem than the National Education Association?

Figure 1. A sixth-grade reading vocabulary item based on a similar one from a nationally standardized achievement test.

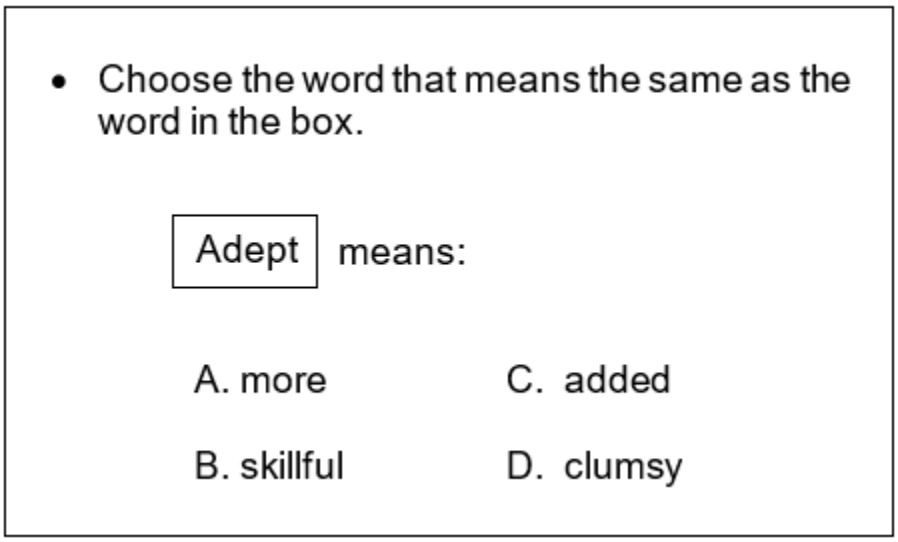


Figure 2. A sixth-grade science item based on a similar one from a nationally standardized achievement test.

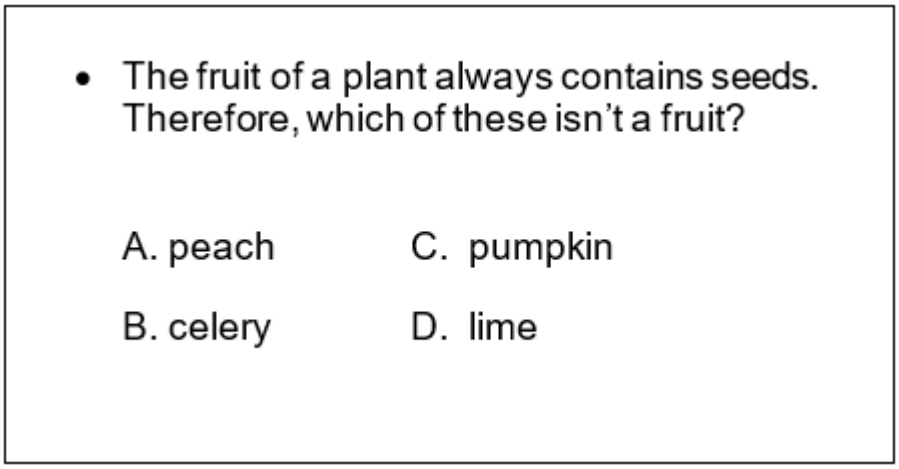


Figure 3. A fourth-grade mathematics item based on a similar one from a nationally standardized achievement test.