

DERELICTION DISCONTINUED: HOW AERA CAN HELP DETER TODAY'S MISUSE OF HIGH-STAKES TESTS*

W. James Popham
University of California, Los Angeles

Over 40 years ago, I taught my first course in educational research. At that time, I had just joined the faculty at San Francisco State College, and I was really enthused about teaching the educational research course. After all, the course dealt with what I wanted to be during my career, that is, *an educational researcher*. I suspect it is an academic truism that the impact of first-taught courses is almost always profound. Just as most people readily recall their first love, the consequences of a first-taught course can be lasting.

The textbook I selected for my educational research course had been authored by Robert M. W. Travers. At that time, I had not yet met Bob (who, I later discovered, was a gentleman and a scholar). Nor, for that matter, had I met anyone who employed two middle initials. But I thought his rather slender textbook about educational research was wonderful. In that book he succinctly set forth what he believed educational research was all about. As he put it, “educational research seeks to determine the nature of relationships among educationally relevant variables.” There it was, nicely spelled out—what I hoped to be doing for the rest of my life. And it sounded so scientific, so scholarly, so sublime. I was going to be a nonpartisan truth-seeker. And the truths I sought would improve the quality of education that children received. What a wonderful career-choice I had made!

An Entire Association of Nonpartisan Truth-Seekers

I attended my first annual meeting of the American Educational Research Association (AERA) in 1958, and haven't missed one since. I loved AERA; still do. For me, the Association offered all sorts of opportunities to learn from folks, far wiser than I, about how to be an educational researcher. The thrust of AERA's activities, without question, fit happily into Travers' definition of what educational researchers ought to be about. The Association's members were diligently digging up data regarding relationships among educationally relevant variables.

And when, during the 1970s, I became active in the Association's Council, first as a divisional vice president and, later, as AERA president, I continued to see how—for the most part—the leaders of the Association strictly adhered to the precept that we ought to be nonpartisan truth-seekers. If someone introduced a proposal to the Association Council that smacked of genuine action, political or otherwise, that proposal was almost instantly knocked down because such proposals were viewed as being inconsistent with AERA's *apolitical* scientific mission.

* Presented at a symposium, “Three Blueprints for a Revolution: How to Halt the Harm Caused by High-Stakes Tests,” during the annual meeting of the American Educational Research Association, Seattle, Washington, April 10-14, 2001.

Lest I imply that I was in there pleading for AERA to take a more active stance regarding real-world educational concerns, this wasn't the case. I had totally bought into the idea that AERA, a plucky band of nonpartisan truth-seekers, should be "above it all." We were researchers, not reformers. I thought, as did most of my Association colleagues, that AERA should not become involved in the gritty, day-to-day world of American schooling. Rather, it was our mission to ferret out research-based pearls, then cast those pearls in front of educational practitioners who, we assumed, would initiate pearl-triggered activities of one sort or another.

Rarely, in those days, did the Association take policy positions or initiate action-based projects that, if relevant to real-world concerns, dealt with anything other than a fairly bland variety of mother-pie and apple-hood causes. AERA, in the seventies—and later, played it safe. AERA was a nonprofit assembly of educational *researchers*. Bob Travers would have been delighted.

An Altered World Spurs Altered Missions

Well, that was then and this is now. If the early mission of educational research was to illuminate the nature of relationships among educationally relevant variables, the implicit rationale for such inquiry was always, in the long run, *to improve educational practice*. If educational researchers really didn't expect their inquiries to enhance educational practice, then why muck around with such inquiries in the first place? The assumption underlying all our noble truth-seeking was that some real-world educators would do a better job with their students by relying on the research-derived insights we dished up.

But if real-world educators were supposed to do a better job because of the activities of AERA's members, what about the flip side? What about educators' doing a *worse* job because of the activities of educational researchers? In that case, is it still appropriate for AERA members to sit on their nonpartisan hands? Is nonpartisan hand-sitting acceptable when we witness a serious erosion in the caliber of our nation's education? I think not.

And that's where we are right now. One of the most prominent data-gathering tools employed by educational researchers, namely, *educational testing*, is being dramatically misused in the real world of American schooling. These misuses are so profound that large numbers of America's children are being educationally harmed—perhaps irreversibly.

I believe, therefore, that AERA must take immediate *action* to halt the harm stemming from the misuse of today's "high-stakes" tests. We should do so not only because educational tests have been a mainstay through the years for so much of our data-gathering, but also because—as a group—our Association possesses the largest collection of professionals who are knowledgeable about educational testing.

I know, of course, that AERA has, through the years, become far more diverse than it originally was. We now have many members who have never in their research efforts ever used a test, even once, as a data-gathering device. Historical researchers might choose to study E. L. Thorndike's measurement contributions and even the actual test items he wrote, but educational historians need not collect data by using tests. And also the Association now contains a number

of first-rate qualitative researchers who can surely snare educational insights without relying on educational tests.

But all that notwithstanding, AERA has heavily relied on tests throughout its entire existence. We have, because of that usage, given educational tests at least our implied imprimatur as suitable measurement tools. Moreover, as noted above, AERA does, in fact, have among its members most of the people today who really understand how educational tests should be used—and how such tests shouldn't.

The time has come for AERA to undertake a serious effort to deter the increasingly prevalent misuse of educational measuring instruments. Serious harm to an entire generation of children hinges on our actions.

An Important Policy Stance

I was delighted to see the AERA position statement concerning high-stakes testing published in the November 2000 *Educational Researcher*. The thoughtful responses of five individuals to the position statement, also published in the same issue, suggests that the AERA statement may, indeed, have a catalytic impact on the dialogue regarding high-stakes testing. The position statement, incidentally, nicely defines a “high-stakes” test as one that carries serious consequences for students or for educators.

The Association's leaders should be lauded for coming forth with such a statement. It is surely an excellent contribution to the discussion of appropriate and inappropriate high-stakes testing programs. But, as will be made clear later in this analysis, I believe AERA needs to do more.

Educational Mischief in a Nutshell

This is not the forum in which to undertake a detailed analysis of the nature or the magnitude of educational harm now being caused by the misuse of high-stakes educational tests. However, a quick summary of the major kinds of problems triggered by such tests will, hopefully, make it clear that serious corrective action is warranted. Here then, briefly, are three sorts of educational difficulties stemming from the misuse of high-stakes tests.

Curricular reductionism. Some teachers, straining mightily to increase their students' test scores, have abandoned curricular content that—only a few years earlier—they had regarded as absolutely indispensable. This shift in teachers' curricular priorities is the direct consequence of the perceived need to boost students' scores on high-stakes tests. Because teachers need more time to address the knowledge and skills included on a locally used high-stakes test, the easiest way to secure the needed test-preparation time is to dump any curricular content not covered on that test. Unfortunately, this kind of tunnel-visioned preoccupation with tested content fails to allow students to learn many important things that, had it not been for the stultifying impact of high-stakes testing, they would surely have learned. Many broad and rich curricula have become narrow and impoverished due to the score-boosting pressures induced by unsound high-stakes testing programs.

From delight to drudgery. Not all children derive joy from schooling. But many surely do. Yet, in classrooms where the pressure to excel on high-stakes tests has taken over, a relentless regimen of drill often extinguishes any positive attitudes children may possess toward education. Test-pressured teachers sometimes literally shut down any form of instruction other than a joy-numbing marathon of practice sessions focused exclusively on the knowledge and skills measured by a designated high-stakes test. After weeks, *literally weeks*, spent plowing the same practice fields, is it any wonder that students think their teachers have abandoned education in favor of test-drills? Is it any wonder that students, who otherwise might actually enjoy school, learn to hate it?

Models of malfeasance. And, finally, it is difficult to read any issue of *Education Week* these days without encountering one or more reports regarding teachers or administrators who have been apprehended in the act of violating test-related security regulations. Teachers have been caught giving their students practice sessions incorporating actual items copied from an operational form of a high-stakes test. Administrators have been caught in the midst of massive erasure campaigns to replace students' wrong answers with right ones. And most of these episodes of educator-cheating have soon been learned about by students. If it is true that teachers are supposed to function *in loco parentis*, then many of today's students are learning that their surrogate parents are cheaters.

Although other harmful consequences from high-stakes testing programs could be described at length, the foregoing examples should suffice to demonstrate that, in many settings, the nation's children are suffering serious educational harm as a result of high-stakes testing programs.

Two Turnabout Tasks

How can the current misuse of high-stakes tests be stopped, or at least significantly reduced? I don't intend to oversimplify my answer, but I believe there are two tasks that, if accomplished, could play a major role in helping turn this situation around. Indeed, I think the accomplishment of each task represents a genuine *sine qua non* for the reduction in the misuse of high-stakes tests. Here, then, are these two, all-important tasks:

Task One: All relevant constituencies must be made to understand why it is that traditionally constructed standardized achievement tests do not provide accurate indications of educational quality.

Task Two: Alternative evidence regarding the effectiveness of education must be provided. Such evidence should contribute to valid inferences about educational quality, yet do so in a way that improves rather than impairs teachers' instructional efforts.

A Closer Look at Task One

Target audiences. The first task, of course, requires that a requisite dose of assessment literacy be swallowed by the folks who have a substantial stake in seeing whether our schools are working well or not. I think of four groups as the targets for this task: (1) educators themselves, (2) parents of school-age children, (3) educational policymakers such as legislators or members of school boards, and (4) citizens in general. By the time that the fourth group (citizens in general) gets tossed into the mix, you'll recognize I actually have most Americans in mind, especially those mature enough to deal with the issue of educational mismeasurement.

And I'm not too sure it wouldn't be wise even to include *students* in the to-be-educated-about-testing pool. Students themselves are clearly impacted by the proper or improper use of educational tests. Moreover, today's students will be tomorrow's adults—adults whose understanding of educational tests can be terribly important.

When might such assessment literacy be introduced for children? Well, when I was growing up as a Catholic, there was much attention paid to “the age of reason.” That was the age at which a young person could distinguish between moral right and moral wrong. If I recall correctly, “the age of reason” arrived when kids were about seven or eight years old. Because assessment literacy, based on our recent experiences, seems to be at least as complicated as the avoidance of sin, I'd probably not introduce the topic of educational assessment until children were in late middle school, but I would definitely do so by the time kids were in their teens. Children have a right to understand the nature of the measurement process that may mess up their lives.

Trouble in River City. All of these target constituencies must understand that three major problems with traditionally constructed standardized achievement tests preclude their use as legitimate indicators of educational effectiveness. Let's look briefly at each of these three reasons.*

First, there is the problem of teaching-testing mismatches. Both formal and informal judgmental analyses demonstrate convincingly that there are typically serious mismatches between (1) what a traditionally constructed test attempts to measure and (2) what a given group of teachers are supposed to be teaching. And these mismatches often exist even when a state-customized test has been specifically constructed to supposedly assess a state's curricular preferences more accurately.

Second, because traditionally constructed standardized achievement tests must create sufficient score-variance, there is a psychometrically imposed tendency in such tests to avoid the inclusion of items covering important, instructionally emphasized curricular content. That's because items answered correctly by too many students will fail to make adequate contributions to the production of the score-variance required by a traditionally constructed standardized achievement test. And such items, of course, tend to be excised whenever such tests are revised.

* A more detailed analysis of these three deficits is available, e.g., see Popham, W. James, “Why Standardized Tests Don't Measure Educational Quality,” *Educational Leadership*, March 1999, 56, 6:8-15.

A final problem is also attributable to the quest of the creators of traditionally oriented tests for sufficient score-variance. There are many items on traditionally constructed standardized achievement tests whose correct answers are substantially dependent on (1) the socioeconomic status, i.e., SES, of a student's family or (2) a student's inherited academic aptitudes, namely, a child's in-born verbal, quantitative, or spatial aptitudes. Such items, of course, lead to a serious interpretational problem. It will always be unclear what the proportion is of students' test performance that's due to (1) what students were taught in school versus (2) what students brought to school in the form of SES and inherited academic aptitudes.

For all three of these reasons, traditionally constructed standardized achievement tests ought not be used to evaluate the caliber of education. All constituencies who are concerned about the quality of schooling must learn about the seriousness of these three reasons. Ideally, members of those constituencies should understand these problems at a depth sufficient to inspire corrective actions.

A Closer Look at Task Two

If traditionally constructed standardized achievement tests don't supply evidence regarding how well our schools are working, then more appropriate evidence of educational quality must be provided. Educational accountability is a good thing. Educational accountability is premised on the notion that the most important entities in the educational game are *children*. And to make sure that children are receiving truly first-rate educations, policymakers have installed accountability programs to supply the evidence needed to say how effective a given educational program is. Properly conceptualized and implemented accountability programs can help make sure that children are being well-taught.

This second task deals exclusively with the *evidence* needed to make accountability programs work properly. If members of key constituencies have learned that traditionally constructed achievement tests do not produce an accurate picture of educational quality, then more defensible indicators of educational success must be made available. If no such alternative forms of evidence are available, then it is easy to see how most people would readily choose traditionally constructed standardized achievement tests to supply accountability evidence. "Some evidence," they would reasonably contend, "is better than no evidence at all." So, Task Two calls for the installation of assessments that can provide evidence leading to accurate judgments about educational quality. Moreover, those assessments must contribute to teachers' instructional decisions rather than distort those decisions.

I have two levels of assessment in mind when I think about this second task. First, we have learned during the past two decades about how to create *large-scale assessments* that supply solid evidence of educational quality, yet do so in a way that helps, not hinders classroom instructional activities. We need, in short, large-scale tests that are more suitable for the evaluation of educational quality.

Second, teachers can be shown how to collect their own *classroom evidence* of instructional effectiveness in a way that yields accurate inferences regarding the success of the teacher's instruction. Teachers will need support to learn how to (1) build suitable classroom

assessments for such evidence-gathering, (2) score students' responses in a credibility-inducing manner such as by having parents blind-score sets of mixed-together pretests and posttests, and (3) report this evidence in an accurate and convincing fashion to those who will be interested in it.

Neither of the two tasks I have described here, by itself, will do the job. But both tasks, if accomplished in tandem, could. And that's what I turn to now, how AERA could help accomplish both of these tasks.

A Task Force With a Mission

There are *task forces* and there are *task farces*. I have served on both kinds. The task force I have in mind, however, is one that would have a meaningful mission plopped squarely on its plate. The mission of this brand new AERA task force would be to bring about changes in what's happening in our nation's schools. This would not be a task force whose charge was to provide a report to the Association Council about what might be done to make things better. This task force's mission would be to *do those things*. Clearly, all significant activities of the task force would need to be cleared, and financially supported, by appropriate Association authorities. But the charge of the task force would be to "get on with it."

The mission of the proposed task force would *not* be to oppose the use of traditionally constructed standardized achievement tests. Such tests can provide both parents and teachers with useful information regarding children. Rather, the task force would try to see that such tests were not used improperly, that is, for making judgments about the quality of schools or teachers.

Action-Option Possibilities

If an AERA task force were created to tackle this important problem, its members would surely explore action-options in more depth than I will here. However, let me toss out a few possibilities to suggest the sorts of activities that such a mission-focused group might consider.

With respect to Task One, *the educative task*, a variety of targeted informational materials might be written, then distributed by AERA to appropriate clientele (e.g., parents or legislators). A speakers' bureau could be created so that educators and/or policymakers could easily identify individuals, perhaps in their geographic areas, who could address the substance of both tasks one and two. The task force could provide guidelines regarding how to conduct item-by-item judgmental analyses of high-stakes tests (either those currently in use or those under consideration) to see whether a particular test's items were sufficiently free from factors that would diminish the validity of inferences about educational quality. Indeed, the task force might conduct one or two such judgmental-review studies itself to provide models for other investigators. All task force activities worthy of dissemination could be made available via the AERA web site.

With respect to Task Two, *the evidence-production task*, the task force could create a *request for proposal (RFP) template* that would be made available to appropriate individuals and/or groups in each state. This RFP template would be a carefully devised document, readily

revisable at the state or district level, calling for the creation of high-stakes tests that could provide not only accurate accountability information about school-level success, but also beneficial instructional guidance for teachers. If a state's officials chose to use the RFP template, they could do so without charge, and possibly, even with some guidance from task force members about how to customize the RFP. At the very least, the task force could devise and distribute suggestions about how to modify the model RFP locally.

I assume that the respondents to a state's RFP would be the usual test-development firms who build high-stakes tests these days. Such firms are in the test-building and money-making business. So, if the financial incentives for bidding on the RFP-stipulated tasks were sufficient, there would be an adequate response to any RFP. But the constraints embedded in these new RFPs would preclude test-building as usual. A traditional approach to the construction of achievement tests would not satisfy the RFP's requirements. Instead, genuinely different sorts of *accountability-focused but instructionally-facilitative* tests would need to be built. They could be built if AERA helps state and district officials learn how to demand that such tests be built.

Finally, for the collection of the kind of classroom-level assessment data foreseen in Task Two, there would need to be ample explanatory materials, and even models, provided to the nation's teachers. These materials would offer step-by-step guidance with respect to how a teacher could collect, analyze, and present *credible* data regarding the teacher's instructional success. Just as was true with the RFP-induced collection of evidence via large-scale tests, the tests to be used for classroom-level evidence must stimulate improved instructional decision-making as well as provide truly believable evidence of instructional quality.*

A Modest But Meaningful Policy Shift

At the outset of this analysis, I indicated that AERA has traditionally been composed of a collection of professionals bent on discovering how educationally relevant variables interact. I really have no desire to change that. Yet, when our association is positioned to make a genuine contribution to the quality of what's going on in the real-world of schooling, or to help stop clearly unsound educational practices in that real world, then AERA should take *action*. I urge the Association's leadership to do so.

* Both of these two action options are treated in greater detail elsewhere, e.g., Popham, W. James, "Educational Assessment: High Quality Testing for a High Stakes World," *Association for Supervision and Curriculum Development*, scheduled for 2001 publication. I assume that task force members would collect analyses such as these, then decide which action-options were most fruitful to pursue.