

# **CONDUCTING INSTRUCTIONAL-SENSITIVITY REVIEWS OF EDUCATIONAL ACCOUNTABILITY TESTS**

W. James Popham  
University of California, Los Angeles  
([wpopham@ucla.edu](mailto:wpopham@ucla.edu))

April 2007

## CONDUCTING INSTRUCTIONAL-SENSITIVITY REVIEWS OF EDUCATIONAL ACCOUNTABILITY TESTS

This document consists of a set of guidelines for carrying out reviews of the instructional sensitivity of educational tests such as the accountability tests so widely used throughout the world these days. The procedures to be recommended here have not been officially sanctioned by any organization. Nor, in candor, have most of the recommended procedures even been carried out in a real-world setting. Rather, the procedures to be suggested here are fundamentally derivative from a variety of judgmentally focused ways that certain aspects of educational tests are currently evaluated. The approach being recommended, therefore, blends together a number of comparable procedures for evaluating tests, but is focused on the appraisal of a test's instructional sensitivity.

The *instructional sensitivity* of an educational test refers to the capacity of a test to determine 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. For example, 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* 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* scores are meaningfully, but not exclusively, attributable to ineffective instruction.

An instructionally *insensitive* test, however, 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.

Because educational accountability programs that employ instructionally insensitive tests not only yield misleading estimates of instructional quality, but also can lead to educationally harmful consequences for students, there is a pressing need to appraise the instructional sensitivity of any test used in an educational accountability program. In a separate analysis, a general introduction to a strategy for gauging an accountability test's instructional sensitivity is set forth (Popham, 2007). In the present set of guidelines, some of the procedural particulars are treated that, had they been included in the aforementioned general introduction, would have been too detailed. The present set of guidelines was written for those who are considering the conduct of an instructional-sensitivity review focused on one or more educational accountability tests. While this exposition will not be exhaustive, it is hoped that it will

supply would-be reviewers of instructional sensitivity with sufficient specifics to get underway.

It should be recognized at the outset that, if even a handful of instructional-sensitivity reviews are carried out—and their results and procedures shared, we will most certainly be able to make substantial improvements in the procedures set forth here. For example, given the current absence of a body of results regarding the manner in which certain evaluative dimensions of a test's instructional sensitivity are apt to be judged, we have no way of determining just how positively a test ought to be rated in order to arrive at sensible conclusions about a test's instructional sensitivity. But, in time, we should be able to assemble collections of evaluative data that will make it easier to reach subsequent decisions about the degree to which a test really must be sensitive to instruction. Comparative results from numerous instructional-sensitivity reviews, even if collected from reviews that were not completely identical, can nonetheless be illuminating. Procedural refinements, of course, will most surely be seen as more and more instructional-sensitivity reviews are carried out.

### **A Judgment-Based Approach**

As you will soon see in the remainder of this document, what will be recommended is an exclusively judgment-based approach for determining the instructional sensitivity of a given educational test. In essence, a panel-focused strategy is being proposed because it represents a practicable, low-cost way to estimate how instructionally sensitive a particular accountability test really is. Four important evaluative dimensions will be recommended for use by an *instructional-sensitivity panel* (ISP) so that, having focused on those evaluative factors in separately, the ISP can arrive at a defensible overall judgment about how instructionally sensitive a particular accountability test seems to be.

There is, however, an alternative approach to the determination of an accountability test's instructional sensitivity, and it involves the collection of test-takers' actual performances under specified conditions. In contrast to the *judgmental* approach being recommended here, an *empirical* strategy for estimating a test's instructional sensitivity has much to commend it. Empirical approaches to this problem, however, typically require far more resources and effort to implement. Thus, I recommend empirical strategies for determining a test's instructional sensitivity only as a *confirmatory* process, namely, as a way of corroborating the judgments reached by an ISP. Whereas ISP-based procedures can be carried out with only modest resources, empirical confirmations typically have a much higher hassle-index. Nonetheless, in selected instances it makes a good deal of sense for those who rely chiefly on a judgment-based approach to find out, via empirical studies, whether their judgment-based approaches to instructional sensitivity are on the mark.

## **Who Initiates an Instructional-Sensitivity Review?**

Two groups of educators are most likely to undertake an instructional-sensitivity review. First, there are staff members of a state education agency (SEA), or even the school board of a local education agency (LEA), who might wish to arrive at a judgment regarding the instructional sensitivity of an accountability test that's currently in use. The same individuals might also wish to appraise the instructional sensitivity of an under-development accountability test. However, in addition to governmental groups, there are many educational organizations such as teachers' unions or professional groups such as the Association for Supervision and Curriculum Development (ASCD). These nongovernmental groups, of course, have a vital interest in the nature of the accountability tests now being used to evaluate the caliber of educators' efforts.

Whereas governmental agencies might carry out instructional-sensitivity reviews with the intent of either reporting their conclusions to a governing board such as a state board of education, a nongovernmental association might conduct an instructional-sensitivity review with a view to reporting not only to the association's membership, but also to the public at large. Although there are apt to be minor differences in the procedures used by the two groups, particularly because governmental agencies will have greater access to secure test-items, the overall approach to this problem by either governmental or nongovernmental groups should be fundamentally similar.

## **What's the First Step?**

An instructional-sensitivity review gets underway soon after a group decides it wants to undertake such a review. In arriving at that decision, the group must surely decide why it wishes to carry out such an endeavor. That is, there must be a good answer to the question: *Why do this?* If, for example, officials of an SEA have seen that their state's educational reform-initiatives have not had a meaningful impact on improving students' scores on the state's accountability tests, is this because the reform-initiatives were ineffectual or because the accountability tests being used were instructionally insensitive? Perhaps the SEA officials involved, if it turns out that the accountability tests are insensitive, are willing to replace those tests with brand new, instructionally sensitive accountability tests. This sort of situation would provide a strong justification for an SEA to conduct an instructional-sensitivity review of the accountability tests currently being used in their state.

As an example of why a nongovernmental group might wish to carry out an independent instructional-sensitivity review, assume that an ASCD chapter in a given part of a state has seen its members often castigated in the local media because students' scores on the state accountability tests have remained essentially stagnant for several years running. The ASCD chapter's officers might decide to collect evidence regarding the state tests' instructional sensitivity with the intent of widely publicizing results of the review. The officers' obvious hope is to inform the

public that the state's educators are being judged on the basis of tests that are essentially impervious to the impact of even first-rate instruction. Clearly, if such is the intent of a nongovernmental organization, then procedural steps must be taken to enhance the credibility of the instructional-sensitivity review, for example, by including on any ISP a meaningful number of non-educators. If the state's test turns out to be instructionally insensitive, but the public regards the instructional-sensitivity review as little more than a home-grown and self-serving exercise by educators who are fleeing accountability, then the review's original intent will not have been realized.

The point here, of course, is that the purpose(s) of an instructional-sensitivity review will often influence the nature of the way that the review itself is conducted. The greater the clarity regarding why an instructional-sensitivity review is being conducted, the more defensible will be the review staff's initial and en route decisions.

### **The Instructional-Sensitivity Panel**

There is no such thing as the "appropriate" size of an ISP. Rather, designers of instructional-sensitivity reviews must take their cues from the sizes of review panels that perform similar functions related to various aspects of a test's quality. For example, during the last two decades when the actual items for high-stakes educational tests are being reviewed by panels of educators to judge, say, the alignment of those items with a state's content standards or, perhaps, to judge the presence of item-bias against particular student subgroups, those panels have typically ranged between 15 and 25 members—sometimes smaller than 15 and sometimes larger than 25. However, because the conclusions reached by such panels have generally been accepted by others, it has become *conventional* to see panels of between 15 and 25 members performing certain judgmentally oriented test-appraisal tasks. The size of an ISP, therefore, probably ought to be, at minimum, about 15 members. If resources permit, and panels of more than 25 members are employed, there are few disadvantages associated with the use of a larger ISP.

The individuals who are invited to take part in an ISP, of course, are critical. It is possible to have a resplendent set of procedures for judging an accountability test's instructional sensitivity using panel-judgments, but if the wrong kinds of persons constitute an ISP, all the procedural finesse in the world will not prevent the instructional-sensitivity review from floundering. Thus, panelists must be chosen who are manifestly capable of rendering the kinds of judgments to be described later in this document. The individuals coordinating the selection of ISP members should first become thoroughly familiar with each of the ratings that members of an ISP will be called on to make during an instructional-sensitivity review, and only then select panelists who possess sufficient expertise to render those ratings. Care should be taken to choose panelists whose institutional affiliations or whose previously expressed opinions will either (1) make it less likely that their judgments will be

objective or (2) make it appear to others that those panelists' judgments may not be objective. To illustrate, if an ISP were formed whose panelists were all members of a local chapter of a teachers' union that has been frequently on record as opposing *all* forms of accountability testing, then results of that ISP might, if negative, be seen as more of a prejudiced repudiation of accountability testing itself rather than as an even-handed judgment of a given test's instructional sensitivity. In short, both the competence and the perceived credibility of panelists should be considered when constituting an ISP. What is being sought is an honest and accurate appraisal of a test's instructional sensitivity. Panelists must be selected with that mission in mind.

### **Time Needed for Panel Deliberations**

As will soon be seen, the instructional-sensitivity review being recommended here revolves around four evaluative dimensions. One of these dimensions calls for a set of item-by-item judgments based on three separate judgments per item. Clearly, if *all* of an accountability test's items are to be reviewed in this manner, a substantial amount of time will be required for an ISP to conclude its work. However, it may be possible for a separate group of item-reviewers to have already engaged in the needed series of item-by-item judgments which, then, can be presented to ISP members in summary form. If such is the case, then it seems certain that all of the necessary judgments to be made by an ISP about a particular test can be rendered in one day. If the members of the ISP must review a substantial number of items themselves, then a meeting time of two or more days (depending on the number of items to be reviewed) would be required.

Because the review of an accountability test's actual items often requires a setting in which security-control procedures can be employed, such as in the offices of an SEA, it is apt to be more manageable if per-item reviews are carried out separately, then presented to the ISP who would be able to use those reviews to arrive at their own judgments regarding this evaluative dimension.

Another possibility for per-item reviews is to present an ISP with a randomly selected *sample* of a test's items for appraisal. To illustrate, if a mathematics accountability test contains 60 items, a truly random sample of items might identify 20 of the items for ISP review. Thus, a one-third sample of the items might be regarded as sufficient for purposes of judging the items. If these sorts of samples are employed, it might still be possible for an ISP to accomplish its tasks in a single day of, perhaps, six or seven serious-work hours.

### **Materials Needed**

Those individuals moderating the instructional-sensitivity review must make certain that all panelists have access to every piece of information they will need during the review itself. Thus, for example, one of the evaluative dimensions to be considered in the review process recommended here involves a panelist's making a judgment regarding the clarity of the descriptive information depicting what a test is supposed

to be assessing. Panelists are to consult whatever descriptive information (about eligible assessment targets) is routinely available to teachers, then render a 1-to-10 judgment about the clarity of such descriptions. Clearly, it would be necessary to make sure that copies—one per panelist—of all routinely available descriptions of test-eligible curricular aims would be made available.

Similarly, those conducting the instructional-sensitivity review would need to think through, in advance, every single step in the upcoming process to make sure that sufficient materials are available. For example, there will be panelist review forms to be completed, and rubrics to be provided that relate to each evaluative dimension for which a panelist's judgment is to be made. Sufficient numbers of response forms and rubrics, therefore, must be on hand. Similarly, all materials necessary during an introductory orientation/training session must be available. If the ISP is to review secure test-items, then some type of security, non-disclosure form is typically required.

Summing up, in advance of the instructional-sensitivity review, those moderating the session must consider every step of the upcoming process, then make certain that sufficient copies of all needed materials are ready to provide.

### **Orientation/Training**

At the outset of any ISP process, care must be taken to orient all panelists to what is expected of them. Not only should a complete description of the process be provided, but panelists should be familiarized with the rating forms and rubrics they will be using. Panelists' questions should be answered, of course, and panelists should feel comfortable with what's ahead of them. It is suggested that, during the orientation, copies of both the rating form to be used and the four separate rubrics should be distributed to the ISP. A copy of a suitable rating form is supplied on page nine of this document. Copies of the four rubrics can be made from the document.

As an overview, the language below might be used as is by a moderator (or modified to mesh better with a specific ISP activity). Typically, each panelist will also be given a written copy of this type of information. [Suggestions to moderators, presented here in brackets, should be removed from any materials distributed to panelists.]

*Today you will be taking part in an important activity, namely, a review of the instructional sensitivity of (name of test). The instructional sensitivity of a test refers to the ability of that test to determine whether test-takers have been effectively or ineffectively taught what's being measured by the test. Tests vary in the degree to which they are instructionally sensitive. That is, few tests are completely sensitive to instruction or completely insensitive to instruction. Your task today is to identify the degree to which the test you will be reviewing is instructionally sensitive or, conversely, instructionally insensitive. It may be helpful for you to think of a test's instructional sensitivity as the sort of 1-to-10 continuum seen in Figure 1.*

(Insert Figure 1 about here.)

*You will be reviewing the test on the basis of four evaluative dimensions, each of which will be described to you shortly. For each of these evaluative dimensions you will individually decide whether the test—on that particular evaluative dimension—should receive a 1-to-10 rating, with a rating of 10 being high and a rating of 1 being low. After panelists' separate ratings have been made for all four evaluative dimensions, each panelist will be asked to consolidate those four ratings in to arrive at a composite 1-to-10 rating representing the test's overall instructional sensitivity.*

*A two-directional rubric will be supplied to assist you in arriving at your 1-to-10 ratings for each of the four evaluative dimensions. Each rubric indicates what would typically be involved for a panelist to assign a top rating, that is, a 10, as well as a bottom rating, that is, a 1, for a particular evaluative dimension. Although a ten-point rating scale is to be used, the scale is intended to elicit only approximate judgments from panelists. Each rubric will be accompanied by four examples of how a panelist might have rated a fictional accountability test on that evaluative dimension.*

*In general, the panel will deal with one evaluative dimension at a time, and the process to be used will follow a multi-step model in which (1) panelists supply an initial rating; (2) panelists discuss, as a group, their individual ratings; (3) panelists then, if they wish, alter their initial ratings; and (4) these final ratings are then summarized for the entire panel on that particular evaluative dimension.*

*This procedure will be repeated for all four evaluative dimensions. Then, in a similar manner, panelists will each provide an initial overall judgment about the test's overall instructional sensitivity based on that panelist's four earlier ratings. After a group discussion of panelists' initial overall ratings of the test's instructional sensitivity, panelists will be given an opportunity to change their initial overall ratings if they wish to do so. At that point, all panelists' final overall ratings will be collected, and a summarization of those individual ratings will represent the panel's judgment about the instructional sensitivity of (name of test).*

*Are there any questions about how the overall panel procedures are to operate? [Answer any questions.] All right, then, it is time to describe the four evaluative dimensions to be used in arriving at your panel's judgment regarding the test's instructional sensitivity. [If possible, list the four evaluative dimensions in the following paragraph on a chalkboard or flip-chart.]*

*The four evaluative dimensions to be used in today's instructional-sensitivity review are (1) the number of curricular aims assessed, (2) the clarity of assessment targets, (3) number of items per assessed curricular aim, and (4) item sensitivity. Although these evaluative dimensions will be considered later, this is what each one means.*

*First, the number of curricular aims assessed focuses on whether there are too many curricular aims being measured by a test—so many, in fact, that teachers are*

*unable to foresee what a test is really going to measure. The second evaluative dimension, clarity of assessment targets, refers to the way in which teachers can discern what is to be tested, that is, whether the descriptive materials accompanying a test communicate clearly about what's to be assessed. Third, the number of items per assessed curricular aim relates to whether there are enough test-items to let teachers and students know whether a student has mastered each skill or body of knowledge being assessed. And the fourth evaluative dimension, item sensitivity, focuses on whether the items on a test are more likely to assess what a student has been taught in school rather than what students might have acquired elsewhere.*

*Let's turn now to a consideration of the first of our four evaluative criteria, namely, the number of curricular aims assessed. After discussing this evaluative dimension in sufficient detail, you will then supply the first of your ratings regarding this particular aspect of a test's instructional sensitivity.*

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

At this point, the persons in charge of the instructional-sensitivity review would distribute any official information indicating how many curricular aims are eligible for assessment by the test under review. For example, if the only information a state's teachers are given regarding what is to be tested turns out to be the state's content standards themselves and the benchmarks that fall beneath each of those content standards, then this is the information that would be supplied to panelists. If there is other information provided to teachers regarding what is eligible to be assessed, then such information would also be made available to the ISP. The idea is to give panelists all readily available information so they can arrive at an accurate judgment regarding how many curricular targets are "fair game" for assessment.

The most useful kind of information to be supplied to panelists would allow panelists to determine, at the "grain-size" of most meaning to teachers related to their instructional decision-making, whether there are relatively few curricular aims eligible for assessment or a relatively large number of curricular aims eligible for assessment. The most appropriate "grain-size" of a curricular aim in this context is one that allows teachers to design instruction aimed directly at students' accomplishment of the skill or body of knowledge embodied in that curricular aim. Curricular aims formulated at such a broad grain-size that teachers cannot make accurate instructional plans for that aim are typically, but not always, spelled out in more detail with curricular aims representing a smaller grain-size. It would be these smaller, more instructionally meaningful curricular aims on which an ISP focuses.

The rating scale for this evaluative dimension, as for all four evaluative dimensions, ranges from a high of 10 to a low of 1. Presented below is the two-dimensional rubric panelists are to use in arriving at their 1-to-10 rating on *the number of curricular aims assessed*. To supply their ratings on this evaluative dimension, panelists must have access to the information routinely distributed to teachers regarding the curricular targets eligible to be assessed by an accountability test. In addition, of course,

# INSTRUCTIONAL-SENSITIVITY PANEL RATING FORM

Your Name \_\_\_\_\_ Date \_\_\_\_\_

Name of Test Being Reviewed \_\_\_\_\_

<b>Evaluative Dimensions</b>	<b>Initial Rating</b>	<b>Final Rating</b>
Number of Curricular Aims Assessed	_____	_____
Clarity of Assessment Targets	_____	_____
Number of Items per Assessed Curricular Aim	_____	_____
Item Sensitivity	_____	_____
_____ SES Influence		
_____ Inherited Academic Aptitudes		
_____ Responsiveness to Instruction		

---



---

	<b>Initial Rating</b>	<b>Final Rating</b>
Overall Instructional Sensitivity	_____	_____

copies of the two-directional rubric and its four “pre-rated” examples must be available to panelists.

After panelists have been given copies of the necessary materials, the individual who is coordinating the instructional-sensitivity review should spend several minutes making sure that panelists have a clear understanding of their rating assignment. It

will be useful to remind panelists that the 1-to-10 rating scale is not an inordinately precise one, and that what is really being sought in panelists' ratings of this evaluative dimension, and all others, is simply the most accurate rating a panelist can make. Those ratings will, of course, often be approximations.

### **Rubric: Number of Curricular Aims Assessed**

**Higher ratings are to be given when the collection of curricular aims eligible to be assessed by a test constitutes a sufficiently modest number so that teachers can effectively promote students' mastery of all those aims in the instructional time available.**

**Lower ratings are to be given when the set of curricular aims potentially assessable by a test is so large that there is insufficient time for teachers to promote students' mastery of all those aims in the available instructional time.**

*Example 1: If a state's accountability test in reading measures students' mastery of only eight curricular aims, that is, eight reading skills, in grades 3-8 and 11, and those skills are stated at an instructionally addressable grain-size, a rating of 9 or 10 might be appropriately given.*

*Example 2: Suppose a state has attempted to winnow the number of curricular aims it regards as suitable for annual testing, and has ended up with 20-25 possible aims in both mathematics and reading, each subject to be assessed on the state's accountability assessments. Because, for elementary teachers, this translates into a range of 40-50 potentially assessable curricular aims—too many for most teachers to address effectively—a rating of 5 or 6 might be suitable.*

*Example 3: If a provincial accountability test is designed to measure students' mastery of only six content standards in mathematics, for instance, geometry, but beneath those content standards there are 59 separate "indicators" representing the six content standards, then the province's teachers will not really know which of the 59 indicators will be measured on a given year's test. This provincial accountability test, therefore, might receive a rating of, perhaps, 3 or 4 on this evaluative dimension.*

*Example 4: For a test alleged to measure students' mastery each year of more than 100 mathematics objectives and 100 reading objectives, a rating of 1 or 2 would be in order.*

## Clarity of Assessment Targets

Let's turn now to the second evaluative dimension for which members of an ISP are to supply ratings, namely, the clarity of the targets to be assessed. To supply their ratings on this dimension, panelists must be given not only the rubric presented below, but copies of all information readily accessible to teachers regarding the nature of the skills and/or bodies of knowledge eligible to be assessed. As before, prior to asking panelists to supply their ratings, time should be allowed for panelists to raise any questions they might have regarding the nature of this rating.

### Rubric: Clarity of Assessment Targets

**Higher ratings are to be given when descriptions of what is to be assessed communicate so clearly to teachers regarding the nature of the skills or knowledge to be measured that teachers can accurately design instruction to promote students' mastery of those to-be-assessed skills and/or bodies of knowledge.**

**Lower ratings are to be given when whatever descriptive information about what's eligible to be assessed is so fragmentary or opaque that teachers would have great difficulty understanding the nature of the skills or bodies of knowledge that might be assessed.**

*Example 1: If a state's annual accountability test is accompanied by explicit descriptions of every skill or body of knowledge to be measured each year, and those descriptions are not only fairly brief, but also communicate unequivocally to teachers about the essence of the test's assessment targets, a rating of 9 or 10 would probably be warranted.*

*Example 2: Imagine that a state had invested substantial energy in laying out a set of "test specifications" for use by the state's test-development contractor in the construction of test-items. State officials then had distributed this set of test specifications to the state's teachers. Although those specifications describe the nature of what's to be assessed, they do so in such elaborate detail that few teachers would have the patience to wade through those highly delineated specifications. In this instance, a rating of 5 or 6 might be properly given.*

*Example 3: If a state annually releases a percent of previously used test items to the state's teachers along with an explanation that "subsequent items are apt to be similar to these released items," the state's teachers will typically be obliged to carefully analyze the nature of those items in order to infer the nature of the skills or bodies of knowledge being measured. Moreover, because future items are only "apt" to mirror the released items, a rating of 3 or 4 might be warranted in this instance.*

*Example 4: When a state supplies no elaboration of what is to be assessed by its annual accountability tests other than the rather general list of the state's numerous*

*content standards, a rating of 1 or 2 would seem appropriate for this evaluative dimension.*

### **Number of Items per Assessed Curricular Aim**

The next evaluative dimension an ISP should address is the number of items per assessed curricular aim. To render an accurate judgment on this dimension, panelists will need to know how many items are assigned to gauge a test-taker's status with respect to *each* skill or *each* body of knowledge being assessed. In some settings, this will mean that panelists must be given whatever test specifications govern the particularized inclusion of items on a given test form. In instances where there are so many assessment targets to be measured, and some of those targets are measured by only one or two items—or none at all on certain test forms—there would be less need for panelists to examine the test specifications (given the paucity of items per assessed curricular aim). Obviously, this evaluative dimension interacts strongly with the initial evaluative dimension regarding the number of curricular aims being assessed.

What is particularly of concern in connection with this third evaluative dimension is the grain-size of the assessed curricular aim in relation to the actual number of items assigned to assess that skill or body of knowledge. Smaller-scope curricular aims can be more satisfactorily assessed by fewer items than can broader-scope curricular aims. However, for a given curricular aim to be assessed with sufficient accuracy that instructional decisions (about a student, for example) can be made, the curricular aim must be *adequately represented* by the items assessing it. For instance, we might think of a curricular aim in geometry dealing with triangles for which there are eight items on each year's accountability test. At first glance, this number of items seems somewhat reasonable. However, on further scrutiny, it turns out that all eight items deal only with isosceles triangles—no other kinds of triangles. Clearly, the triangle terrain would not have been properly represented by those isosceles-only items.

An ISP must have access to descriptions of the nature of the items being used to assess each curricular aim or, preferably, be given an opportunity to review sets of current or previously used items that are being employed to measure a student's mastery of a curricular aim. Such materials should be made available to panelists prior to their tackling the assignment of ratings related to this evaluative dimension.

As always, the ISP should have access to the necessary rating form as well as the rubric related to this evaluative dimension. An opportunity for panelists to seek task-clarification should be provided after panelists have had an opportunity to review the rubric.

## **Rubric: Number of Items per Assessed Curricular Aim**

**Higher ratings are to be given when there appears to be a sufficient number of items per curricular aim for instructional decision-making purposes *and* when those items, as a group, appear to satisfactorily represent the nature of the curricular aim being measured.**

**Lower ratings are to be given either when there are too few items to provide a reasonable estimate of a student's mastery of the assessed curricular aim or when the items, if sufficiently numerous, do not satisfactorily represent the curricular aim being measured.**

*Example 1: Suppose that an accountability test contains between 6 and 11 items per assessed curricular aim, and that the number of items per aim was determined by a panel of experienced teachers who suggested different numbers of items for different aims, but who insisted that each collection of items adequately represent the content of the curricular aim being assessed. Assuming that the directives of these experienced teachers were followed, this test would probably get high marks of, say, a 9 or 10 on this evaluative dimension.*

*Example 2: Assume that a state has identified approximately 15 "indicators" subsumed by the state's content standards in a given subject, and has insisted that each indicator be measured by five items so that, insofar as possible, the content of the indicator is well represented by the five items. Although there is clearly an attempt in this instance to provide a certain number of items per assessed curricular aim, it is unlikely that genuine representativeness of all curricular aims can be based on only five items per aim, so this test might warrant a rating of 7 or 8 on this dimension.*

*Example 3: Suppose a state has no requirement for a curricular aim to be assessed by a certain number of items, but there is a stipulation that every assessment-eligible curricular item be measured by at least one item. If this requirement leads to at least half of the assessed curricular aims being measured by only a single item, this test would probably be given a rating of 3 or 4.*

*Example 4: If a provincial accountability test contains no stipulation that every one of its 47 curricular aims be assessed by a minimum number of items, and the result is that many of the 47 aims are not assessed at all or, possibly, are assessed by only one or two items, a rating of 1 or 2 might be appropriate here.*

### **Item Sensitivity**

Turning now to the final evaluative dimension, item sensitivity, a somewhat different procedure is involved. Panelists either must themselves review a representative set of items from an accountability test, or they must have access to the results of item reviews made by another group of individuals (referred to here at the item-

reviewers). If item-reviewers are used, instead of the ISP, then the same procedures would be used by the item-reviewers as are described here.

For this evaluative dimension, the focus is on the items themselves, and the task is to discern how sensitive to instructional impact a set of items actually is apt to be. This means, of course, that actual items must be appraised by panelists (or their item-reviewer surrogates) in order to arrive at a conclusion about item sensitivity. Panelists can either review actual items from an operational form of a test or can review released items from earlier forms of the test. The former approach, of course, would require the use of security-monitored procedures.

For each item reviewed by an instructional-sensitivity panel (or by a different, designated group of item-reviewers) three separate judgments must be rendered. These judgments take the form of a Yes, No, or Not Sure response to each of the following three questions:

- *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?*

Thus, each item being reviewed by an ISP would have a three sensitivity indices computed, one each for the three factors cited above, that is, (1) SES influence, (2) inherited academic aptitudes, and (3) responsiveness to instruction. These three indices would simply be the percentage of panelists' No responses to the first two questions and the percentage of panelists' Yes responses to the third question.

Before beginning this phase of the work of an ISP (or of item-reviewers), clarification should be provided regarding the meaning of each of the three questions involved. To illustrate, each question should be dissected, word-for-word, so that a panelist understands the importance of the modifier "dominantly" in the first two questions. Clearly, a student's SES background and inherited aptitudes will have an influence on the way that student responds to many test questions. What this question attempts to get at, however, is the degree to which one or both of those factors is the *overriding* reason a student's response to a given item is likely to be correct or incorrect.

Similarly, the third question needs to be clarified so that panelists recognize why an item might not be instructionally sensitive. For instance, perhaps the item being

reviewed measures students' understanding of an esoteric rather than a central element of the assessed curricular aim. Thus, even if students were well-taught regarding the heart of this curricular aim, those students might not be apt to respond correctly to an item dealing with this exotic aspect of the aim.

Having spent sufficient time clarifying the meaning of each of the three review questions, and allowing panelists to raise any questions they have, then panelists are to engage in a series of two-step review of sets of, say, ten items at a time. Each panelist answers all three questions (Yes, No, or Not Sure) for ten items at a time. At that point, the moderator goes briefly through that set of ten items, asking for one panelist to indicate aloud why that panelist had given a Yes or No response to one of the three questions. Panelists having a different view would be asked to say why they disagreed with the first panelist. Then, if a *brief* discussion seems warranted, the other two questions for that item would be treated in a similar manner. Obviously, if all panelists agree, there will be no need for any discussion. Panelists will be encouraged to change their original answers to any of the three questions for any of the items if the comments of other panelist have inclined them to do so.

After all items have been reviewed, in groups of approximately ten items, then each panelist counts the number of Yes, No, and Not Sure responses for the entire set of items. The percent of No responses to the first two questions and the percent of Yes responses to the third question are then calculated for the entire ISP. These three total-panel percentages should then be considered by panelists as they render their final judgments regarding the fourth evaluative dimension, that is, item sensitivity. In this instance, individual panelists are apt to be influenced not only by the total panel's judgments, but also by their own personal responses to the three per-item questions. Ratings on the fourth evaluative dimension, then, are to be made using the following rubric:

#### **Rubric: Item Sensitivity**

**Higher ratings are to be given when responses to all three item-review questions indicate that a test's items are apt to detect the presence of effective instruction.**

**Lower ratings are to be given if panelists believe, based on one or more of the three item-review questions, a test's items are unlikely to detect the presence of effective instruction.**

*Example 1: If a test's items contain few items dominantly influenced by SES or inherited academic aptitudes (for instance, fewer than 10 %) and the test also has many items (for instance, more than 75 %) that seem likely to detect the presence of effective instruction, the test might be assigned ratings of 9 or 10 on this evaluative dimension.*

*Example 2: Suppose a test's item-reviews indicated that only a small percent of the test's items are dominantly influenced by SES or inherited academic aptitudes, but that only about half of the items seemed responsive to instruction, that test might be given a rating of 7 or 8.*

*Example 3: If a substantial portion of a test's items, say, 25-30 %, have been rated as being dominantly influenced by SES or inherited academic aptitudes, and only about the same percentage of items have been judged to be responsive to instruction, then a rating of 4 or 5 might be warranted.*

*Example 4: If more than half of a test's items have been judged to be such that a student's responses will be dominantly influenced by SES, inherited academic aptitudes, or both—and only a small proportion of the test's items have been judged to be responsive to instruction, then a low rating or 1 or 2 would seem appropriate for this evaluative dimension.*

### **Combining Panelists' Ratings for the Four Evaluative Dimensions**

At this point in the instructional-sensitivity review, it is necessary for panelists to coalesce their separate 1-to-10 ratings in order to arrive at an *initial* overall rating of the test's instructional sensitivity. If ISP members wish to do so, and if no directives have been given to weight any of the evaluative dimensions more heavily than others, the most straightforward approach to this is to simply compute an arithmetic average of the four separate 1-to-10 ratings by adding all four together, then dividing by four. The overall rating is also to be supplied on a 1-to-10 scale, so panelists should round their average ratings to the nearest whole number.

Next, panelists are to submit their completed rating forms to the session moderator so that all panelists' initial overall ratings can quickly be compiled, then revealed to the entire panel. Any clear representation method can be applied such as a simple numeric listing of all panelists' overall ratings on a chalkboard or flip-chart. Similarly, a 1-to-10 graph such as seen in Figure 2 can be used to display all panelists' ratings by simply placing an X or dot at the score point where each panelist's overall rating fell.

(Insert Figure 2 about here.)

After all panelists have seen the entire array of panelists' initial overall ratings, the moderator should encourage one or more panelists whose ratings are the highest *and* the lowest to speak briefly in favor of higher or lower ratings. A brief discussion may ensue. Thereafter, all members of the ISP are to provide their second and *final* overall ratings of a test's instructional sensitivity. A panelist's final rating may be the same or different than that panelist's initial rating.

These final ratings, then, are reported as an ISP-determined representation of the instructional sensitivity for the accountability test under consideration.

### **Confirmatory Empirical Evidence**

We can turn now, briefly, to the second general category of evidence capable of contributing to a determination of the degree to which an accountability test is instructionally sensitive. If resources permit, such evidence can be collected as a way of confirming the accuracy of any judgmental approaches to determining a test's instructional sensitivity. Empirical evidence must be collected, that is, data derivative from students' actual test performances. Three ways of collecting such evidence will be briefly described here, although other data-gathering strategies are certainly possible (e.g., D'Agostino, et al., 2007). Given sufficient experience in assembling such data, we might—over time—reach the point where we could translate this sort of evidence into indices analogous to those employed in the assembly of judgmental evidence.

*“Taught” and “untaught” students.* In all three kinds of empirical evidence to be considered here, there will be references to “taught” and “untaught” students. In order to understand the data-gathering approaches to be described, it is important to clarify how a reasonable distinction can be drawn between taught and untaught students.

Clearly, from the moment children enter kindergarten (and well before), those youngsters have already learned things. All students have, at least to some extent, been taught. The notion of “taught” to be used here, 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 conclude whether test-related teaching has actually transpired. Let's consider 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 that the end-of-year accountability test will contain items on this topic, one could infer that, by the end of the school year, the state's fifth-graders will have been taught about the mechanics of writing. This seems to be, however, a remarkably lenient way of demonstrating that 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 a painstakingly devised 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 that reasonable 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 "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, thereafter, be possible to establish required levels of instructional emphases so that only those students would be regarded as having been taught if they were enrolled in the classes of teachers whose reports of test-related instruction exceeded stipulated minima.

More rigorous tactics for ascertaining that students have truly been taught are patently preferable, but practicality often precludes the use of the most stringent procedures for identifying students who've been 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's turn, now, to the first of the three sorts of empirical data that 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. Consider, for example, 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 had been 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 seem better suited for biology courses). 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 well the test-takers have been taught.

The first variety of empirical evidence, 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 such items. 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, one ready interpretation 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 could routinely 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 chronological year's worth of developmental differences 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 that it is possible to collect the relevant test-score evidence immediately rather than being obliged to wait for an extended time period, as is necessary in the next kind of empirical evidence to be described. Comparing different taught and untaught students, of course, will depend on a hefty dose of serendipity 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 a contrast of the untaught performances with the taught 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 flopped 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., by using the accountability test routinely administered at that grade level), so this assessment could function as the next year's *pretest* for those students. Then, at the close of the subsequent school year, those same students would complete the accountability test specified for the end of their new 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 in-between 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 sufficiently similar content (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) had outperformed untaught students (in the fall).

Summing up, the three kinds of empirical data described here, depending on the particular circumstances involved, can contribute to the determination of an accountability test's instructional sensitivity. Thus, comparisons could be made between the judged instructional sensitivity of a test and the estimate of that same test's instructional sensitivity as determined by one or more empirical procedures.

As indicated earlier, however, the primary strategy for ascertaining the degree to which an accountability test is instructionally sensitive should be judgmentally based. With few exceptions, empirical approaches to instructional sensitivity should be strictly confirmatory.

#### References

- D'Agostino, J.V., M. E. Welsh, and N.M. Corson, "Instructional Sensitivity of a State's Standards-Based Assessment," *Educational Assessment*, Volume 12, Number 1, 2007, 1-22.
- Popham, W.J., *Instructional Insensitivity of Tests: Accountability's Dire Drawback*, a presentation at the annual meeting of the American Educational Research Association, Chicago, Illinois, April 9-13, 2007.

Figure 1. A Continuum of Instructional Sensitivity

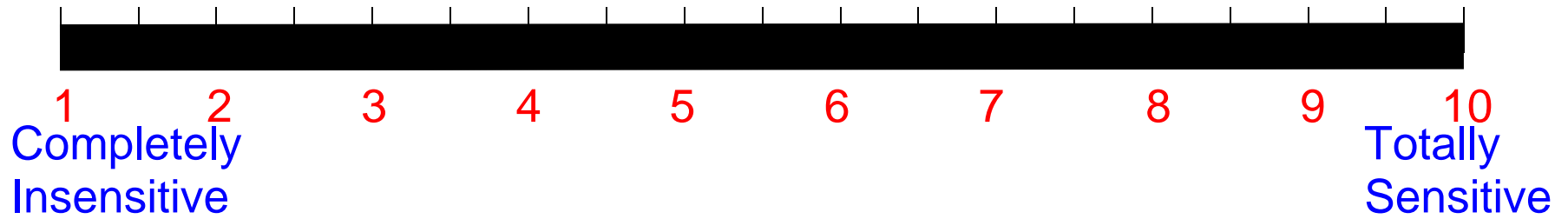


Figure 2. Panelists' Initial Ratings

