

## DETERMINING THE INSTRUCTIONAL SENSITIVITY OF ACCOUNTABILITY TESTS\*

W. James Popham  
University of California, Los Angeles

Educational accountability tests exist because of doubts regarding the caliber of schooling. Typically, accountability tests are mandated by policymakers who rely on a rationale that, once educators discover their instructional effectiveness is to be routinely appraised via students' performances on achievement tests, those educators will try to boost their students' test scores by teaching better. Yet, for such a rationale to make sense, the accountability tests being employed must be able to determine the impact of instruction on students' test scores. Accountability tests incapable of distinguishing between effective and ineffective instruction are, therefore, patently unsuitable for use in a sound 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 whatever instruction was provided to promote students' mastery of the skills and/or bodies of knowledge assessed by the test. To illustrate, an *instructionally sensitive test* would be capable of distinguishing between terrific and tawdry instruction by allowing us to validly conclude that a set of *high* students' scores are meaningfully, but not exclusively, attributable to effective instruction. Similarly, such a test would allow us to accurately infer that a set of *low* students' 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 of this nation's 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 success with which a school's students have been taught.

Instructionally insensitive tests, therefore, render irrational the basic rationale underlying educational accountability testing. 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 for students?

---

\* A presentation at the annual Large-Scale Assessment Conference, Council of Chief State School Officers, San Francisco, California, June 25-28, 2006

Indeed, there is ample evidence available these days to indicate that ill-conceived accountability programs can seriously diminish the quality of instruction, not improve it. Too often do we find teachers engaging in curricular reductionism whereby they give scant, if any, instructional attention to content not assessed by accountability tests. Too often do we learn of teachers who impose excessive test-preparation drills on their students and, thereby, extinguish the genuine joy those students ought to experience as they learn. Too often do we hear of teachers and administrators blatantly portraying students' test-scores' as improved when, in fact, no such improvement has transpired.

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 all too apparent, it accomplishes little when educators carp about policymakers' reliance on the wrong kinds of accountability tests. Educators who complain 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 our current calamity in which thousands of American 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 educational accountability test really is.

## **A Proposed Framework**

In the remainder of this analysis I intend to lay out the chief features of an evidence-based framework capable of supplying an accurate and credible answer to the question of how instructionally sensitive a given accountability test is. I believe the proposed framework is capable of being implemented immediately and, if used, would lead to reasonable judgments regarding the instructional sensitivity of any accountability test. Nonetheless, what the framework needs, as you will soon see, is much more refinement and, in particular, a series of carefully crafted rubrics in order for certain kinds of evidence to be quantified. I will attempt to spell out in modest detail the chief features of the framework, but I readily concede there's some heavy thinking needed about how best to assemble the variety of evidence I will soon be identifying.

To illustrate, one of the essential attributes of a well-formed rubric is that its use permits qualitative differentiations to be made regarding whatever is being judged. The rubrics I have in mind would need such differentiations, probably the more fine-grained the better. However, all I intend to do at this point is set forth a rough, two-directional quality definition for each evaluative factor I suggest might be used when appraising a given accountability test. I realize much more attention must be given to these rubrics, but I hope that a two-directional

definition of quality (that is, what should be rated high and what should be rated low) will suffice at this point.

Before digging into the framework's specifics, let me first lay out briefly the overall nature of what I am proposing. I believe there are two main categories of relevant evidence that can contribute to the determination of an accountability test's instructional sensitivity, namely, *judgmental evidence* and *empirical evidence*. Judgmental evidence can be collected regarding several relevant evaluative dimensions regarding a test's instructional sensitivity by using panels of experienced teachers. Empirical evidence can be provided from students' actual test scores, but these data might also be transformed into similar evaluative scales. Although it would be possible to arrive at a conclusion regarding a specific test's instructional sensitivity based only on judgmental evidence or only on empirical evidence, ideally both categories of evidence would be employed when deciding whether a given test is instructionally sensitive enough for use in an accountability program.

From the foregoing comment, it should be apparent I think instructional sensitivity is best 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. Rather, in the real world, one finds accountability tests ranging in their ability to gauge instructional quality. However, if we possess a methodology for ascertaining an accountability test's relative sensitivity to instruction along such a continuum, it will then be easier to decide whether to use that test in a particular accountability program.

To repeat, the framework I will be laying out needs substantial refinement and further explication. I hope that, should there be sufficient interest, I can join with concerned colleagues to move this model further along in its clarity, ease of use, and power.

### **Judgmental Evidence**

There is nothing sacrosanct about what I am about to suggest regarding the four kinds of judgmental evidence I think appropriate for appraising an accountability test's instructional sensitivity, or how to go about assembling such evidence. However, for openers, I suggest that panels of seasoned, content-knowledgeable teachers serve as the suppliers of this judgmental evidence. I am thinking of panels containing 10-15 teachers who would be asked to render individual ratings, on 10-point scales, for as many as possible of the evaluative dimensions I will describe 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, if necessary, illustrations) so panelists would be approaching their tasks using similar evaluative perspectives. Such rubrics must be carefully worded, pilot tested, and revised until they function satisfactorily. However, once this sort of rubric-

generation and rubric-refinement has been accomplished, panelists should be able to supply accurate judgments regarding each of the evaluative dimensions I'll soon propose.

How would these panels function? Well, there are all sorts of procedural permutations possible, but if I were designing an activity of this sort, I'd probably lean toward a process similar to the iterative models commonly employed in standard-setting procedures for the past couple of decades. In those approaches, panelists make individual judgments, then these judgments are shared with the entire panel, either panelist-by-panelist or in the form of group averages. Thereafter, an open discussion of panelists' judgments occurs, followed by another set of individual panelist-rendered judgments. As many iterations of this rate-solo-then-discuss 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.

Let me turn now to four evaluative dimensions I believe should be incorporated in the assembly of judgmental evidence regarding an accountability test's instructional sensitivity. For each evaluative dimension identified, I'll supply a brief justification as well as indicate the heart of a two-directional quality definition underlying the rubric for this evaluative dimension.

*Number of curricular aims assessed.* The first of my four evaluative dimensions 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 too many curricular aims. In many states, a plethora of officially approved curricular aims often obliges those states' teachers to *guess* about what will be assessed on a given year's accountability tests. There are far too many 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 measuring the impact on instruction, all teachers should be pursuing the same curricular aims, not teacher-guessed 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 I am *not* referring to the *worth* of those curricular aims. Indeed, 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 the test is designed to measure.

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. For example, some states have approved very general sets of "content standards" such as, in the field of mathematics, content standards like "measurement" or "algebra." This sort of grain-size is much too large to make sense of when applying this evaluative dimension. Instead, our 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," "expectancies," "objectives," or something similar. Judgmental ratings regarding the number of curricular aims being assessed by an accountability test should be based on curricular aspirations described at an understandable grain-size—a grain-size matched to the way teachers think about their instructional decisions. To illustrate, the focus of teacher-panels should be on a state's approved curricular *objectives* rather than the broad-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 quality definition in which higher ratings would be given to a set of assessed curricular aims whose numbers would be regarded by teachers as readily addressable in the instructional time available. In other words, teachers who supplied positive ratings on this evaluative dimension would believe they 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 to sets of assessment-eligible curricular aims regarded by teachers as too many to teach in the available instructional time. Teachers who supplied low ratings on this evaluative dimension would typically think the numbers of curricular aims were so numerous that teachers would typically be obliged to guess regarding which of the aims would be assessed on a given year's accountability test.

*Clarity of assessment targets.* The second evaluative dimension to be judgmentally considered 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 stuff. An instructionally sensitive accountability test, therefore, should be accompanied by "assessment descriptions" laying out not only the types of items eligible to be used on the test but, more importantly, describing the essence of the skills and/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.

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 provided 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 is only the state's official curricular aims, then that's the information to be used. If a state's tests are delineated in the form of assessment descriptions, then this is the information to use. The descriptive information to be employed in this instance must be routinely accessible to teachers, not hidden in the technical reports associated with an accountability test.

The rubric for this evaluative dimension should revolve around teachers' perceptions regarding the clarity with which they understand the nature of the skills and/or bodies of knowledge to be assessed. Higher ratings would be supplied when teacher-panelists really understand the essential nature of the skills and/or knowledge to be assessed—that is, understand what's to be assessed well enough to design appropriate instructional activities related those outcomes. Lower ratings would be supplied by teachers who, having read whatever descriptive information accompanied the test, were 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 panelists, an activity could be carried out in which panelists were first given the descriptive material available for the accountability test, asked to read it, then *in their own words* write out—*independently*—what they understood to be the essence of the skill(s) or knowledge to be assessed. The degree to which such independently authored descriptions were homogeneous would then be revealed, thus supplying panelists with an idea of just how much ambiguity appears to be present or absent in the test's descriptive materials.

*Items per assessed curricular aim.* The third evaluative dimension on which an accountability test's instructional sensitivity can be ascertained deals with whether there are enough items on the tests to allow teachers (and students) to determine if *each* assessed curricular aim has been satisfactorily addressed. The reason underlying the choice of 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 remedy 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 can't tell whether additional instruction appears needed on that skill. The

whole notion of instructionally diagnostic assessment falls flat 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 (or term)—are given meaningful information regarding their incoming students' skills and bodies of 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 it's certain there will be scant instructional dividends from an accountability test.

The rubric to appraise this evaluative dimension (*items per assessed curricular aim*) should be fashioned around teachers' judgments with regard to the *number and representativeness* of the sets of items being used to assess students' status regarding the curricular aims the test purports to measure. Teacher-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 excellent content representativeness based on 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.

*SES/Aptitude contamination.* The fourth and final evaluative dimension for judgmental evidence deals with the degree to which an accountability test's items favor (1) students from more affluent backgrounds or (2) students born with higher academic aptitudes. Clearly, students from higher-SES families will—*in general*—tend to perform better on achievement tests because those students will have typically had access to a richer range of stimuli than will their lower-SES counterparts. Students who, from birth, possess stronger verbal, quantitative, and spatial aptitudes will—*in general*—tend to perform better on achievement tests because those can employ their innate capacities when responding to a test's items. We'll never, of course, be able to *completely* excise the potential influence of SES or inherited academic aptitudes on students' test scores. However, if a student's performance on an accountability tests is *substantially* influenced by that student's SES and/or inherited academic aptitudes, then such

influence will surely diminish the degree to which we can ascribe the student's test scores to instructional impact.

It is important to note that many 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 dominantly comparative measurement mission of traditional achievement tests is to permit comparisons between test-takers' scores. But in order for those comparisons to work properly, there must be a reasonable degree of score-spread among students' tests scores, that is, students' test results must be meaningfully different. Because students' SES and inherited academic aptitudes are both nicely distributed variables, and ones that do not change very rapidly, test items linked to either of these variables do an efficient job in spreading out students' test scores. Accordingly, builders of traditional achievement tests often end up putting a number of these items into achievement tests—including those 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 teacher-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, I think we have a reasonable chance to ferret out the magnitude of such contaminants.

The question we might pose to panelists regarding this evaluative dimension (*SES/Aptitude contamination*) would ask them to rate the degree to which there is "no substantial influence" on students' tests performances of those students' socioeconomic status *and/or* their inherited verbal, quantitative, or spatial aptitudes. At the most positive end of a two-directional quality definition, we could ask panelists to supply high ratings if the contaminating impact of students' SES and/or inherited academic aptitudes was *minimal*. At the least positive end of this evaluative dimension's rubric, low ratings would be assigned to tests where there was thought to be *substantial* impact of SES and/or inherited academic aptitudes on students' test scores.

To review, I have wheeled out four evaluative dimensions I believe teachers, properly oriented, could address with reasonable objectivity. For each of these dimensions, teacher panels would employ carefully crafted 10-point rubrics so that, in the end, four average ratings ranging from a low of 1 point to a high of 10 points would be generated. It is possible, thereafter, that these separate ratings could be amalgamated into a single, overall panelists' rating, but such an amalgamation is not necessary. Were such an amalgamation to be undertaken, it would be possible—if warranted—to differentially weight the importance of the four categories of evidence.

## Empirical Evidence

Let me turn now to the second general category of evidence capable of contributing to a determination of the degree to which an accountability test is instructionally sensitive. I refer to empirical evidence, that is, data derivative from students' actual test performances. I will describe three ways of collecting such evidence. Given sufficient experience in assembling such data, we might—over time—reach the point where we could translate this sort of evidence into indices (on a 10-point scale) analogous to those employed in the assembly of judgmental evidence. That sort of transformation of evidence would make it easier to arrive at any sort of composite estimate of an accountability test's instructional sensitivity.

*“Taught” and “untaught” students.* In all three kinds of empirical evidence I'll be considering, I will be referring to “taught” and “untaught” students. In order for you to make much sense out of the data-based indicators I'll be proffering, I need to describe what I have in mind when distinguishing between taught and untaught students.

Clearly, from the moment children enter kindergarten (and well before), those youngsters have been taught things. All students have, at least to some extent, been taught. The notion of “taught” I will be using, however, refers to the degree to which students have been taught to master the specific skills and/or bodies of knowledge assessed by a particular accountability test. The procedures we might employ to determine whether such teaching has taken place are myriad, especially regarding the rigor with which we ascertain whether test-related teaching has actually transpired. Let me illustrate two such procedures now.

Suppose a state's official curricular guidelines call for the state's fifth-grade students to receive substantial instruction dealing with the mechanics of writing, i.e., spelling, punctuation, grammar, and word usage. Indeed, the state's official fifth-grade language arts accountability test, administered in the late spring near the end of the school year, always includes 8-10 items dealing with the mechanics of writing. In a sense, therefore, especially because teachers know the end-of-year accountability test will contain items on this topic, one could argue that by the end of the school year the state's fifth-graders will have been taught about the mechanics of writing. This strikes me as a very soft way of demonstrating whether students have been taught something.

A more stringent approach to determining whether certain students have been taught what's to be tested might require teachers to indicate, on some form of self-report device, the degree to which they had actually supplied instruction on the skills or bodies of knowledge to be assessed on an accountability test. This information could be gathered from teachers on an overall basis for the entire accountability test, or collected separately for each skill and/or body of knowledge assessed on the test. All sorts of self-report ploys might be used so

increasing confidence could be ascribed to teachers' reports about the extent to which they had taught students on test-assessed curricular aims. Obviously, there would be a self-serving tendency on the part of at least some teachers to allege they had provided meaningful instruction for all state-approved curricular aims. Careful wording of any self-report inventories employed in such inquiries would need to address and, hopefully, minimize such tendencies. It might be possible to establish required levels for teachers' reports so that only those students would be regarded as having been taught if enrolled in the classes of teachers whose reports of test-related instruction exceeded stipulated minima.

I prefer more rigorous tactics for ascertaining that students have truly been taught, but I recognize practicality often precludes the use of the most stringent procedures for identifying students taught well enough so they have a decent chance to master what's measured by an accountability test. Obviously, the manner in which one defines what constitutes a "taught" student will reciprocally define what constitutes an "untaught" student. Let me turn, now, to the first of the three sorts of empirical data I believe bear directly on the instructional sensitivity of an accountability test.

*Item p-values for untaught students.* A test item's  $p$ -value indicates the proportion of students who answered that item correctly. While often erroneously characterized as a reflection of an item's "difficulty," the size of a  $p$ -value is actually dependent on both the difficulty of the item *and* the degree to which test-takers have been taught what the item measures. Think, for example, of an esoteric principle in advanced physics, a principle so abstruse that if everyday citizens who knew no physics were asked to respond to a test item about that principle, the item's resultant  $p$ -value might hover around zero. Yet, suppose at the end of a superb course in advanced physics, a course in which the esoteric principle was skillfully taught, students in the class responded to the very same item. Now the item's  $p$ -value might be .92 (and would have been 1.00 had it not been for several inattentive students who appear to have been better suited for biology classes). Does the zero  $p$ -value make the item a tough one or does the .92  $p$ -value make the item an easy one? Clearly, in addition to the content of the item *per se*, there is the matter of how test-takers have been taught.

The first variety of empirical evidence I recommend, therefore, consists of the  $p$ -values earned by untaught students on an accountability test's items. Obviously, if those  $p$ -values are very high, then there will be little room for students to improve on items that are very easy. Remember, we have defined these students as untaught ones. Hence, it would appear be the intrinsic difficulty of the items that has led to a flock of high  $p$ -values. If untaught students are scoring too well on an accountability test's items, then the test has little chance of detecting improved instruction. In truth, we rarely encounter a setting in which untaught students are scoring excessively well on achievement tests. In such atypical instances, my suspicion is that the supposedly untaught students have actually been taught, and it would seem they've been taught rather well.

Nonetheless, a routine verification that there is sufficient growth-room on an accountability test's items would seem to be the first sort of empirical data we should consider regarding an accountability test's instructional sensitivity.

*Comparing different taught/untaught students.* A second sort of empirical evidence regarding an accountability test's instructional sensitivity consists of the test-performance levels of two groups of students, one of whom has been taught and one of whom has not been taught. What we are looking for, in order to help ascribe instructional sensitivity to an accountability test, is meaningfully higher test scores—on the very same test—by the taught students. If there's no practical difference between the test performances of taught and untaught students, then this is evidence of an accountability test's instructional insensitivity.

The challenge in collecting such evidence, of course, is to find taught and untaught students who are similar in other important respects. It would not do, for example, to contrast the scores of fifth-graders who had been given a solid dose of instruction regarding the mechanics of writing with the scores of fourth-graders who hadn't received such instruction. A full-year developmental difference between the two groups would certainly confound any conclusions about whether instruction had boosted the fifth-graders' test performances.

One advantage of using different groups of students is it is possible to collect the relevant test-score evidence immediately, rather than being obliged to wait for an extended time, as is necessary in the next kind of empirical evidence I'll describe. Use of this approach might depend on a bit of serendipity, however, to find two groups of *similar* students whose chief difference is the degree to which they have received instruction directed toward the skills and/or bodies of knowledge assessed by a particular accountability test. But, if the right kinds of student groups can be located, the resultant evidence from such a data-gathering effort can be compelling.

*Pre-instruction versus post-instruction contrasts.* A final form of empirical evidence regarding an accountability test's instructional sensitivity calls for us to a contrast of the untaught and taught test performances of the very same group of students. This sort of data-gathering, of course, must take place over an often extended period of time. Still, if an accountability test's items indicate that, prior to having received focused instruction on the accountability test's assessed skills and/or knowledge, a group of students floundered but, after such instruction, those same students flew, then this is persuasive evidence supporting the test's instructional sensitivity.

More often than not, a group of students would be assessed at the end of a given grade (e.g., the accountability test for that grade), so this assessment could function as the pretest for those students. Then, at the close of the subsequent

school year, those same students would complete the accountability test specified for their grade level. To the degree that certain of the skills and/or bodies of knowledge are measured on both tests, and intensified instructional attention is given to this curricular content during the subsequent (second) school year, sensible contrasts can be made between the same student's pre-instruction and post-instruction test performances. Higher post-instruction performances, of course, would reflect favorably on the test's instructional sensitivity. Clearly, there would need to be careful attention given to whether an end-of-instruction accountability test administered one year earlier contains sufficient content similarities (and numbers of items) to an end-of-instruction accountability test administered one grade higher (and one year later).

It would also be possible to administer a specially designed test on a pre-instruction basis. To illustrate, we could selectively sample items from an end-of-year accountability test in order to create a pretest to be administered at the beginning of the school year in the fall. Then, after administering the regular accountability test in the spring of the same school year, it would be possible to see if taught students (in the spring) outperformed untaught students (in the fall).

Summing up, the three kinds of empirical data I have described, depending on the particular circumstances involved, can contribute to the determination of an accountability test's instructional sensitivity. It would always be necessary to decide whether those circumstances are sufficiently favorable to go to the trouble of collecting such data. What's required is some good, hard thinking about the meaningfulness of any data collected in a particular context. If the situation at hand realistically precludes the assembly of sensible empirical evidence of regarding a given accountability test's instructional sensitivity, then I recommend total reliance on the sorts of judgmental evidence described earlier.

### **Amalgamation Time**

*A graphic depiction.* I've been describing several sorts of evidence that, separately or in concert, might be used to adjudge the instructional sensitivity of a specific accountability test. Ultimately, a judgment must be made about the degree to which that test is sufficiently sensitive to instruction so its use in an accountability program ends up helping rather than harming students. Ideally, I think that any final judgment should be based on an amalgam of as many of the sorts of evidence I've set forth here that (based on the particular procedures involved) seem contributory to a defensible judgment about a given accountability test's instructional sensitivity.

In Figure 1, I have attempted to portray graphically how the various kinds of evidence addressed here might function in arriving at an amalgamated judgment regarding the accountability test's instructional sensitivity. As you can see, I

---

Insert Figure 1 about here.

---

am suggesting that a weighting operation be carried out prior to the synthesizing of each set of evidence (both judgmental and empirical) prior to its being coalesced in a final, overall judgment regarding an accountability test's instructional sensitivity.

*A reiterated entreaty.* Early on in this analysis, I indicated I was hoping to provide a framework, and most likely a fairly primitive one at that, for arriving at a determination of an accountability test's instructional sensitivity. This is tricky terrain, and I recognize what I've put forth here is in serious need of careful scrutiny and, without question, scads of sprucing up. Near the start of this paper, I invited interested colleagues to join with me in a collaborative attempt to improve this sort of evaluative framework. I now do so again.

The use of instructionally insensitive accountability tests is currently harming hoards of American children. We need to increase our awareness of this problem, and we need to figure out how to provide usable tools for educators and others to tell whether the accountability tests being used are sufficiently sensitive to instruction. I welcome allies in this quest.

FIGURE 1. A FRAMEWORK FOR DETERMINING AN ACCOUNTABILITY TEST'S INSTRUCTIONAL SENSITIVITY

