

## **INSTRUCTIONAL INSENSITIVITY OF TESTS: ACCOUNTABILITY'S DIRE DRAWBACK\***

W. James Popham  
University of California, Los Angeles

Today's educational policymakers rely increasingly on the results of large-scale accountability tests to influence their deliberations. The role of accountability tests is currently influential, however, not only in shaping educational policies, but also in influencing day-by-day classroom events. Any educator who doesn't know that accountability tests have a profound impact on teachers' behaviors must have been doing some serious hibernating.

The pivotal premise underlying the use of accountability tests is that students' test-scores will be indicative of the quality of instruction those students have received. If students score well on accountability tests, it is concluded that those students have been well taught. Conversely, if students score poorly on accountability tests, it is believed that those students have been badly taught.

One key assumption made by accountability-testing advocates is that if teachers realize they are going to be routinely judged by their students' accountability test-scores, those teachers will try to do a better instructional job. Another important assumption of accountability testing is that if test-results indicate ineffective instruction is taking place in, say, a school or a district, then higher-level authorities can take action to bolster the quality of instruction in those low-performing settings.

Yet, for either of these assumptions to make sense, the accountability tests being employed must actually be able to determine the impact of instruction on students' test-scores. Accountability tests that are incapable of distinguishing between effective and ineffective instruction are, therefore, patently unsuitable for use in any sensible educational accountability program. But, at this moment in time, all but a few of the accountability tests now having such a profound impact on our nation's schools are *instructionally insensitive*.

### **Instructional Sensitivity**

A test's *instructional sensitivity* represents the degree to which students' performances on that test accurately reflect the quality of instruction specifically provided to promote students' mastery of whatever is being assessed. To illustrate,

---

\* A presentation at the annual meeting of the American Educational Research Association, Chicago, Illinois, April 9-13, 2007.

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 between strong and weak instruction. Currently, for example, students' performances on most accountability tests are more heavily influenced by students' socioeconomic status (SES) than by the quality of teachers' instructional efforts. 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, therefore, render untenable the two aforementioned assumptions underlying a test-based educational accountability strategy. 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 ineffective instruction on the basis of low test scores if those low scores really aren't a consequence of ineffective instruction?

Indeed, there is ample evidence available these days to indicate that ill-conceived accountability programs can seriously diminish instructional quality, not improve it. Too often we find teachers engaging in curricular reductionism whereby they give scant, if any, instructional attention to content not assessed by accountability tests. Too often we learn of teachers who impose excessive test-preparation drills on their students and, thereby, extinguish the genuine joy those students should experience as they learn. 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 policymakers' reliance on the wrong kinds of accountability tests. Educators who simply carp about accountability tests are usually seen as individuals eager to escape evaluative scrutiny. 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 whereby teachers are being asked to improve students' performances on tests patently unable to detect improved instruction. Clearly, we need a credible procedure to determine how instructionally sensitive a given accountability test truly is.

The following paragraphs describe the main features of a practical procedure for ascertaining the instructional sensitivity of any test. It is a procedure that can be used to identify the instructional sensitivity of an existing accountability test or the instructional sensitivity of an under-development accountability test. Because the instructional sensitivity of an accountability system's tests is the dominant determiner of whether that system educationally helps or harms students, it is hoped the approach to be described here will be used widely. The kind of instructional-sensitivity review described here can be carried out either by governmental agencies such as a state department of education or by nongovernmental groups such as educators' professional organizations. Although the chief ingredients of the approach will be described here, as in most such endeavors, devils hide in details. Thus, a separate, but readily accessible document provides a more detailed description of the procedural particulars of such a system (available from [wpopham@ucla.edu](mailto:wpopham@ucla.edu) or [www.ioxassessment.com](http://www.ioxassessment.com)). Let's turn now to the two basic kinds of evidence that can illuminate an accountability test's instructional sensitivity.

### **Opting for One Evidence Category**

Two main categories of relevant evidence can contribute to the determination of an accountability test's instructional sensitivity, namely, *judgmental evidence* and *empirical evidence*. Judgmental evidence can be collected regarding relevant evaluative dimensions of a test's instructional sensitivity 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, when differences are compared between "taught" and "untaught" students' test-scores.

Whether the instructional sensitivity of a test is determined by reliance on only judgmental evidence, only empirical evidence, 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 *totally* insensitive to instruction. Instead, a test's instructional sensitivity can be thought of as a continuum in which, for example, a 1 represents a test completely *insensitive* to instruction and a 10 represents a test completely *sensitive* 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.

Although both judgmental and empirical evidence can help us establish the degree of a test's instructional sensitivity, I recommend that—for practicality's sake—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, I would be delighted merely to see the collection of judgmental evidence regarding the instructional sensitivity of an accountability test. The assembly of confirmatory empirical evidence is a luxury surely to be desired but not absolutely requisite 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 based on judges' reviews of a test's items.

There is nothing sacrosanct about what is about to be suggested regarding the kinds of judgmental evidence deemed appropriate for appraising an accountability test's instructional sensitivity, or how to go about assembling such evidence. However, for openers, it is suggested that *instructional-sensitivity panels* composed chiefly of content-knowledgeable teachers or curriculum specialists serve as the suppliers of the necessary judgmental evidence. I am thinking of panels containing 10-15 individuals, most of whom are experienced educators. For credibility's sake, especially if the results of an instructional-sensitivity review are to be released to the public, it is often sensible to also include several non-educators as panelists.

After receiving ample orientation and training, panelists would be asked to render individual ratings, on 10-point scales, for four evaluative dimensions to be described shortly. For each evaluative dimension, panelists would be given a rubric whose potential scores could range from a high of 10 to a low of 1. Each rubric would contain sufficient explanatory information (and, as necessary, previously judged exemplars) so panelists would be approaching their tasks using similar evaluative perspectives.

How would these panels function? Well, there are all sorts of procedural permutations possible, but most would probably be similar to (1) the iterative models commonly employed in standard-setting procedures for the past couple of decades and (2) 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, then these judgments are shared with the entire panel. Thereafter, an open discussion of panelists' judgments occurs, followed by another set of individual panelist-rendered judgments. As many iterations of this procedure are carried out as seem necessary. A group consensus can be sought or, absent such a consensus, the average of individual panelists' final ratings can serve as a panel's overall judgment.

Here, then, are four evaluative dimensions that should be incorporated in the assembly of judgmental evidence regarding an accountability test's instructional sensitivity. Those evaluative dimensions are (1) *the number of curricular aims assessed*, (2) *the clarity of assessment targets*, (3) *number of items per assessed curricular aim*, and (4) *the instructional sensitivity of items*. Panels of judges would be given sufficient information to allow them to arrive at a 1-to-10 judgment on each of these dimensions, and those four separate ratings would then be combined to arrive at an overall 1-to-10 judgment regarding a test's instructional sensitivity. Those individuals who were supervising a given instructional-sensitivity review would

determine whether to assign equal weight to panelists' judgments on the four dimensions or, instead, to assign different weights to certain evaluative dimensions.

### **Number of Curricular Aims Assessed**

The first evaluative dimension deals with the number of curricular aims an accountability test attempts to measure. Experience makes it all too clear that teachers can't 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, therefore, many teachers simply abandon any reliance on the state's sanctioned curricular aims. If an accountability test is to be genuinely sensitive to measuring the impact of instruction, all teachers should be pursuing the same curricular aims, not seeking 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 a set of curricular aims should be a separate, albeit indispensable, activity. We are looking here at an accountability test's ability to detect instructional impact on whatever curricular aims the test is designed to measure. A test's instructional sensitivity is not dependent on the grandeur of the curricular aims being measured.

To carry out an appraisal of the number of curricular aims assessed, it is necessary to deal with those curricular aims at a grain-size that meshes with teachers' day-to-day or week-to-week instructional decisions. Some states have approved very general sets of "content standards," for instance, such mathematics content standards as "measurement" or "algebra." This grain-size is typically much too large for panelists to make sense of when using this evaluative dimension. Instead, a panelist's judgmental focus needs to be on the smaller-scope curricular aims typically subsumed by more general content standards. These smaller grain-size curricular aims are often labeled "benchmarks," "indicators," "objectives," or something similar. Judgmental ratings about the number of assessed curricular aims should be based on curricular aspirations described at a grain-size matched to the way teachers think about their instructional decisions regarding what students should be learning. If the grain-size of a curricular aim is so large that its breadth prevents a teacher from devising instructional activities sensibly targeted toward that curricular aim, then the curricular aim's grain-size is too broad. Thus, for instance, the focus of panels would typically be on a state's smaller-scope curricular *objectives* rather than on the broader-scope content standards under which those objectives are grouped.

The rubric dealing with this evaluative dimension (*number of curricular aims assessed*) should be organized around a two-directional definition in which higher 1-to-10 ratings would be given to a set of assessed curricular aims whose numbers would be regarded by teachers as sufficiently addressable in the instructional time available. In other words, instructional-sensitivity panelists who supplied positive ratings on this evaluative dimension would believe teachers have enough instructional time to teach students to achieve *essentially all* of the to-be-assessed curricular aims. In contrast, lower ratings would be given by panelists to sets of assessment-eligible curricular aims regarded as too numerous to teach in the available instructional time. Panelists who supplied low ratings on this evaluative dimension would typically believe the numbers of assessment-eligible curricular aims were so profuse that 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 and/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. An instructionally sensitive accountability test, therefore, should be accompanied by descriptive information laying out not only the types of items eligible to be used on the test but, more importantly, delineating the essence of the skills or bodies of 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 to-be-assessed skills and/or bodies of knowledge rather than toward specific test items. A test consisting of items measuring teacher-understood instructional targets 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, of course, can vary all over the lot. Sometimes state officials supply no descriptive information at all—other than the curricular aims themselves. In other instances, a state's educational authorities have provided 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 information turns out to be only the state's official curricular aims, then that's the information to be used when instructional-sensitivity panelists render their judgments about this second evaluative dimension. If a state's tests are delineated in the form of more detailed assessment descriptions, then this is the information to use. The descriptive information to be reviewed by instructional-sensitivity 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 revolve around panelists' perceptions regarding the clarity with which teachers are apt to understand the nature of the skills and/or bodies of knowledge to be assessed. Higher ratings would be supplied when panelists believe teachers can readily comprehend the essential nature of the skills and/or knowledge to be assessed—that is, when teachers understand what's to be assessed well enough to design instructional activities appropriately promoting students' achievement of those outcomes. Lower ratings would be supplied by panelists who, having read whatever descriptive information accompanies the test, believe teachers would be unclear about the actual nature of the skills and/or bodies of knowledge to be tested.

Ideally, if time permits, before ratings on this evaluative dimension (*clarity of assessment targets*) were collected from instructional-sensitivity panelists, a separate data-gathering activity would be carried out in which a half-dozen or so teachers were first given copies of whatever materials are routinely available describing the accountability test's assessment targets, asked to read it carefully, then 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 authored descriptions were homogeneous would then be supplied to instructional-sensitivity panelists prior to the panelists' rendering a judgment on this second evaluative dimension. This information would help supply panelists with an idea of just how much ambiguity appears to be present or absent in the test's descriptive materials. Although not necessary, this optional activity would clearly strengthen the conclusions reached by any instructional-sensitivity panel about the clarity with which an accountability test's assessment targets are described.

### **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 surely be unable to repair 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.

The whole notion of instructionally diagnostic assessment stumbles if teachers and students can't tell, reasonably well, whether students have mastered each of an

accountability test's assessed skills and/or bodies of knowledge. How can teachers sharpen their instruction from year to year if they don't know which elements of their instruction need amelioration? Similarly, if teachers—at the beginning of a school year—are given meaningful information regarding their incoming students' skills and knowledge, then more appropriately tailored instruction can surely be provided for those new students. In short, the whole instructional game becomes more sensible to play if teachers and students know *which* skills and bodies of knowledge have been mastered. If there are enough items per assessed curricular aim, and those items appear to satisfactorily represent the skill and/or body of knowledge being measured, then teachers can effectively employ students' test results. If there are too few items per skill and/or body of knowledge, then there will be scant instructional dividends from an accountability test.

Thus, if an accountability test has sufficient numbers of items to provide diagnostic insights for teachers, and if those diagnostic dividends are acted on by teachers to provide instruction more suitably tailored to students, it is likely that such a test will, over time, contribute to the caliber of the instruction whose effects are being measured.

Although not strictly related to a test's instructional sensitivity, the reporting of students' status with respect to each curricular aim assessed can transform an *instructionally sensitive* test into one that is also *instructionally supportive*. When teachers know how well students have mastered each assessed curricular aim, they can not only better teach individual students, but can also identify which segments of their instruction are working well and which segments need fixing. From an instructional perspective, the availability of evidence regarding students' mastery of each assessed curricular aim is enormously beneficial.

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 grain-size of the curricular aim being measured. Thus, certain broad-scope curricular aims would surely require more items than would curricular aims of a narrower scope. The numbers of items on a given test measuring students' mastery of different curricular aims, therefore, might vary from curricular aim to curricular aim. Instructional-sensitivity panelists, thus, would need to arrive at their final 1-to-10 ratings on this evaluative dimension by reviewing the general pattern of a test's distribution of items per assessed curricular aims after taking into consideration the grain-sizes of the particular outcomes being assessed.

The rubric to appraise this evaluative dimension (*items per assessed curricular aim*) should be fashioned around panelists' judgments with regard to the *number and representativeness* of the sets of items being used to assess student's status regarding the curricular aims the test purports to measure. 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 (intended to measure a particular skill or body of knowledge) satisfactorily provides an estimate of a test-

taker's status with respect to what's being assessed. High ratings by teachers would reflect both excellent content representativeness as well as 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 body of knowledge being measured. Low ratings would be based either on too few items, insufficient representativeness of the items, or both.

### **Item Sensitivity**

The fourth and final evaluative dimension by which panelists can appraise an accountability test is the degree to which the items on the test are judged to be sensitive to instructional impact. To arrive at a 1-to-10 rating on this dimension, either the panelists must either (1) be able to render item-by-item judgments themselves on a substantial number of actual items from the test being scrutinized or (2) 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 (either those items currently being used or, possibly, released items from earlier test-administrations) so that a defensible conclusion can be drawn about the instructional sensitivity of a test's items. Sometimes, because of test-security considerations, it may make more sense for these item-by-item judgments to be made in security-controlled situations by individuals other than the regular instructional-sensitivity panelists. Ideally, the panelists would personally review a test's items—one item at a time.

There are three aspects of this evaluative dimension that, in concert, can allow panelists to arrive at a 1-to-10 rating of a test's item sensitivity. First, per-item judgments can be made about the *absence of dominant SES influence*, that is, the degree to which a student's likelihood of responding correctly to an item is *not* chiefly due to the student's socioeconomic status. A second consideration is the *absence of dominant influence by inherited academic aptitudes*, that is, the degree to which a student's likelihood of responding correctly to an item is *not* chiefly due to the student's genetically acquired verbal, quantitative, and/or spatial aptitudes. A third aspect of this evaluative dimension is the judged *responsiveness to instruction* of an item, that is, the degree to which item-reviewers believe that—if instruction related to what's measured by the item *has* been genuinely effective—most well-taught students would be likely to respond correctly to the item.

Accordingly, for each item reviewed by an instructional-sensitivity panel (or by another, designated group of item-reviewers) three separate judgments would 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 such as the following:

- *SES Influence: Would a student's likelihood of responding correctly to this item be dominantly determined by the socioeconomic status of the student's family?*

- *Inherited Academic Aptitudes: Would a student's likelihood of responding correctly to this item be dominantly determined by the student's innate verbal, quantitative, and/or spatial aptitudes?*
- *Responsiveness to Instruction: If a teacher has provided reasonably effective instruction related to what's measured by this item, is it likely a substantial majority of the teacher's students will respond correctly to the item?*

Assuming that item-reviewers have been given the option of supplying a Yes, No, or Not Sure response to each of these three questions for every item they review, 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, a percent of item-reviewers' judgments indicating the degree to which the item is instructionally sensitive would be reported on all three of these sub-dimensions. Thereafter, the instructional-sensitivity panel would use a rubric for this fourth evaluative dimension (*item sensitivity*) by using such per-item data to arrive at a 1-to-10 judgment.

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 comparative measurement 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 among students' test scores. That is, students' test results must be meaningfully different so that fine-grained contrasts among 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.

### **The Need for Instructional-Sensitivity Reviews**

As noted earlier, instructionally insensitive accountability tests fundamentally prevent well-intentioned accountability programs from accomplishing what the architects of those programs had hoped. It was also contended that the vast majority of today's

educational accountability tests are fundamentally insensitive to the detection of instructional quality. Thus, if educators find themselves functioning in a setting where the quality of their instructional efforts is being determined by students' scores on accountability tests that are inherently incapable of detecting either effective or ineffective instruction, steps should be taken to review the instructional sensitivity of the tests. The judgmental procedures set forth here provide the framework for a practicable process by which such a review can be carried out—and without exorbitant cost.

If the review of an accountability test reveals it is substantially *sensitive* to instruction, then it is likely other test-influenced elements of the accountability program are also apt to be acceptable. If, however, 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, in fact, sensitive to instruction. If that replacement-effort fails, however, it is imperative to inform the public, and especially educational policymakers, that the accountability tests being used are unable to detect successful instruction even if it were present. If this second course of action is contemplated, then it becomes particularly important to involve non-educators as instructional-sensitivity panelists. Parents and members of the business community can be readily prepared to function effectively as members of an instructional-sensitivity panel. Instructional-sensitivity reviews dare not be seen by the public as educators' escapism from accountability.

The need to have the nation's accountability tests undergo instructional-sensitivity scrutiny is long overdue. We must discover whether the key data-gathering tools of the accountability movement, just as was the case with a fictional emperor who is said to have paraded in the buff, have been claiming to do something they simply can't pull off.