
Instructional Insensitivity of Tests: Accountability's Dire Drawback


If we plan to use tests for purposes of accountability, we need to know that they measure traits that can be influenced by instruction. Mr. Popham offers a model procedure for judging our tests.

By W. James Popham

LARGE-SCALE accountability tests have become increasingly important. They influence the deliberations of policy makers and affect the day-by-day behaviors of teachers in their classrooms. The premise underlying the use of these accountability tests is that students' test scores will indicate the quality of instruction those students have received. If students score well on accountability tests, we conclude that those students have been well taught. Conversely, if students score poorly on accountability tests, we believe that those students have been poorly taught.

Furthermore, advocates of these tests make two assumptions: 1) that teachers who realize they are going to be judged by their students' test scores will try to do a better instructional job and 2) that higher-level authorities can take action to bolster the quality of instruction in schools or districts where test results indicate ineffective instruction is taking place. For either of these assumptions to make sense, the accountability tests being employed must actually be able to determine the effect of instruction on students' test scores. However, all but a few of the accountability tests now

■ W. JAMES POPHAM is an emeritus professor at the University of California, Los Angeles, and founder of IOX Assessment Associates. This article is adapted from a presentation to the annual meeting of the American Educational Research Association, Chicago, April 2007.



having such a profound impact on our nation's schools are *instructionally insensitive*. That is, they are patently unsuitable for use in any sensible educational accountability program.

INSTRUCTIONAL SENSITIVITY

A test's *instructional sensitivity* represents the degree to which students' performances on that test accurately reflect the quality of the instruction that was provided specifically to promote students' mastery of whatever is being assessed. In other words, an *instructionally sensitive* test would be capable of distinguishing between strong and weak instruction by allowing us to validly conclude that a set of students' *high* test scores are meaningfully, but not exclusively, attributable to effective instruction. Similarly, such a test would allow us to accurately infer that a set of students' *low* test scores are meaningfully, but not exclusively, attributable to ineffective instruction. In contrast, an *instructionally insensitive* test would not allow us to distinguish accurately

ly between strong and weak instruction.

Students' performances on most of the accountability tests currently used are more heavily influenced by the students' socioeconomic status (SES) than by the quality of teachers' instructional efforts. That is, such instructionally insensitive accountability tests tend to measure the SES composition of a school's student body rather than the effectiveness with which the school's students have been taught.

Instructionally insensitive tests render untenable the assumptions underlying a test-based strategy for educational accountability. How can the prospect of annual accountability testing ever motivate educators to improve their instruction once they've realized that better instruction will not lead to higher test scores? How can officials accurately intervene to improve instruction on the basis of low test scores if those low scores really aren't a consequence of ineffective instruction?

There is ample evidence that, instead of improving instructional quality, ill-conceived accountability programs can seriously diminish it. Teachers too often engage in curricular reductionism and give scant, if any, instructional attention to content not assessed by accountability tests. Too often teachers impose excessive test-preparation drills on their students and thereby extinguish the genuine joy those students should experience as they learn. And too often we hear of teachers or administrators disingenuously portraying students' test scores as improved when, in fact, no actual improvement has taken place.

Yet, while the distinction between instructionally sensitive and insensitive accountability tests may be readily understandable and the classroom consequences of using instructionally insensitive accountability tests are all too apparent, it accomplishes little when educators complain, even profusely, about policy makers' reliance on the wrong kinds of accountability tests. Educators who simply carp about accountability tests are usually seen as individuals eager to escape evaluation. Only when we can convincingly *demonstrate* that an accountability program is relying on instructionally insensitive tests will we be able to remedy the current absurdity. Clearly, we need a credible procedure to determine the instructional sensitivity of a given accountability test.

This article describes the main features of a practical procedure for ascertaining the instructional sensitivity of any test, whether it is already in use or is under development. Because the instructional sensitivity of an accountability system's tests is the dominant determinant of whether that system helps or harms students, this approach should be used widely. Although the chief ingredients of the approach are described here, devils

hide in details, and thus a more detailed description of the procedures is available from wpopham@ucla.edu or at www.ioxassessment.com.

GATHERING EVIDENCE

There are two main categories of evidence for determining the instructional sensitivity of an accountability test: *judgmental evidence* and *empirical evidence*. Judgmental evidence can be collected by using panels of trained judges to rate specified attributes of a test. Empirical evidence can be provided by students' actual test scores, but these test scores must be collected under specific conditions — for instance, by comparing differences between the test scores of “taught” and “untaught” students.

Whether the instructional sensitivity of a test is determined by reliance on judgmental evidence alone, empirical evidence alone, or a combination of both, instructional sensitivity should be conceived of as a continuum rather than a dichotomy. Rarely will one encounter an accountability test that is *totally* sensitive or insensitive to instruction. The task facing anyone who wishes to determine an accountability test's instructional sensitivity is to arrive at a defensible estimate of where that test falls on such a continuum.

For practical reasons, the chief evidence to be routinely gathered about a test should be judgmental, not empirical. If resources permit, empirical studies should be used to confirm the extent to which judgmental data are accurate. But in today's busy world of education, the collection of even judgmental evidence regarding instructional sensitivity would be an improvement. The assembly of confirmatory empirical evidence is desirable but not absolutely necessary when embarking on an appraisal of an accountability test's instructional sensitivity. A number of key test-appraisal procedures currently rely only on judgment-based approaches, for instance, studies focused on content-related evidence of validity are based on judges' reviews of a test's items.

There is nothing sacrosanct about the kinds of judgmental evidence for appraising instructional sensitivity or how to go about assembling such evidence. One practical method is to create panels of 15 to 20 curriculum specialists and teachers who are knowledgeable about the content. If the results of an instructional sensitivity review are to be released to the public, it is sensible to include several noneducators as panelists for the sake of credibility.

After receiving ample orientation and training, panelists would use 10-point scales to rate the tests on four evaluative dimensions. For each evaluative dimension,

panelists would be given a rubric that contains sufficient explanatory information and, as necessary, previously judged exemplars so that all panelists would use similar evaluative perspectives.

Panelists could use a variety of procedures for their tasks. But most likely their procedures would be similar to either the iterative models that have been commonly employed in setting standards for the past couple of decades or the judgmental methods used in recent years to ascertain the alignment between a state's accountability tests and the content standards those tests are ostensibly assessing. In both of those approaches, panelists typically make individual judgments and then share them with the entire panel. After that, an open discussion of panelists' judgments occurs, followed by another set of individual judgments. As many iterations of this procedure are carried out as are necessary for the group to reach a consensus. Another method uses the average of the panelists' final ratings as the overall judgment.

The four evaluative dimensions that should be used are 1) the number of curricular aims assessed, 2) the clarity of assessment targets, 3) the number of items per assessed curricular aim, and 4) the instructional sensitivity of items. As noted above, panelists would be given sufficient information to allow them to rate each dimension on a 10-point scale. Then the four separate ratings would be combined to arrive at an overall rating of a test's instructional sensitivity. Those who are designing an instructional sensitivity review need to determine whether to assign equal weight to each of the four dimensions or to assign different weights to each dimension.

NUMBER OF CURRICULAR AIMS ASSESSED

Experience makes it all too clear that teachers cannot realistically focus their instruction on large numbers of curricular aims. In many states, lengthy lists of officially approved curricular aims often oblige teachers to guess about what will be assessed on a given year's accountability tests. More often than not, there are far too many "official" curricular aims to be tested in the available testing time (or, in truth, to be taught in the available teaching time). After a few years of guessing incorrectly, many teachers simply abandon any reliance on the state's sanctioned curricular aims. If an accountability test is to be genuinely sensitive to the impact of instruction, all teachers should be pursuing the same curricular aims, not teacher-divined subsets of those aims.

Clearly, therefore, one evaluative dimension to be considered when determining an accountability test's

instructional sensitivity should be the number of curricular aims assessed by the test. Note that there is no reference here to the *worth* of those curricular aims. Obviously, the worth of a set of curricular aims is extremely important, but the appraisal of that worth should be a separate, albeit indispensable, activity. A test's instructional sensitivity is not dependent on the grandeur of the curricular aims being measured.

To evaluate the number of curricular aims assessed, it is necessary to deal with those curricular aims at a grain size (that is, degree of breadth) that meshes with teachers' day-to-day or week-to-week instructional decisions. Evaluators must be wary of aims that are too large. If the grain size of a curricular aim is so large that it prevents a teacher from devising activities sensibly targeted toward that curricular aim, then the curricular aim's grain size is too broad. For example, some states have very general sets of "content standards," such as "measurement" or "algebra" in mathematics. This grain size is much too large for panelists to make sense of when using this evaluative dimension. Instead, a panelist's focus needs to be on the smaller curricular aims typically subsumed by more general standards. These smaller curricular aims are often labeled "benchmarks," "indicators," "objectives," or something similar.

The rubric for this evaluative dimension should be organized around a definition in which higher ratings would be given to a set of curricular aims whose numbers would be regarded by teachers as easily addressed in the instructional time available. In other words, if teachers have enough instructional time to teach students to achieve *all* of the curricular aims to be assessed, panelists would give the highest ratings. In contrast, lower ratings would be given to sets of curricular aims regarded as too numerous to teach in the available instructional time, because teachers would be uncertain about which of the aims would be assessed on a given year's accountability test.

CLARITY OF ASSESSMENT TARGETS

The second evaluative dimension revolves around the degree to which teachers understand what they are supposed to be teaching. If teachers have only a murky idea of what constitutes the knowledge or skills they are supposed to be teaching — as exemplified by what's measured on an accountability test — then those teachers will often end up teaching the wrong things. Thus an instructionally sensitive accountability test should be accompanied by descriptive information that describes not only the types of items eligible to be used on the test but, more important, the essence of the skills or

knowledge the test will be measuring. If teachers have a clear understanding of what's to be measured, then their instructional efforts can be directed toward those skills and bodies of knowledge rather than toward specific test items. A test consisting of items that measure instructional targets that teachers understand is surely more apt to accurately measure the degree to which those targets have been hit.

The manner in which an accountability test describes what it's supposed to be measuring can, of course, vary considerably. Sometimes state officials supply no descriptive information beyond the curricular aims themselves. In other instances, a state's educational authorities provide explicit assessment descriptions intended to let the state's teachers know what's to be measured by the state's accountability tests. And, of course, there are many other ways of describing what's to be assessed by an accountability test. Thus, in carrying out a judgmental appraisal of an accountability test's descriptive clarity, the material under review should be *whatever descriptive information is readily available to teachers*. If this turns out to be only the state's official curricular aims, then that's the information to be used when panelists render their judgments about this second dimension of instructional sensitivity. If a state's tests have more detailed assessment descriptions, then this is the information to use. The descriptive information to be reviewed by panelists must be routinely accessible to teachers, not hidden in the often fugitive technical reports associated with an accountability test.

The rubric for this evaluative dimension should emphasize the teachers' likely understanding of the nature of the skills and knowledge to be assessed. Higher ratings would be supplied when panelists believe teachers can readily comprehend what's to be assessed well enough to design appropriate instructional activities.

Ideally, before ratings on this evaluative dimension are collected, a separate data-gathering activity would be carried out in which a half-dozen or so teachers are first given copies of whatever materials are routinely available that describe the accountability test's assessment targets, are asked to read them carefully, and then are directed to put that descriptive information away. Next, in their own words and without reference to the previously read descriptive material, the teachers would be asked to write, *independently*, what they understood to be the essence of each skill or body of knowledge to be assessed. The degree to which such independently written descriptions are homogeneous would then be supplied to the panelists before they render a judgment. This information would supply panelists with an idea of just how much ambiguity appears to be present in

the test's descriptive materials. Although not necessary, this optional activity would clearly strengthen the conclusions reached by the panel.

ITEMS PER ASSESSED CURRICULAR AIM

The third evaluative dimension on which an accountability test's instructional sensitivity can be judged deals with whether there are enough items on a test to allow teachers (as well as students and students' parents) to determine if each assessed curricular aim has been satisfactorily achieved. The rationale for this evaluative factor is straightforward. If teachers can't tell which parts of their instruction are working and which parts aren't, they'll be unable to improve ineffectual instructional segments for future students. Moreover, if there are too few items to determine a student's status with respect to, say, a specific skill in mathematics, then a student (or the student's parents) can't tell whether additional instruction appears to be needed on that skill. Similarly, if teachers are given meaningful information regarding their incoming students' skills and knowledge at the beginning of a school year, then more appropriately tailored instruction can be provided for those new students. Although not strictly related to a test's instructional sensitivity, the reporting of students' status on each curricular aim can transform an instructionally sensitive test into one that is also instructionally supportive.

The number of items necessary to arrive at a reasonably accurate estimate of a student's mastery of a particular assessed skill or body of knowledge depends, of course, on the curricular aim being measured. Broad curricular aims require more items than do narrower ones. Thus the number of items on a given test might vary for the different curricular aims to be measured. Panelists need to make their ratings on this evaluative dimension by reviewing the general pattern of a test's distribution of items per assessed curricular aim after taking into consideration the particular outcomes being assessed.

The rubric to appraise this evaluative dimension should take into account the *number* and *representativeness* of the sets of items being used. Panelists would first be asked to review any materials describing what the test is supposed to measure, then consider the degree to which a designated collection of items satisfactorily provides an estimate of a test-taker's achievement. High ratings would reflect both excellent content representativeness and sufficient numbers of items. In other words, to get a high rating on this evaluative dimension, there would need to be enough items to assess a

given skill or body of knowledge, and those items would need to satisfactorily sample the key components of the skill or knowledge being measured. Low ratings would be based on too few items, insufficient representativeness of the items, or both.

ITEM SENSITIVITY

The fourth and final evaluative dimension is the degree to which the items on the test are judged to be sensitive to instructional impact. The panelists must either be able to render judgments themselves on a substantial number of actual items from the test or have access to item-by-item judgments rendered by others. In either scenario, the item reviewers must make judgments, one item at a time, about a sufficiently large number of actual items so that a defensible conclusion can be drawn about the instructional sensitivity of a test. Sometimes, because of test-security considerations, these judgments may be made in controlled situations by individuals other than the regular panelists. Ideally, the panelists would personally review a test's items one at a time.

There are three aspects of this evaluative dimension that, in concert, can allow panelists to arrive at a rating of a test's item sensitivity. First, three separate judgments need to be rendered about each item. These judgments might take the form of Yes, No, or Not Sure and would be made in response to three questions:

1. *SES influence.* Would a student's likelihood of responding correctly to this item be determined mostly by the socioeconomic status of the student's family?

2. *Inherited academic aptitudes.* Would a student's likelihood of responding correctly to this item be determined mostly by the student's innate verbal, quantitative, or spatial aptitudes?

3. *Responsiveness to instruction.* If a teacher has provided reasonably effective instruction related to what's measured by this item, is it likely that a substantial majority of the teacher's students will respond correctly to the item?

An instructionally sensitive item should receive a flock of No responses for the first two questions and a great many Yes responses for the third question. For each item, then, the reviewers' judgments indicating the degree to which the item is instructionally sensitive would be reported on all three of these questions. Then the panel would use the per-item data to arrive at a judgment on the test as a whole.

It should be noted that many current accountability tests, especially those constructed along traditional psychometric lines, contain numerous items closely

linked to students' SES or to their inherited academic aptitudes. This occurs because the mission of traditional achievement tests is to permit comparisons among test-takers' scores. In order for those comparisons to work properly, however, there must be a reasonable degree of score spread in students' test scores. That is, students' test results must be meaningfully different so that fine-grained contrasts between test-takers are possible. Because students' SES and inherited academic aptitudes are both widely dispersed variables, and ones that do not change rapidly, test items linked to either of these variables efficiently spread out students' test scores. Accordingly, builders of traditional achievement tests often end up putting a considerable number of such items into their tests, including those tests used for accountability purposes.

To the extent that accountability tests measure what students bring to school rather than what they are taught there, the tests will be less sensitive to instruction. It is true, of course, that SES and inherited academic aptitudes are themselves substantially interrelated. However, by asking panelists to recognize that either of those variables, if pervasively present in an accountability test, will contaminate the test's ability to gauge instructional quality, we have a reasonable chance to isolate the magnitude of such contaminants.

INSTRUCTIONAL SENSITIVITY REVIEWS

The vast majority of today's educational accountability tests are fundamentally insensitive to instructional quality. If these tests cannot indicate whether students' scores are affected by the quality of a teacher's instruction, then they prevent well-intentioned accountability programs from accomplishing what their architects had hoped. If educators find that the quality of their instructional efforts is being determined by students' scores on accountability tests that are inherently incapable of detecting effective instruction, they should take steps to review the tests' instructional sensitivity. The judgmental procedures set forth here provide the framework for a practical process for carrying out such a review.

If the review of an accountability test reveals it to be substantially sensitive to instruction, then it is likely that other test-influenced elements of the accountability program are acceptable. However, if a review indicates that an accountability program's tests are instructionally *insensitive*, then two courses of action seem warranted. First, there should be a serious attempt made to replace the instructionally insensitive tests with those that are

(Continued on page 155)

Instructional Insensitivity

(Continued from page 150)

sensitive to instruction. If that replacement effort fails, it is imperative to inform the public, and especially education policy makers, that the accountability tests being used are unable to detect successful instruction even if it is present. In that case it is particularly important to involve noneducators as review panelists so that the public does not see the instructional sensitivity review as the educators' attempt to escape accountability. Parents and members of the business community can be readily trained to function effectively as members of an instructional sensitivity panel.

An evaluation of the instructional sensitivity of the nation's accountability tests is long overdue. We must discover whether the key data-gathering tools of the accountability movement have been claiming to do something they simply cannot pull off. **■**

File Name and Bibliographic Information

k0710pop.pdf

**W. James Popham, Instructional Insensitivity of Tests:
Accountability's Dire Drawback, Vol. 89, No. 02, October 2007, pp.
146-150.**

Copyright Notice

Phi Delta Kappa International, Inc., holds copyright to this article, which may be reproduced or otherwise used only in accordance with U.S. law governing fair use. MULTIPLE copies, in print and electronic formats, may not be made or distributed without express permission from Phi Delta Kappa International, Inc. All rights reserved.

Note that photographs, artwork, advertising, and other elements to which Phi Delta Kappa does not hold copyright may have been removed from these pages.

Please fax permission requests to the attention of KAPPAN Permissions Editor at 812/339-0018 or e-mail permission requests to kappan@pdkintl.org.

For further information, contact:

Phi Delta Kappa International, Inc.
408 N. Union St.
P.O. Box 789
Bloomington, Indiana 47402-0789
812/339-1156 Phone
800/766-1156 Tollfree
812/339-0018 Fax

<http://www.pdkintl.org>

Find more articles using PDK's Publication Archives Search at
<http://www.pdkintl.org/search.htm>.