DOCUMENT RESUME

ED 379 337 TM 022 716

AUTHOR Bell, Gregory
TITLE The Test of Testing: Making Appropriate and Ethical Choices in Assessment.
INSTITUTION North Central Regional Educational Lab., Oak Brook, IL.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 94
CONTRACT RP91002007
NOTE 30p.
AVAILABLE FROM Publications Department, North Central Regional Educational Laboratory, 1900 Spring Road, Oak Brook, IL 60521 ($4.95, order number RPIC-TT-94).
PUB TYPE Reports - Evaluative/Feasibility (142)

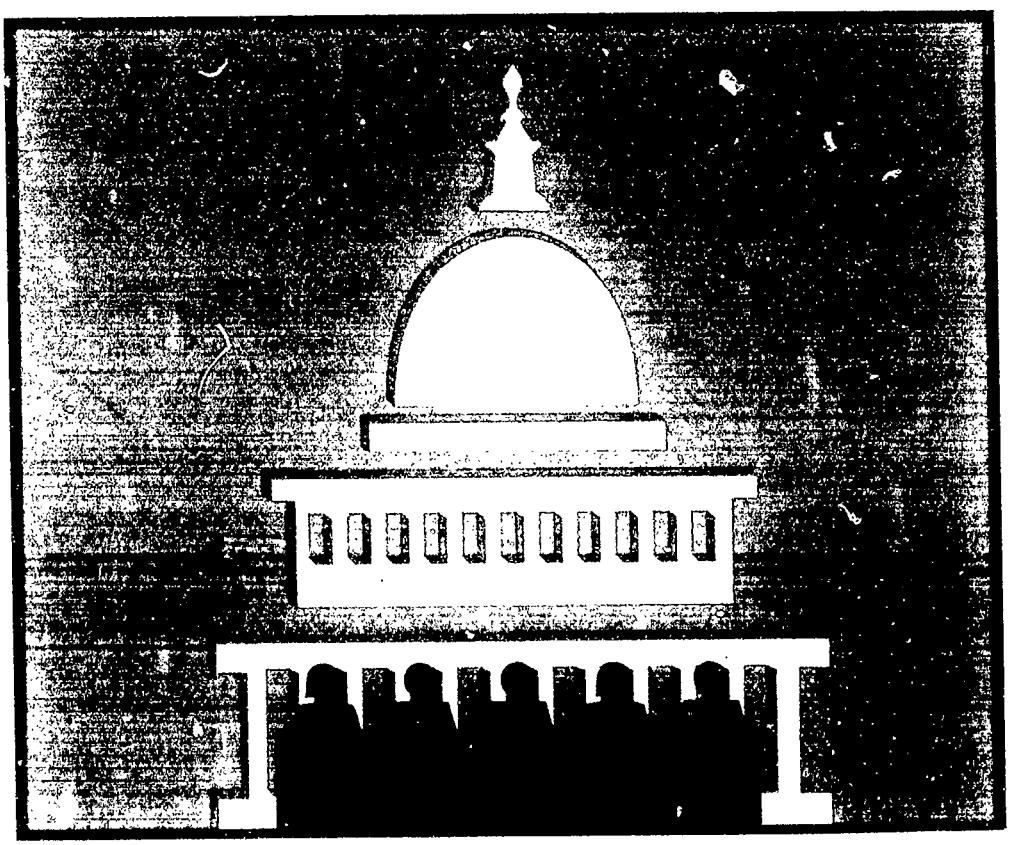
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Assessment; *Ethics; Selection; *Student Evaluation; Test Bias; Test Coaching; Test Construction; *Testing; Test Interpretation; *Test Use; Test Wiseness

ABSTRACT

The realities of inappropriate and sometimes unethical testing practices must be confronted to make sure that current assessments are as reliable and effective as possible. While this paper does not attempt to provide practical guidelines for ethical and appropriate testing, it does draw a picture of appropriate and ethical testing practices. All persons involved in testing programs should try to maintain their focus on the fundamental reasons for testing which is the education of the students being assessed. It is vital that the selected test be appropriate for specific purposes and intended populations, and that all intended and possible unintended uses be considered. Because preparing students to take the test is the source of many problems with assessment, a continuum of appropriate test preparation practices is suggested. Issues that must be considered in administering the test are reviewed, from disclosure through improper use and interpretation and test bias. (Contains 27 references.) (SLD)

The Test of Testing: Making Appropriate and Ethical Choices in Assessment

REGIONAL POLICY INFORMATION CENTER



By Gregory Bell

NCREL

North Central Regional Educational Laboratory
1900 Spring Road, Suite 300
Oak Brook, IL 60521
(708) 571-4700, Fax (708) 571-4716

NCREL is one of ten federally supported educational laboratories in the country. It works with education professionals in a seven-state region to support restructuring to promote learning for all students—especially students most at risk of academic failure in rural and urban schools.

The Regional Policy Information Center (RPIC) connects research and policy by providing federal, state, and local policymakers with research-based information on such topics as educational governance, teacher education, and student assessment policy.

This publication is based on work sponsored wholly or in part by the Office of Educational Research and Improvement (OERI), Department of Education, under Contract Number RP91002007. The content of this publication does not necessarily reflect the views of OERI, the Department of Education, or any other agency of the U.S. Government.

RPIC-TT-94 ($4.95)

The Test of Testing: Making Appropriate and Ethical Choices in Assessment

by Gregory Bell

Table of Contents

The Test of Testing: Making Appropriate and Ethical Choices in Assessment

by Gregory Bell

I. Introduction

Millions of American children take tests every school day. Most tests are the kind familiar to all of us—the quizzes and in-class tests that always have been important tools in teaching. However, another type of testing has come to exert a pervasive and profound influence on American education. Ranging from commercial norm-referenced exams and assessment programs used at the district or state level, to state-developed assessment programs and the National Assessment of Educational Progress (NAEP), to newly created performance assessments, these tests have been called upon to serve many functions. The mandate for these tests originates outside of the classroom, at the district, state, and federal levels.

The results of these assessments are used primarily to inform decisions made by the policymakers and administrators who shape and direct the provision of education. They often help decision-makers monitor the ability and achievement of students and determine eligibility for access to special programs and resources. They also are used to illustrate the state of American education at the school building, district, state, national, and even international levels. And they are increasingly used to drive education reform and policy.

In some cases, test results become the principal—if not only—measure by which institutions and educators are judged in this "age of accountability." They often play a pivotal role in important decisions about the future of programs, the level and allocation of funding, salary increases, and whether administrators or teachers are praised or sanctioned. Even when they do not form the foundation of official decisions, the results of these tests are likely to be used by the public to assign blame or praise, to distinguish schools or teachers as "good" or "bad," and to advance political and educational agendas.

While motivated by a sincere desire to increase the quality of education, the infusion of high stakes into the administration of large-scale state or national testing programs and into the use of the results has put intense pressure on educators to improve the test scores of their students. This pressure can create a climate in which test performance drives what is—and, just as important, what is *not*—taught, and how it is taught. This atmosphere may distort the teaching process, restrict the scope of the education that schools provide, and influence the results of the test itself.

Some critics—the most conspicuous being John Cannell—charge educators with purposefully misleading the public on the subject of student achievement through inappropriate or unethical testing practices. Cannell has argued that these abuses have resulted in a seemingly absurd "Lake Wobegon effect," under which all 50 states report scores above the national average (Cannell, 1988, 1989).[1] Of course, when confronted with the pressure to improve student performance on tests, certain individuals will choose to engage in clearly unethical or inappropriate efforts to raise the scores. However, it is likely that many educators who

become involved in inappropriate testing practices do not realize that their conduct is improper.

Categorizing practices as "unethical" or "inappropriate" should take into account the atmosphere or context in which assessments are administered. Some test administrators view over-reliance on any high-stakes assessment for making decisions that affect students or institutions to be inherently fallible (Smith, 1991). Such individuals may make assessment decisions without adequate regard for factors over which educators and schools have little control, such as poverty, the educational aspirations and backgrounds of parents, and medical or mental disabilities. These educators may experience a dissonance between their daily, intimate knowledge of a student's potential and that student's performance on the assessment. It is also possible that the assessment does not adequately sample the curriculum that is taught in the classroom. The resulting skepticism and sense of inequity may seem to justify testing practices that could be characterized as inappropriate or unethical. Although such a response cannot be condoned, the factors surrounding it must be acknowledged and overcome. By dealing with these factors and limiting the connection between high-stakes decisions and a single assessment, policymakers and administrators can reduce inappropriate and unethical testing practices.

Of course, it is easy to assess "blame" when inappropriate or unethical practices are employed to raise scores on tests or when test results are misused. By using such practices to raise test scores, teachers may believe that they are "saving their jobs," administrators may be attempting to "promote" their schools or districts, a politician may be attracting attention, or others may be acting out of self-interest. However, the roots of inappropriate or unethical testing practices can be more complex. The causes can originate with anyone involved in assessments, including those who develop the tests, the policymakers and administrators who choose assessments and interpret or act upon the results, and the educators who prepare students for tests and administer them.

When testing programs are selected, administered, and used appropriately, they can make a valuable contribution to American education. And until we develop a more comprehensive means of student assessment to provide the information needed in making informed educational decisions, testing will continue to play a dominant role in these decisions. For this reason, we must confront the reality of inappropriate and sometimes unethical testing practices in order to make sure that current assessments are as reliable and effective as possible.

The misuse of testing is probably unintentional in many instances. Individuals may not completely understand their proper roles or the range of appropriate practices involved in standardized testing. A better understanding may help educators make informed decisions. To that end, this paper will attempt to identify the chief responsibilities of educators and administrators at the classroom, district, and state levels when conducting assessments. After discussing some of the consequences associated with inappropriate or unethical testing

practices, this paper will look at the specific responsibilities of educators and decision-makers in testing.

These responsibilities include selecting an assessment instrument, preparing students for testing, administering the test itself, and interpreting or using the results. This discussion is derived from various codes, guidelines, and suggestions advanced by the professional assessment community and others. This paper does not attempt to provide a practical "how-to" resource or a comprehensive set of guidelines for ethical and appropriate testing practices. The complexity and scope of this task is best left to the professional assessment and research communities and educational policymakers, who continue to do a substantial amount of work in this area. This paper also does not address the ethics of implementing a high-stakes assessment program. Some would argue that high-stakes tests result in an over-reliance on test results, which in turn encourages unethical testing practices. This paper merely attempts to draw a clearer picture of appropriate and ethical testing practices for those educators "in the trenches."

II. The Effects of Inappropriate and Unethical Testing Practices

> A high stakes outcome . . . 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. (Madaus, 1990, p. 34)

The most damaging effects that inappropriate and unethical testing practices have on American education, regardless of their source, are their impact on the value of the test itself and the scope of the education that students receive. The following is a short excursion into basic testing theory, which is the foundation of this discussion.

Tests, whether high stakes or not, are instruments that take a *sample* of questions or tasks from some content domain to *represent* the more important broader whole. What makes a test valuable is the degree to which a student's performance on the sample supports an inference concerning whether the student understands or has mastered the larger domain. The correctness of this inference—its validity—is the single most important concept in testing, preserving the reliability of the measurement. If an unethical practice inflates test scores, one can no longer infer that a good score indicates mastery of the larger content domain being sampled by the test. Therefore, the validity of the test—that is, whether or not the test results can be interpreted as intended—is jeopardized (Mehrens, 1984).

Test validity is the extent to which an inference and any resulting decisions about or characterizations of individual students, teachers, or institutions based on test performance can be considered appropriate and meaningful (Madaus, 1990). However, test validity must

not be measured in a vacuum—when measuring test validity, we should consider whether test results are meaningful for particular populations and uses.

When a testing program leads to important decisions or outcomes—either actual or perceived—a process may begin that corrupts the test's ability to represent the relevant domain. And a test that no longer represents the relevant domain undermines the validity of any inferences drawn from that test and any decisions based on those inferences. The more a testing program influences such decisions, the more it distorts what it is intended to measure. This effect is one of perception—if students, teachers, or administrators believe that the results of the test will have important consequences for them as individuals or for their institution, it does not matter whether their impression is true (Madaus, 1990). An intense pressure to "teach to the test" develops. This pressure does not justify a clearly unethical reaction—educators can choose to resist it—but their choices in preparing and conducting the assessment could be influenced by it. The influence may be subtle, but as it grows it can lead to harmful effects. Subject areas and intellectual activities that the assessments do not measure may receive less attention. Rote memorization skills may eclipse higher-order thinking skills. The particular format of the assessment may force instruction to focus on that format and measure only those skills that will help students find "right" answers within that context. Finally, and most important, all involved—administrators, teachers, parents, and students—may believe that improved test results are the primary goal of education, not merely a useful indicator of student learning.

The point is that focusing on the test may corrupt the validity of any inferences made from it about the wider domain as the domain of the student's knowledge approaches that of the test's sample. In the end, narrowing the scope of what is taught may leave students better prepared for increased performance on the tests, but perhaps no better equipped for the challenges that they face later in life.

III. Roles and Responsibilities

The validity of testing programs is undermined by inappropriate or unethical test preparation or administration, inappropriate use of the tests, and factors beyond the control of schools and their personnel. These effects have been called "test score pollution" (Haladyna, et al., 1991; Messick, 1984). Indeed, even appropriate testing practices may "pollute" the validity of testing programs, since the individuals and institutions that the tests measure and compare do not always follow the same practices (Haladyna, et al., 1991).

Test pollution must be minimized in order to preserve the integrity of the assessment process and the validity of the decisions made from that assessment. To assist those involved in the development and execution of assessment policy, this paper outlines "ethical" and "unethical" assessment practices and explains how the various players in assessment planning can minimize the test pollution that can be caused by unethical practice.

The remainder of this paper will be devoted to addressing some of the specific roles and responsibilities that educators and administrators should assume while conducting assessments. How these actors conduct themselves will have a profound effect on whether an assessment effectively serves its purposes and is fair to all who may be affected by its results. Although each of these roles carries with it specific issues and resolutions, some general responsibilities may be applied to anyone involved in educational assessment.

A. General Ethical Assessment Responsibilities[2]

Regardless of their role, all persons involved in testing programs should strive to maintain their focus on the fundamental reason for any testing program—the education of the students being assessed. The most effective way to achieve this goal is to ensure that the tests used are reliable and that no practice used in preparing for and administering the tests detracts from their validity.

Anyone who takes on a role in assessment must have the experience and competence to play that role effectively and should try to promote appropriate and ethical assessment practices. Those involved in assessment programs also should maintain and continually improve their competence, serving as an example to others. In addition, these individuals should work to increase the literacy of other educators, administrators, parents, and the general public in sound testing practices.

B. The Selection and Development of Testing Programs[3]

Developing and selecting assessments require developers, policymakers, and administrators to make important choices that can have a significant impact on the usefulness of the assessments and the validity of the results. Those involved in developing and selecting assessments are responsible for ensuring that the instruments are well crafted and are suited to the students being assessed.

1. Appropriate for Specific Purposes and Intended Populations

An assessment should be selected only when it satisfies the specific purposes for which it is to be used and is appropriate for the intended population(s). Therefore, the first responsibility of those who select tests is to define clearly the purpose(s) for testing and understand the characteristics of the population that they wish to assess. An assessment selected for one purpose or population may not be effective for another. An assessment should be selected only after all potential assessment strategies or instruments have been thoroughly and objectively evaluated within the context of the intended use(s). In order to promote a dialogue during the selection process, all prospective users of assessments under consideration should be informed of the strengths and weaknesses of the various assessments available, their relative costs, and their appropriateness for the intended use(s).

Unfortunately, some test selectors accept the title of a test as an accurate and complete explanation of what the test measures. Those who select tests should evaluate them based on *documented* evidence of their technical quality and utility, rather than on unsubstantiated claims by the test developer or others. Besides reading the materials provided by the developer, those who select tests should investigate other potentially useful sources of information, such as *The Eleventh Mental Measurements Yearbook* (Buros Institute of Mental Measurements), to corroborate the claims of test developers and testing materials. This appraisal of an assessment could include reading independent evaluations of the test, interviewing others who have used the assessment, and conducting a careful evaluation of specimen sets, disclosed tests or sample questions, directions, manuals, answer sheets, and score reports. The appraisal should address the validity and reliability of the instrument, the age and adequacy of the norms used in its development, and any evidence of bias detection. Assessments recommended for use should be independently substantiated and qualified, regardless of how much praise is heaped upon them.

Testing programs should be developed in such a way that they minimize possible bias based on gender, ethnicity, socioeconomic status, religion, age, or other characteristics. Those responsible for the selection of an assessment should seek evidence from an assessment's developers and others in order to substantiate claims that the instrument minimizes potential bias. In addition, all prospective assessments should be evaluated to determine whether sufficient, documented evidence exists to indicate that they may be validly administered, interpreted, and used for the population(s) to be assessed, including both content and norms or comparison group(s) used.

Those who select tests should ensure that the individuals who are assigned to evaluate the content and technical quality of assessments are competent. They should have a thorough understanding of the *Standards* developed by the American Educational Research Association, the National Council on Measurement in Education, and the American Psychological Association for educational and psychological testing. Assessments should not be selected if no one who is competent to administer the test or interpret its results is available to potential users, especially when the population to be assessed includes individuals subject to physical and educational disabilities, limited English proficiency, or other special conditions.

Finally, the test selection process should be thoroughly documented in order to provide a record that can be consulted by all users of the assessment and those who represent the students assessed. This documentation also will be helpful for evaluation of future assessments.

2. Information Is Needed from Test Developers

Test developers should provide the information that test users need in order to select appropriate tests. They should accurately represent their instrument—its purposes, characteristics, uses, and limitations. Developers should ensure that the assessments they

produce meet professional standards. They should seek independent evaluations to determine whether the tests are appropriate for the uses and populations that they purport to serve. Developers should completely and objectively report data on the pretesting, standardization, validation, and other steps taken in producing the instrument, including both positive and negative consequences that these data may have on the use of the assessment.

Of course, supporting materials should not make inaccurate or misleading claims about the instruments, their uses, or the interpretation of the results. Test developers should not withhold information concerning their assessments, even when the disclosure could adversely affect the use of the instrument. When inaccuracies in testing instruments or their supporting materials become known, they should be corrected as soon as feasible.

Designers also should attempt to minimize the re-use of test formats, items, or tasks when such variations will not interfere with reliable and efficient measurement. The expense, efforts, and technical problems associated with the development of multiple test forms ought to be weighed against the possibility that the results derived from a test will be contaminated due to familiar and often-used tasks.

Test developers should identify any special skills needed to administer and interpret the results of their test. They should emphasize the importance of ensuring that those who use and interpret the selected instrument are competent to do so effectively. When necessary, however, they should explain relevant concepts at a level of detail appropriate for those needing guidance.

3. Considering the Uses and Consequences of Testing

During both the development and selection process, those involved should consider all of the intended and unintended uses to which the assessment might be put by those who will have access to its results, including both the positive and negative consequences of those uses. They also should consider how policymakers and the public might use or interpret the results. Prospective users should be informed about the potential ways in which an assessment's results could be misused or over-interpreted, since prior disclosure and understanding of these possibilities and their consequences may lessen the potential for such problems.

4. Mixed Motives and Conflicts of Interest

Because many standardized assessments are purchased from commercial vendors, it is important for those who select tests to keep in mind that a developer's motives may be mixed. The developer may suggest a particular assessment for reasons having little to do with its appropriateness for the goal or task at hand. Those who select tests must consider the possibility that an assessment's developer has not met its responsibilities. In order to minimize the potential for selecting an inappropriate test, test selectors should ensure that the developer has met its responsibilities before selecting the instrument.

Possible conflicts of interest by test selectors should be disclosed to all involved with the assessment, particularly if the selectors have any associations or affiliations with the authors, publishers, or others involved in developing the testing programs being considered. Attempts by any party to exert undue influence on the selection process should be disclosed prior to selection. Potential problems may be reduced when assessments are independently evaluated by a disinterested third party.

5. Test Security

When selecting an assessment, it is necessary to preserve the security of the assessments being reviewed in order to undermine efforts to raise test scores through inappropriate preparation practices. For the same reason, tests should be kept secure during the development process. Such steps should at least include signed agreements to protect security by those with access to the tests, limited access to the location in which the test is being developed and to the test development materials themselves, collecting and destroying notes and drafts, making sure that no copies are lost or stolen during the printing process, and accounting for all test development materials before they are distributed.

C. Preparing Students for an Assessment

Whether the specific means are imposed from above or develop out of a fear of the impact of bad scores, preparing students to take the test is the source of many of the problems with assessments. Preparation is the point at which the pressure to raise test scores is highest, having its greatest effect on the individuals who have the least input or power within the process. Teaching to the test is a human—although sometimes inappropriate or unethical—response to this situation. The logic is understandable: If educators are to be held accountable for their students' learning a particular type of knowledge or set of skills, it is in their best interest to teach that type of knowledge or set of skills. Indeed, tests often are used by policymakers or administrators to drive the kind of instruction that they believe students should receive. The problem is in defining what the "specific set of things" is—a particular sample of the content tested, actual test questions, the domain of objectives from which the test objectives are sampled, or the domain of items from which the test questions themselves are sampled (Mehrens, 1991).

Since test preparation may affect a test's validity—resulting in test score "pollution" or limiting what students actually learn to the content of the test—the question of whether specific preparation practices are appropriate is vital. Unfortunately, the line dividing appropriate and inappropriate practices is not always clear. Although several surveys indicate that educators prepare their students using practices that the professional assessment community considers inappropriate or unethical, none of the national *Codes* or *Standards* have directly addressed issues of test preparation (Hall & Kleine, 1990; Nolen, Haladyna, & Haas, 1990).

1. Some General Guidelines: Accuracy of Inferences, Ethical Role Models, and Educational Indefensibility

One universal guideline in test preparation is to avoid any activities that could undermine the accuracy of the inferences drawn from the test scores (NCME Task Force, 1991). Popham has offered two other general standards to determine whether a particular practice is appropriate (Popham, 1991):

(1) Test preparation should not violate the ethical standards of the education profession. Educators should not violate general ethical standards concerning theft, cheating, lying, and the like. In addition, educators must realize that because they act *in loco parentis*—in place of the parent—they have an ethical obligation to serve as models of behavior for their students. This ethical foundation should inform choices made when preparing students.

(2) Test preparation should have "educational defensibility." Under this concept, a test preparation activity that raises student test scores is inappropriate unless it simultaneously increases student mastery of the content domain tested. Test preparation, as with any instructional activity, should be employed in the best interest of the students. Accordingly, because inappropriate test preparation practices deprive students of a portion of their education and deceive them (and others) about their true mastery of a subject, such activities are educationally indefensible. Some think that Popham's "educational defensibility" standard suggests that any test preparation activities that increase student mastery of the content domain should be considered appropriate or at least defensible (NCME Task Force, 1991).

Since testing is sometimes consciously aimed at improving instruction, test results are sometimes properly used by educators to adjust curriculum and instruction. To account for these adjustments in instruction, the possibility of the score inflation that results should be acknowledged when scores are reported and used, rather than taking action that could undermine any improvement in instruction that the test may generate. Finally, changing the test content from year to year also will help focus instruction on the underlying domain rather than specific test content.

2. Test-Specific Instruction

To ensure that test preparation activities do not undermine the accuracy of the inferences drawn from an assessment, some test selectors look to a test's content and format domain to derive more specific guidelines (NCME Task Force, 1991). For most standardized tests, where the domain of interest is larger than the set of objectives tested, it is inappropriate to limit instruction to the objectives actually sampled on the test. It is therefore inappropriate to use commercially or locally prepared instructional guides that claim to provide students with focused practice and review of only the skills necessary to perform well on the current edition of the test (Mehrens, 1991). Some criterion-referenced tests, however, cover all of

the objectives in the domain of interest. In such cases, when the domain objectives and test objectives are the same, it is appropriate to teach the particular objectives.

Preparing students for an assessment using the actual questions or tasks found on the test or using the current test itself is almost universally considered to be inappropriate (NCME Task Force, 1991; Mehrens, 1991). This practices teaches the students the sample of the domain, not the domain itself.

How students perform when they are asked questions about a domain only in a particular way (e.g., multiple choice or short answer, or always using the same phrasing, terms, or manner of presentation) is of little use to educators. Rather, educators must be able to show that their students' performance on one format indicates how they would perform in other formats and indicates their overall mastery of the content domain. For these reasons, it is inappropriate to limit preparation activities to questions that are framed in the format used on the test. However, it is appropriate to teach test-taking skills by spending a small amount of time teaching students how to work with various types of formats (NCME Task Force, 1991).

3. Preparation Activities: Where Do We Draw the Line?

Several efforts have been made to organize test preparation activities, arranging them along a continuum and attempting to draw a line between activities that are appropriate or ethical and activities that are not. Since these attempts address practices somewhat differently than the above guidelines, it may be helpful to review them. These various attempts will be explained separately from one another, however, because they identify the practices on the continuum differently and reach somewhat different conclusions regarding where the line should be drawn.

Mehrens and Kaminski arrange the following seven test preparation activities along a continuum from the most ethical to the most unethical (Mehrens & Kaminski, 1989):

(1) General instruction on objectives that were not determined by looking at any set of published test objectives

(2) Teaching test-taking skills

(3) Instruction on objectives generalized from objectives measured on a variety of tests

(4) Instruction based on objectives that specifically match those on the test to be taken

(5) Instruction based on objectives that specifically match those on the test to be taken following the same format as the test questions

(6) Practice or instruction on published "parallel" forms of the current test

(7) Practice or instruction on the same test

According to this analysis, the practice described at (1) is always ethical, while the practices described at (6) and (7) are never ethical. Teaching test-taking skills generally may be considered acceptable. The "point where one crosses over from a legitimate to an illegitimate practice" lies somewhere between (3) and (5) (Mehrens & Kaminski, 1989, p. 16). However, Mehrens and Kaminski suggest that the acceptability of test preparation practices—the place where the line should be drawn—may vary depending on what is intended to be measured with the assessment.

Other researchers have evaluated a slightly different range of practices (Haladyna, et al., 1991). Their conclusions concerning where to draw the line are much more exacting, based on their concerns about test score pollution and equity, since they believe that "scores are and will continue to be used to compare the educational effectiveness of teachers, administrators, classes, schools, districts, states, and nations" (Haladyna, et al., 1991, p. 4). They place test preparation activities along the following continuum:

(1) Training in testwiseness skills

(2) Checking answer sheets to make sure that each has been properly completed (only to the extent that the test developer recommends it or all units that are being compared engage in the same practice)

(3) Increasing motivation for improved performance through appeals to students, parents, and teachers

(4) Developing a curriculum based on the content of the test

(5) Preparing objectives based on items on the test and teaching accordingly

(6) Presenting items similar to those on the test

(7) Using commercially prepared score-boosting materials, such as *Scoring High*, or other activities aimed specifically at boosting scores

(8) Dismissing low-achieving students on testing day to boost scores artificially

(9) Presenting items verbatim from the test to be given

Haladyna, et al., find only the practices described in (1) through (3) to be ethical. The practices described in (4) through (7) are considered unethical, while those in (8) and (9) are found to be highly unethical (Haladyna, et al., 1991, p. 4). Their review of other reports

and surveys revealed evidence of the "staggering" degree to which score polluting practices are used (Haladyna, et al., 1991, p. 4-5). It is important to recall, however, that they also assert that even ethical preparation practices can lead to test score pollution if the institutions or individuals that are to be compared prepare for the test differently.

Popham presents five common test preparation activities and assesses their appropriateness through reference to his two evaluative standards (Popham, 1991):

(1) *Previous form preparation:* special instruction and practice based directly on students' use of a previous form of the test

(2) *Current form preparation:* special instruction and practice based directly on students' use of the form being employed

(3) *Generalized test-taking preparation:* special instruction covering test-taking skills relating to a variety of test formats

(4) *Same-format preparation:* regular classroom instruction dealing directly with the content of the test, employing only practice items in the same format as on the test

(5) *Varied-format preparation:* regular classroom instruction dealing directly with the content of the test, but employing practice items that represent a variety of formats

Popham rejects the use of "previous form preparation." This activity is not educationally defensible, he argues, because it is more likely to raise test scores without bringing about a corresponding increase in the students' mastery of the content. This type of preparation also may be unethical because it may appear to the public and others to be coaching merely to raise scores. Popham suggests that this conclusion also applies to the use of commercial test preparation materials that are based chiefly on newly created "parallel" forms of the test that is being used.

"Current form preparation" is a clear loser on both standards. Popham considers the use of actual test items when preparing for an assessment to be outright cheating.

Popham sees "generalized test-taking preparation" as appropriate, as long as it is brief and does not seriously stray from the students' ongoing education. Indeed, insofar as such preparation equips the students for coping with a number of formats, Popham believes that the performance of students prepared in this manner may in the end more closely reflect their true mastery of the content.

"Same format preparation" may be ethical, but Popham concludes that it is educationally indefensible. If, during their regular instruction, students are only exposed to the same item format that will appear on the test, their ability to generalize what they have learned is seriously undermined.

"Varied-format preparation" satisfies both standards for Popham. If students are prepared during their regular classroom instruction and are provided with instruction on the test content not only as it is conceptualized or formatted on the test, but also in other ways, an increase in test scores for students will likely correspond to an expansion in their mastery of the content. Unlike some other researchers, Popham perceives little problem in dealing directly with the content of the test, as long as it is done during regular classroom instruction and the exercises involved vary in format.

Although inappropriate test preparation can be subtle or blatant, it always will affect the validity of the inferences that can be drawn from an assessment. Due to the complexity of testing—the interaction of its objectives, format, and uses—it is unlikely that a clear line can be drawn between appropriate and inappropriate practices. The approaches above, and those to come from further research, should be considered within that context. Those responsible for the definition of ethical assessment practice will need to consider their circumstances and the consequences of their testing results to determine which types of preparation will be in the best interest of their students.

D. Administering the Test[4]

Everyone involved with a testing program expects it to be implemented with appropriate care. Those who have a stake in the results must be sure that they can trust the accuracy of the data that the assessment provides. To that end, all efforts should be made to see that the administration of an assessment does not undermine its reliability and validity. In addition, the importance of test security needs to be continually emphasized. Breaches in security or deliberate attempts to manipulate the test results are serious and should be treated as such.

Uniformity and security during the administration of a testing program is a key component in ensuring that the assessment is reliable and useful. If security is lacking, many doors may be opened to those whose response to pressure is to cheat or otherwise inappropriately raise the test scores of their students. If a test is not administered uniformly, the inferences and uses for which test developers have validated their assessments may become meaningless, since these inferences and uses are often inherently linked to the way in which the test is administered.

1. Disclosure

Prior to a test, all involved in an assessment, including those who are to be tested and their representatives, should be told why the information is being collected, how it will be judged or scored, and how the results of the test will be reported and used. They should know who will have access to the results and how the testing results will be distributed and kept on file. Such disclosure will help to identify potential inappropriate practices by providing additional "eyes and ears" who possess enough understanding of the assessment to recognize its abuse.

2. Developing a Written Assessment Policy

Those responsible for an assessment should ensure that all who administer it are instructed in appropriate test administration practices. To that end, a written testing policy should be developed and disseminated that explicitly spells out practices to be followed and avoided, clearly outlining the responsibilities of students, teachers, and administrators in the process.[5] This policy should include security regulations to be followed at the school level, including information about how and where tests will be stored, how they will be distributed and collected, and which practices are and are not permissible. Those who will administer the assessment should be familiar with these policies in a standard way throughout the area in which the assessment is given.

3. Limiting Access to the Assessment

Appropriate security precautions should be taken before, during, and after assessments are administered. One security precaution that will minimize inappropriate testing practices—especially those involving teaching actual test items, photocopying test questions, and tampering with answer documents—is to limit access to the test outside of the time that actual testing takes place. Except for those portions of teachers' manuals that do not contain actual test items, testing materials should be delivered to schools shortly before testing begins and kept secure until they are needed. In order to ensure that the materials have not been tampered with prior to the administration of the test, testing materials can be kept in sealed boxes or shrink-wrapped, and testing booklets can be closed with gummed labels. At each testing location, an individual should be given the responsibility to ensure that security precautions are strictly followed and to report any breach of policy.

For most tests, materials should be distributed immediately before the test is administered and collected and returned to locked storage as soon as the testing is completed.[6] Records should be maintained of the number of test booklets and answer documents distributed to each individual administering a test and accounted for on return. Testing materials should be collected both from those administering the tests and from the testing sites as soon as practicable after the testing is finished. If practicable, assessments should be administered by personnel who have little or no stake in their outcome.

Finally, conditions that permit or foster the ability of individual students to achieve inappropriate scores by fraudulent means should be minimized. These efforts can include, where appropriate or practicable, simultaneous administration to all who are taking the same form of test, identification procedures, seating charts and assignments, space between seats, and continuous monitoring. Of course, many of these precautions are expensive, but the expense must be balanced against the potential negative consequences of "test pollution."

4. Monitoring the Administration and Scoring of an Assessment

Those who direct testing programs should provide for monitoring of the administration of the test. This supervision can include unannounced observations and interviews of those administering the assessment. It can also include secondary information obtained from other sources, such as teachers who have observed what they believe are inappropriate practices, students who complain that other students were "cheating," or information that parents were given by their children after the test. In addition to identifying security breaches, monitoring can help to identify the strengths and weaknesses of particular testing procedures, which can then be addressed before the next round of testing.

When answer sheets are scored, the procedures used should be documented to ensure the accuracy of scoring. Those engaged in scoring should monitor the frequency of error and report it upon request. In addition, auditing procedures should be developed to review answers and results. Score processing should be audited to make sure that the data are processed correctly and the materials are securely maintained. Auditing also can include computer studies of the test results to determine whether unusual patterns of responses exist that may indicate possible cheating or other unethical practices. These studies could include erasure counts by class or school, analysis of patterns of responses from students seated in close proximity, or analysis of unusual gains as compared to predicted scores or the previous year's performance.

Some reports indicate that students with language or other obstacles to performance have been excluded on the day of an assessment—sometimes even being sent on field trips or home—in order to elevate the overall scores of a school or district. Since this practice is obviously inappropriate, excluding students from an assessment should require outlining a clearly articulated and appropriate reason based on the best interests of the individual students. In addition, any exclusions and their grounds should be disclosed to those who interpret and use the results of the assessment so that these exclusions can be accounted for, if necessary.

5. Dealing with Breaches

Those in charge of an assessment should be notified of any conditions during its administration that may limit the effectiveness or validity of the test, whether or not the conditions could be controlled by those who supervised the administration. When apparent breaches in appropriate practice are found, a more formal investigation should be initiated to determine the nature of the activity, its impact on the overall assessment, possible solutions, and further administrative or legal action that may be appropriate. These investigations must be undertaken with care so that the rights of all individuals involved are protected. Statutory or administrative rules, as well as professionally trained personnel at the state level, should give testing officials the authority and means to conduct appropriate investigations and act upon their results. On the other hand, breaches of security involving students are best

handled at the local level, in the same manner as any ordinary academic or disciplinary action.

6. Testing Conditions and Environment

It is important to follow the conditions or procedures prescribed by the developer of an assessment. These measures are often closely tied to the legitimacy of the claims made about what and how the assessment measures. Without the uniformity that these measures provide, the validity of comparisons made on the basis of the assessment may be questionable. Therefore, specific directions regarding instructions to be given to test takers, time limits, the form of presentation or response, and the testing materials or equipment to use must be strictly observed, with exceptions based only on carefully considered professional judgment. A reasonable opportunity should be provided before a test is administered for those involved to clarify their understanding of the directions.

Differences in environmental conditions can undermine the validity of the test's results to the extent that those who are tested are unevenly affected by these variations. The testing environment should be reasonably comfortable and have minimal distractions. Examples of the types of conditions to avoid are noise or other disruptions in the testing area, extremes of temperature, inadequate working space, and illegible instructions or test questions.

7. Appropriate Accommodations for Students with Special Needs

The test also should be administered in such a way that sources of potential bias are eliminated. For example, all reasonable accommodations should be made to ensure that the scores of disabled students or individuals with limited English proficiency are not prejudiced by the way in which the test is administered. A test should assess a student's achievement, not the disability or its effect upon demonstrating that achievement. For example, students with visual impairments should have instructions and questions read to them, or large-print or Braille copies of the test provided, if necessary; students with a hearing disability may need written instructions; and students whose primary language is not English should not be expected to read English on any test that is not meant to measure their ability to read English, which will often include tests to assess mathematics, science, or social studies concepts.

8. "Correcting" Answer Sheets

Because students sometimes make mistakes when they fill out answer sheets, teachers or test administrators sometimes "correct" them. This practice should be strictly limited, since it presents a high potential for abuse. For example, when a student mistakenly fills in two "bubbles" on an answer sheet, a decision to erase one of them may reflect the administrator's or the teacher's answer, not the student's, and thus undermine the test's validity. Most states allow changes only to demographic and identification information. Nevertheless, checking answer sheets to make sure that they are properly filled out (e.g., students have filled in

bubbles rather than drawn an "X" through them) may be acceptable, but only if clearly allowed by the assessment's developers. It is best to discourage any changes to students' answer sheets, either directly or indirectly, unless called for in the developer's instructions.

E. Interpretation and Use of Test Results[7]

Much of the criticism of testing programs is aimed at the way in which tests are used and the types of inferences that are drawn from their results. There is widespread agreement among educators, the assessment community, and test publishers that tests are often used for purposes for which they were neither designed nor validated; in addition, their results are often misinterpreted. When an assessment's results are interpreted and used inappropriately, the validity of the entire exercise is undermined, or sometimes destroyed, even if the assessment was selected, prepared for, and administered appropriately.

1. Promote Valid Inferences

Those who interpret and use the results of assessments should promote valid inferences that are likely to produce positive outcomes and minimize negative outcomes for the individuals and programs involved. The most universal guideline governing the interpretation and use of test results is that these activities should not be conducted in a vacuum. Scores must be considered within the full context of the educational environment surrounding the students. A test score can only attempt to describe a level of performance achieved by a particular person at a particular time. The score alone reveals nothing about the causes of the student's performance. In addition, every test score contains a certain amount of error in measurement—the test score should not be interpreted as a fixed and unchangeable index of a student's performance. A particular test score must not be seen to reflect a lack of ability without considering the many other examples of student performance, such as class assignments, other tests, or additional factors. It is therefore vital that no decisions that may have important effects on the lives of individuals or institutions are based solely on the results of a single assessment.[8] However, using an assessment as the end point in decision-making is considered appropriate (Mehrens, 1993) as long as multiple indicators are used leading up to the summative examination, and as long as the examination accurately assesses those skills and competencies considered most important to master. However, concerns abound when the test result is in conflict with other indicators (e.g., a low test score from an average or above average student), and it is the relative weight given the assessment that causes concern.

Before the results of an assessment are interpreted or used, the test should be evaluated to ensure that any inferences drawn from it are valid and reliable for the specific uses intended. If substantial changes in format, content, instructions, language, or administration of the test are made, the intended uses of the test should be reevaluated and validated in view of those changed conditions. If this step is not taken, a coherent and documented rationale must be given to explain why such a reappraisal is unwarranted.

2. Avoiding Improper Uses and Interpretations

Those who interpret and use the results of testing should avoid using the results for purposes that have not been specifically recommended by the test developer, unless they have first evaluated and obtained research evidence to support the intended use of the results. Of course, test results should not be used as a basis for claims that cannot be substantiated or to support false or misleading statements about those assessed or the institutions involved. Nor should they be used to justify a decision made primarily on other grounds, such as political and other pressures, funding considerations, or other noneducational factors.

All of those who are involved with or affected by the assessment should join in efforts to avoid or discourage such inappropriate interpretations and uses of testing results and report instances of misuse and misinterpretation.

3. Identify and Educate the Audiences for Testing Results

It is necessary to identify all of the potential audiences that may receive the results of an assessment, as well as their likely level of background knowledge of testing theory and practice, so that the results can be reported clearly and effectively. By evaluating the audiences to which test results will be communicated, those responsible for an assessment program can limit the possibility of misinterpretation and misuse of results by those who merely misunderstand what the numbers mean and how they can be compared.

Those involved with interpreting and using test results should be provided with information about the assessment and its intended purposes and uses, so that they can properly understand the meaning of the results. They should be given information on how the assessment results were derived, including how scores and other summaries were developed and may be interpreted. The potential direct and indirect consequences that test results may have on individuals or programs also must be understood and evaluated.

When reporting test results, test developers should provide simple score reports that describe test performance clearly and accurately, in order to deal with problems in the use and interpretation of test results. They should explain the meaning and limitations of the scores reported. The populations representing any norms used should be described, as well as the process used to select the samples of test-takers and the dates on which the data were gathered. Test developers also should warn test users of reasonably anticipated misuses of the scores. Finally, test users should be given information that will help them outline reasonable procedures to be used for setting passing scores (if necessary).

When reporting test results to students, parents, legal representatives, teachers, and the media, those responsible for testing programs should provide their audiences with clear descriptions of what the test measures, what the scores mean, common misinterpretations of the scores, and how the scores will be used. Misinterpretations, invalid comparisons, and other misuses of testing results should not be left unchallenged.

It is also important to tell students and their representatives where the test results will be kept on file and to advise them of any individual rights that they may have concerning access to the information or contesting scores and how those rights may be exercised. All should strive to protect the rights of privacy of both individuals and institutions.

4. Understand the Limitations of the Assessment

Those who interpret and use the results of an assessment should understand and, if appropriate, communicate the results within the context of the test's limitations. They should disclose the shortcomings of the particular testing instrument, including the shortcomings that are related to the type of assessment used and its quality, the content assessed, the effects that the characteristics of those examined could have on the test's validity, and any other factor that might influence the proper interpretation of the results. They also should evaluate and, if appropriate, communicate the adequacy of the norms or standards used in interpreting the results, providing information on the scale used for reporting scores, the characteristics of any norms or comparison group(s), and the other limitations of the scores. Scores should be interpreted and used only after evaluating the differences between the norms or comparison groups and the actual population assessed. The date(s) of norms used should be identified and taken into account. The effects of differences in test preparation and administration practices or students' familiarity with the specific questions on the test should be accounted for. These steps are especially important when scores between schools, districts, etc., will be compared.

5. Address Potential Bias

Finally, it is very important not to discount the possibility that the test scores were affected by bias—cultural or otherwise—in the content or format of the testing instrument. The implications of these influences should be included in the technical report of an assessment's results. It is also important that those who interpret and use test results recognize the sometimes hidden influence that bias may have on the educational opportunities of those individuals who may have been affected by prejudice.

IV. Conclusion

Due to the loud cries of a public that demands results from our schools, those who guide American education have increasingly turned to testing for answers and direction. Because of the pressures inherent in the political and social climate of education and the complexity involved in assessment, inappropriate or unethical choices have too often been made when tests have been selected, prepared for, and administered, and when their results have been interpreted and used. Such choices undermine the foundation for conclusions drawn from an assessment's results and the decisions made about students or institutions based on those inferences. While educators and policymakers attempt to outline and agree upon the best means of assessing students, the widespread use of over-interpreted and "polluted" tests will remain. One cannot discount the possibility that the assessment strategies proposed for

replacing our reliance on standardized tests, including performance-based assessment, portfolios, and the like, may be subject to many of the same influences and abuses.

It should be noted in closing that the questions surrounding appropriate and inappropriate testing practices remain open to discussion among *all* who are involved in assessment. The potential for inappropriate or unethical testing practices may be greatly reduced if those developing assessment policy seek the involvement of all of those who have roles in testing. A continuing conversation will at least foster greater understanding of where testing practices cross the line at all levels—from assessment professionals to policymakers to the classroom teacher.

Sources Cited or Used

Cannell, J.J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average.* Daniels, WV: Friends for Education.

Cannell, J.J. (1989). *The "Lake Wobegon" report: How public educators cheat on standardized achievement tests.* Albuquerque, NM: Friends for Education.

Cannell, J.J. (1990). *Testing ethics model legislation.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

National Council on Measurement in Education Ad Hoc Committee on the Development of a Code of Ethics. (1993, February). *Code of ethical assessment practices in education.* (Draft 1).

Cuban, L. (1991, September). *The misuse of tests in education.* Paper prepared for the Office of Technology Assessment.

Haladyna, T.M., Nolen, S.B., & Haas, N.S. (1991, June-July). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher,* 2-7.

Hall, J.L., & Kleine, P.F. (1990). *Educators' perceptions of achievement test use and abuse: A national survey.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education.* Washington, DC: Author.

Koretz, D. (1988). Arriving in Lake Woebegone: Are standardized achievement tests exaggerating advancement and distorting instruction? *American Educator, 12,* 2.

Kramer, J.J., & Conoley, J.C. (1992). *The eleventh mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska, Lincoln.

Linn, R.L. (1991). *Test misuse: Why is it so prevalent?* Boulder, CO: Center for Research on Evaluation, Standards, and Student Testing.

Linn, R., Graue, M., & Sanders, N. (1989, March). *Comparing state and district test results to national norm: Interpretation of scoring "above the national average."* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Madaus, G.F. (1985). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan*, 66, 611-17.

Madaus, G.F. (1990). The distortion of teaching and testing: High-stakes and instruction. *Peabody Journal of Education, 65*, 29-46.

Mehrens, W.A. (1991, April). *Defensible/indefensible instructional preparation for high stakes achievement tests: An exploratory trialogue.* Revision of presentation given in a symposium of the same title at AERA/NCME Annual Meeting, Chicago, IL.

Mehrens, W.A. (1984). National tests and local curriculum: Match or mismatch. *Educational Measurement: Issues & Practice*, 3(3), 9-15.

Mehrens, W.A. (1993). Review of Greg Bell's *The test of testing: Making appropriate and ethical choices about assessment*. Personal communication to Linda Bond and David Braskamp, North Central Regional Educational Laboratory, October 25, 1993.

Mehrens, W.A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice, 8*, 14-22.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*, 215-37.

NCME Task Force. (1991, April). *Regaining trust: Enhancing the credibility of school testing programs.* Report by a task force of the National Council on Measurement in Education.

Nolen, S.B., Haladyna, R.M., & Haas, N.S. (1990). *A survey of actual and perceived uses, test preparation activities, and effects of standardized achievement tests.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Office of Technology Assessment. *Testing in American schools: Asking the right questions.*

Phillips, G.W., & Finn, C.E. (1988). The lake woebegone effect: A skeleton in the testing closet? *Educational Measurement: Issues and Practice, 7*, 2.

Popham, W.J. (1991, Winter). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice*, 12-15.

Shepard, L. (1989, March). *Inflated test score gains: Is it old norms or teaching to the test?* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Smith, M.L. (1991). Meanings of test preparation. *American Educational Research Journal, 3*, 521-42.

Suarez, T.M., & Gottovi, N.C. (1992). The impact of high stakes assessment on our schools. *NASSP Bulletin, 76*, 82-88.

Endnotes

1. Cannell's study reported that 48 of the 50 states and 90 percent of the nation's 15,000 school districts asserted that they were testing "above the national norm" on commercial elementary achievement tests. He also found that outright cheating was common on norm-referenced and criterion-referenced tests. Cannell's second report charged that educators in many states were blatantly cheating on standardized tests. The effects outlined in Cannell's original "Lake Wobegon" findings have been explained in several different ways, including the possible effects of dated norms, content familiarity, and teaching to the test (Phillips & Finn, 1988; Linn, Graue & Sanders, 1989; Shepard, 1989; Koretz, 1988).

2. Derived from a draft of the *Code of Ethical Assessment Practices in Education*, which was developed by the National Council on Measurement in Education's Ad Hoc Committee on the Development of a Code of Ethics.

3. The suggestions in this section were derived from the *Code of Fair Testing Practices*, the *Code of Ethical Assessment in Education*, and the report of a National Council on Measurement in Education task force entitled *Regaining Trust: Enhancing The Credibility of School Testing Programs*.

4. The suggestions in this section were derived from the *Code of Fair Testing Practices*, the *Code of Ethical Assessment in Education*, the *Standards of Educational and Psychological Testing*, and the NCME task force report cited above. The *Standards* are the most detailed set of guidelines for testing practices. However, they are very technical and specific. The other *Codes* cited have to a great extent been based on their concepts and requirements.

5. For an example of one such policy, see the *Testing Code of Ethics for North Carolina Testing Personnel, Teachers and School Administrators*.

6. On tests that need to be read to students (excluding reading tests and those in which students assessed have language or reading difficulties) a balance should be maintained between the necessity to rehearse reading the test items (especially when young students are involved) and limiting access to the test.

7. The suggestions in this section were derived from the *Code of Fair Testing Practices*, the *Code of Ethical Assessment in Education*, and the *Standards of Educational and Psychological Testing*.

8. However, it should be noted that in some situations it may *appear* that a single assessment is the sole criterion, due to its being placed as a "gateway" at the end of a sequence of consideration of and decisions about other relevant criteria. This use is entirely appropriate.

North Central Regional Educational Laboratory
1900 Spring Road, Suite 300
Oak Brook, IL 60521-1480
(708) 571-4700
Fax (708) 571-4716