

DEALING WITH THE SEDUCTIVE ALLURE OF DATA

W. James Popham
University of California, Los Angeles

Data, at least to most educators, is a word that simply reeks of goodness. Although the term is probably less heart-warming than "children," "smaller-classes," and "summer vacation," *data* inclines most educators to think good thoughts laced with notions of evidence, science, and rigor. Indeed, the current issue of *Educational Leadership* can surely be considered a paean to the role of data in improving student achievement. In any educational lexicon these days, the term data is inarguably one of our most positively loaded nouns.

Data Scored?

But *data* shouldn't elicit automatic obeisance from right-thinking educators. Indeed, some data should be spurned. And that's precisely what I intend to do in the following analysis. I intend to dismiss data—at least 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 truly need for making 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?

One way to combat this growing disregard for data, is to help teachers learn how to distinguish between instructionally delightful and instructionally dismal data. I'm hoping this brief essay will help in that regard.

Of Singulars and Plurals

Before turning to an analysis of the problem at hand, however, I need to expend a few sentences in an effort to justify the five years of Latin classes that I completed while in high school and college.

"Data," even though these days it is commonly employed as a singular noun, is actually a plural Latin noun. Its singular form is "datum." Thus, to be grammatically correct, one should say, "The data are positive." rather than "The data is positive."

Although I have, over time, become inured to the common misuse of "data" as a singular noun, I cannot bring myself to personally do so—especially in

* A version of this essay appeared in the February 2003 issue of *Educational Leadership*, Vol. 60, No. 5, pp. 48-51.

print. Were I to commit such a syntactical sin, my now deceased Latin teacher (Professor Henry Geuss) would surely do some serious turning over in his grave. So, I beg your indulgence when I use data as a plural noun in the remainder of this analysis. I just can't break the habit!

At the Top of the Heap—*Test Data*

Although there are all sorts of data out there that might be helpful to American educators, the overridingly most important data in America these days are *test data*. Of particular importance, of course, are the data describing students' performances on achievement tests. That's because those data are being increasingly employed to evaluate educators' effectiveness.

State-determined achievement tests, such as those called for in the *No Child Left Behind* act signed by President Bush last January, are increasingly serving as the centerpieces of state accountability systems intended to distinguish between successful and unsuccessful schools.

But the data (test results) provided by most states' accountability tests are, unfortunately, of little value instructionally—despite the fact that such data are constantly touted as being instructionally serviceable to teachers.

The simple truth is that the data provided by most state-level accountability tests are of almost no value to educators. More dangerously, however, such tests lull educators into believing that they are being given appropriate data when, in fact, they are not. As a consequence, many educators fail to pass for more meaningful, instructionally valuable data. Most seriously, in the absence of instructional valuable data, the nation's students typically are less well taught than they should be.

Instructionally Beneficial Data

Instructionally beneficial data, in a nutshell, are only provided by *instructionally useful tests*. And that's what I now want to describe, namely, five attributes of an instructionally useful test. These five attributes, incidentally, apply both to large-scale assessments as well as to teacher-made classroom assessments.

Significance. The first attribute of an instructionally useful test is that it must measure students' attainment of a *worthwhile* curricular aim, for instance, a high-level cognitive skill or a substantial body of important knowledge. There is no sense in assessing students' mastery of trifling outcomes.

Teachability. A second attribute of an instructionally useful test is that the test measures something that is truly *teachable*. Teachability, in this sense, 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 is measured. For instance, an instructionally useful test should not measure students' innate, patently unteachable intelligence. Similarly, certain high-level inference skills are extraordinarily difficult to teach because the cognitive processes that are central to those skills are usually dependent on the idiosyncratic nature of a particular student's prior experiences. There is simply no sense in assessing students' mastery of essentially unteachable outcomes.

Describability. A third attribute of an instructionally useful test is that it must provide, or be directly based on, sufficiently clear descriptions of the skills and/or knowledge being measured so that teachers can design properly focused instructional activities. Furthermore, the descriptions of what's being measured must not only be provided in plain language, but must be sufficiently succinct so that those descriptions will not be off-putting to busy teachers.

If a test is based on an already *clearly described* set of content standards, and if it is apparent *which* of those content standards are to be assessed, then there is no need for further descriptive information. If, however, the content standards are *not* clear enough to unambiguously let teachers know what those curricular targets actually are, then an instructionally useful test must be accompanied by lucid, teacher-palatable descriptions of what's going to be assessed.

There is no sense in assessing students' mastery of ill-defined curricular targets or in forcing a state's teachers to play an annual guessing game about which of the state's content standards will be assessed by a given year's statewide accountability tests.

Reportability. An instructionally useful test's fourth attribute is that its results are reportable at a level of specificity sufficient to inform teachers about the effectiveness of the instruction they provide. In October 2001 a national commission urged that any educational accountability test report its results *on a standard-by-standard basis for individual students* (Commission on Instructionally Supportive Assessment, 2001). If such per-standard reporting of results were provided, then teachers would be able to identify those parts of their instruction that—based on students' post-instruction test data—were successful or unsuccessful.

It makes no sense to provide teachers with data so general that those teachers cannot evaluate, hence improve, their own instructional efforts. Similarly, it makes no sense for assessors to contend that the complete array of a state's content standards have been assessed when, in fact, some standards

have been measured either by only a handful of items or, in many instances, by no items at all.

Nonintrusiveness. The fifth and final attribute of an instructionally useful test is that it shouldn't take too long to administer, that is, it should not be excessively intrusive on a teacher's instructional activities. In clear recognition that *testing time* takes away from *teaching time*, the intrusiveness of instructionally helpful tests should be kept to a minimum. Thus, for instance, if a state-level test of students' reading skills is to be administered each spring, such a test 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 so that, in a given year, several weeks of precious instructional time end up being diverted to assessment.

In review, then, *instructionally useful data* are most likely to be obtained via 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 kids in the first place.

Detecting Dismal Data

As suggested earlier, a test that's not instructionally useful can disincline educators to demand data that *are* instructionally beneficial. Let me briefly consider three assessment situations in which 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. All nationally standardized achievement tests have been constructed according to a traditional measurement approach that's aimed at providing a comparative picture of students' relative performances. As a consequence, nationally standardized achievement tests fall way short of what's needed for a test to be instructionally useful. One reason is that these tests must attempt to assess students' achievements in many settings whose curricular preferences do not coincide. What's emphasized curricularly in Michigan or Wyoming may be different from what's emphasized curricularly in Kansas or California. As a consequence, the developers of national achievement tests try to devise a "one-size-fits-all" test and to describe its assessment targets in a manner that will make a test attractive to many potential purchasers. As a result, the clarity with which these tests describe what they are assessing falls well short of what teachers actually

need for on-target classroom instructional planning. Other than some extremely general descriptions of what's being measured, nationally standardized tests are not accompanied by properly tied-down descriptions of what they assess. But, of course, teachers can't aim their instruction accurately if they only have murky assessment targets.

A second problem with nationally standardized tests is that, in order to produce the score-spread on which comparative score interpretations are so dependent, there are many items in those tests that turn out to be instructionally insensitive. Such items are those linked to a student's socioeconomic status or to a student's inherited academic aptitudes. It is particularly difficult for teachers to increase students' performances on such items. With respect to the attribute of *teachability*, then, nationally standardized achievement tests are especially deficient.

In addition, the results of nationally standardized achievement tests are almost always reported at levels of generality altogether unsuitable for teachers' day-to-day instructional decision-making. Some national tests, of course, are better than others when it comes to reporting students' results. But in no case are the data provided by these tests truly useful to teachers when appraising their own instructional effectiveness.

I believe that there is a definite role for nationally standardized achievement tests in education. Both parents and teachers can benefit from data indicating that a child is relatively strong in one subject, yet relatively weak in another. But the genuine *instructional* yield of nationally standardized tests is much more modest than the publishers of these tests would have us believe.

Summing up, then, based on the five attributes of an instructionally useful test, today's nationally standardized tests typically don't look too bad with respect to the *significance* of what they assess or their *nonintrusiveness* (especially for the "short form" of such tests). However, nationally standardized achievement tests miss the mark dramatically with respect to *teachability*, *describability*, and *reportability*. Nationally standardized achievement tests, in my view, are not instructionally useful. Thus, the data they provide will be of little utility to classroom teachers.

"Standards-based" tests. There is a charade currently going on in the way our nation carries out its educational assessment activities. And its name is "standards-based assessment." A standards-based test is supposed to measure students' mastery of a state's officially approved content standards, that is, the skills and knowledge constituting the state's curricular aims. Yet, because the content standards adopted by most states are too numerous, and are often

stated too vaguely, most states' standards-based tests just don't do a decent job in determining a student's mastery of those content standards.

It is patently hypocritical to pretend that a one-hour or two-hour state test can provide a meaningful fix on a student's mastery of myriad, often fuzzy content standards. 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.

There's another equally serious shortcoming in the data yielded by today's standards-based tests. Those data don't provide any indication of *which* content standards a student has/hasn't mastered. In the absence of such data, how are teachers to tell which parts of their instruction need to be modified? It is of little instructional value for teachers to learn from a standards-based test 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 have been mastered by their students (hence, have been well taught) and *which* of the 17 content standards have not been mastered (hence, 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 turn out to be weak regarding *significance* because, in a fruitless effort to measure *all* of a state's sprawling content standards, many of the most important content that students should master are simply not assessed. Standards-based tests also put low grades on *describability* because they usually fail to describe their assessment targets satisfactorily. That's because those tests are "based" on a plethora of too many, insufficiently clear content standards. A particular weakness in most of the nation's standards-based tests is that they fall down on *reportability*, that is, fail to provide standard-by-standard reports to teachers, students, or students' parents. Thus, with respect to *significance*, *describability*, and *reportability*, current standards-based tests strike out. These tests, contrary to the way they are promoted, are not instructionally useful. The data they provide, therefore, will be of limited utility to classroom teachers.

Teachers' classroom assessments. Let me turn, now, to the sorts of classroom tests that teachers create themselves. Given the enormous pressure that teachers are under these days to boost their students' scores on external exams, there is an understandable inclination by teachers to give less attention to their own classroom assessments. That would be a mistake—but *only* if a teacher's classroom tests are instructionally useful.

What I'm suggesting is that teachers judge the instructional utility of their classroom assessments by using the very same five attributes of an

instructionally useful test I just applied to large-scale external exams. Teachers need to ask themselves the following questions:

Significance: “Do my classroom assessments measure genuinely worthwhile skills and/or knowledge?”

Teachability: “Will I be able to promote my students’ mastery of what’s measured in my classroom assessments?”

Describability: “Can I describe what skills and/or knowledge my classroom tests measure in language sufficiently clear for my own instructional planning?”

Reportability: “Do my classroom assessments yield results that allow me to tell which parts of my instruction were effective or ineffective?”

Nonintrusiveness: “Are my classroom tests taking up too much time away from my instruction?”

Clearly, the answers to these questions will vary from teacher to teacher. Generally, I find that teachers who employ their classroom assessments most appropriately adopt a “less is more” approach wherein they focus on measuring only a modest number of curricular aims, but make certain that those aims deal with genuinely significant student outcomes. A dividend of focusing on a smaller number of significant outcomes is that those outcomes can then be well described, hence, better understood by the teacher.

There’s one additional consideration that teachers must deal with if they intend to use their classroom data to supplement results from external exams. I refer to the need for the data from classroom tests to be *credible*. There’s too much likelihood that skeptics will dismiss the results of teachers’ classroom testing as “self-interested home cooking.” I’m not talking about tests that teachers use only to inform themselves about their on-going instruction. Rather, I am referring to the more significant sorts of data that could be used to reflect 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 identical assessments are given at, say, the start of a semester and 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 are directed not to date their responses. The teacher, first, secretly codes the pretests and posttests so they can be subsequently identified, then mixes them all together so that it is

impossible for a scorer to discern which responses are pretests and which ones are posttests.

At this point the teacher calls on a *nonpartisan* (for instance, a parent) to *blind-score* the students' responses. Only *after* the mixed-together papers have all been scored does the teacher sort them into pretests and posttests. If the bulk of the high-scoring papers are posttests, then those data constitute credible evidence of the teacher's instructional success (Association for Supervision and Curriculum Development, 2001).

What's To Be Done?

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

Earlier, I concluded that most of today's standards-based tests are not instructionally useful. But that need not be the case. A national commission had recently described how to create accurate accountability tests that are simultaneously instructionally useful (Commission on Instructionally Supportive Assessment, 2001). If you live in a state where such tests do not exist, lobby aggressively for their introduction.

I also indicated that nationally standardized achievement tests are not instructionally useful. If you live in a state where such instructionally insensitive tests are used for accountability purposes, try your hardest to get them replaced with more appropriate accountability tests—tests that are instructionally useful.

Teachers also need to bring common sense to the scrutiny of their own classroom assessments. Although classroom assessment is a good thing, teachers need to remain somewhat sane while assessing their students. In general, a quest for assessment sanity will lead to teachers to adopt a less-is-more measurement approach. However, if the resultant data are to be used for purposes of instructional evaluation, then those data must be collected in a sufficiently credible manner that even non-believers will be persuaded of the data's meaningfulness.

I commenced this analysis with the promise that I would be scorning data. That's because, to educators, the wrong data can often be seductively appealing. But some data will, in fact, help teachers do a better job with kids. *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. Available online at www.aasa.org, www.naesp.org, www.principals.org, www.nea.org, www.nmsa.org.
- Popham, W. James. (2001). *The Truth About Testing: An Educator's Call to Action*. Alexandria, Virginia: Association for Supervision and Curriculum Development.