ARE YOUR STATE'S NCLB TESTS INSTRUCTIONALLY *INSENSITIVE*? HERE'S HOW TO TELL!

W. James Popham
University of California, Los Angeles

Abstract

Contending that if a state's officials implement the *NCLB Act* with instructionally *insensitive* tests, that state's students will be educationally harmed, three qualities of instructionally *sensitive* tests are identified. Such tests will provide the state's teachers with (1) clear descriptions of what is to be assessed, (2) an intellectually manageable number of assessment targets, and (3) instructionally informative results. In contrast, instructionally *insensitive* tests will lack one or more of these qualities. Suggested ways of determining whether state NCLB tests are deficient on each of these three counts were offered. Concluding the analysis was a plea for instructionally sensitive NCLB tests to assess truly significant curricular aims.

The No Child Left Behind (NCLB) Act has imposed the need for a state's educators to promote substantial annual improvements in students' scores on state-determined NCLB achievement tests. If those improvements do not occur, then any deficient schools and/or districts will be identified as having failed to achieve adequate yearly progress (AYP). After two consecutive years of AYP failure, schools or districts receiving federal NCLB funds are placed on an improvement track linked to increasingly severe sanctions.

If a state's educational decision-makers have chosen to install *instructionally insensitive* NCLB tests, that is, tests incapable of detecting improvements in teachers' instructional activities, then the undergirding rationale for test-based AYP obviously becomes senseless. How can educators be expected to promote students' substantial improvement on tests essentially unable to gauge the impact of either good or bad teaching?

Instructionally insensitive NCLB tests, therefore, will almost certainly lead to the inaccurate evaluation of a state's schools and districts. And, just as surely, instructionally insensitive tests will subvert the praiseworthy educational aspirations of the congressional lawmakers who crafted the NCLB Act. Instructionally insensitive NCLB tests will certainly cause rampant labeling of a state's educators as ineffectual. But what if those labels are wrong?

Because a state's teachers will be under enormous pressure to increase students' scores on their state's NCLB tests, but will be unable to do so because of the instructional insensitivity of those tests, some teachers will be driven to reprehensible classroom practices such as excessive drilling or blatantly dishonest test-preparation

and test-administration. Students who are on the receiving end of such tawdry classroom practices will surely be the recipients of a lower-quality education. Indeed, the serious educational harm visited on those students may be irreparable.

If instructionally *sensitive* NCLB tests are in place, however, then a state's teachers can design and deliver the best instruction of which they are capable. That's because high-quality instruction will lead to higher scores on instructionally sensitive tests. Effective instruction will be identified so it can be applauded. Ineffective instruction will be identified so it can be improved.

In short, if the NCLB Act is to have a positive, rather than negative, impact on the education of a state's students, the state *must* use NCLB tests that are instructionally sensitive. If your state is currently using instructionally *insensitive* NCLB tests, or is planning to install such tests, you should strive to alter that situation. Only instructionally *sensitive* tests will allow the NCLB Act to benefit your state's children.

Detecting Instructional Insensitivity: Three Required Attributes

In order for state-level NCLB achievement tests to be instructionally *sensitive*, they must possess three attributes. First, the skills and/or bodies of knowledge the tests assess must be sufficiently well described so that teachers have a clear idea of what skills and/or bodies of knowledge their students must master. Second, the assessed skills and/or bodies of knowledge must be sufficiently small in number so that teachers are not overwhelmed by too many assessment targets. Third, the tests' results must permit teachers to identify whether each assessed skill and/or body of knowledge has been mastered by a student.

If a state's NCLB tests fail to possess one or more of these three attributes, those tests will be instructionally insensitive. And, as indicated previously, the use of instructionally insensitive tests to implement a state's NCLB-required accountability program will result in a reduction of educational quality.

Thus, in the remainder of this brief analysis, I intend to describe how it is possible to determine if a state's NCLB tests fall short on each of the three aforementioned attributes of instructionally sensitive tests. I will conclude the analysis by identifying one additional, but significant, consideration to employ when appraising the quality of a state's NCLB tests.

Attribute One: Clear Descriptions of Assessment Targets

Today's statewide achievement tests are typically supposed to assess, at least in some general fashion, a state's content standards (that is, the skills and knowledge designated as state-approved curricular aims). But in many states, those content standards are far from unambiguous. In fact, a number of states have chosen to make their content standards nothing more than general labels, for instance, "measurement."

Such general designations are intended to describe a bevy of more specific curricular outcomes referred to as “benchmarks,” “performance indicators,” and so on.

To determine whether a state test’s assessment targets are sufficiently well described, you need to focus on the descriptive level that teachers will typically employ in order to focus on their instructional activities. Typically, that descriptive level will be one step below the most general descriptive level. In other words, if a state’s curricular aims are set forth as sets of somewhat general content standards that subsume more specific benchmarks, the descriptive level to which you must pay attention is the benchmark level, not the content-standard level. To illustrate how an appraisal of a state’s curricular aims might work, let’s focus on benchmark-level assessment targets.

The key question, then, becomes the following:

> *Is each benchmark stated with sufficient clarity that almost all of the state’s teachers can identify what the benchmark really means?*

One straightforward way to answer this question is to assemble a small group of teachers, perhaps a half-dozen or so, then find out to what extent they *independently* concur in their interpretation of the nature of any assessment target. Many NCLB tests supply no descriptive information beyond the state’s benchmarks. For other state tests, descriptive information about test-coverage is provided in addition to the benchmarks.

So, after providing a group of six teachers with whatever descriptive materials are available, the teachers should then be asked to write, *in their own words,* what the meaning is of a set of randomly selected benchmarks. After the teachers have independently written their descriptions, those teacher-authored descriptions for each benchmark should then be compared to determine how similar they are. The more alike the teachers’ independently created descriptors are, the more confidence one can have in the clarity of descriptions regarding the assessed skills and knowledge. In short, if the teacher-generated descriptions are not homogeneous, this deficit in descriptive clarity will contribute to a test’s instructional insensitivity.

If, because of inadequate descriptions of what’s to be measured by a state’s NCLB tests, the tests turn out to measure one set of skills and knowledge, but the state’s teachers aim their instruction at a different set of skills and knowledge, then students’ test performances will surely not be indicative of how well the state’s teachers have been teaching.

There’s an even more simple way to secure a determination of the clarity with which a state test’s assessment targets are described, and that’s simply to ask teachers to give you their opinions. For instance, you could interview several teachers individually so you could ask each of them, in turn, to respond to a question such as the following:

Based on the descriptive materials you have been provided regarding the skills and/or knowledge to be measured by our statewide tests, how clearly do you understand—for instructional planning purposes—the skills and/or knowledge to be assessed?

What you are apt to receive in response to such a question is a range of different answers. You might even probe further by asking a teacher to clarify. What you are looking for are honest opinions, from teachers themselves, regarding the clarity of a test's assessment targets.

Finally, you could secure teachers' anonymous responses to a questionnaire that includes items such as the following:

For purposes of a teacher's instructional planning, how clearly are this state test's assessment targets (the skills and knowledge it measures) described? Respond by circling one number from the number line given below:

10 9 8 7 6 5 4 3 2 1 0
Extremely Clear ... *Totally Unclear*

It is possible to probe the descriptive adequacy of any state's tests by employing several data-collection efforts akin to the three examples presented here. What you are trying to do is reach a defensible judgment regarding the clarity with which a state tests (with or without reference to a state's content standards or benchmarks) describe the skills and/or knowledge that they assess. If those descriptions are ambiguous, the tests will be instructionally *insensitive*.

Attribute Two: A Manageable Number of Assessment Targets

Too many targets can overwhelm. That's true with hunters. That's true with teachers. If an NCLB test contends that it measures 30 or 40 curricular targets, teachers will be unable to focus their instructional plans properly. An NCLB test that purports to measure too many content standards (or benchmarks, etc.) is certain to be instructionally insensitive.

In the first place, because there is a limited amount of time that can be used to administer an achievement test to students, especially young students, it is patently impossible to measure children's mastery of large numbers of curricular aims. That's because it is typically impossible to measure a student's mastery of a particular curricular aim with only one or two test items. And if there are numerous curricular aims to be assessed, you can be sure that the test simply cannot measure all of them *well*—or even measure some of them *at all*. Thus, what's going to be assessed on each year's NCLB test frequently turns into a guessing game—a game in which teachers

often guess wrong guesses. It is difficult for a state's teachers to focus their classroom instruction accurately if they are forced to conjecture about what's to be assessed.

From an instructional perspective, it is impossible for most people to keep too many targets in mind. Most teachers can intellectually handle six or seven instructional emphases. Few teachers can intellectually handle several *dozen* instructional emphases. This is a classic instance in which *more* turns out to be *less*. An NCLB test that purports to assess myriad curricular aims will serve as a dysfunctional framework for teachers' instructional planning.

Typically, of course, NCLB tests are based on a state-approved curriculum that frequently contains an excessive number of content standards, benchmarks, etc. Thus, the underlying reason that NCLB tests attempt to measure too many curricular targets is often the state's curriculum itself. Nonetheless, a state NCLB test that asserts it will assess a large number of curricular targets has, by its acquiescence to the state's too-numerous curricular targets, nonetheless presented the state's teachers with an onerous and unhelpful set of assessment targets. Such tests are sure to be instructionally insensitive.

One way to determine whether a state's NCLB tests are based on too many assessment targets is simply to count those targets. There may be no "magic maximum number" of suitable assessment targets for an NCLB test, but sensible people can typically recognize too many targets when they see them. For instance, if you learned that the fifth-grade NCLB tests in your state supposedly measure students' mastery of 14 reading benchmarks and 33 mathematics benchmarks, you'd have little difficulty in concluding that there were too many assessment targets being measured. For a fifth-grade teacher to focus sensibly on 47 assessment targets in reading and math (not to mention the teacher's need to deal with curricular aims on other subject areas) is impossible. The number of assessment targets is obviously unmanageable.

In some states, NCLB tests are being used in which teachers at a given grade level are supposed to focus their instructional attention on a hundred or more math and reading benchmarks. This just can't be done.

Another way to find out if your state's NCLB tests attempt to address too many assessment targets is simply to ask teachers to tell you, without allowing them to consult any reference materials, what curricular aims are currently being measured by the state's tests. If a reasonable number of assessment targets are being measured, the teachers will usually be able to tell you what those assessment targets are. If there are too many assessment targets, you can bet that the teachers will be vague and/or uncertain. If you receive vague and/or uncertain responses about what's currently being measure by your state's tests, odds are that there are too many assessment targets being measured.

If an NCLB test attempts to assess many more than a half-dozen or so curricular aims in any subject area it assesses, then it is likely that the test is aimed at an

intellectually unmanageable number of assessment targets. Such tests will be instructionally *insensitive*.

Attribute Three: Instructionally Informative Results

If NCLB tests do not supply teachers with students' results in a form that permits teachers to determine which segments of their instruction have been effective or ineffective, those tests will be instructionally insensitive. Test results cannot be used by teachers to improve the quality of their instruction. Over the long haul, students won't be taught better because teachers won't know which parts of their instruction need to be improved. Thus, the third factor by which the instructional sensitivity of NCLB tests should be judged hinges on the manner in which the tests' results are reported.

If the tests are organized around a modest number of assessment targets, for instance, six or seven, then the tests' results should be reported in such a way that a teacher can determine *which* of the six or seven targets were mastered (by the teacher's students) and *which* ones weren't mastered. Typically, to provide such per-outcome reporting, a test must include at least a reasonable number of items per outcome. Otherwise, there is insufficiently reliable evidence to determine whether students have mastered a given outcome. If teachers don't know which parts of their instruction have been working, they won't be able to improve that instruction. An instructionally sensitive test must contribute, over time, to improved instruction.

To find out whether an NCLB test's reporting structure is satisfactory in the way its results are reported, once more it makes sense to talk to teachers who have been on the receiving end of the test's reports. You might interview a number of teachers, individually if possible, to have those teachers describe, in specific language, how they would use test-results to judge the quality of their own teaching. Try to get very explicit during those interviews. Don't accept generalities but, instead, ask the teachers to spell out just how a given set of test results could lead them to make a specific decision about the instructional effectiveness of particular parts of their teaching. It is usually helpful during these interviews to have copies available of the actual score-reports that are provided to the state's teachers (so that interviewees can refer to those reports, if necessary).

If your state's NCLB tests have not been administered yet, then simply provide teachers with mocked-up examples of the score-reports that will be provided in the future. Then seek teachers' reactions to the *instructional* relevance of those illustrative reports. Clearly, if the tests' results are reported at too general a level, teachers will be unable to determine whether specific parts of their instruction have worked or not. What you need to do is push the teachers to spell out, in specific terms, just how a set of test-results will help them, if at all, in judging the caliber of particular segments of their teaching. NCLB tests that fail to provide meaningful per-standard or per-benchmark feedback to teachers will be instructionally *insensitive*.

Three Strikes and You're Out

To review, if tests are deficit in any *one* of the three attributes just described, that is, (1) clarity of assessment targets, (2) a manageable number of assessment targets, and (3) instructionally informative results, the tests are quite likely to be instructionally insensitive. If the test falls short on two, or even three, of these attributes, it will most certainly be instructionally insensitive. Use of instructionally insensitive tests to implement the NCLB Act will lower the quality of education a state provides to its students. Accordingly, instructionally insensitive tests should definitely not be used to implement the NCLB Act.

Is Instructional Sensitivity Sufficient?

An NCLB test can be instructionally sensitive, yet not benefit a state's students. Here's why. Instructionally sensitive tests can be built to assess *low-aspiration* curricular aims. You see, an instructionally sensitive test could satisfy all three attributes I've just described, yet deal with truly trifling curricular outcomes. Putting it another way, instructional sensitivity is simply not enough. Instructional sensitivity is a necessary, but not sufficient condition for an NCLB test that's going to benefit students. An instructionally sensitive test, for the well-being of a state's children, must be aimed at assessing genuinely *defensible* curricular targets.

Disturbingly, to avoid the appearance that too many state educators will appear to be ineffective, officials in some states have decided to have their state's NCLB tests assess skills and knowledge that are not genuinely challenging (or they have set their state's "proficiency" levels much lower than will be good for the state's children). Such actions may lead to fewer identifications of failing schools, but serious educational harm will be done to the state's children by trivializing the state's assessment aspirations.

So, *after* your state's NCLB tests' instructional sensitivity has been assured, be certain the tests measure significant skills and knowledge that children ought to be mastering. If a state's NCLB tests assess students' mastery of only low-level curricular outcomes, all the instructional sensitivity in the world won't make those tests contribute to children's well being. Instructional sensitivity constitutes one critical requirement for an educationally sensible NCLB test. The significance of the assessment targets measured by those tests is an equally imperative requirement.

Prepared for the National School Boards Association
February 2003[*]

[*] Additional information regarding the isolation of defensible curricular content is available from the Commission on Instructionally Supportive Assessment: *Building Tests That Support Instruction and Accountability: A Guide for Policymakers* (www.aasa.org) and "Crafting Curricular Aims for Instructionally Supportive Assessment," available from the National Center for Educational Outcomes at http://education.umn.edu/NCEO/Presentations/CraftingCurricula.pdf