DOCUMENT RESUME

ED 349 826 FL 020 645

AUTHOR Haladyna, Thomas
TITLE Test Score Pollution: Implications for Limited English Proficient Students.
PUB DATE Aug 92
NOTE 49p.; In: Focus on Evaluation and Measurement. Volumes 1 and 2. Proceedings of the National Research Symposium on Limited English Proficient Student Issues (2nd, Washington, DC, September 4-6, 1991); see FL 020 630.
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Tests; Elementary Secondary Education; *Limited English Speaking; *Scores; *Standardized Tests; Test Format; Test Validity

ABSTRACT

The topic of this paper is the second of a two-faceted problem involving achievement testing in the United States. The first facet is the lack of correspondence between test content and intended student outcomes in school districts, and the second facet is "test score pollution." Test score pollution describes instances where test scores for a unit of analysis, such as a class or school, are systematically inflated or deflated without corresponding changes in the content domain that a test is supposed to represent. Test score pollution is associated with standardized achievement tests; however, authentic assessments may be even more susceptible to test score pollution. First, the concept of validity is examined, and second, particular attention is focused on the meaning of school achievement. Third, test score pollution is described and research on the problem is evaluated. To conclude, the effects of test score pollution on limited-English-proficient students are discussed. Responses to the paper by Gary Hargett and Maria Pennock Roma are appended. (VWL)

Test Score Pollution: Implications for Limited English Proficient Students

Thomas Haladyna
Arizona State University, Tempe

Introduction

Standardized tests have a multitude of interpretations and uses. Test score pollution is a condition that affects the validity of these interpretations and uses. This paper presents the problem of test score pollution in the context of achievement testing, speculates about its origins, provides evidence of its complexity and severity, and addresses the implications of test score pollution for limited English proficient students.

Test Score Pollution: Implications for Limited English Proficient Students

Current reform in the organization of schooling has been accompanied by significant reform in testing (Toch, 1991). Standardized achievement tests have been under siege for many years (Hoffman, 1964; Fair Test Examiner, 1987), and "authentic assessment" has recently been proposed as an alternative or replacement for the standardized achievement test. Baker (1991) summarized the prevailing attitude behind this test reform when she stated that the authentic assessment is more holistic and realistic of what real teaching represents, while the standardized testing is more molecular and facts-based.

Part of the testing reform movement can be attributed to persistent criticism that standardized achievement tests fail to measure the important outcomes of schooling or that it only partially measures these outcomes (Berk, 1988; Brandt, 1989; Frederiksen, 1984; Haertel 1986; Haertel and Calfee, 1983; Linn, 1987; Madaus, 1988; Messick, 1987; Shepard, 1989).

The topic of this paper is the second of a two-faceted problem involving achievement testing in the United States. The first facet is the lack of correspondence between test content and intended student outcomes in school districts, and the second facet is "**test score pollution.**" This term describes instances where test scores for a unit of analysis (such as a class or school) are systematically inflated or deflated without corresponding changes in the content domain

that a test is supposed to represent (Haladyna, Nolen, and Haas, 1991). Whether we use a standardized test or an authentic, assessment is probably irrelevant. Because standardized achievement tests have been used for many years, test score pollution is associated with this type of test, but authentic assessments may be even more susceptible to test score pollution (Canner, 1991).

First, we examine the concept of validity. Second, we look carefully at the meaning of school achievement. Third, we define test score pollution and then evaluate the research bearing on this problem, and finally we speculate about the effects of test score pollution on limited English proficient (LEP) students.

Construct Validity

Traditionally the topic of validity has been treated in three categories (construct, criterion-related, and content), but recently Messick (1989) has presented a unified approach to validity under the rubric "construct validity." In this conceptualization, validity refers to interpretations as well as uses of test results.

For instance, Haladyna, et al. (1991) presented 29 different uses of standardized achievement test scores. Table 1 summarizes these interpretations and uses. Dorr-Bremme and Herman (1986) offer findings from their national survey illustrating the variety of uses of test results.

Table 1
Consumers and Uses of Standardized Achievement Test Information

Consumer: National Level	*Units of Analysis*
Allocation of Resources to Programs and Priorities	Nations, States
Federal Program Evaluation (e.g., Chapter 1)	States, Programs
Consumer: State Legislature/State Department of Education	
Evaluate State's Status and Progress Relevant to Standards	State
State Program Evaluation	State, Program
Allocation of Resources	Districts, Schools

Consumer: Public (Lay persons, Press, School Board Members, Parents)

Evaluate State's Status and Progress Relevant to Standards	Districts
Diagnose Achievement Deficits	Individual, Schools
Develop Expectations for Future Success in School	Individuals

Consumer: School Districts--Central Administrators

Evaluate Districts	Districts
Evaluate Schools	Schools
Evaluate Teachers	Classrooms
Evaluate Curriculum	District
Evaluate Instructional Programs	Programs
Determine Areas for Revision of Curriculum and Instruction	District

Consumer: School Districts--Building Administrators

Evaluate School	School
Evaluate Teacher	Classrooms
Grouping Students for Instruction	Individuals
Placement into Special Programs	Programs

Consumer: School Districts--Teachers

Grouping Students for Instruction	Individuals
Evaluating and Planning the Curriculum	Classroom
Evaluating and Planning Instruction	Classroom
Evaluating Teaching	Classroom
Diagnosing Achievement Deficits	Classroom, Individuals
Promotion and Graduation	Individuals
Placement into Special Programs (e.g., Gifted, Handicapped)	Individuals

Consumer: Educational Laboratories, Centers, Universities

Policy Analysis	All units
Evaluation Studies	All units
Other Applied Research	All units
Basic Research	All units

While many observers do not support these interpretations and uses, little doubt should exist that researchers, evaluators, policy analysts, and lay persons (including legislators and the press) are interested in interpreting and using test results in these ways.

The Standards for Educational and Psychological Testing (American Psychological Association, 1985) are very explicit about the need to validate any interpretation or use. Standard 1.1 on page 13 states:

> *"Evidence of validity should be presented for the major types of inferences for which the use of a test is recommended. A rationale should be provided to support the particular mix of evidence presented for intended uses."*

In a national survey by Hall and Kleine (1990), 90 percent of the respondents reported that tests are used to evaluate teacher effectiveness. Berk (1989) and Haertel (1986) have offered strong criticism against such use. Another example is the use of state-by-state comparisons to draw inferences about a state's success at educating its students, a practice that has received much criticism (Guskey, & Kifer, 1990; Koretz, 1991).

A storm of protest about the misinterpretation and misuse of test scores has existed for years within the community of testing specialists education (e.g., Brandt, 1987; Frederiksen, 1984; Haertel, 1986; Haertel and Calfee, 1983; Linn, 1987; Madaus, 1988; Messick, 1987; Shepard, 1989). As test users, we must be vigilant about misinterpretation and misuse of test results for purposes of evaluation and policy making affecting our jurisdictions.

Construct validation calls for the collecting of evidence to support any of the 29 different uses or interpretations of test results that we desire. Messick (1989) provides a very comprehensive discussion of construct validation and the logical and empirical types of evidence necessary to validate test interpretations and uses. Without such evidence, we should question the ethics of those within the profession of education making unsupported claims based upon test results. Seldom do we see evidence presented to support any of the interpretations and uses found in Table 1. Consequently, we should resist attempts to interpret or use test results in ways unintended and unsupported by validating evidence.

School Achievement

School achievement is the main construct of education. Hypothetically, we can define school achievement in terms of many subject matter areas, using instructional objectives, and organize these

objectives by content and by a level of cognitive behavior, such as found in the Bloom taxonomy. An explicit, national curriculum does not exist, but the belief that the standardized achievement test reflects this general national curriculum has been expressed at various times by various writers (e.g., Freeman, Belli, Porter, Floden, Schmidt, & Schwille, 1983; Leinhardt & Seewald, 1981; Phillips & Mehrens, 1987). In general mixed evidence exists on this issue of whether the test represents a national curriculum, but staunch advocates of systematic instruction argue that no standardized achievement test is likely to be interchangeable and represents specific classrooms, curricula, and instruction (Cohen, 1987; Nitko, 1989).

The Arizona Department of Education learned recently that only about 27 percent of its essential skills could be found on a standardized achievement test (Noggle, 1988). The Department of Education changed its testing program to provide a closer alignment to its state-mandated essential skills curriculum. Other states, like Missouri, have already accomplished this. School achievement is going to have to be redefined by a jurisdiction, and carefully measured, if reform in testing is to be effective.

Several researchers have questioned the kinds of inferences we can draw from standardized achievement test data (Nolet and Tindal, 1990; Wardrop, Anderson, Hively, Hastings, Anderson, and Muller, 1982). They claim that only general interpretations can be made about standardized achievement test results. Test companies have never claimed that their tests measure school curricula, instructional practices in school districts, schools, or classrooms (Mehrens & Kaminski, 1989). Koretz (1989, p. 33) stated it succinctly:

> *"Put simply, an achievement test is typically a brief and incomplete proxy for a more comprehensive, but less practical, assessment of some domain of achievement."*

Teachers generally believe that standardized test results do not reflect their teaching and they tend to rely on their own observations (Dorr-Bremme & Herman, 1986; Haas, et al., 1989).

Causal Attribution

Part of the problem of achievement is the strong desire to know what or who has caused students to achieve or not achieve. Accountability requires that we make causal statements about achievement. School achievement is the result of many influences existing over a child's lifetime and even prior to a child's birth. Some of these factors, such as family and home influences, parental education, socioeconomic status, family mobility, and neighborhood exist outside the

influence of schooling. Other factors, such as learning environment, motivation and attitude, and quality and quantity of instruction, are under the influence of school personnel. While we have trouble measuring school achievement, we have even more trouble with causal attribution. We have not yet completely understood the influence and interactions of these variables on school learning, although models like Walberg's productivity model (Walberg, 1980) provide a workable framework for our understanding of causes of learning. Lay persons tend to oversimplify education by using test results as the operational definition of achievement and the teacher as the singular cause of school learning.

Higher Level Thinking

A common distinction among all educators is that student learning comes in various forms of mental complexity, ranging from recall to various types of higher level thinking, often expressed in the Bloom taxonomy. Many critics and researchers alike have concluded that curricula, teaching, and testing have focused on lower level thinking, such as recall, at the expense of hard-to-measure higher level thinking outcomes. Nickerson (1989) leaves little doubt that American education will focus on making its students thinkers, and therefore higher level thinking will become a strong feature of new standardized achievement tests.

A dilemma presents itself (Haas, Haladyna, and Nolen, 1990; Nolen, Haladyna, and Haas, in press; Smith, 1991): Teachers are forced to give standardized tests, which they believe measure lower level thinking. Some teachers promote higher level thinking in their classrooms at the expense of preparing students for the standardized tests, while other teachers faithfully drill students on the kinds of outcomes known to be tested. Who is the more effective teacher? This dilemma is part of the problem of test score pollution.

The problem of testing higher level thinking is further complicated by recent reports that teachers are either reluctant or unable to develop classroom tests to measure higher level thinking (e.g., Stiggins, Griswold, & Wikelund, 1989), while standardized tests are equally at fault for failing to measure higher level thinking. Nonetheless, the new thrust in performance testing (euphemistically referred to as "authentic assessment") promises to give greater emphasis to the measurement of higher level thinking through the development of multi-step exercises.

Multiple-Choice versus Performance

A current opinion held in education is that performance tests measure higher level thinking outcomes while multiple-choice tests measure recall, and other trivial forms of behavior (Baker, 1991).

Recent and past reviews of research on the equivalence of open-ended versus selected-response formats reveals their equivalence (Bennett, Rock, and Wang (1990). Further these researchers submit that the stereotype that multiple-choice tests measure trivial content and factual recall while open-ended tests measure higher level thinking is FALSE.

Measurement specialists have consistently maintained that multiple-choice items can be used to measure higher level thinking outcomes, admitting that it is difficult to do via any format. For instance, the context-dependent item set that contains a stimulus and a set of test questions can be used to measure various types of higher level thinking outcomes via a multiple-choice format (Haladyna, 1991, in press a, in press b).

Conclusion

School achievement is a complex constellation of knowledge and skill that is difficult if not impossible to measure with a single test. Therefore, no current test seems to be adequate toward the end of measuring the complete domain represented by a school district's curriculum. Further, we lack many technologies in item writing and scoring to measure adequately many aspects of human behavior.

The variety of purposes listed in Table 1 are not served by using a standardized achievement test. That is why many observers call for significant reform in testing where multiple indicators are used and where achievement is better defined in terms of its many aspects.

Test Score Pollution

Test score pollution is any influence that affects the accuracy of achievement test scores. Messick (1984) called these influences "contaminants" but did not specify exactly what these contaminants are. Haladyna, Nolen, and Haas (1990) identified three sources of contamination and reviewed the research bearing the seriousness of each. These are: (1) test preparation, (2) situational factors, and (3) external conditions. Table 2 provides a list of 21 specific sources of test score pollution organized by these three categories, adopted from Haladyna *et al.* (1991).

Table 2
21 Documented Sources of Test Score Pollution

Test Preparation Activities

Testwiseness Training
Increasing Motivation
Curriculum Matching
Changes in the Instructional Program
Specific Inappropriate Instruction (*Scoring High*)
Presenting Items Similar to Those Found on the Test
Presenting Items Identical to Those Found on the Test
Excusing Low-achieving Students From Taking the Test
Cheating

Situational Factors

Test Anxiety
Stress
Fatigue
Speededness of the Test
Motivation
Recopying and Checking Answer Sheets
Test Administration Practices

Context

Language Deficits
Socioeconomic Context
Family Mobility
Family and Home Influences
Prenatal/Early Infant Influences

Origins of Test Score Pollution

Undoubtedly, the range of uses of standardized test scores has changed drastically from the 1950s to the 1990s (Haertel and Calfee, 1983). The current overuse and misuse of test results, coupled with the "high stakes!" nature of many uses has badgered superintendent, principals, and teachers to prepare students to perform on these tests. According to Haas et al. (1990), although the preparation forces teachers to depart from regular instructional practices and teachers almost uniformly dislike the test and disagree with the public's misuse of test results, the pressure to produce high test scores is unbearable. One teacher commented:

...I feel that if I am pressured any more to do well on the TEST, I will do everything I can to make sure my kids do well...even cheat. I have a family to support and I would be stupid not to do this. My job is more important than my values. (Haas, et al., 1990, p. 128).

Test Preparation

A variety of school activities falls into the category of test preparation. Haladyna et al., (1990), Mehrens and Kaminski, 1989) and Smith (1991) present a continuum of test preparation activities. The following is Smith's conceptualization.

The first is **no special preparation**. Nolen et al., (in press) reported that 12 percent of teachers surveyed did no special preparation. The fact that 88 percent did introduces a form of pollution.

The second is to **teach test-taking skills**. Nolen et al., (in press) reported that over 60 percent of teachers surveyed did this. Test taking skills (or "testwiseness" as it is sometimes referred to) is well defined in the extant literature, and Bangert-Drowns, Kulik, and Kulik (1983) and (Sarnacki (1979) reported that indeed testwiseness training does work. Comparisons between those teaching test-taking skills and those not teaching test-taking skills introduce test score pollution.

A third method is **exhortation**. This includes advice on eating and sleeping before the test, pep rallies, the principal's announcements and words of encouragement, and other measures designed to "motivate" students to do their best on the "test."

A fourth method is the **design of instruction to match the test content**. Some materials, such as *Scoring High in Math* (Foreman & Kaplan, 1986), appear designed to identify the exact content of a standardized test and to provide specific instruction on this material (Mehrens & Kaminski, 1989). Toch (1991) presents a more comprehensive description of the extent of the industry for producing materials to prepare for standardized achievement tests. Haas *et al.* (1990), Nolen et al., (in press) and Smith, Edelsky, Draper, Rottenberg, and Cherland (1989) report extensive use of these materials in elementary school classrooms as well as disenchantment with this practice. A national survey conducted by Hall and Kleine (1990) revealed that 69 percent of the sample reported changes in the curriculum to match the standardized achievement test, 39 percent reported changes in the curriculum to match particular questions on these tests, and 82 percent reported teaching material because it is on the test. Several critics of these practices have stated that the curriculum, in effect, is narrowed, that time for instruction on non-test related and other important content is lost, that instruction is

very test like, and that both teachers and students suffer in many ways (Smith & Rottenberg, in press). Popham (1990), among others, criticized the ethics of this narrowing of curriculum and instruction.

A fifth method is "**stress inoculation.**" Teachers report helping students boost test scores for the purpose of increasing the students' collective self-respect. Since the improvement or maintenance of self-respect is so important, the achievement of high test scores is viewed as a vehicle for this worthy goal.

A sixth method is **practicing on items of the test itself or a parallel form**. Both Nolen, et al., (in press) and Mehrens and Kaminski (1989) stated that about 10 percent of teachers reported doing this. While these researchers believe that this is blatantly dishonest, some teachers believe that since the tests are so inherently misused and misinterpreted, this practice is done to "play the game" with administration and the school board.

A seventh method, **cheating,** refers to giving answers to students, providing hints to students, and changing answer sheets after the test.

Table 3 provides a list of test preparation activities from Haladyna, et al., (1991), and their judgments regarding how ethical these test preparation practices are. Mehrens and Kaminski (1989) offer a similar set of judgments, and Cannell (1988) also provides his appraisal of the ethics of various test preparation practices. Haladyna et al., (1990) also make the point that despite whether a test preparation activity is ethical or not, all test preparation activities are polluting if one class, school, or school district does it while others do not.

Table 3
A Continuum of Test Preparation Activities

Test Preparation Activity:	Ethical Degree
Training in testwiseness skills	Ethical
Checking answer sheets to make sure that each has been properly completed.	Ethical[1]
Increasing student motivation to perform on the test through appeals to parents, students, and teachers.	Ethical
Developing a curriculum based on the content of the test.	Unethical
Preparing objectives based on items on the test and teaching accordingly.	Unethical
Presenting items similar to those on the test.	Unethical
Using *Scoring High* or other score-boosting activities.	Unethical
Dismissing low-achieving students on testing day to artificially boost test scores.	Highly Unethical
Presenting items verbatim from the test to be given.	Highly Unethical

[1]Ethical to the extent that the test publisher recommends it or to the extent that all schools, classes, and students being compared have the same service.

Another aspect of undesirable test preparation is that by raising test scores, there is no correlated gain in the general domain of achievement that each test is supposed to represent. Recently, Koretz (1991) presented some evidence to support this suspicion, and more research results are expected to further support the polluting influence of many forms of test preparation. Linn Graue, and Sanders (1990) concur with Cannell's findings (Cannell, 1988), that achievement scores are higher than ever, but they assert that the problem may indicate (1) teaching too specifically to the test while at the same time the norms are not keeping up with this specific form and (2) questionable forms of test preparation.

Situational Factors

Haladyna, et al., (1990) in their review of research on test score pollution have documented many factors that are specific to the administration of the test and are also very polluting. Some of these may have saliency for LEP students and these will be addressed more fully in another section of this paper.

Test anxiety. Kennedy Hill and his colleagues (Hill, 1979; Hill & Wigfield, 1984; Hill & Sarason, 1966) have extensively studied test anxiety and estimate that over 25 percent of the school age population have some debilitating form of this disorder. Test anxiety is treatable, but it is also exacerbated by stress-producing conditions in the classroom and school. If an explicit or implied threat exists, test anxiety can be increased (Zatz and Chassin, 1985). Mine, and others (1987) noted that some Japanese families actually promote high test anxiety through parental restriction, blame, inconsistency, overprotection, and rejection. They also state that praise has the *same* effect on test anxiety instead of the opposite effect.

Stress. Children experience many stress-provoking situations in life, many of which are related to school or affect school life (Karr and Johnson, 1987). Oddly, little is known about stress in the classroom. Recent reports give some credence to the role of stress in standardized testing situations (e.g., Nolen, et al., in press; Paris, Lawton, Turner, & Roth, 1991).

In the Paris et al., study, they specifically asked children questions about the effects of the testing experience. Three aspects of why stress may be increased under the condition of the standardi. ł testing experience are that (1) students become increasingly skepti cal about the value of test results as they become older, (2) the purposes or uses of the test are not clearly revealed, (3) there is a socia' impact on students based on their test score status.

Fatigue. Reports of fatigue during the testing process, particularly with younger children, have been reported (Dorr-Bremme & Herman, 1986; Haas et al., 1990; Nolen, et al., in press; Smith et al., 1989). In sun belt states, such as Arizona, temperatures during May testing may reach into the 90s or low 100s, a condition that increases this potential source of pollution. Interestingly, there is no research that specifically addresses the problem of test fatigue.

Timed testing. One condition of all standardized tests of this type is the time limit, which must be strictly followed to provide standardized test results. Reports of plodders and sprinters in timed tests reveal a possible source of test score pollution (Wright and Stone, 1979). This factor is particularly significant to LEP learners

and, it will be treated more extensively in another section of this paper. In addition, timed testing seems particularly harmful to test anxious children (Plass and Hill, 1986). Wodtke, Harper, Schommer, and Brunellia (1990) report liberal violations of time limits in tests administered by teachers. Hall and Kleine (1990) reported that 9 percent of the teachers surveyed in their national study felt pressured to extend time limits and commit other nonstandard testing practices. If the stakes for test results are indeed very high, this should come as no surprise.

"Blowing off the test." Motivation to perform on the test is very important to test performance. Some school districts expend considerable effort in motivating its students, while other districts do not. Haladyna, et al., (1990) identify a host of factors known to increase or decrease performance, all of which are in some way related to motivation. Widespread reports exist that younger students are likely to be more attentive to the test but that older students, seeing the lack of consequence for their test performance, will often resort to random marking (Paris, Turner, & Lawton, 1990). Dorr-Bremme (1986) also reported anecdotal evidence from interviews suggesting that many students do not give much effort to performing well on these tests.

Teacher attitudes may have something to do with test performance. When teachers are highly motivated to get high test scores, student performance may be maximal. With poorly motivated teachers, students merely go through the motions, knowing that the results mean nothing to the teacher. While this hypothesis about teacher attitude is very speculative, anecdotal reports in Haas, et al., (1990) reveal widespread discontent with the standardized test and with the motivation of students to perform on these tests. Smith (1991) also discusses the discouraging climate that standardized testing creates for teachers and the dilution of their professionalism.

Recopying, checking, and repairing mismarked answer sheets. Some school districts have policies that allow the checking of answer sheets for stray marks and light marks, or mismarked answers. Parents, other volunteers, or paid classroom aides are asked to check answer sheets in some schools. The fact that some schools or districts have policies and procedures for this practice while others do not creates another possible source of pollution.

Summary. This section has provided a brief overview of possible test score polluting practices that reside in the test administration or events preceding test administration that do not include test preparation. While many of these practices exist in schools, we know very little about the importance of each as a test score pollutant. Still, indications from this limited research suggest that our concern is warranted and further study is needed.

External Factors

Anyone close to the educational process knows the many factors that underlie poor test performance: inadequate prenatal care, low mental ability, poor early childhood nutrition, lack of social capital in the family and home, disintegrating family social structure, poor motivation, LEP, low socioeconomic status, high family mobility, and lack of education of parents. While this list is brief and hardly all inclusive, it represents factors *outside* the influence of schools and school personnel that are believed to affect school performance. In various evaluation and policy studies at national, state, and school district levels, seldom is reference given to the influence of these variables on test scores. In actuality, schools and school personnel are often given the "blame" or "praise" for test scores that were obviously influenced by these external factors. Therefore, these factors, when unnoticed or not considered, are a source of test score pollution because they affect the accuracy of test score interpretations and uses.

Acting on a state law, Arizona's Department of Education has to report all standardized test scores in the context of two external factors, language proficiency and socioeconomic status (as determined by frequency of use of the school lunch program). Model reporting systems such as this one attempt to reduce the severity of pollution from these external factors.

Implications for Limited English Proficient Children

This section of the paper addresses implications for LEP educators arising from the problem of test score pollution. This section also suggests some fruitful areas for research on the role or influence of test score pollution on LEP students. Finally, recommendations are offered to protect LEP children from negative consequences due to using polluted test scores.

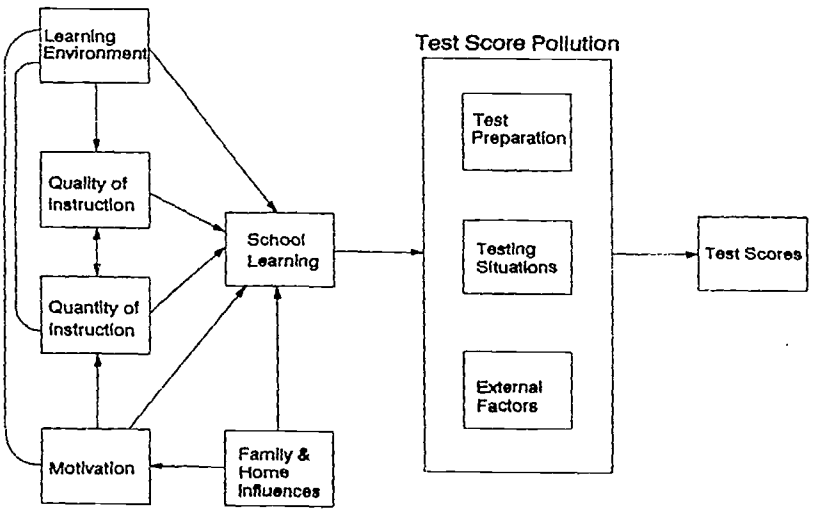
This section of the paper is loosely based on a working model of school learning that includes test score pollution. The following review of research is not very comprehensive but helps build a working hypothesis about why we should be very cautious about test scores obtained from LEP children.

A Causal Model of School Learning Modified to Accommodate Test Score Pollution

To begin this section, a causal model of test performance is offered that is loosely based on the Walberg productivity model

(Walberg, 1980). Figure 1 provides an illustration of the model. The elements are familiar to most educators, and various studies and meta-analyses speak of the potential influence of such constructs as family and home as causal determinants of children's motivation and their learning--as inferred from a non polluted standardized achievement test. Quality and quantity of schooling is also positively and causally related to learning. Learning environment contributes to a high quality of instruction and increases learning time, quantity of instruction, which, in turn, leads to better learning. Learning is demonstrated in many ways in schools, grades being one indicator. The standardized achievement test, at best, provides a gross, general measure of school learning (Nolet and Tindal, 1990; Wardrop, Anderson, Hively, Hastings, Anderson, and Muller, 1982), but as Figure 1 shows, all test performance is mediated by the three possible forms of test score pollution. Therefore, no test score interpretation or use, for any unit of analysis (class, school, district, state, or nation) is valid until we can eliminate the influences of test score pollution.

Figure 1
The Role of Test Score Pollution in Interpreting School Achievement



Facts About LEP Children

As a prelude to the following discussion, several facts about LEP children should be stated. For instance, in a recent publication from the National Center for Education Statistics (Rock, Pollack, & Hafner, 1991), the performances of LEP children as well as other demographics are well documented.

First, and most obvious, LEP children have the handicap of reading, writing, speaking, and listening in a foreign language. Levels of facility in English vary and handicap these children's test performance. Another source of evidence comes from Arizona state testing (Bishop, 1988), which contains information about the test performances of LEP and English proficient children in Arizona. The typical range of LEP children's performance on the state's mandated standardized achievement test ranges between the 14th and 43rd percentiles, while the English primary language students' performance level is near the 62nd percentile. Rock, et al., (1991) report from their national sample of LEP and non LEP students in reading, mathematics, science, and history/citizenship/government that language facility is indeed an important factor in test performance. Effect sizes ranged from .58 for reading to 1.07 for the social studies factor. These are substantial differences.

Second, most LEP children are below average in terms of socio economic status.

Third, the majority of LEP children are from ethnic groups, and each has its distinct culture (Rock et. al., 1991). More than one half of the LEP children in their national sample are Spanish-speaking, and they are more handicapped than those LEP children who speak other languages.

Fourth, LEP education programs offer a "non mainstream" experience designed to help LEP students become mainstream students, but the process of being in LEP programs socially distinguishes these students from mainstream students in social and intellectual ways.

If these assumptions are tenable, the following review of research and discussion bears on test score pollution for LEP children.

Standards

The Standards for Educational and Psychological Testing (American Psychological Association, 1985) are explicitly concerned about LEP students, and it seems worthwhile to review several standards in relation to this problem of test score pollution. Standard 13.1 (page 74) states:

"For non-native English speakers or for speakers of some dialects of English, testing should be designed to minimize threats to test reliability and validity that may arise from language differences."

Studies cited in the next section of this paper give some evidence for potential bias against LEP students. Standard 13.3 (p. 75) states:

When a test is recommended for use with linguistically diverse test takers, test developers and publishers should provide the information necessary for appropriate test use and interpretation.

If the test manual lacks this information, we should submit that the test is probably NOT suitable for LEP persons, since the potential for polluted test scores is too great to risk using the score for any important educational decision. Standard 13.5 (p. 75) states:

"In employment, licensing, and certification testing, the English language proficiency level of the test should not exceed that appropriate to the relevant occupation profession."

This is a serious threat to the validity of professional licensing examinations and tests used to make personnel decisions. Since LEP persons typically have a significant handicap in reading, the existence of unnecessarily difficult reading levels in "high-stakes" tests creates a significant yet subtle form of bias. It would be easy to challenge an examination that has high reading demand on examinees as an example of adverse impact on LEP students.

Test Interpretations and Use

As Table 1 attests, we have witnessed a steady increase in the number and variety of interpretations and uses of achievement test scores. The issue is validity. Some of these interpretations and uses have serious consequences on the extent of education and futures of all children. For instance, test scores are used for placement into special programs (for handicapped or gifted) and for placement in achievement tracks (for example, in courses ranging from beginning to advanced mathematics). Such tests are also used for minimum competency decisions, for example, for high school graduation or promotion.

The first point about test interpretation and use is that it behooves test users to ensure that these scores are unpolluted before using test results. A second point is that the placement of children in programs strictly based on test scores should be questioned. If LEP children's test performances are lower due to test score pollution, then the system that misuses these scores for these various assignments is at fault.

Test Preparation

All children should be experienced test takers. They should have comprehensive test-taking courses and be equally skilled in test-taking. Popham (1990) also submits that practice testing on content related to the test is reasonable if the test formats are varied to encompass a wide range of possible test formats, since focused practice on the actual format of the test may lead to spuriously high results.

Since LEP students typically lack testing experience of this type, they also may lack test-taking skills. Without the experience of test-taking coupled with test-taking skills, they suffer a significant handicap. This inexperience may contribute to other test pollution problems, such as test anxiety. All other forms of test preparation should be viewed as contradictory to effective teaching and fair uses of standardized test results. Any attempt to promote high test performance through other means should be viewed the way the public views the use of steroids for body building, a dangerous and unhealthy shortcut. Moreover, the spurious increase in test performance due to these test preparation activities does not represent significant learning. LEP children have enough handicaps in school and in life without having them suffer through activities designed to produce spuriously inflated test scores that do not represent true learning.

Situational Factors

Test anxiety. The most pervasive and insidious test score depressant is test anxiety. It has been most extensively measured and researched, and though more research is needed, particularly with LEP children, a strong case in the form of a working hypothesis can be built around this prior research and the assumptions we made about LEP children. In a comprehensive review of test anxiety in the schools, Eccles and Wigfield (1989) submit that text anxiety increases over time and negatively affects school performance. Some factors that seem to contribute to test anxiety are:

1. High stakes tests,
2. Severe time limits on tests,
3. Use of letter grades,
4. Transition from elementary to junior high schools,
5. Poor quality of instruction,
6. Unstructured learning environment, and
7. Negative learning histories.

Given our assumptions about LEP students, the seven conditions cited as contributing to test anxiety seem prevalent in this population. LEP students have more negative learning histories. Negative learning history is also associated with low letter grades, another contributor to test anxiety. Their typically low socioeconomic status creates social conditions by which comparisons with mainstream students leads to lower self-image and lower motivation. If instruction is loosely organized, their test anxiety is heightened. If the learning environment does not fit the culture and the work habits of its LEP students, then the learning environment may serve to increase anxiety. The fact that tests are timed and that LEP students are taking tests in a foreign language must increase their test administration time and reduce their test performance. Besides increasing test anxiety, stress is believed to be a potent factor that also affects test performance (Duran, 1983).

One interesting exception to the above line of reasoning and evidence can be found in a review of American Indian children's test performances by Neely and Shaughnessy (1984). They cite research showing that anxiety is actually lower, so low that it may lead to low test performance.

Timed testing. Some research reports the phenomenon of fast and slow test-taking styles. Knapp (1960) submitted that Mexicans are disadvantaged on timed test because their culture does not promote a fast test-taking style, therefore Mexican children may be disadvantaged in timed tests. The argument and research extends to Native American children. However, as Bridgman (1980) points out, there is very little research to report on the test-taking speed of LEP children.

Examiner effect. Part of test performance can be attributed to the learning environment of the classroom. The role of the examiner on Puerto Rican children was studied by Thomas, Hertzig, Dryman, & Fernandez (1971). They found that performance on an IQ test was increased when the examiner was similar to the child in terms of gender, ethnic background, and fluency in Spanish. Such a study raises an issue that the social context for the test may have some bearing on how hard children try on these tests. Having a teacher who is similar to his or her children may have a positive effect on test performance, and, conversely, differences between teachers and students may have opposite effects.

Setting. Seitz, Abelson, Levine, and Zigler (1975) contend that the site for the test has some effect on children's performances. Their study dealt with disadvantaged children instead of LEP children. However, since LEP children are often disadvantaged, these findings may equally apply to both sub-populations.

Context Factors

Language handicaps. The barrier of learning English and at the same time performing on an achievement test written in that language has to be significant in light of assumptions made earlier about LEP students. As pointed out previously in this paper, huge differences exist between the test scores of LEP and monolingual students in Arizona (Bishop, 1988) and with a national sample (Rock et al., 1991). As one teacher explains (Haas, et al., p. 124):

> Iowa Test of Basic Skills testing regulations discriminate against ESL students. As it takes four to seven years for students to truly become proficient in a second language, especially "academic" language, testing them at grade level after one year on the same level as native speakers is inane.

Fortunately, significant research has been done and is further needed on language proficiency (Duran, 1988). The implication is that before students from diverse educational, ethnic, and social backgrounds can perform on published standardized achievement tests in a mainstream environment, they must first qualify by proving to have a satisfactory level of mastery in the English language. Without such proven proficiency, it would be easy to invalidate test results for LEP children.

Cultural influences. Little research has been reported on the influence of culture on test scores. Nonetheless, there is enough logical and some empirical evidence to suggest that culture plays an enormous role on the success of children. For instance, as previously reported in this paper, in the study by Mine, with others (1987), Japanese parents were shown to negatively influence test anxiety through child-rearing patterns. The study by Knapp (1960), while outdated and about IQ testing, suggests that Hispanic students generally have a different approach to standardized testing. The study by Thomas et al., (1971) shows that the ethnic background and language facility of the examiner may have an influence on test results.

Neely and Shaughnessy (1984) reported that over 300 tribes and 250 languages exist within American Indian culture. These researchers conclude that within this population, and probably other populations, the existence of a different culture is a serious deficit with respect to schooling. For instance, native American children are typically noncompetitive, and do not want to be singled out for recognition. These researchers also point out that most American Indian children speak English only in the schools, therefore the language facility is a serious handicap in a testing situation, because most tests deal with American life that is foreign to tribal children. Such disparities between American Indian children and mainstream

children are often cited by teachers as reasons for invalidating standardized achievement test scores (Haas, et al., 1990).

Socioeconomic status. While this fact is obvious to most educators, in evaluation and policy studies, the socioeconomic status of school districts, schools, and children is unnoticed in the reporting of test scores. A considerable relationship exists between family income and test scores (Test Scores and Family Income, 1980). Since LEP children are often of low socioeconomic status, test scores need to be reported in this context so interpretations and uses can be made with the understanding of the handicapping condition presented by low socioeconomic status.

Another factor is *social capital*, a term coined by sociologist James Coleman (1987) that refers to money, other forms of support, and opportunities available to children both inside and outside the home for their growth and development. Coleman believes that social capital is eroding and affecting children's progress in schools. Thus in the interpretation of test scores and the formulation of policy regarding schooling, social capital should be considered as part of the context of the test scores. To fail to consider social capital pollutes test score interpretations and uses.

Summary and Recommendations

1. **Test uses and interpretations should be based on multiple rather than a single indicator.**

   The mindless use of a single score or a set of test scores from a single test is indefensible.

2. **Test results should not be used in ways unintended by its publishers.**

   As indicated in numerous references in this paper, there is gross overreliance, overuse, and misuse of test scores.

3. **Causal interpretations relating to schools and teachers are invalid without considering the full context of causes, and particularly with a test that fails to measure the full scope of school achievement.**

   The need for accountability forces us to make causal attributions about the influences of school on school learning. However, the meaning of any test score, if unpolluted, reflects a lifetime of school and non school learning and a myriad of influences, which partially include, prenatal care, infant stimulation, nutrition, parental support for education, education levels of parents, number of parents in the home, amount of television viewing, degree to

which parents read to children, mental ability of parents, economic status, English language facility, developmental status, mental health, family mobility, social capital, motivation, attitude, academic self-confidence, fatalism (locus of control), self-esteem, learning environments in home and school, and quality and quantity of learning in home and school. Many of these factors reside outside of schools.

4. **Interpretations and uses of standardized test scores are often polluted. Extreme caution should be used in interpreting and using test scores for important decisions.**

   We have gained invaluable understanding in the process of aligning curriculum and instruction with testing. The sensible application of this process will lead to better instruction and better outcomes, but all educators and laypersons must understand that outcomes must come fairly and not through deceptive practices such as exemplified in the litany of test score pollution.

5. **We need more wisdom in the definition and measurement of school achievement and sensible, defensible interpretations and uses.**

   As many observers have pointed out, school achievement is not well defined, and therefore its measurement cannot be entirely successful. Also, the general concept of school achievement is changing toward problem solving and other forms of higher level thinking.

6. **Test scores from LEP students appear to be invalid for many interpretations and uses listed in Table 1.**

   While research is woefully inadequate on this topic, enough information exists to suggest that scores obtained from LEP students are going to be very low and language facility blocks both performance and efforts to learn. We need to make certain that test scores are used in ways we can defend and avoid unwise uses of test scores of LEP children.

7. **We need more research to understand the context and motivational factors influencing test performance of LEP students, particularly those students with test anxiety.**

   Sufficient evidence exists to suggest that other factors interfere with the test performance of LEP students. These factors may substantially include motivation.

This paper has identified a problem with the interpretation and use of test scores. The problem has become so serious that standard-

ized achievement tests are being abandoned in favor of "authentic assessment." Unfortunately, the problem is not with the type of test. The problem appears to stem from unwise uses of test results as well as attempt to improve test results through questionable means. The implications for the education of LEP students are significant, because test score pollution may be exacerbated in this context. The recommendations offered here express the concern that the role of testing in instructional programs needs to be more focused around alignment of curriculum, instruction, and tested outcomes. Also, laypersons will need to be better instructed in this role of testing in instructional programs.

Note

[1] A phrase (p. 145) coined by Popham (1987) to describe test results with severe consequences, such as non promotion, the funding of schools or districts, or the awarding of merit pay to teachers or principles on the basis of high test scores.

References

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

Baker, E. L. (1991). Expectations and evidence for alternative assessment. Authentic Assessment: The rhetoric and the reality. Symposium conducted at the annual meeting of the American Educational Research Association, Chicago, IL.

Bangert-Drowns, R. L., Kulik, R. L., & Kulik, C. C. (1983). Effects of coaching programs on achievement test scores. Review of Educational Research, 53, 571-585.

Bennett, R. E., Rock, D. A., & Wang, M. (1990). Equivalence of free-response and multiple-choice items Journal of Educational Measurement, 28(1), 77-92.

Berk, R. A. (1988). Fifty reasons why student achievement gain does not mean teacher effectiveness. Journal of Personnel Evaluation in Education. 1(4), 345-364.

Bishop, C. D. (1988). Statewide report for Arizona pupil achievement testing. Phoenix, AZ: Arizona Department of Education.

Brandt, R. (1989). On misuse of testing: A conversation with George Madaus, Educational Leadership, 46(7), 26-30.

Bridgman, B. (1980). Generality of a "fast" or "slow" test-taking style across a variety of cognitive tasks. Journal of Educational Measurement, 17(3), 211-217.

Cannell, J. J. (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.

Cannell, J. J. (1989). How public educators cheat on standardized achievement tests. Albuquerque, NM: Friends for Education.

Canner, J. (Chair) (1991). Task Force on Standardized Testing. Washington, DC: National Council on Measurement in Education.

Cohen, S. A. (1987). Instructional alignment: Searching for the magic bullet. Educational Researcher, 16, 16-20.

Coleman, J. S. (1987). Families and schools. Educational Researcher, 16, 32-38.

Dorr-Bremme, D. W. & Herman, J. L. (1986). Assessing student achievement: A profile of classroom practices. CSE Monograph Series in Evaluation, (Number 11). Los Angeles, CA: Center for the Study of Evaluation.

Duran, R. (1983). Hispanics' education and background. New York, NY: College Entrance Examination Board.

Duran, R. (1987). Validity and language skills assessment: Non-English background students. In H. Wainer & H. I. Braun (Eds.) Test validity. Hillsdale, NJ: Erlbaum, pp. 105-127.

Educational Testing Service (1980). Test scores and family income. Princeton, NJ: Author.

FairTest Examiner (1987). 1(1), 1-16.

Foreman, D. I. & Kaplan, J. D. (1986). Scoring high in math. (Subject-centered book B). New York, NY: Random House.

Freeman, D. J., Belli, G. M., Porter, A. C., Floden, R. E., Schmidt, W. H., & Schwille, J. R. (1983). The influence of different styles of textbook use on instructional validity of standardized tests. Journal of Educational Measurement, 20(3), 259-270.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39, 193-202.

Guskey, T. R. & Kifer, E. W. (1990). Ranking school districts on the basis of statewide test results: Is it meaningful or misleading? Educational Measurement: Issues and Practices, 9, 11-16.

Haas, N. S., Haladyna, T. M., & Nolen, S. B. (1990). Standardized testing in Arizona: Interviews and written comments from teachers and administrators. (Technical Report 89-3). Phoenix, AZ: Arizona State University West.

Haertel, E. (1986). The valid use of student performance measures for teacher evaluation. Educational Evaluation and Policy Analysis, 8(1), 45-60.

Haertel, E. & Calfee, R. (1983). School achievement: Thinking about what to test. Journal of Educational Measurement, 20(2), 119-132.

Haladyna, T. M. (in press a). Context dependent item sets. Educational Measurement: Issues and Practices.

Haladyna, T. M. (in press b). The effectiveness of several multiple-choice formats. Applied Measurement in Education.

Haladyna, T. M. (1991). Generic questioning strategies for linking teaching to testing. Educational Technology: Research and Development, 39(1), 73-82.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. Educational Researcher, 20(5), 2-7.

Haladyna, T. M., Haas, N. S., Nolen, S. B. (1990). Test score pollution. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Hall, J. L. & Kleine, P. F. (1990). Educator 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.

Hill, K. & Wigfield, A. (1984). Test anxiety: A major educational problem and what can be done about it. The Elementary School Journal, 85, 105-126.

Hoffman, B. (1964). The tyranny of testing. New York, NY: Collier.

Karr, S. K. & Johnson, P. L. (1987). Measuring children's stress: An evaluation of methods. Paper presented at the annual meeting of the National Association of School Psychologists. (ERIC Document Reproduction Service No. ED 285 072)

Koretz, D. M. (1991). The effects of high stakes testing on achievement: Preliminary findings about generalization across tests. In R. L. Linn (Chair), Effects of High-Stakes Educational Testing on Instruction and Achievement. Symposium conducted at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, IL.

Koretz, D. M. (1989). The new national assessment: What it can and cannot do. NEA Today, 7(6), 32-37.

Koretz, D. M. (1991). State comparisons using NAEP: Large costs, disappointing benefits. Educational Researcher, 20(3), 19-21.

Knapp, R. R. (1960). The effects of time limits on intelligence tests of Mexican and American subjects. Journal of Educational Psychology, 51, 14-20.

Leinhardt, G. & Seewald, A. M. (1981). Overlap: What's tested, what's taught? Journal of Education Measurement, 18(2), 85-96.

Linn, R. L. (1987). Accountability: The comparison of educational systems and the quality of test results. Educational Policy, 1, 181-198.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of the claims that "everyone is above average." Educational Measurement: Issues and Practices, 9(3), 5-14.

Madaus, G. F. (1988). The influence of testing on 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, IL: University of Chicago Press.

Mehrens, W. A. & Kaminski, J. (1989). Methods for improving 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, 215-237.

Messick, S. (1987). Assessment in the schools: Purposes and consequences (Research Report RR-87-51). Princeton, NJ: Educational Testing Service. Also appears as a chapter in P. W. Jackson (Ed.) (1988) Educational change: Perspectives on Research and Practice. Berkeley, CA: McCutchan.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) Educational Measurement (3rd ed, 13-103). Washington, DC: American Council on Education.

Mine, H., with others. (1987). A study of the effects of child rearing patterns on test anxiety in late adolescents. Paper presented at the biennial meeting of the International Society for the Study of Behavioral Development. Tokyo, Japan. (ERIC Document Reproduction Service No. ED 289 634)

Neely, R. & Shaughnessy, M. F. (1984). Assessments and the Native American. (ERIC Document Reproduction Service No. ED 273 889).

Nickerson, R. S. (1989). New directions in educational assessment. Educational Researcher, 18, 3-7.

Nitko, A. J. (1989) Integrating teaching and testing. In R. L. Linn (Ed.) Educational Measurement (3rd ed.). Washington, DC: American Council on Education.

Noggle, N. L. (1988). Report on the match of standardized tests to Arizona essential skills. Tempe, AZ: College of Education, Arizona State University, School Personnel Evaluation and Learning Laboratory.

Nolen, S. B., Haladyna, T. M., Haas, N. S. (In press). A survey of actual and perceived uses, test preparation activities, and effects of standardized achievement tests. Educational Measurement: Issues and Practices.

Nolet, V. & Tindal, G. (1990). Evidence of construct validity in published achievement tests. Paper presented at the annual meeting of the American Educational Research Association, Boston MA.

Paris, S., Lawton, T. A., Turner, J. C. & Roth, J. L. (1991). A developmental perspective on standardized achievement testing. Educational Researcher, 20(5), 12-20, 40.

Paris, S., Turner, J. C., & Lawton, T. A. (1990). Students views of standardized achievement tests. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Plass, J. A. & Hill, K. T. (1986). Children's achievement strategies and test performance: The role of time pressure, evaluation anxiety, and sex. Developmental Psychology, 22, 31-36.

Phillips, S. E. & Mehrens, W. A. (1987). Curricular differences and unidimensionality of achievement test data: An exploratory analysis. Journal of Educational Measurement, 24(1), 1-16.

Popham, W. J. (1987). The merits of measurement driven instruction. Phi Deltan Kappan, 68, 679-682.

Popham, W. J. (1990, January). Appropriateness of teachers' test preparation practices. Paper presented at a Forum for Dialogue Between Educational Policymakers and Educational Researchers. UCLA Graduate School of Education and the California School Boards Association, University of California, Los Angeles, CA.

Rock, D. A., Pollack, J. M., & Hafner, A. (1991). The tested achievement of the National Educational Longitudinal Study of 1988 Eighth Grade Class. Washington, DC: US Department of Education.

Sarnacki, R. E. (1979). An examination of testwiseness in the cognitive test domain. Review of Educational Research, 21, 252-279.

Seitz, V., Abelson, W. D., Levine, E., & Zigler, E. (1975). Effects of place of testing on the Peabody Picture Vocabulary Test scores of disadvantaged Head Start and non-Head Start children. Child Development, 46, 481-486.

Shepard, L. A. (1989). Why we need better assessments. Educational Leadership, 46(7), 4-9.

Smith, M. L. (1990). The meanings of test preparation. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. Educational Researcher, 20(5), 8-11.

Smith, M. L. & Rottenberg, C. (In press). Unintended consequences of external testing in elementary schools. Educational Measurement: Issues and Practices.

Smith, M. L., Edelsky, C., Draper, K., Rottenberg, C., and Cherland, M. (1989). The role of testing in elementary schools. Los Angeles, CA: University of California at Los Angeles, Center for Research on Educational Standards and Student Tests, Graduate School of Education.

Stiggins, R. J., Griswold, M. M., & Wikelund, K. R. (1989). Measuring thinking skills through classroom assessment. Journal of Educational Measurement, 26, 233-246.

Thomas, A., Hertzig, M., Dryman, I., Fernandez, P. (1971). Examiner effect in IQ testing of Puerto Rican working-class children. American Journal of Orthopsychiat, 41(5), 809-821.

Toch, T. (1991). In the name of excellence: The struggle to reform the nation's schools, why its failing, and what should be done. New York, NY: Oxford Press.

Walberg, H. J. (1980). A psychological theory of educational productivity. In F. H. Farley & N. Gordon (Eds.) Perspectives on educational psychology. Chicago, IL & Berkeley, CA: National Society for the Study of Education & McCutchan Publishing.

Wardrop, J. L., Anderson, T. H., Hively, W., Hastings, C. N., Anderson, R. I., Muller, K. E. (1982). A framework for analyzing the inference structure of educational achievement tests. Journal of Educational Measurement, 19(1), 1-18.

Wodtke, K. H., Harper, F., Schommer, M. & Brunellia, P. (1990). How standardized is school testing? An exploratory study of standardized group testing in kindergarten. Educational Evaluation and Policy Analysis, 11(3), 223-236.

Wright, B. D. & Stone, M. (1979). Best test design. Chicago, IL: University of Chicago MESA Press.

Zatz, S. & Chassin, L. (1985). Cognition of test-anxious children under naturalistic test-taking conditions. Journal of Consulting and Clinical Psychology, 53, 393-401.

Response to Thomas Haladyna's Presentation

Gary Hargett
University of New Mexico

My discussion of Dr. Thomas Halaydna's paper will be in two parts. First, I want to talk about a kind of test score pollution that Dr. Thomas Halaydna mentioned which has to do with the public perception of test scores and put my discussion in the context of the debate on educational reform. Secondly, I will offer some thoughts that are more specific to implications of testing for LEP students in the context of Title VII program evaluation which is originally what I was asked to do for this symposium anyway.

Dr. Thomas Halaydna and his paper listed several sources of test score pollution that invalidate test scores. The sources include tailoring curricular to specific tasks, coaching students for tests, teaching test wiseness, even excluding low achieving students from taking standardized tests. All of these we know happen, but I would suggest that maybe the greatest source of test score pollution behind these other sources is one that he has alluded to and I think needs further attention and that is the disproportional importance attached to tests by policy makers, editorialists and other commentators based on misconceptions about the role of tests and misinterpretations of the meaning of test scores.

I can illustrate this with a recent example. As you know, just last week the SAT scores were released. This year students in Oregon where I live had the highest average among the states, which got favorable local press. But I was listening to an editorial on television and the commentator thought it was really shameful that our average was only somewhat over 400 on the SAT which he said was only fifty-some percent of the maximum possible score of 800. He clearly does not have any concept of what SAT scores are -- about standard scores and that kind of thing. And I wondered if he really thinks that students take the SAT, sit there and answer 800 questions on each section of the test. But I think that his misunderstanding exemplifies the thinking of many of the loudest critics of the schools who just don't know what normal test scores are all about, and I have even seen this type of misunderstanding at the level of school superintendents who really should know better.

I think like most members of the public, policy makers and public commentators want information on how much this nation's students know, whether they are achieving at grade level, which is itself a not very well-motivated construct; and in fact standardized test scores that are most commonly reported, stanines, percentiles, grade equivalent scores, and NCEs, really don't tell how much a student

knows about the subject or what specifically a student knows even if we can assume a high degree of content and construct validity, which we know is not a safe assumption. The only kind of information these scores really convey is the degree to which a student is at, above, or below the average of the test norming group. These tests were developed on the assumption that I was taught on my own measurement training which is that we are interested in differences among people and we want to obtain reliable measurements of true differences among people. We want to know who's best, who's worst and who's in between.

These scores can be construed to represent content only if we share a common concept of content at each grade level. Yet, Dr. Thomas Halaydna gave the example of his own state, Arizona, where they found that the content of the test that had been used did not correspond to what the state curriculum was mandating be taught, and I think most of us have had this kind of experience; even among curriculum text publishers you have a very large variation of what a publisher construes to be grade level kind of work. So the very logic of norm referenced standardized tests may be inconsistent with the kinds of interpretations that most policy makers and editorialists try to impose upon them. These people want to know what our students know. The tests really only tell who knows more and who knows less. What worries me is the implications of the continued use of these kinds of tests in educational reform where there is emphasis on competition and, in my opinion, not a very healthy emphasis.

Any norm referenced test free of score pollution will find half the students above average and half below average. Now we are facing proposals for universal national testing and in an atmosphere of academic competition I think there will be a lot of hand wringing over who is below average and laying a lot of blame usually at the steps of the school. I think this will be true even if subsequent generations of students do achieve more than their predecessors. This approach to testing does not promote the principle of excellence for all. It only invites comparisons which some policy makers do want and which norm referenced tests can provide. But it still doesn't really tell what our students have learned.

I think there are plenty of examples of how test scores are perceived. The August 25 issue of Parade Magazine, which as you know enters millions of homes every Sunday, headlined an article about the schools with a statement about declining test scores. Yet we know, and as Dr. Thomas Halaydna mentioned a few minutes ago, achievement test scores have not declined, they have risen, and presumably for reasons due to test score pollution as he pointed out. It has not been suggested to my knowledge that in some cases the scores may have risen because schools are really doing a good job. The current orthodoxy and, in fact, it's almost a national policy now,

is that our public schools are a failure and there aren't any real criteria for that judgment, and it's contrary to much of the objective evidence that does exist. Dr. Thomas Halaydna referred to the Cannell report that came out a few years ago and that was discussed in the Fall 1990 issue of Educational Measurement: Issues and Practice. In that issue, I found it interesting that many explanations were offered as to why standardized achievement test scores were inflated, due mostly to pollution, but I still did not see any evidence for the decline of American education. Laurie Shepard's article in that issue cited data from the National Assessment of Educational Progress that showed modest gains and cited findings from the congressional budget office with figures that also showed improved achievement, just not as dramatic as the gains that are shown on the standardized achievement tests.

I don't mean to suggest that we are not facing real serious educational problems and I certainly don't suggest that we should not be seriously discussing educational reform. I think, of course, we can do better -- we should be doing better. But I think we should take a hard look at our expectations for student achievement and I don't think we should base our discussions on the a priori premise that the schools have failed without any solid evidence to that effect. The evidence as far as I can tell is pretty much anecdotal. I'd like to give an example of my own state of Oregon which has recently been nationally praised for taking the lead in educational reform. You may have heard about our reform package that was passed by the legislature just this summer. I think the point of view of most Oregon educators is that our legislators enacted a reform package without any clear statement of what the problems were or any compelling linkage of the reforms to those problems.

At the Seattle hearings on the national goals, I heard Dr. Ramsey Seldon, who is a member of the National Goals Panel Resource Group, remark that at this point we really don't know what's going on in the schools. He says, for example, we don't even know how many teachers and how many schools are using skills based as opposed to whole language reading approaches and to what degree they are using them. In other words, we're clamoring for reform without necessarily knowing what it is we are trying to change.

Dr. Thomas Halaydna's conclusions about the problems of LEP students taking standardized tests are certainly valid and they point up certain problems associated with recent proposals for universal testing. I refer to the proposal that every student should take an achievement test or a series of such tests at certain points in his or her educational career. I think we have to look at the implications of this kind of universal testing and I would suggest that we do not need universal testing to assess the attainment of educational goals

assuming, that is, that tests can be calibrated to those goals or if tests are the most desirable measure of goal attainment.

I think we can accomplish that through well-applied matrix sampling, which is what the California Assessment Program does. The only reason for obtaining test scores on every individual is if there are individual consequences and implications based on the individual's test score. I recently heard a spokesman for a group called Educate America, advocate testing of all high school students in the fall of their senior year. In his comments he said that at first this would be low stakes testing but, then, when it was pointed out that students are not motivated to do well on low stakes test, he said students will be motivated to do well because these scores might be considered in college admissions or looked at by potential employers. Well, at this point these become high stakes test scores.

I agree with the observation that, for many or most purposes, test scores for LEP students tend to be invalid. I think they're valid in one sense, in the logic of norm referenced tasks that LEP students don't know as much or don't have the same kind of skills as the norming group on whatever it is the standardized tests measure. Whether that's important or whether LEP students have academic talents that are not measured by the tests is a separate issue.

I personally do not advocate large scale high stakes testing, but I am worried about certain implications of the exclusion of LEP students from such tests even out of benign concern for the invalidity of their test scores. My most important concern is that this sends a message that marginalizes LEP students, that since we cannot test them they're marginal to education. If a point of tests is to drive excellence in education, they should drive excellence for LEP students as well. My other concern is that scores from large scale high stakes test may become another kind of credential. Rightly or wrongly, the high school diploma is widely perceived as not necessarily representing the mastery of academic skills. That is part of the reason for the demand for new tests such as the minimum competency tests we have seen in many states. If LEP students are excused from tests because their test scores are invalid due to language, they will be leaving school without an important credential.

I think we are seeing the possibility of this in Oregon where part of our reform package is that, at tenth grade, students will take a test for a Certificate of Initial Mastery -- whatever that means. And after that, they go into either a college prep track or a vocational track, which has many of us sort of in horror. But if the LEP students cannot take these tests for the Certificate of Initial Mastery, then you wonder, well, what options are open to them after the tenth grade? I don't mean to imply that I favor testing LEP students with tests based on English only norms because I certainly don't. I don't

even favor developing alternative norms because I think that would be pointless and probably impossible. I am only pointing out some logical consequences for LEP students in the context of large scale testing. My personal preference is that we back away from the imposition of high stakes testing for all students.

This brings me to the second part of my discussion dealing with the use of tests, particularly standardized achievement tests and Title VII program evaluation. The Title VII regs do not require standardized tests. They require reports of educational progress measured as appropriate by tests of academic achievement and they require that the evaluation instruments that are used consistently and accurately measure progress toward the project objectives, that they be appropriate considering several factors including language proficiency, and that they be administered at twelve month testing intervals. I think that many people have construed this to mean that standardized tests are required because of the key terms "academic achievement" and "twelve month intervals". They may also think that since the tests they use have to be reliable and valid, they should use standardized tests because, after all, these have technical manuals that report their validity and reliability.

However, as Dr. Thomas Halaydna has pointed out, these are not reliable and valid tests for LEP students for a number of reasons, including lack of content validity for a typical Title VII project curriculum. What they most reliably do is show that LEP students perform much lower than other students measured by these tests, which is not surprising since part of the definition of LEP is that they are not able to learn successfully in classrooms for the language of instruction and the testing is in English. What Title VII evaluation and regulations call for is a measure of progress toward accomplishing the objectives of the project. It's not uncommon to see Title VII project objectives written in terms of bringing the LEP student up to grade level.

But I think we need to think about the implications of this kind of project objective and how to test it. It would seem on the face of it that standardized tests would be a logical measure of that kind of objective. But I see two problems apart from the obvious question of what grade level even means. We lose sight of the fact that grade level is not a point but a range of abilities. The first problem is whether this kind of objective is reasonable for many projects, especially if you consider projects that are serving some older students -- upper elementary and high school students who may be coming into the schools with very weak academic preparation in their own native languages. It's probably not reasonable to expect them to perform comparably to the norm group on the standardized achievement test or in other measures as well.

Where I have seen these tests most effectively used has been with projects that work with early elementary students and are able to give sustained service over a period of years and, in fact, a service that is actually mainstreaming from the very beginning. It's not the model where first we give them Title VII and then we give them the real curriculum. I think that any bilingual education program should strive to help the students advance as much as possible in language and academic abilities, but if the measure of gain is performance on a standardized achievement test and the goal performance is comparable to the norming group, that may be an elusive goal. I would like to see Title VII projects experiment with some of the alternative assessment approaches that are being discussed in this symposium. One reason for this is something I've learned during my experience with program evaluation both through the EAC-WEST and other evaluation roles I've played. I've learned that evaluation issues become a focal point, maybe even a lightning rod, for the discussion and clarification of many other issues.

We've seen this in the national debate on education. Unfortunately, this debate is murky because the evaluation issues are not well understood. But I think there is a great potential for the role of performance or authentic assessments in Title VII evaluation. I think first of all that many, maybe most Title VII project curricula, really are not built around the kind of things standardized tests are intended to tap into. Therefore, the projects need assessments that are built around the curricula, and we hope that those curricula are targeting levels of excellence and meaningful tasks and applications. I think that the development of performance assessments provides the form for articulating expectations, thereby setting standards of excellence to teach toward. I think that's a more exciting educational concept than either grade level or minimum competency. By the way, this is not an easy process, as the people who have been working on performance assessment can tell you. From my own experience in many of the workshops I have given, one of the hardest things to do is to get teachers to articulate the outcomes they expect for their students, and this is true of many kinds of teachers, not just teachers in Title VII programs.

But this is what teachers and other educators have to do in order to meet the kinds of standards of excellence that AMERICA 2000 is supposed to be about. I'm afraid, that if educators don't articulate the expectations, then politicians will, and I personally have more confidence in the educators than the politicians to do a good job of that. Developing performance assessments can have several advantages because by their very nature they set standards of excellence, and I think that's an attitude that Title VII programs must assume, and move away from the deficit model. We know that all students, including LEP students, tend to meet expectations, so we should have expectations that embody excellence. Other potential advan-

tages of performance assessments are high curricular validity and the communicability of test findings. As I suggested earlier, there's nothing really wrong with standardized test scores, but the way they are interpreted miscommunicates the content of the scores, whereas performance assessments are couched in terms of actual performance, what students really can do and how well they communicate.

I don't want to give the impression that performance assessment is an automatic panacea. For one thing, it is a supplement to, not a replacement for, other kinds of assessment. They still do have their place and they do have pitfalls which Dr. Eva Baker went into yesterday. By the way, I think the biggest pitfall is trying to impose performance assessments in the traditional setting. I think the performance assessment only makes sense in an atmosphere where students are performing, and problem solving is part of their everyday educational experience. So if you do want to develop performance assessments for your Title VII project evaluations, I would encourage you to do so but look for guidance and, of course, the EACs are a good place to start looking for that guidance.

To summarize my remarks, I agree with Dr. Thomas Halaydna that test scores have become polluted. The proliferation of test scores might itself be said to be polluting, but the most dangerous pollution is the over-interpretation and the misinterpretation of test scores which I think leads to many of the other sources of pollution that Dr. Thomas Halaydna listed. And I also think that the standardized test should be used cautiously with Title VII evaluations and the Title VII project should consider alternative methods of assessment that promote excellence.

Response to Thomas Haladyna's Presentation

María Pennock-Román
Educational Testing Service, Princeton

Overall, I concur with Haladyna on many points, and I agree with most of his recommendations about proper and improper uses of tests. Nevertheless, I find that his application of the labels "pollution" and "contaminants" obscures the issues at hand, beginning with the title. If one looks closely at most of Haladyna's criticisms, it is evident that he disapproves of common uses of tests by state policy makers, school administrators, and teachers. For this reason, I believe it would be more appropriate that his paper be titled "Test Use Pollution."

My reaction to his major points are summarized in three tables in order to conserve space and time. Most of the entries in the tables are self-explanatory so that only selected rows will be discussed.

Desirable and Undesirable Test Practices

Table 1 presents a contrast between Haladyna's opinions and mine concerning what testing practices are desirable or inappropriate. Next to each testing practice that Haladyna considers a "contaminant" in test scores is his classification as to whether the practice is ethical (E) or unethical (UE). In the adjacent column are my views concerning this classification and comments to explain my rationale.

As shown in Table 1, the author in some ways contradicts himself as he applies the negative label of "contaminants" to testing practices that he himself considers "ethical." Haladyna on the one hand considers training test wiseness or increasing student motivation as contaminants of test scores but, on the other, he classifies training in test wiseness and increasing motivation as ethical practices. Later, he makes a recommendation that LEP students be trained to take tests properly.

Happily, there is a fairly easy way to resolve this inconsistency by changing the form in which the "contaminant" is described. For example, I believe that in the case of students outside the mainstream it is inexperience with tests or test naïvete that may add unnecessary noise to scores. In a study of test-taking skills of Hispanic junior and high school students in California (Pennock-Román, Powers, & Perez, 1991), I was appalled to find that even filling out answer sheets presented problems for some students. Certainly, test naïvete may reduce the validity of the test for inexperienced test tak-

Table 1
Juxaposition of Haladyna's and Pennock-Roman's Views of Testing Practices

Haladyna's Views	Pennock-Roman's Views			
Testing Practices Considered "Contaminants"	Ethical/ Nonethical	Ethical/ Unethic.	Contaminant?	Rationale
1) Preparing students for tests				
a) Training in testwiseness	E	E	No	Extreme test naivete may introduce irrelevant variance. See Maspons & Llabre (1985).
b) Checking answer sheets for proper completion	E	E	No	See above
c) Presenting items similar to those on the test	UE	E	No	See above
d) Using test preparation books such as Scoring High	UE	E/UE	Some Cases	Depends on test preparation book. Good if reviews fundamental math, etc.
e) Presenting items verbatim from the test	UE	UE	Yes	Introduces variability in scores irrelevant to domain of achievement.
f) Cheating (giving out answers or hints, correcting answer sheets)	UE	UE	Yes	Introduces variability in scores irrelevant to domain of achievement.
2) Promoting student motivation for the test	E	E	No	It is lack of motivation that is irrelevant to the measurement of academic achievement.
3) Teaching to the test				
a) Preparing objectives based on items on the test and teaching accordingly	UE	E/UE	Usually	Depends on completeness of test. In some instances, (e.g., Advanced Placement Tests), objectives are comprehensive.
b) Developing a curriculum to match the test	UE	E/UE	Usually	Same as above.
4) Dismissing low-achieving students on testing day to artificially boost test scores.	UE UE	E/UE	Some Cases	Occasionally justified, but criteria for exclusion should be uniform and well defined.

ers. A small-scale study by Maspons and Llabre (1985) lends support to this view, but a lot more research needs to be done in this area.

He and I also agree in our disapproval of adapting curricula and "teaching to the test," under most circumstances. However, I can think of exceptional cases where especially comprehensive tests can serve as good guides to curricula. At the risk of sounding as though I'm putting in a "plug" for my company, consider the Advanced Placement (AP) Tests which are college-level achievement tests in various subjects. Students attaining high grades on a given Advanced Placement Test receive college credit for that course. These tests have been carefully designed to cover a domain area quite rigorously and thoroughly under the guidance of college professors from representative universities. Because of the care in its construction, curricula designed to encompass the material of an AP test may indeed be a good one to follow.

However, tests such as the AP tests are the exception rather than the rule. In general, "teaching to the test" is not a good idea because most achievement tests are not linked to specific, well-defined courses of study.

Haladyna and I also concur in disapproving of the practice of dismissing low-achieving students on testing day to artificially boost test scores. One exception mentioned later on by Haladyna are LEP students who should be excused from standardized achievement tests until their competency in English is sufficiently high to make test scores meaningful. Of course, defining the point at which there is enough proficiency in the language of the test is a difficult task. More research is needed in this area. Besides LEP students, there are other groups of exceptional children who are learning disabled or physically handicapped for whom traditional tests may be invalid. These students ought to be excluded from analyses of summary statistics for a given school.

In any case, the criteria for exclusion of special children from public reports of test results need to be well-defined. Results will be comparable across school districts only when such criteria are applied consistently on the districts that are contrasted. It would be desirable if such criteria could be defined on a national basis to make norms on widely used achievement tests more useful.

Table 2
Juxaposition of Haladyna's and Pennock-Roman's Views Concerning Relevance of Variables to Domain Tested

Haladyna's Views		Pennock-Roman's Views	
Variables Considered "Contaminants"	Contaminant?	Rationale	
1) Situational Factors			
Test Anxiety	Sometimes	Excessive anxiety or nonchalent attitude are problems.	
Stress	Yes		
Speededness of the test	Usually	But in rare cases, speedyness part of construct.	
Examiner Effects	Sometimes	Probably little effect with multiple-choice tests	
2) Context			
Language deficits	Usually	Influence and relevance of language deficit varies by test content. Quantitative tests less of a problem.	
Socioeconomic context	No	For achievement tests, these sources are related to underlying content area.	
Family Mobility	No	See above.	
Family and Home Influences	No	See above.	
Prenatal/Early Infant Influences	No	See above.	

Contextual and Situational Factors

As shown in Table 2, Haladyna and I are also largely in agreement with regard to the issue of speediness and language deficits as a contaminant in standardized test scores. He should consider adding to his paper some recent reviews of the literature on speediness for non-native speakers of English which support this point of view (Llabre, 1991; Pennock-Román, in press). There is evidence from many sources, that non-native speakers of English have great difficulty in completing selective-admissions tests, particularly the verbal portions of tests such as the SAT, GRE, and GMAT (see review by Pennock-Román, in press.)

In Pennock-Román (1990) and in the aforementioned review (Pennock-Román, in press), there is also a discussion of language proficiency in the language of the test as a factor that interferes with the measurement of ability and achievement. However, one finding of special interest is that some curriculum-specific achievement in subject areas are somewhat less influenced by language proficiency than more global types of ability tests. Naturally, quantitative tests are less influenced by language factors, but more verbal types of tests show this effect also. One explanation is that non-native speakers of English may be on more equal footing with mainstream students in regard to academic vocabulary (e.g., technical terms in science) than they are with language terms learned mostly outside of the school environment (e.g., names of fruits, furniture).

In contrast, Haladyna and I differ in our positions concerning family background and other contextual influences on test performance. He is somewhat ambivalent in this position concerning the classification of these variables as contaminants or meaningful variance. Whereas, he lists socioeconomic context, family mobility, family and home influences as "documented sources of test score pollution," on page 31 he states that "Any test score, if unpolluted, reflects a lifetime of school and non school learning and a myriad of influences." Hence, his position is not clear -- are home influences pollution or not?

From my perspective, background factors affect the quality of training a student has had, which for the most part is a valid source of variance because it does affect the content domain to be measured (academic achievement) in our society where educational resources are unevenly distributed. On the other hand, these sources do limit the uses that test scores can serve. And it is not proper that teachers and schools should be evaluated without taking into consideration these factors. Thus, there are problems with using student test performance to evaluate teacher effectiveness because teachers are only one of many influences on those scores. Multiple indicators are nec-

essary to evaluate schools and teachers. I believe that we need to make a distinction here in terms of the different uses of tests and relevance of these variables to the purpose for which the test serves.

Construct Validity Should Refer Only to Intended and Recommended Uses of Tests, Not to All Other Uses That Occur

While the Haladyna uses the APA Standards definition for what is proper evidence of validity, he states that "construct validation calls for the collecting of evidence to support any of the 29 different uses [referred to in his Table 1]," whether it is recommended or not. This is clearly not the intent of the standards. Key words in the Standards are "Evidence ...presented for the major types of inferences FOR WHICH THE USE OF A TEST IS RECOMMENDED... Support the particular mix of evidence presented for INTENDED uses."

He implies that performance tests, alternative testing, and "authentic" measures will provide a future solution, because they are free of the problems that multiple-choice tests have. However, as he points out, the main problems stem from misuse and misapplication of multiple-choice tests. Won't future, performance and alternative tests be subject to misuse also? And, given the many problems in scoring such tests because of subjectivity in grading, won't the potential for misuse be even greater?

As long as we continue to blame the test rather than school and state policies for improper test use, problems will not be corrected, and they will recur with any kind of test that is devised, standardized or not.

His Recommendations Are Mostly Points of Agreement between Us

My points of agreement or disagreement on recommendations are presented in Table 3; you can see that there are few disagreements with the recommendations, and most are self-explanatory. I'd like to suggest that he repeat in the latter pages (pp. 30-31) some of the points referred to earlier in the manuscript, because many are worth reiterating.

Table 3
Evaluation of Haladyna's Recommendations

Pennock-Roman's View of Haladyna's Recommendations	Haladyna's Recommendations
1) Testing Practices	
Mostly Agree	"All children should have comprehensive test-taking courses and be equally skilled in test-taking." (p. 24)
Agree	Should avoid test activities designed to produce spuriously inflated test scores that do not represent true learning. (p. 25)
2) Situational Factors	
Agree	Test anxiety may be reduced through ethical test preparation activities. (p. 25)
3) Context	
Mostly agree	One must first prove that students have a satisfactory level of mastery in the English language before being tested in English. (pp. 28-29)
Mostly agree	Scores should be reported by SES. (p. 19)
Agree	One should take into consideration students' SES and other factors when evaluating teachers and schools using scores. (p. 31)
Summary Recommendations	
Agree	"We should resist attempts to interpret or use test results in ways unintended and unsupported by validating evidence." (p.5)
Agree but issue not well developed in his paper.	"Need more wisdom in the definition and measurement of school achievement." (p. 31)
Agree	"Test scores obtained from LEP students appear to be invalid for many [typical] interpretations and uses." (p. 31)
Agree	"Use multiple indicators for evaluating schools and teachers." (p. 30)
Disagree	"Standardized test scores are often polluted. We should avoid using such test results." (p. 31)

Conclusion

In general, most of the criticisms and recommendations that Haladyna makes are sensible; many have been suggested before by measurement specialists and other educators, so there is relatively little new here. The majority of his criticisms do not address test content, format, or test construction. However, by using the term "test score pollution" he puts the blame for many wrong uses of tests on the instruments themselves, rather than on test users. Furthermore, there are some contradictions introduced by grouping too many things under the label of pollutants.

Problems with the use of the terms "pollution" and "contaminants" arise for two reasons. First, these terms, which are loaded with negative connotations, are applied in an overinclusive manner to a variety of practices considered both ethical and unethical according to Haladyna himself. Hence, the label of "contaminants" tends to obscure his distinctions between appropriate and inappropriate uses of tests, thus making his policy recommendations unclear. Second, I find that, in this controversial area, the use of inflammatory language is counterproductive. It interferes with the constructive dialogue among test specialists, educators, and advocates of LEP children that is necessary for positive solutions to measurement problems.

References

Llabre, M. M. (1991). Time as a factor in the cognitive test performance of Latino college students. In Deneen, J., Keller, G., & Magallan, R. (Eds), pp. 95-104, Assessment and access: Hispanics in higher education, NY: SUNY Press.

Maspons, M. M., & Llabre, M. M. (1985). The influence of training Hispanics in test taking on the psychometric properties of a test. Journal for Research in Mathematics Education, 16, 177-183.

Pennock-Román, M. (1990). Test validity and language background: A study of Hispanic American students at six universities. New York: College Entrance Examinations Board.

Pennock-Román, M. (in press). Interpreting test performance in selective admissions for Hispanic students. In Geisinger, Kurt (Ed.) Psychological Testing of Hispanics, Washington, D.C.: American Psychological Association.

Pennock-Román, M., Powers, D. E., & Perez, M. (1991). A preliminary evaluation of Testskills: A kit to prepare Hispanic students for the PSAT/NMSQT. In Deneen, J., Keller, G,., & Magallan, R. (Eds)., pp. 243-264, Assessment and access: Hispanics in higher education, NY: SUNY Press.