Educational Leadership

All About Accountability / Instructional Quality: Collecting Credible Evidence

W. James Popham
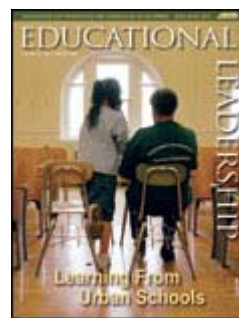
March 2005

After peeling away lots of lofty rhetoric regarding the societal role of education accountability programs, what remains is a simple truth: The mission of such programs is to see that schools and teachers provide excellent instruction to help kids learn better. That's a worthwhile aspiration.

Current accountability approaches, however, usually rely on large-scale standardized test scores as the sole indicator of teachers' effectiveness. Unfortunately, the scores on most of today's large-scale accountability exams reflect students' socioeconomic status much more accurately than they do teachers' instructional skills. No wonder many teachers yearn for more valid ways to tell whether they're doing a good job.

It's not that today's teachers merely wish to assemble evidence to show the world that they're effective—although that's surely a motive in some instances. Rather, most teachers recognize that because large-scale accountability tests don't provide a true picture of their classroom effectiveness, those tests don't give them the information they need to evaluate and, if necessary, alter their instructional strategies.

Yet ranting against tunnel-vision reliance on standardized test scores as the only criterion of teachers' effectiveness doesn't help much. What educators really need to do is come up with other, more useful and valid ways to provide evidence about their classroom competence.

One straightforward way for teachers to secure evidence of their instructional effectiveness is to collect defensible data showing whether students have made substantial progress in mastering significant cognitive skills. Using this method, teachers collect and contrast pre-instruction and post-instruction evidence of students' skill mastery. But the evidence must be collected in an atypical manner to ensure that students' performances can then be objectively scored by nonpartisan judges.

Let's look at how a fictitious middle school language arts teacher (I'll call her Carmen) might use this classroom evaluative design. One of the most important skills that Carmen's students must master is writing short persuasive essays. Carmen first identifies a suitable assessment task—for instance, having students write a letter to the mayor urging an expansion of public parks. She asks her students to write this kind of essay early in the school term and again at the term's conclusion (although she does not forewarn her students of the upcoming end-of-term essay-writing task).

Because Carmen has decided to measure her students' skill mastery using an on-demand task, she distributes a sheet of background information about the city's parks and asks students to write their essays—on paper that she has supplied—during a single class period. She directs students not to place the day's date on the essays. Although Carmen uses these pre-tests to gain instructional insights about her students' initial essay-writing skills, she does not make any marks

on the pre-tests.

Then, after a full term's worth of instruction dealing with this significant skill, Carmen gives her students the same task once more, supplying them with the same background information and type of paper. Again, students do not date their essays.

At that point, Carmen codes the essays so that only she can tell which ones were pre-tests and which ones were post-tests, and she mixes them all together. Next, she enlists the assistance of a small panel of nonpartisan scorers, perhaps parents or members of the business community, to judge the essays—without knowing when each essay had been written. Usually, these nonpartisan judges need a dollop of guidance regarding how to score such essays.

At the conclusion of this blind scoring, Carmen divides the coded essays into those written prior to and those written following instruction. If the blind-scored post-test essays are decisively better than the pre-test essays—and they should be, if Carmen's instruction was effective—then they provide credible evidence of Carmen's instructional prowess.

If, however, students' post-test essays weren't a whale of a lot better than their pre-test essays, then Carmen needs to take a hard look at what she did instructionally. Did her students need more guided or independent practice? Had the students really mastered the subskills necessary to write a powerful persuasive essay? Credible evidence of instructional effectiveness can, of course, reveal both spiffy and spotty teaching.

The most important consideration in the assembly of such evidence is to collect it in a consummately credible fashion. In devising a plan for obtaining such evidence, teachers should imagine that the data they collect must be able to withstand the skeptical scrutiny of an adroit attorney. If the evaluative evidence can stand up to such a careful review, odds are that it accurately reflects a teacher's instructional success.

More than ever these days, teachers need to buttress students' scores on large-scale accountability tests with evidence of their own classroom competence. I encourage a school's educators to collectively determine what kinds of credible evidence, other than NCLB test scores, can help provide an accurate and honest picture of their school's success. Then, without delay, they should collect and disseminate this evidence—no matter what it reveals.

W. James Popham is Emeritus Professor in the UCLA Graduate School of Education and Information Studies; wpopham@ucla.edu.