

## QUESTIONS TO ASK WHEN EVALUATING A HIGH-STAKES TESTING PROGRAM

As tests become an increasingly common administrative mechanism to drive educational reform, controversy grows over what constitutes appropriate and inappropriate test use. This document is a [distillation] compilation of the major questions and concerns that have been raised by critics of high-stakes standardized testing, by advocate groups who see their constituents unfairly and disproportionately affected by some of the present uses of tests, and by plaintiffs in litigation involving test use.

The document provides educational writers with a tool to evaluate the use of standardized test results by state departments of education or school systems, to make important decisions automatically about students, teachers, schools and school districts . Important decisions that are commonly made on the basis of student test performance include, among others whether to:

- : promote a student from one grade to the next;
- award a high school diploma;
- assign a student to a remedial class;
- allocate funds to a school or district; award merit pay to teachers on the basis of their students' test performance;
- certify a school or district;
- place a school system in “educational receivership”.

When test results are used to make decisions like these the testing program can be characterized by the adjective high-stakes. This document is primarily concerned with high-stakes testing programs.

The adverb automatically describes the common situation of using a pre-determined test score to trigger, without review or exception, a high-stakes decision. The decision cannot be appealed on the basis of other indicators, including teacher judgment, that may contradict or moderate it. In short, this automatic decision feature of high-stakes testing programs is deliberately designed to eliminate input from educators relative to the decision in question.

It is difficult to list all the factors that are equally applicable across all high-stakes testing programs because they differ in terms of such important characteristics as :

- what they purport to measure;
- who controls the program;
- who is tested;
- when students are tested and under what conditions;
- and how the results are used;

Thus, the relevance of the suggestions that follow, where you would go to get data, and the data themselves, will differ according to the context of the program.

### **1. What are the Stated Goals of the Testing Program?**

In order to properly evaluate a high-stakes testing program it is necessary to inquire about its goals. Explicit program goals are those described in the formal language of the statute or board policy mandating the testing program and in the public language used by the policy-makers to describe the program after adoption. Explicit goals are generally laudatory and positive. For example few would argue with the following goals often associated with high-stakes testing programs:

- to assure the public and employers that all students possess basic literacy and numeracy skills before graduating from high school or being promoted from grade to grade;
- to diagnosis weaknesses in basic skills so that students can receive appropriate remediation;
- to allocate compensatory funds and other resources on a more rational basis;
- to restore public confidence in the high school diploma;
- to introduce real accountability into the schools;
- and to improve the quality of the teaching corps.

These are but a few examples of the type of laudatory goals associated with high-stakes testing programs. Few could argue against the

realization of such goals. The question you must ask is whether this testing program can in fact lead to the realization of such goals.

In addition to the explicit goals of the program often there unstated, implicit goals that are sometimes more powerful than the explicitly stated goals. The implicit goals of the program include the goals that people who adopted or advocated for the program discussed behind the scenes. For example, many programs were adopted for the explicit purpose of “fostering educational accountability”; however, many of these same programs were pushed by politicians and the business community because they felt that new businesses were not coming to the state or city because the schools were perceived as inferior. It was felt that the initiation of the testing program would be viewed by potential developers or new businesses as an attempt to upgrade standards and improve the quality of the school system. Another often mentioned implicit goal of high-stakes programs is that they give the outward appearance of seriousness about educational standards, and the reform of education; that is, at the very least the high-stakes testing program bolsters the public image of the schools, or of the teaching profession.

The explicit and implicit, stated and unstated goals of a testing program should be the starting point of one's evaluation of the program. The question to ask is whether the testing program can in fact realize explicit and implicit goals such as these or whether it is nothing more than an attempt at a quick public relations fix that does not address serious systemic problems. Ask yourself the following questions about the goals of a program:

**1a. Is it really possible to attain program goals through the use of a multiple choice test, or are the goals and the testing program merely window dressing to assuage negative public perceptions about the state of education?**

Ask whether a high-stakes testing program can really service as a guarantor of educational standards, excellence and quality or is it mainly an ephemeral evanescent public relations gesture.

For example, some of the explicit goal language used to justify most minimum competency tests programs does not (and cannot) describe the reality of the program. Such programs are initially justified, and then touted as providing diagnostic information about students. In fact such tests do not give diagnostic information. They can only classify students as falling above or below some predetermined cut-score. The results do not tell a teacher or parent why a student scored low on the test; there is no information that helps teachers to identify the nature or underlying cause of the student's problem. Without knowing the nature of the student's educational "illness" prescribing appropriate remedial treatment requires additional diagnosis, often expensive and not built into the program. Further, a low score by itself generally is not sufficient grounds for inferring the absence of the skill or knowledge purportedly measured by the test. That is, the student may actually possess the skill or knowledge measured by the test, but for any number of reasons (some personnel e.g., health, problems at home, lack of interest, test anxiety; others related to the mode of testing e.g., multiple choice format, lack of test taking skills etc.) was unable to demonstrate the skill or knowledge at the time of testing. Thus, the medical diagnosis metaphor used by policy makers when describing the purpose of minimum competency tests may, in some cases, be inappropriate; the student may not be deficient relative to the ability measured by the test despite his or her low test score, and in fact does not need remediation. Unfortunately, most programs are structured so that some students who actually

possess the skills measured are automatically forced into remediation classes -- sometimes with attendant stigma -- and out of their regular classrooms, until such time as they can "pass" the test.

**1b. Do the schools already possess the information about student performance without the imposition of a new testing program?**

Are there test scores already available (from district level testing programs for example) that give the same information about student performance? Ask for a list of all the tests already being routinely administered in the school district.

To what extent do these existing test scores agree with, or contradict those from the high-stakes program?

Can't teachers identify those students who are having difficulty reading or in arithmetic and need remediation?

If not, why not?

In fact, the research evidence clearly shows that teachers already know who most of the students are who will eventually fail high-stakes tests, and that the test information rarely gives information that differs from that already available from teacher appraisals. Given this fact ask whether one of the motives policy makers have in mandating a high-stakes testing program isn't a distrust of teacher judgments. If this is so -- and in many program it is -- how does this negative attitude toward teachers' appraisals of students square with the oft stated desire of policy makers to upgrade the teaching profession? Particularly when the testing program is explicitly designed to ignore the teachers' professional judgments an automatically make a high-stakes decision about a student?

**1c. On whom do the sanctions and rewards associated with the testing program primarily fall?**

This question relates to the previous one concerning motives. Ask question such as:

Do the sanctions fall directly only on the students -- those with the least power?

If the answer is yes, ask will students already at risk because of economic, linguistic, cultural or handicapping conditions most likely be those adversely affected?

If the sanctions are directed at teachers, will all teachers in the system be affected, or will it only be teachers at the grade levels covered by the test, or those who teach students at risk?

In a more general sense ask whether the testing program would have been mandated absent large numbers of at risk students, or large numbers of urban schools servicing predominately minority populations.

Ask:

How are special education or learning disabled students following Individual Educational Plans (IEP) treated in the testing program?

Are they exempt?

If they included what accommodations are made for their disability (see # 4 below) and to what extent does the test match their IEP's (see #2a below)?

**1d. Have the policy makers allocated the resources necessary to attain the goals of the program?**

For example:

If the goal of the testing program is to identify those in need of remediation have sufficient funds been provided to staff the remedial programs?

Will the numbers accepted for remediation be based on actual need, or on how many students the program can afford to remediate?

What kinds of remedial programs are planned?

Will the remedial program simply coach students to “pass” the test -- a real danger associated with many so called “remedial” programs -- or is the program designed to identify, and then overcome the reasons underlying poor test performance?

Do we know how to design new, effective ways of enhancing basic skills attainment that we have not previously used, and which hold promise of helping students identified by the testing program?

If the answer to this question is yes ask;

what these programs are;

why they have not previously been use and;

why it has taken a high-stakes testing program to implement them?

**1e Are there potential negative costs associated with achieving the stated goals of the program?**

Nothing in this world is free. In attempting to achieve the stated goals of the testing program one should always inquire about the unintended, negative costs associated with it. For example, the goal of restoring meaning to the high school diploma might be realized at the unintended expense of driving students at risk out of school before they graduate; students who previously might have persisted and received a diploma. There is anecdotal evidence that many poor, minority, bilingual students who take minimum competency graduation tests and fail become discourage, never re-take the test, and eventually drop out of school without the diploma. Having a high school diploma today is no big deal, but not having one is a disaster. Thus, this push-out phenomenon -- if it exists -- may very well be the most serious unintended negative outcome associated with such programs.

The relationship between high-stakes tests and the drop-out rate is an area that needs to be investigated. If the high-stakes tests are used for promotion decisions from grade to grade the impact of

these tests on eventually dropping-out my take years to surface. For example, if a student is retained in grade twice on the basis of low test performance during the elementary years he or she may reach the school leaving age and drop-out before he or she has to sit for a graduation test.

Consider another example. If the students' test scores are used to hold teachers or administrators accountable will the multiple-choice test eventually drive what is taught, how it is taught, what is learned, and how it is learned? Will material within a curriculum area not covered by the test, or curriculum areas not included in the testing program receive less attention because of the very human tendency to try to optimize test performance -- the principal indicator of "excellence" or "quality" in assessing accountability? Already there is evidence that publishers are providing teachers with multiple choice drill material to use in their classes. One company markets material called *Scoring High* which can be used to prepare students for most commercially available achievement tests. Ask about the degree to which the multiple choice mode associated with high-stakes testing is driving instruction in an all important subject like reading. That is, is the reading instruction being reduced to having students read short paragraphs and answer multiple choice questions .

**1f. Has the agency in control of the testing program set aside resources for an independent audit of the positive *and* negative aspects of the program?**

Policy makers, and civil servants changed with implementation of high-stakes testing programs have a vested interest in demonstrating that the program is realizing its goals. Therefore, there is a reluctance -- bordering on avoidance -- to investigate possible negative, disconfirming, contradictory outcomes associated with the program.

## 2. What Inferences or Decisions Will Be Made From the Test Score?

This is the fundamental question that must be asked of all tests. Its answer determines the type of validity evidence that must be provided by the agency using the test to support the accuracy of the inference, or correctness of the decision made on the basis of a person's test score. Validity is a question of the degree to which the inference or decision made from the test score about a test taker is correct or accurate. It is not the test per se that is validated but the correctness of the inference or decision made on the basis of the test score.

Try to determine both the explicit and implicit inferences and decisions about test takers made on the basis of the test, not only by the state agency or the school system, but also by the media and by the general public. Then ask the agency in charge of the program for the evidence that supports these inferences or decisions.

### 2a. Does the agency try to limit the inferences to how well a person has mastered a particular aspect of the curriculum?

If this is the case then to use a testing term, **content-related evidence** of validity must be provided. That is, the agency must show that the test measures the correct domain of achievement -- important skills and knowledge actually taught to the students -- and that the test questions (items) accurately reflect not only the domain but also how the material was taught. Federal court cases have held that, when a test is used to make a critical decision about a student, such as the award of a high school diploma, students have a constitutional right to receive instruction in the skills and knowledge covered on the test. (See Debra P. v. Turlington, 644 F. 2d 397.)

Collecting content-related evidence by asking educators to make decisions about item/domain match is the most common approach

taken by test developers in validating most high-stakes tests. This judgmental approach is relatively cheap, and can be design by the contractor to assure a positive outcome -- that it allows them to be able to affirm validity. Ask if the procedures used to gather content-related evidence are design not only to confirm, but also to possibly disconfirm validity.

Further, ask if content-related evidence is a sufficient basis on which to justify the high-stakes decisions being made on the basis of the multiple choice test. In fact, it is extremely unlikely that most inferences or decisions associated with such high-stakes tests can be supported exclusively by content-related evidence; although this is precisely what happens in most high-stakes testing programs. This leads us to ask another question related to test validity:

**2b. Do the decisions involve inferences (explicit or implicit) about some future performance by the test takers?**

For example:

- Are inferences being made about the person's ability or lack thereof to do the work at the next grade level?
- Is the decision to place the test taker in a remedial program based on the belief that the person will benefit from the program?
- Is the decision not to admit the student to a course or sequence of study (for example, the use of the Pre Professional Studies Test to screen college sophomores for admission to teacher training) predicated on the belief that the person does not have the necessary skills or knowledge to succeed?

If the answer to these and similar questions is yes, then ask if the agency has what is known in testing as **criterion-related evidence** which supports such futuristic inferences. That is, is there any evidence to indicate that test performance actually correlates with the future performance in question? In the case of remedial

placement what is the evidence that placement in the program actually improves the skill being remediate and not simply the person's test score?

Criterion-related evidence is seldom gathered in most testing programs. Criterion-related evidence is difficult and expensive to obtain. An excuse often given for avoiding gathering it is that we lack a good measure of the criterion (the future performance) in question. Nonetheless if criterion-related inferences or decisions are being made then lack of a proper criterion measure is no excuse for not being able to validate the test use. A more plausible reason for avoiding collecting criterion-related evidence is that often it calls into question the correctness of the inferences or decisions user wish to make. Given the threat of litigation associated with high-stakes test programs states are reluctant to gather evidence that might not support the future related inference or decisions being made. Regardless, in many situations criterion-related evidence, in addition to content-related evidence, is necessary in order to support the ways in which the test scores are being used.

There is still another validity question that needs to be asked:

**2c. Are inferences being made about the extent to which the test taker possesses a certain trait or characteristic that cannot be directly seen but only inferred from some observable performance or behavior on the part of the test taker?**

Are inferences being made about such things as the person's:  
competence,  
functional literacy,  
numeracy,  
reading comprehension,  
ability to function as a citizen, etc.?

If so then, in testing parlance, **construct-related evidence** is necessary to support such inferences.

Construct-related evidence is the most fundamental aspect of test validity, and almost every test use involves implicit or explicit construct related inferences. While construct-related evidence consists of content, and depending on the construct (i.e., competence), criterion-related evidence, it also includes the testing of hypotheses about the nature of the construct or trait. That is, predictions about how someone possessing the trait or construct should(or shouldn't) behave or act in certain situations, or how performance on the construct measure should correlate with performance on measures of other, different constructs. Again keep in mind that a true validation study tests hypotheses that might not only confirm but also disconfirm (cast doubt about) the ability of the test to measure the construct or trait it purports to measure.

Proper validation consists of a series of studies involving a mix of the three kinds of validity evidence described above: content-related; criterion-related; and construct-related evidence.

Unfortunately as noted above, most state and local testing programs limit validation to gathering only content-related evidence, and then in ways which are guaranteed to confirm or affirm validity. If this is the case in the program you are evaluating ask why aren't others types of validity evidence being gathered?

**3. Is a cut-score used to trigger a decision or inference?**

In most high-stakes testing programs a single test score -- called a cut-score -- is selected as the point at which an automatic decision, or disposition is made about the test taker. For example, if the cut-score on a 100 item test is 70 and an examinee gets 70 or more questions correct, he or she is automatically promoted, is eligible for a diploma (if other criteria such as sufficient Carnegie credits, passing all course etc. are also met), or is classified as not needing remediation. However, if the examinee gets 69 or fewer questions correct then he or she is automatically retained in grade, denied a diploma (regardless of meeting the other criteria), or is

automatically placed in a remedial program. Thus the cut-score is the axis around which all decisions and inferences associated with the testing program revolve.

Given the cut-score's importance ask the following questions related to it:

**3a. What evidence is there to support correctness of the inference or decision made on the basis of the particular cut-score ?**

Most testing programs fail to recognize that the validity issue cannot be separated from the choice of the cut-score which automatically triggers the decision or inference. Most cut-scores are chosen by having individuals make judgments about the importance of each question and then summarizing and aggregating those judgments across raters. Therefore, ask, whether the procedures used to arrive at a cut-score provide any validity evidence to support its eventual use. This is a particularly important question to ask when, as is generally the case, construct-related and criterion-related evidence is required to support test use.

**3b. Can the trait, attribute, knowledge or skill which underlies and supports the decision about the examinee be dichotomized?**

Most skill levels, knowledge levels, or competencies are continuous in nature. To pick a single point on that continuum and label those at or above it as competent, or in possession of sufficient skill or knowledge, and those below it as incompetent, or not possessing sufficient skill or knowledge, is to ignore the continuous nature of most of the attributes of interest. As noted above this dichotomization also raises issues about the validity of the decision made on the basis of the selection of the single point on the

continuum. Therefore, ask if the constructs, abilities or traits purportedly measured by the test can meaningfully be dichotomized.

**3c. How stable are the classifications made on the basis of the cut-score?**

That is, if students were to be tested and thus classified more than one time, how many examinees would receive the same classification (pass or fail) and how many would switch positions (from pass to fail and vice versa)? This information is necessary in order to estimate the number of false positives (i.e., classified as having the skill, knowledge or competency but really not possessing it) and false negatives (i.e., classified as not possessing the necessary skills and competencies when in fact they do possess them). Programs generally err in trying to reduce the number of false negatives, but if remediation is really the goal then false positives are also important because they are -- if the test is valid -- truly in need of remediation.

**4. Are there provisions to accommodate the special needs of some students?**

Ask if the high-stakes tests are ever administered in other languages or other formats, or different versions for handicapped or limited-English-proficient students?

There are two different types of test alterations that can be made. One would be a change in the type of knowledge or information measured on the test; if this type of change is made, ask if the test is still measuring the things intended when the program was adopted. In other words is the altered form measuring the same ability, trait, construct, knowledge or skill as the primary version? The other type of change would not be a change in content but in the way the test is administered. For example, students with physical impairments may not be able to respond to a paper and pencil format, and are given the test orally or by some other means;

a Braille version may be used; some students may be given additional time, etc. Federal law (Section 504 of the Rehabilitation Act) requires that tests be administered in a way that accommodates for handicapping conditions.

In both situations ask what evidence there is that the inferences or decisions made from the alter form are correct and compatible to those made from the original version of the test. Modifications to the test require that additional validity evidence be gathered. The agency administering the program cannot assume that validity evidence produced to justify the original version also justifies the inferences or decisions made when the test is translated into another language, or the mode of administration or the format of the test is modified.

**5. What judgmental and empirical steps have been taken to insure that the test is free of cultural and gender bias and stereotyping?**

How does the agency handle the original screening of questions?

Are the items reviewed by representatives of various groups for offensiveness or stereotyping?

Is there a tryout where the pass rates for different groups are compared on each item?

If not why not?

If so what happens to items which on the tryout or field test are shown to be disproportionately more difficult for one group then for another?

**6. How are the results reported?**

Question the nature of the performance data released by the agency:

Are pass/fail data broken out by district, school, grade, classroom, race, sex, ethnicity?

Do certain types of students seem more likely to fail or score less well on the tests, i.e., black or other minority

students, TitlesI/Chapter I students, students in low income neighborhoods, etc.?

Does the agency report changes in test results over time?

If so are more students passing and if score there particular groups of students who are more successful on the test over time?

To what extend is the agency able to track the progress of individuals over time?

What is the relationship between performance over time for individual students?

What is the turn over, mobility or change of residence rate for students over time?

How does this mobility rate effect the agency's ability to make inferences about changes in test performance over time?

If the pass rates are going up ask why. Is the test itself easier?

Is the same test used year after year?

If a different form of the test is used each year ask if the forms have been equated so that they are of equal difficulty over the years.

Has the cut score been lowered?

How does student test performance compare to other indicators of educational achievement (other tests, grade point average, attendance, drop-out rates, transfers to other schools etc.)

**7. Can you look at an actual test and the test key or do you have to rely on “representative questions” when evaluating the test?**

Can copies of the tests be reviewed by the press, parents, students?

If not why not (and isn't there some accommodation that can be made so that some members of the public can review the test? If

there is an open review provision examine the test and look for the following kinds of things:

Do the items seem to measure the types of skills and objectives intended by those who initiated the testing requirement?

Do all of the items have a single best correct answer and only one? On some tests -- particularly initial teacher certification tests -- there is not enough contextual information to pick the keyed response or the response depends on the emphasis placed on the topic in the school attended.

Does the test seem too easy or too difficult?

Are there items on the test that would be unfair or offensive to certain segments of the population (racial or ethnic minorities, women particular religious groups)?

If criterion-related or construct-related inferences are being made examine each item in that light.

Does someone really need to answer the item correctly in order to succeed at some future task i.e., classroom teaching, the next grade level, teacher preparation, adult citizenship.

Be critical of each and every item. Test items or questions are the essential building blocks of the total test score. Each item is itself a mini test, and contributes equally (i.e., one point) to a person's total score. A single bad, faulty, ambiguous or miskeyed item can be the difference between passing and failing for some individuals.

**8. Related to 7 above ask if there is a “truth in testing” provision built into the program were by parents, guardians, students can examine their test, answer sheet and the test key?**

Most states and districts do not have a “truth in testing” provision, arguing that it violates test security and would prohibitively

increase the cost of the program because a new test would need to be constructed each year. Ask why such a provision is not feasible for the high-stakes testing program you are evaluating. In Europe high-stakes test are routinely released each year so that students, teachers, parents and the public can examine the test questions. In this country the New York State Regents Examinations were routinely released each year, and the SAT survived the New York State Truth in test requirement despite the cries of doomsayers that it would destroy the test's validity and the ability to equate from year to year.

Also ask:

What kinds of information, support and guidance are provided to parents of students who must take the test?  
What provisions are made for parents to appeal or ask for a review of a decision made on the basis of the test?  
What if the student fails a high-stakes reading test and the parent feel and could produce evidence that the child can in fact read?

**9. How do parents teachers, students, administrators, employers and the public feel about the high-stakes testing program?**

Talking to various groups can shed a great deal of light on the positive and negative aspects of any high-stakes testing program. However, great care must be taken to protect the anonymity of students to shield them from the stigma or labe associated with failure, and of educators who might be critical of the program but fear reprisals from the sponsoring agency.