

*Proceedings of the Second National Research Symposium on Limited English Proficient Student Issues:
Focus on Evaluation and Measurement. OBEMLA, 1992.*

Test Score Pollution: Implications for Limited English Proficient Students

Thomas Haladyna
Arizona State University, Tempe

[Table of Contents](#) * [Next Paper](#) * [Previous Paper](#) * [Discussants](#) * [References](#)

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	Units of Analysis
Allocation of Resource 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	Program
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, & Schville, 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, socio-economic 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, & Wiklund, 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 \(1989\)](#) 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 <i>Scoring High</i> 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 student 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 standardized testing experience are that (1) students become increasingly skeptical 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 social 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 Meine \(1990\)](#) 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

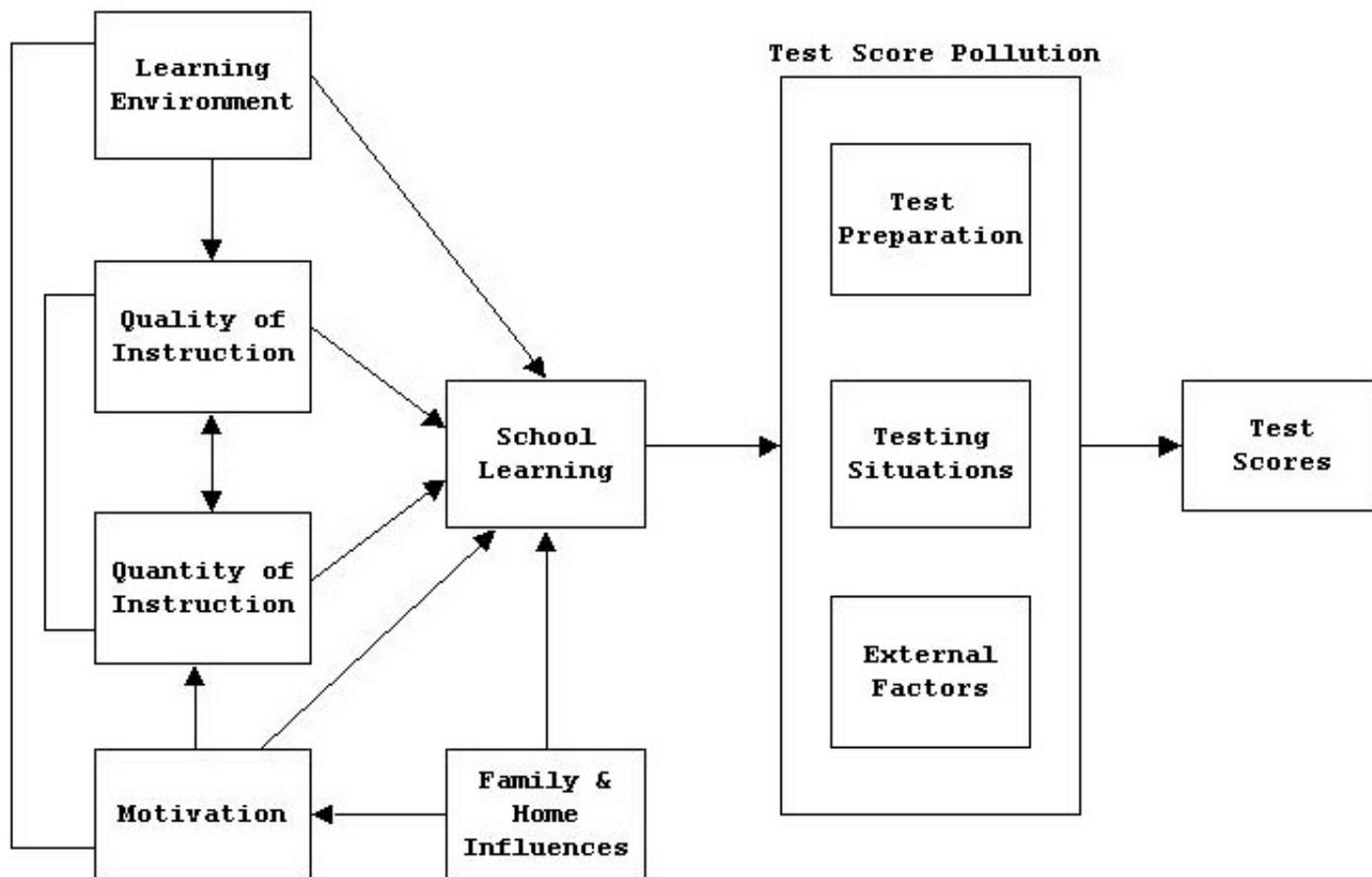
This section of the paper addresses implications for LEP educators arising from the problem of test score pollution. This section also suggests some fruitful areas for research on the role or influence of test score pollution on LEP students. Finally, recommendations are offered to protect LEP children 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 I](#) 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 I 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, Hertzog, 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, 1987](#)). 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 lay-persons 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 standardized 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, lay-persons will need to be better instructed in this role of testing in instructional programs.

Note

1A 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.

[Table of Contents](#) * [Next Paper](#) * [Previous Paper](#) * [Discussants](#) * [References](#)

Use Back Button to Return to Page
