The Seductive Allure of Data

Most state accountability tests fail to produce the kinds of data that will improve teaching and learning. Teachers can get the data they need from classroom assessments—if they know how to design instructionally useful tests.

W. James Popham

The word *data*, at least to most educators, simply reeks of goodness. Although probably less heart-warming than *children, smaller classes,* and *summer vacation,* the term *data* inclines most educators to think good thoughts laced with notions of evidence, science, and rigor. Indeed, the theme of this issue of *Educational Leadership* reflects educators' belief that data play a central role in improving student achievement. In any education lexicon these days, the term *data* is inarguably one of our most positively loaded nouns.

Data Scorned?

But *data* shouldn't elicit automatic obeisance from right-thinking educators. Indeed, we should spurn some data. In the following analysis, I intend to dismiss certain sorts of data. I want educators to realize that the wrong kinds of data, even if warmly applauded by many, can actually stifle teachers' pursuit of accurate evidence regarding their students' achievement.

Currently, teachers are buffeted by messages that the often undecipherable test results they receive are, in fact, the data they need to make instructional decisions. Is it any wonder when, after trying in vain to make sense of such opaque test data, many teachers simply quit believing in the instructional utility of data? To avoid becoming disillusioned with all data, teachers must learn how to distinguish between instructionally delightful and instructionally dismal data.

At the Top of the Heap: Test Data

Although all sorts of data might help to improve instruction, the most important data in the United States these days are *test* data—particularly data describing students' performance on achievement tests. That's because schools increasingly employ those data to evaluate educators' effectiveness.

State-determined achievement tests increasingly serve as the centerpieces of state accountability systems. But data from most states' accountability tests, unfortunately, have almost no value for improving teaching and learning. More dangerously, such tests lull educators into believing that they have appropriate data when, in fact, they do not. As a consequence, many educators fail to ask for more meaningful, instructionally valuable data that would help them teach students better.

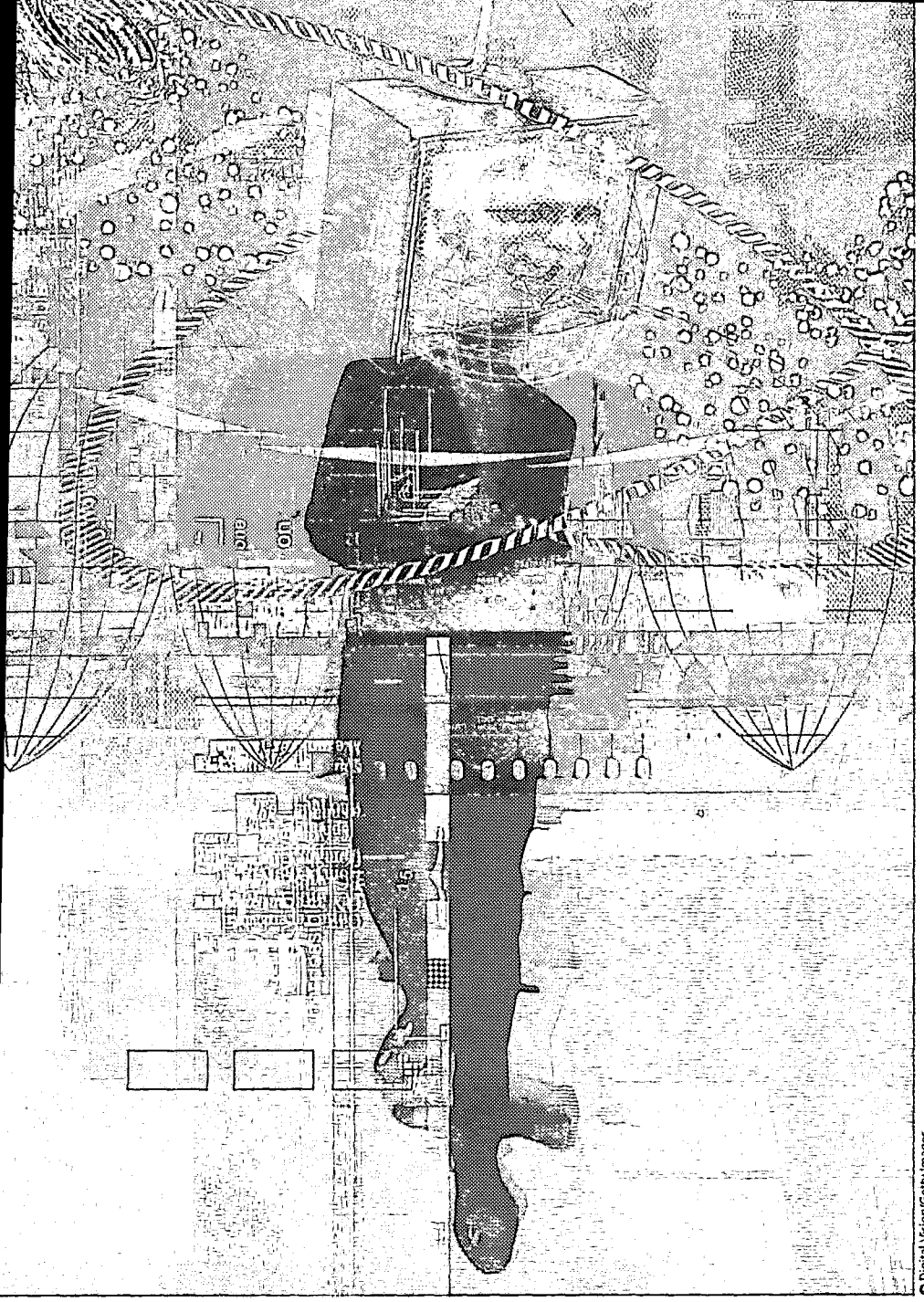
Instructionally Beneficial Data

Instructionally beneficial data can only come from instructionally useful tests. Here are five attributes of an instructionally useful test, which apply to large-scale assessments as well as to teacher-made classroom assessments.

Significance. An instructionally useful test measures students' attainment of a worthwhile curricular aim—for instance, a high-level cognitive skill or a substantial body of important knowledge. It makes no sense to assess students' mastery of such trifling knowledge as esoteric scientific terms or dates associated with obscure historical events. (I suppose that someone might come up with a cogent argument for asking students to memorize state capitals. I've never been able to.)

Teachability. An instructionally useful test measures something teachable. Teachability means that most teachers, if they deliver reasonably effective instruction aimed at the test's assessment targets, can get most of their students to master what the test measures. For instance, an instructionally useful test should not measure students' innate intelligence. In standardized achievement tests, we frequently encounter items requiring students to engage in such spatial visualization tasks as mentally "folding" letters or geometric shapes into two equal halves. Such tasks clearly depend on a student's inherited visualization aptitude.

Similarly, certain high-level inference skills are extraordinarily difficult to teach because the cognitive processes central to those skills usually depend on the idiosyncratic nature of a particular student's prior experiences. It simply



© Digital Vision/Getty Images

makes no sense to assess students' mastery of essentially unteachable outcomes.

Describability. A useful test provides or is directly based on sufficiently clear descriptions of the skills and knowledge it measures so that teachers can design properly focused instructional activities. These descriptions must not only be provided in plain language, but must also be sufficiently succinct so that they are not off-putting to busy teachers.

If a test is based on an already clearly described set of content standards, and if teachers can tell which of those content standards the test will cover, then no further descriptive information is needed. But if the content standards are not clear enough to unambiguously let teachers identify those curricular targets, then lucid descriptions of what the test will assess must accompany any instructionally useful test. A content standard such as "Students will read a variety of different types of texts" communicates little of instructional value to the teacher.

It makes no sense to assess students' mastery of ill-defined curricular targets or to force teachers to play an annual guessing game about which of the state's content standards the statewide accountability tests will assess.

Reportability. An instructionally useful test yields results at a specific enough level to inform teachers about the effectiveness of the instruction they provide. A national commission has urged that any education accountability test report its results on a standard-by-standard basis for individual students (Commission on Instructionally Supportive Assessment, 2001). Such per-standard reporting of results would enable teachers to identify those parts of their instruction that were successful or unsuccessful on the basis of students' post-instruction test data.

It makes no sense to provide teachers with data so general that those teachers cannot evaluate and improve their own instructional efforts. Similarly, it makes no sense for assessors to contend that they have assessed the complete array of a state's content standards when, in fact, they have measured some standards either by only a handful of items or by no items at all.

Nonintrusiveness. In clear recognition that testing time takes away from teaching time, an instructionally useful test shouldn't take too long to administer—it should not intrude excessively on instructional activities. For instance, if a state-level test of students' reading skills is administered each spring, it should be administrable in one, or at most two, class periods. Longer tests simply soak up too much instructional time. It makes no sense to test students interminably, diverting several weeks of precious instructional time each year to assessment.

In review, we are most likely to obtain instructionally useful data through the use of instructionally useful tests. The five attributes of an instructionally useful test are its significance, teachability, describability, reportability, and nonintrusiveness. The data derived from an instructionally useful test will enable teachers to do a better job of instructing their students. And that, after all, should be the reason we test students in the first place.

Detecting Dismal Data

As suggested earlier, tests that don't produce instructionally useful data can disincline educators to demand data

that are instructionally beneficial. In the following three common assessment situations, the wrong kinds of data—provided by the wrong kinds of tests—have diminished the quality of education that we provide to our students.

Nationally Standardized Achievement Tests

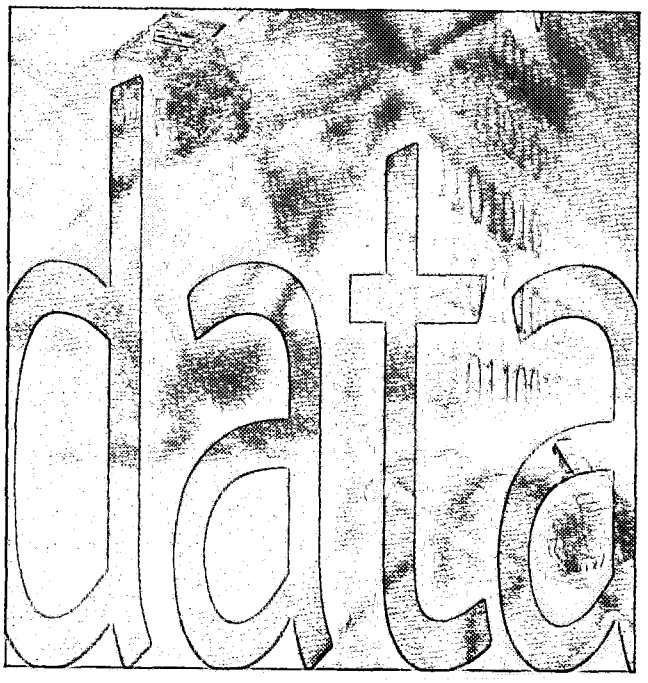
Today's nationally standardized tests miss the mark dramatically with respect to three of the attributes of instructionally useful assessment:

- *Describability.* All nationally standardized achievement tests have been constructed according to a traditional measurement approach aimed at providing a comparative picture of students' relative performances. The developers try to devise a "one-size-fits-all" test and describe it in a manner that will make it attractive to many potential purchasers. As a result, nationally standardized tests don't include properly tied-down descriptions of what they assess. Teachers can't aim their instruction accurately if they have murky assessment targets.
- *Teachability.* In order to produce the score spread on which comparative score interpretations depend, nationally standardized tests contain many instructionally insensitive items that are linked to students' socioeconomic status or inherited academic aptitudes. It is particularly difficult for teachers to increase students' performance on such items.
- *Reportability.* Nationally standardized achievement tests almost always report their results at levels of generality altogether unsuitable for teachers' day-to-day instructional decision making. Some national tests do a better job than others when it comes to reporting students' results. But in no case do these tests provide data that teachers can easily use to appraise their own instructional effectiveness.

I believe that nationally standardized achievement tests have a role in education. Both parents and teachers can benefit from data indicating a student's relative strengths and weaknesses. But the genuine instructional yield of nationally standardized tests is much more modest than the publishers of these tests would have us believe.

Standards-Based Tests

There is a charade currently going on in the way the United States carries out its education assessment activities. Its name is "standards-based assessment." Standards-based tests supposedly measure students' mastery of a state's officially approved content standards—the skills and knowledge constituting the state's curricular aims. Yet because most states have adopted too many content standards and stated them too vaguely, most states' standards-based tests just don't do a decent job determining a student's mastery of those standards.



Pretending that a one- or two-hour state test can provide a meaningful fix on a student's mastery of myriad, often fuzzy content standards is patently hypocritical. Today's standards-based assessments constitute a serious violation of any sort of truth-in-advertising precept. Standards-based tests don't measure what they pretend to measure.

The data yielded by today's standards-based tests have another equally serious shortcoming. Those data almost never provide any indication of *which* content standards a student has or hasn't mastered. In the absence of such data, how can teachers tell which parts of their instruction they need to modify?

Teachers don't learn much of instructional value when the standards-based test results tell them that Johnny is "not proficient" with respect to his mastery of a set of 17 language arts content standards. Teachers cannot discern which of the 17 content standards their students have mastered (hence, which standards have been well taught) and which of the 17 content standards their students have not mastered (hence, which standards have not been well taught).

So most of today's standards-based tests fall down seriously on several attributes of an instructionally useful test. They often lack significance because, in a fruitless effort to measure all of a state's sprawling content standards, they simply do not assess students' mastery of the most important content. Standards-based tests also get low grades on describability—they usually fail to describe their assessment targets satisfactorily, because these tests are based on a plethora of too many, insufficiently clear content standards. And perhaps most seriously, standards-based tests often lack reportability—they fail to provide standard-by-standard reports to teachers, students, or students' parents.

Teachers' Classroom Assessments

Given the enormous pressure placed on teachers these days to boost their students' scores on external exams, teachers understandably tend to give less attention to their own classroom assessments. That's a mistake—but only

if the teacher's classroom tests are instructionally useful.

Teachers can judge the instructional utility of their classroom assessments by using the same five attributes of an instructionally useful test that I just applied to large-scale external exams. Teachers should ask themselves the following questions:

- Do my classroom assessments measure genuinely worthwhile skills and knowledge?
- Will I be able to promote my students' mastery of what's measured in my classroom assessments?
- Can I describe what skills and knowledge my classroom tests measure in language sufficiently clear for my own instructional planning?
- Do my classroom assessments yield results that allow me to tell which parts of my instruction were effective or ineffective?
- Do my classroom tests take up too much time away from my instruction?

Clearly, the answers to these questions will vary from teacher to teacher. Generally, teachers who employ their classroom assessments most appropriately adopt a "less is more" approach. They focus on measuring only a modest number of curricular aims, but make certain that those aims deal with genuinely significant outcomes that students can master with adequate instruction. As a dividend of focusing on a smaller number of significant outcomes, those outcomes can then be clearly described to help the teacher target instruction and assessment.

Teachers must deal with one additional consideration if they intend to use their classroom data to supplement results from external exams: Unless classroom tests provide credible data, skeptics will rush to dismiss the results as "self-interested home cooking." I'm not talking about tests that teachers use only to inform themselves about their ongoing instruction, but rather about the more significant sorts of data that schools use to judge a teacher's instructional effectiveness.

One straightforward way for teachers to collect credible evidence of their own effectiveness is to use a pretest/posttest design in which they give identical assessments at the start of a semester and again at its conclusion. Students must use the same kind of paper if the test calls for a constructed response (such as writing an essay). Students do not date their responses. The teacher codes the pretests and posttests so they can be subsequently identified, and then mixes them all together so that a scorer cannot discern which responses are pretests and which are posttests.

At this point the teacher calls on a nonpartisan scorer (for instance, another teacher or a parent) to blind-score the students' responses. Only after all the shuffled papers have been scored does the teacher sort them into pretests and posttests. The improvement between the pretests and posttests constitutes credible evidence of the teacher's instructional success (Popham, 2001).

What Can Educators Do?

In response to today's increasingly important assessment concerns, I suggest a two-stage course of action. First, educators should disregard data from any test that isn't instructionally useful. Second, they should push for the installation of instructionally useful tests so that the data that those assessments yield will lead to better-taught students.

Although most of today's standards-based tests are not instructionally useful, that need not be the case. A national commission recently described how to create accountability tests that are both accurate and instructionally useful (Commission on Instructionally Supportive Assessment, 2001). Many states assess students' written composition competence by requiring students to generate original writing samples, which are then evaluated according to scoring guides (rubrics) based on teachable criteria. Almost all of today's writing samples are instructionally beneficial. If you live in a state where such instructionally useful tests do not exist, lobby aggressively for their introduction.

If you live in a state that uses nationally standardized achievement tests for accountability purposes, try your hardest to get them replaced with more appropriate, instructionally useful accountability tests.

Data from most states' accountability tests, unfortunately, have almost no value for improving teaching and learning.

Teachers should also bring common sense to the scrutiny of their own classroom assessments. In general, a quest for assessment sanity will lead teachers to adopt a less-is-more measurement approach. However, if the resultant data will be used for instructional evaluation, then teachers must collect those data in a manner sufficiently credible to persuade even non-believers of the data's validity.

To educators, the wrong data can often be seductively appealing. But the right data will, in fact, help teachers do a better job with students. *Those* are the data we need. ■

References

Commission on Instructionally Supportive Assessment. (2001). *Building tests that support instruction and accountability: A guide for policymakers.* Washington, DC: Author. [Online]. Available: www.nea.org/accountability/buildingtests.html

Popham, W. J. (2001). *The truth about testing: An educator's call to action.* Alexandria, VA: ASCD.

W. James Popham is a professor emeritus at the University of California, Los Angeles, Graduate School of Education and Information Studies; wpopham@ucla.edu.