The Distortion of Teaching and Testing: High-Stakes Testing and Instruction

George F. Madaus

Testing has grown enormously and has changed substantially since the 1960s, affecting what is taught, how it is taught, what is learned, and how it is learned. Several factors have contributed to this growth and change; among them was federal legislation which created new markets for the testing industry. For example, the *National Defense Education Act of 1958*, the *Elementary and Secondary Education Act of 1965*, and the *Education for All Handicapped Children Act of 1975* (P.L. 94-142) all contained provisions supporting or requiring the use of standardized tests. Actions at the state level as well have given test sales and use another big boost. State level policy makers first turned to test results for information about schools and populations of at-risk students, but soon realized that test results could also be used as an administrative mechanism to drive or implement policy (Madaus, 1985). As a result, state mandated assessment programs rose from 1 in 1960 to 32 by 1985, and state level basic skills and minimum competency programs went from 1 in 1972 to 34 by 1985 (Office of Technology Assessment, 1987).

The state programs represented not only an increase in the amount of testing, but also effected a fundamental change in the nature of testing programs. Specifically, and increasingly, many school district programs, are characterized as high-stakes testing programs. That is, the test results are directly linked to important rewards or sanctions for students, teachers, or institutions. Anyone close to education knows that the use of test results for accountability purposes is widespread, and has grown enormously over the past 8 years. For example, a recent issue of *Education Week* contained a chart showing 45 states which had mandated an accountability system of some sort. As part of this accountability process, 18 states use achievement test results, 1 uses competency

GEORGE F. MADAUS *is Professor of Education and Director of the Center for Study of Testing Evaluation and Educational Policy, Boston College, Chestnut Hill, MA.*

test results, and 29 use both types of test results ("State Accountability Systems," 1988). High-stakes tests include those used for the certification or recertification of teachers, promotion of students from one grade to the next, award of a high school diploma, assignment of a student to a remedial class, allocation of funds to a school or district, award of merit pay to teachers on the basis of their students' test performance, certification or recertification of a school or district, and placement of a school system into "educational receivership."

In this article, I consider how a high-stakes test can directly and powerfully influence how teachers teach and students learn. I will argue that this process corrupts the test's ability to serve as a valid indicator of the knowledge or skill it was originally intended to measure. A central theme of this article is that increasingly instruction is driven by the testing process, and that the negative effects associated with this state of affairs far outweigh any short-term benefits touted by advocates of "measurement-driven instruction" (Popham, 1981; Popham, Cruse, Rankin, Sandifer, & Williams, 1985; Millman, 1981). The implications of the metaphor of testing, rather than professional discernment and judgment, as the engine or drive train of instruction, deserve much more debate than they presently receive. To understand the consequences of measurement-driven instruction on teaching, learning, and the test itself, I first develop the concept of what a test is—a concept that is frequently misunderstood. Next, I offer a set of six principles that describe the consequences of measurement-driven instruction and show how they affect teacher and student behavior and the test itself. Finally, I discuss what teachers who do not agree with a measurement-driven instructional philosophy can reasonably do when they are faced with a high-stakes testing situation.

What Is a Test?

All teachers are familiar with tests. Teachers regularly give their own tests and administer a host of standardized achievement tests as part of district-wide and state mandated testing programs. Nonetheless, precisely because testing is so familiar and ubiquitous, it is easy to lose sight of the basic concepts behind a test. Many educators have a hard time answering the simple question, "What is a test?" Yet, the correct answer to this seemingly straightforward question underpins the most essential concept of test use—validity.

How should one answer the question, "What is a test?" The first and most basic concept behind a test is that it is a *sample* of questions or situations—frequently called items—from some content domain or uni-

verse of interest. A content domain is a familiar concept to teachers. It is a body of knowledge, skills, or abilities defined so that you can decide whether a particular piece of knowledge, or a particular skill or task is part of the domain.

Figure 1 illustrates this concept of sampling from a content domain. The amorphous closed area shown in the figure represents the content domain of fourth grade arithmetic problems. There are a very large number of questions that one could theoretically ask a student about the domain shown in Figure 1. One way to reduce the large number of potential questions associated with a content domain is to define it much more precisely. This particular content domain could be divided into four sections representing the basic operations of addition, subtraction, multiplication, and division. In the figure, these sub-domains, or facets of the domain as they are sometimes called, are numbered 1 to 4. (Not all domains can be divided into subdomains or facets in this way.) We might limit the addition facet of the arithmetic domain to tasks involving three or fewer digits, with no carrying.

Even with this more limited content domain, there are still an enormous number of questions one could conceivably ask students to answer. The way around the problem is to draw a *sample* of questions to

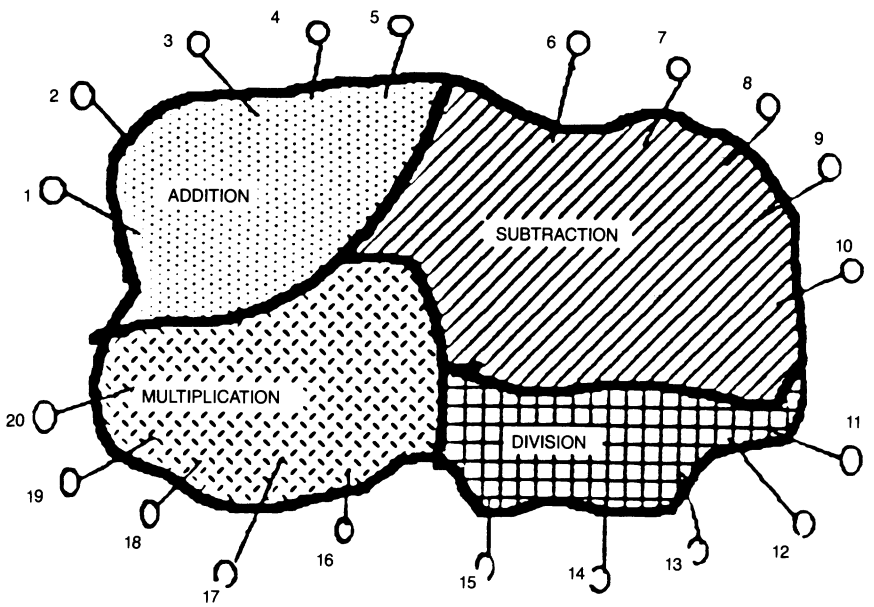


FIGURE 1: Universe or Domain of a Test with a Sample of Questions

represent the important parts. The small circles (shown in Figure 1) that lie outside the domain with the lines running back to it represent a sample of arithmetic questions drawn from each of the domain's four facets. The sample of questions constitutes the test.

A second basic concept, closely related to the concept of a test being a sample of items from a domain, needs to be included in an answer to the question, "What is a test?" That is, what we are really interested in is a student's performance in terms of the domain rather than his or her performance on the particular small sample of questions that make up the test. Figure 2 illustrates this second basic concept. The sample of 20 questions, now called a test, is represented by the 20 circles enclosed in the rectangles on the right. Based on the student's performance on the test (that is, the small sample of questions) the test user makes an *inference* about the student's performance on the entire domain. This inference is represented by the broad arrow in the middle of the diagram. Thus, test performance, a sample of behavior from the domain, is used to make a broader, more general inference about a student's performance relative to the entire domain of interest.

The correctness of an inference made on the basis of test performance, about a student's performance relative to the domain is the central and most important concept in testing. It goes by the technical name of test validity. Test validity refers to the degree to which a particular inference,

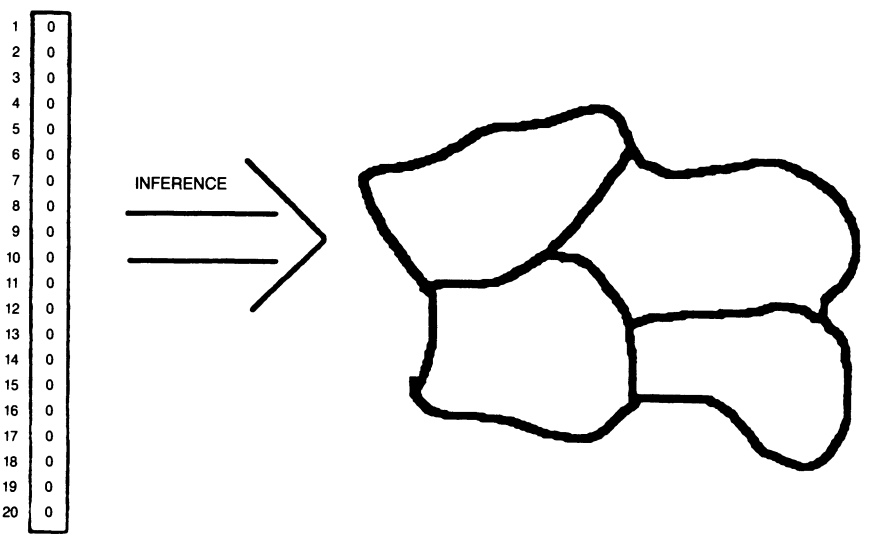


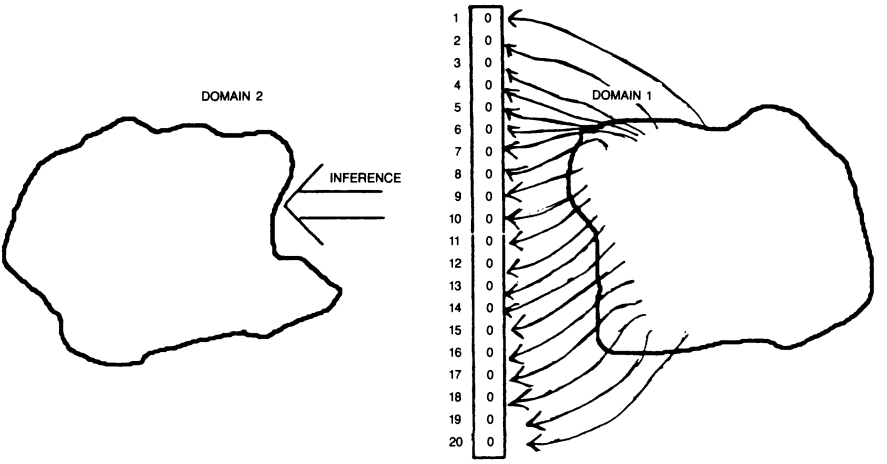
FIGURE 2: Using a Test to Make Inference about Performance in the Domain

and any resultant description or decision about an individual or institution, made on the basis of test performance is appropriate or meaningful. There is no such thing as a generically valid test, nor is a test valid in the abstract. In other words, it is incorrect to broadly and simply assert "This is a valid test" without any further clarification. Instead, when someone talks about test validity, you have to ask the question, "Valid for what?" The answer must always be in terms of the correctness of particular inferences, decisions, or descriptions that are made on the basis of a test score about particular populations.

Validity is a unitary concept, but there are three commonly referred to types of validity evidence. Users of achievement tests make inferences about the degree to which test takers have attained the knowledge or skills of interest. Thus, when using an achievement test it is essential to have evidence that the sample of questions (the test) adequately represents the content, skills, or behaviors of the domain. Evidence that the test properly represents the domain permits the user to affirm that the test is *content valid*. In addition, all tests—*including achievement tests*—involve inferences concerning the degree to which an examinee possesses certain constructs or traits. A *construct* or *trait* is a "theoretical idea developed to explain and to organize some aspect of existing knowledge" (American Educational Research Association, 1985, p. 29). Examples of such constructs are intelligence, motivation, competence, functional literacy, musical aptitude, mathematics problem solving ability, reading comprehension ability, and spatial ability; tests have been developed that purport to measure these and many other constructs as well. Inferences made from the scores on such tests concern the degree to which a person possesses the construct or trait in question (e.g., how much competence, spatial ability, or musical aptitude a person "possesses"). Evidence about the correctness of this type of inference goes by the name of *construct validity.*

Figure 3 illustrates another type of inference that is often made on the basis of a person's performance on a test. Here the test, as usual, is a sample of questions drawn from a domain, but it is used to make inferences about how a student might perform on another, and often quite different domain. For example, the SAT and ACT represent samples of questions from the domains of verbal and quantitative skills. The scores on these tests, however, are used to predict how well a student might perform on the domain of academic tasks required in college. Teacher certification tests sample questions from a domain of professional knowledge or basic skills; the scores are then used to draw inferences about the likelihood of a person *not* being successful in the quite different domain of classroom performance. It is this second, or criterion domain, that gives the correctness of this type of inference its name: *criterion-related validity.*

FIGURE 3: Using a Test from One Domain to Make Inferences about a Different Domain



In summary, then, a test is a sample of questions or tasks from a domain that is used to make inferences about a person's, a group's, or an institution's performance on that domain, or on a different "criterion" domain of interest. The test's validity refers to the correctness of the inference, description, or decision made from the test score. We shall now consider how measurement-driven instruction, particularly in the form of high-stakes testing, can destroy a test's ability to represent the domain of interest, thereby abrogating the validity of any inferences, decisions, or descriptions made from test performance.

The Impact of Measurement-driven Instruction: General Principles

Like any test, a high-stakes test is a sample of items used to make inferences, decisions, or descriptions about people or institutions relative to some domain. A high-stakes outcome such as graduation or teacher accountability grafted to test performance is the fuel of measurement-driven instruction. While the instructional engine is propelled by the high stakes linked to test performance, the equity and fairness of the reward or sanction for an individual or institution depend entirely on the degree to which the inference, decision, or description made from test performance is correct. The consequences of measurement-driven instruction in a high-stakes testing situation can be discussed in terms of the following six principles (Madaus, 1988).

Principle 1

The power of tests and examinations to affect individuals, institutions, curriculum, or instruction is a perceptual phenomenon; if students, teachers, or administrators believe that the results of an examination are important, it matters very little whether this is really true or false—the effect is produced by what individuals perceive to be the case.

When people perceive a phenomenon to be true, their actions are guided by the importance perceived to be associated with it. Thus, testing programs that have the greatest impact on instruction and learning are those that students, teachers, administrators, parents, or the general public believe, rightly or wrongly, are used to make important decisions that immediately and directly affect them.

Many testing programs have sanctions or rewards directly and overtly linked to test performance through legislation or a state or local board mandate. In other testing programs, however, teachers, administrators, parents, or students perceive that sanctions or rewards are associated with test performance when no explicit high-stakes mandate or legislation exists. For example, state-wide test results may not be tied directly to rewards or sanctions by legislation (the results are meant only to inform policy makers), but when the results are used by newspapers to rank schools or districts across the state, educators and parents can perceive the results to be an indicator of effectiveness. When this happens the test results take on unintended important consequences. The SAT or ACT offers another example of how perceptions create a high-stakes test. Many parents and students see the SAT or ACT as the crucial determiner of admission to college. While the SAT and ACT may be a principal admission criterion in some very select colleges, these tests are not a critical determiner of admission at many colleges, particularly as the applicant pool shrinks due to demographics. Nonetheless, commercial companies play on the high-stakes perception of the SAT and ACT, and in so doing reinforce the belief that these tests play a critical role in the college admissions process. For example, a recent ad in *The Boston Globe* (8/8/88) by Britannica Learning Centers raises the stakes associated with the SAT to the theological level of an absolute in life:

> If you choose to believe just one absolute in life;
>
> ---
>
> **Believe this:**
>
> ---
>
> If your child does not do well on the SAT he will not be accepted by the college of his choice.

The text of the ad then boldly asserts, "Based primarily on the results of that *one exam* [the SAT] your child will be flatly denied or openly accepted to the college of his or her choice" (emphasis in the original). Ads like this not only play on parental anxiety about the importance of the SAT, they also help reinforce an aura of criticality around the test that in many cases simply is not deserved.

Thus, perceptions that a particular test has high stakes associated with it, whether true or false, are extremely powerful in defining the reality of how a test is used. Perceptions that a test has high-stakes associated with it are the ignition for test preparation and measurement-driven instruction.

Principle 2

The more any quantitative social indicator is used for social decision-making, the more likely it will be to distort and corrupt the social processes it is intended to monitor.

This Principle comes directly from Donald Campbell's work on social indicators. Principle 2 is a social version of Heisenberg's uncertainty principle: You cannot measure either an electron's position or velocity without distorting one or the other. Any measurement of the status of an education institution, no matter how well designed and well intentioned, inevitably changes its status. This principle reminds us that while, historically, testing is seen as a relatively objective and impartial means of correcting abuses in the system, the negative effects eventually outweigh the early benefits.

A recent report of the Friends for Education, a Working Group for Public Accountability in Education, illuminates this second principle (Cannell, 1987). The report's author, John Jacob Cannell, a West Virginia physician, discovered that "no state is below average at the elementary level on any of the six major nationally normed, commercially available tests" (1987, p. 1). This finding is popularly portrayed as the "Lake Wobegon effect," since in that mythical town all students are above average. Why is it that the Lake Wobegon news is the cause of serious concern rather than national elation? Why don't we trust the apparent good news of Lake Wobegon?

One powerful explanation for this distrust of current standardized test results is the implicit recognition that test results have become so important for students and teachers, so central in the accountability process for administrators, that teaching has been inordinately skewed toward test preparation. A longtime supporter of standardized tests, Albert

Shanker, President of the AFT, recently reassessed his support for such tests in observing:

> Since the reputation of a school, its principal, its teachers and the school board and superintendent depends largely on [standardized] test scores, schools are devoting less time to reading real books, writing essays, and discussing current events and more and more time teaching kids strategies for filling in blanks and choosing the answers to multiple-choice questions. This destroys much of the value of these tests, which only tell you something if they are an independent measure of what the student knows. (1988a, p. 7)

Shanker is describing the corrupting effects of our second Principle. The test drives teaching; this emphasis on test preparation in turn distorts the test's ability to validly portray the skill level of students. The sample of items from the original skill domain has been corrupted, and no longer represents the domain to which we want to make inferences.

The essay component of the Irish Primary Certificate (IPC) examination (a high-stakes test given in Ireland at the end of elementary school between 1943 and 1967) illustrates the cycle of how the test first distorts teaching and the teaching subsequently distorts inferences made from the test. Irish students sitting for the IPC had to write a short essay on a topic which changed each year (e.g. A bicycle ride, 1946; A day in the bog, 1947; A bus tour, 1948). Because the IPC was considered very important in the lives of students and teachers, teachers taught generations of Irish children to memorize a series of stock sentences that could be used with any prompt (e.g. "I awakened early, jumped out of bed and had a quick breakfast. My friend was coming to our house at nine o'clock as we were going for a—*fill in the prompt*—"). As a result of this type of test preparation, a high score on the writing exam was no longer a valid indicator of well-developed writing skills, but only of the students' ability to memorize, recall, and use the stock responses with that year's prompt. The IPC changed teaching, and the ensuing test taking strategies drilled into students vitiated the validity of inferences made from the exam about student's ability to write (Madaus, 1988; Madaus & Greaney, 1985).

How Principle 2 works, and its effects, are described in the remaining Principles.

Principle 3

If important decisions are presumed to be related to test results, then teachers will teach to the test.

Principle 3 simply describes the reality that if the test is important, teachers will teach to it. Why does this happen? First, there is tremendous social pressure on teachers to see to it that their students acquit themselves well on the high-stakes tests. Second, the results of some tests are so important to students, teachers, and parents that their own self-interest dictates that instructional time focus on test preparation.

Proponents of measurement-driven instruction argue that teaching to the test is a virtue. After all, they opine, if the test measures basic skills, preparing students for the skills measured by the test can only improve basic skills. There is no doubt that high-stakes tests focus instruction on specific goals that students must attain. However, their argument is circular. The principal evidence to support the conclusion that basic skills have improved is the undeniable fact that when teachers teach to the test, the test scores, which had been low, rise. However, there is little in the way of independent evidence that the skill of interest necessarily improves simply because the test scores go up. Proponents of measurement-driven instruction ignore two facts: first, that a test is only a secondary indicator of a skill, and second, the ruinous effects of Principle 2.

If rising test scores automatically translate into improvement of basic skills, then why is it that people somehow don't believe the "good news" of rising scores. Why don't they embrace the "Lake Wobegon effect?" As Cannell ruefully observes:

> In 1982, the current administration made a case that our national security depended on reversing "the rising tide of mediocrity" in America's public schools. Recently the U. S. Department of Education made another good case—that few, if any, real improvements have occurred since that *Nation at Risk* assessment. The public seems to agree, judging from how quickly presidential candidates promise to fix our public schools and from sales of books like *Cultural Literacy* by E. D. Hirsch. (Cannell, in press)

If the test is specific to a more specialized curriculum area where higher level cognitive outcomes are the goal—for example, college preparatory physics—then the examination will eventually narrow instruction and learning. Instruction will eventually focus only on those things measured by the tests. Indeed, this narrowing of the curriculum has been one of the enduring complaints leveled at external examinations used for the important functions of certifying the successful completion of elementary or secondary education, and admission to tertiary education or to certain jobs (Madaus, 1988).

We would be remiss if we failed to point out that Principle 3 also has implications for both students and the curriculum. Students adjust their

behavior to a test just as their teachers do. In fact, Garrison Keillor, in his book *Lake Wobegon Days*, describes a second Lake Wobegon effect, a first cousin of Cannell's effect:

> For years, students of the senior class were required to read ["Phileopolis"] and answer questions about its meaning, etc. Teachers were not required to do so, but simply marked according to the correct answers supplied by Miss Quist, including: (1) To extend the benefits of civilization and religion to all peoples, (2) No, (3) Plato, and (4) A wilderness cannot satisfy the hunger for beauty and learning, once awakened. The test was the same from year to year, and once the seniors found the answers and passed them on to the juniors, nobody read "Phileopolis" anymore. (1985)

Principle 3 works its effect on the curriculum in that teachers are much less apt to emphasize material not covered on high-stakes tests. Stake, McTaggart, and Munski (1985), in the conclusion to a case study they conducted of art in two Illinois school districts, offer the following description of the power of tests to mold perceptions about what is important in the curriculum:

> By and large teachers believe that *tests* designed to indicate scholastic aptitude and to measure attained-competence *are* targeted on knowledge and skills that all children should command. Few are troubled that it might be a disservice to many to spend an enormous amount of time getting all learners "proficient." . . . Only socialization and objectives manifest in test items are treated by most teachers as having high priority.
>
> This emphasis on "basics" and testing, and the consequent diminishment of art education, are closely related to cut backs in general funding for education. (p. 56)

A review of the effects of high-stakes tests on the curriculum over many years and in a number of countries indicates that, faced with a choice between objectives which are explicit in the curriculum and a different set of objectives that are implicit in the test, teachers and students generally choose to focus on the latter. In many countries with high-stakes certification tests the amount of instructional time spent on various aspects of the official syllabus is seldom higher than the likelihood of their occurrence on the test (Madaus, 1988).

Principle 4

In every setting where a high-stakes test operates, a tradition of past tests develops, which eventually de facto defines the curriculum.

Given Principle 3, the question remains: "How do teachers teach to the test?" The answer is relatively simple. Teachers see the kind of intellectual activity required by previous test questions and prepare the students to meet these demands. The Irish PC writing exam discussed under Principle 2 illustrates this point.

Proponents of measurement-driven instruction argue strongly that if the skills are well chosen, and if the tests truly measure them, then coaching is perfectly acceptable (Millman, 1981; Popham, 1981; Popham et al., 1985). This argument sounds reasonable, and in the short term, it may even work. However, it ignores a fundamental fact of life: When the teacher's professional worth is estimated in terms of test success, teachers will corrupt the measured skills by reducing them to the level of strategies in which the examinee is drilled. Further, the expectations and deep-seated primary agenda of students and their parents for test success will further corrupt the process. The view that we can coach for the skills apart from the tradition of test questions embodies a staggeringly optimistic view of human nature that ignores the powerful pull of self-interest. It simply doesn't consider the long-term effects of the sanctions associated with the test scores (Madaus, 1988).

Teaching to the high-stakes test is easy. It can quickly become a comfortable form of pedagogy, and students will become proficient at passing tests by mastering the tradition of past tests. Teaching becomes a defensive act. This type of teaching is nonetheless acceptable, because it satisfies "superficially at least, the observable demands of teaching. Yet, in fact, [defensive teaching serves] to control the teaching process and to maintain authority and efficiency for the teacher at the same time" (McNeil, 1988, p. 434). Albert Shanker quotes from Elinor Duckworth's book *The Having of Wonderful Ideas and Other Essays on Teaching and Learning* to describe the effects of this type of teaching on how students are educated:

> The only difficulty is that teachers are rarely encouraged to [help children to come honestly to terms with their own ideas]—largely because standardized tests play such a powerful role in determining what teachers pay attention to. Standardized tests can never, even at their best, tell us anything other than whether a given fact, notion, or ability is already within the child's repertoire. As a result, teachers are encouraged to go for right answers, as soon and as often as possible, and whatever happens along the way is treated as incidental. (Duckworth cited in Shanker, 1988b, p. 9)

Principle 5 describes how the form of the test question can narrow instruction, study, and learning to the detriment of other skills.

Principle 5

Teachers pay particular attention to the form and format of the questions on a high-stakes test (e.g., short answer, essay, multiple-choice) and adjust their instruction accordingly.

This fifth principle describes another weakness inherent in measurement-driven instruction. As we have seen, proponents of measurement-driven instruction argue that if the test is designed to measure important, well-chosen objectives, and the test actually measures those objectives, then teaching for the test is legitimate. However, this argument fails to take into account the power of the form of the test questions to drive instruction and learning. Individual test questions are the building blocks of the test. It is important to keep in mind that by and large the form the questions take in high-stakes programs in our schools is that of the multiple-choice question; students are required to select, not supply, answers. Given the power of our third and fourth principles, teachers when faced with high-stakes tests, will tend to emphasize selection exercises in their instruction and in the practice work they give students. Thus, it it not the skill or objective that drives instruction, it is the type of item used—a rather silly, but expedient and understandable, criterion on which to base instructional decisions.

Given our free enterprise system, publishers have begun to look at state-mandated minimum competency or high-stakes basic skills tests in order to design materials to train pupils to take them. Thus, in the highly commercial milieu of educational testing and textbook publishing, teachers no longer have to figure out the tradition of past exams, because commercially available test preparation materials do it for them.

A hidden aspect of the growing test preparation market is the drill and practice ditto masters and workbooks that present students with multiple-choice questions. In fact, reading instruction in some districts can center on having students read short passages and answer multiple-choice questions. Included in this hidden dimension of test preparation materials are end-of-chapter tests which textbook publishers provide schools on adoption. Because of the power of high-stakes tests, children are apt to find themselves spending more and more time filling out ditto answer sheets or workbooks. The Massachusetts Advocacy Center recently interviewed teachers and students in the Boston Public Schools about the effects of that system's extensive high-stakes testing program on teaching and learning. The Center found that much of the curriculum in "reading" could not qualify as reading at all. Worksheets, vocabulary drills, and answering multiple-choice questions about short passages often took the place of real reading. The stories students read were very

short, often easy, and boring; rarely were students given entire books to read. Consider the description of how one student named "Monica" saw her reading instruction:

> We have U.S.S.R., twenty minutes at the end of the day, but mostly it's "packs" [packets of worksheets]—phonics, spelling, that stuff. It's so boring! Reading is supposed to be reading. But we never really do reading. We do "packs"—vocabulary words, sounds, spelling and dictionary work. But I don't think our reading class should be called "reading." (Massachusetts Advocacy Center, 1988, p. 11)

The Center also found that just as classroom reading instruction had come to imitate reading tests, remediation programs have also evolved into more test preparation. Another student, 14-year-old "Carlos," describes his remedial class as:

> . . . much too easy. They give you baby stuff—books with big words [i.e., typeface]. And so short! Three sentences would make a whole story. I get mad because I read fast and they [are] so short. It's boring! And the questions—so simple! Big words [typeface], you know, like for old people—or for baby books. (Massachusetts Advocacy Center, 1988, p. 12)

Another growing segment of the commercial test preparation market is outright test preparation material. Some of this material is designed to familiarize students with test format, scoring rules, guessing strategies, and test anxiety reduction techniques. These materials are certainly appropriate and are not destructive of either instruction or test validity (Mehrens & Kaminski 1988, Shepard & Kreitzer, 1987). However, as the material begins to resemble the test material more closely, the effects of Principle 2 begin to appear. For example, some of the commercial material gives teachers instructional material that is:

> (1) based on objectives (skills, subskills) that specifically match those on the standardized test to be administered,
> (2) specifically matched to objectives (skills, subskills) and cast in the same format as the test questions,
> (3) built around a published parallel form of the actual test, and
> (4) designed as practice (instruction) on the actual test. (Mehrens & Kaminski, 1988)

Mehrens and Kaminski argue that (3) and (4) are always illegitimate and that the ill-defined line between legitimate and illegitimate test preparation is somewhere between practices (1) and (2). As the commercial material begins to resemble points (3) and (4), its use prior to high-

stakes testing will jeopardize any inferences one might make from the test to a larger domain. Apart from distorting instruction and destroying the validity of the test, the use of test preparation materials, whether commercial or prepared by school districts, has a network of consequences associated with it.

Unfortunately, this network of consequences remains largely unexplored. For example, we don't know how extensive the practice of test preparation is; how much of it is inappropriate; what the associated dollar cost, instructional costs, and teacher costs are; what types of districts and teachers are most apt to use test preparation material; and what types of students receive this type preparation. These and other questions need answers. The network of consequences flowing from test preparation is potentially too serious for students, teachers, administrators, and the general public to ignore any longer.

Principle 6

When test results are the sole or even partial arbiter of future educational or life choices, society tends to treat test results as the major goal of schooling rather than as a useful but fallible indicator of achievement.

This Principle is best summed up by the following observation from a 19th century British school inspector who observed firsthand the negative effects of linking teacher salaries in England and Ireland to pupil examination results:

> Whenever the outward standard of reality (examination results) has established itself at the expense of the inward, the ease with which worth (or what passes for such) can be measured is ever tending to become in itself the chief, if not sole, measure of worth. And in proportion as we tend to value the results of education for their measurableness, so we tend to undervalue and at last to ignore those results which are too intrinsically valuable to be measured. (Holmes, 1911, p. 128)

Of all of the effects associated with high-stakes tests those typified by this final Principle and beautifully described by Holmes, may well be the most injurious. It seems that Holmes' observation that measurableness becomes equated with educational worth and excellence is as apropos today as it was 80 years ago. High-stakes test scores have become the principal criterion used by policy makers, the business community, the general public, and unfortunately, many educators when evaluating systems, schools, teachers, and children. When this happens, of course, both instruction and the test are injured in the ways discussed above.

The irony is that the use of the test as the principal measure of worth destroys the test's ability to serve as an accurate indicator of student attainment.

What Can Be Done?

If the description of the effects of the six principles discussed above is correct, a question remains: what can teachers do when faced with a high-stakes testing program and implicit or overt pressure to let measurement drive instruction? It's not an easy question to answer. The direct and indirect pressures associated with the use of high-stakes test results may be very difficult to resist. Each teacher will have to arrive at his/her own decision on how to cope. What teachers can do collectively, however, is lobby their professional organizations and state and district level policy makers to lower the stakes associated with testing. Shanker's recent column pointing out the dangers of high-stakes testing is certainly a step in the right direction. Further, there must be a concerted effort to educate policy-makers and business leaders to the dangers associated with high-stakes testing.

Teachers also need to lobby for school-based management and a reduction in the bureaucratization of teaching. However, if school-based management is to work, schools must receive waivers from high-stakes testing programs. Without such waivers, the power of the high-stakes testing could destroy authentic school-based management. Waivers can only be obtained through collective bargaining and negotiation when teacher contracts are up for renewal. This does not mean that testing stops. Policy-makers, the business community, and public will demand, and have a right to expect, that the schools are accountable. What needs to be negotiated is an agreement that a host of indicators of student achievement will be developed and used, and that no one indicator by itself will automatically trigger a high-stakes decision.

Teacher organizations need to work toward building a consensus among all the parties interested in school improvement on issues related to testing. Such a consensus might include the following points. First, tests used for accountability should be given on a sampling basis. Second, the tests should measure the skills of interest as directly as possible. Third, when traditional standardized multiple-choice tests are used, they should be given unannounced, on a sampling basis, and the results used to assess curricular impact—not individual teacher or student merit. Fourth, standardized test results must be interpreted in light of other direct indicators of the domain of interest (e.g., student portfolios, having students read directly from a book and explain what has

been read, student writing samples, teacher evaluations). Fifth, teachers need to be intimately involved in the development of these direct indicators. Finally, while these direct indicators of student achievement may very well support the standardized test results, when they contradict such results, a determined effort must be made to look at the direct indicators alongside the standardized results and ask why such a discrepancy exists.

If such a consensus could be developed we would be well on the road to schools where instruction is shaped by teacher discernment and judgment rather than driven by our measurement tools. Testing will then be a tool of instruction rather than the end of instruction.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Cannell, J. J. (In press). Cannell's response to the responders: All 50 states cannot be above the national average. *Educational Measurement: Issues and Practice*.

Cannell, J. J., & Friends for Education. (1987). *Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average*. Daniels, WV: Friends for Education.

Holmes, E. G. A. (1911). *What is and what might be: A study in education in general and elementary in particular*. London: Constable.

Keillor, G. (1985). *Lake Wobegon Days*. New York: Viking.

Madaus, G. F. (1985, Winter). Public policy and the testing profession. *Educational Measurement: Issues and Practice, 4*, 5-11.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh yearbook of the National Society for the Study of Education* (pp. 83-121). Chicago: University of Chicago Press.

Madaus, G. F., & Greaney, V. (1985). The Irish experience in competency testing: Implications for American education. *American Journal of Education, 93*(2), 268-294.

Massachusetts Advocacy Center. (1988, September). *Status report: The way out: Test score patterns and reading in the Boston Public Schools*. Boston, MA: Author.

McNeil, L. M. (1988, February). Contradictions of control, Part 2: Teachers, students, and curriculum. *Phi Delta Kappan, 69*, 432-438.

Mehrens, W. A., & Kaminski, J. (1988, April). *Using commercial test preparation materials: Fruitful, fruitless, or fraudulent?* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Millman, J. (1981, Fall). Protesting the destesting of PRO testing. *NCME Measurement in Education, 12*, 1-6.

Office of Technology Assessment, U.S. Congress. (1987, December). *State educational testing practices*. Washington, DC: Author.

Popham, W. J. (1981, October). The case for minimum competency testing. *Phi Delta Kappan, 63*, 89-92.

Popham, W. J., Cruise, K. L., Rankin, S. C., Sandifer, P. D., & Williams, P. L. (1985, May). Measurement-driven instruction: It's on the road. *Phi Delta Kappan, 66*, 628-635.

Shanker, A. (1988a, April 24). Exams fail the test. *New York Times*, E, p. 7.

Shanker, A. (1988b, November 13). The road to educational diversity. *New York Times*, E, p. 9.

Shepard, L. A., & Kreitzer, A. E. (1987). The Texas teacher test. *Educational Researcher, 16*, 22-31.

Stake, R. E., McTaggart, R., & Munski, M. (1985). *An Illinois pair: A case study of school art in Champaign and Decatur.* (Reprinted from *Art History, Art Criticism, and Art Production: An Examination of Art Education in Selected School Districts*. Rand Corporation.)

State Accountability Systems. (1988, October 12). *Education Week*, p. 14.