DOCUMENT RESUME

ED 304 454 TM 012 816

AUTHOR Gomez, Mary Louise
TITLE Testing Policies and Procedures for the At-Risk Student Program Area.
INSTITUTION National Center on Effective Secondary Schools, Madison, WI.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 86
GRANT OERI-G-86-0007
NOTE 41p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adolescents; Attendance; Discipline Problems; Dropout Rate; Evaluation Methods; *High Risk Students; Low Achievement; *Program Evaluation; *Reading Tests; *School Effectiveness; Secondary Education; Secondary School Students; Special Programs; *Testing Programs; Truancy; *Writing Evaluation
IDENTIFIERS *At Risk Student Program Area; Degrees of Reading Power

ABSTRACT

The At-Risk Student Program Area, part of the National Center on Effective Secondary Schools, will study how at-risk secondary school students are affected by special programs, efforts toward school improvement, and policies of admission and attendance. This paper describes the testing procedures that will be used. Data will be collected in six ways: (1) direct program observation; (2) interviews of students, teachers, administrators, and other persons knowledgeable about the programs; (3) a survey of adolescent social and personal orientations; (4) measures of schools' effectiveness in reducing truancy, dropout rate, and disciplinary problems; (5) a standardized test in reading; and (6) samples of student writing. The original plan called for a mathematics test, but research has not yielded an appropriate instrument. The standardized reading test chosen is the Degrees of Reading Power, to be administered to students in its grade 7 through 9 form in the fall and spring of 1986-87 and 1987-88. Data from this test, in conjunction with evaluation of student writing samples, and direct program observation, interviews, and surveys that will include mathematics, will be used to: (1) help inform program designers, teachers, and administrators about the effects of special programs; (2) highlight the skills and weaknesses of students within a program; and (3) allow researchers to compare and contrast programs. (SLD)

Testing Policies and Procedures for the At-Risk Student Program Area

by

Mary Louise Gomez, Ph.D.
National Center on Effective Secondary Schools
University of Wisconsin-Madison

This paper was prepared at the National Center on Effective Secondary Schools, School of Education, University of Wisconsin-Madison which is supported in part by a grant from the Office of Educational Research and Improvement (Grant No. OERI-G-86-0007). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of this agency or the U.S. Department of Education.

Introduction

The At-Risk Program Area is one component of the National Center on Effective Secondary Schools, funded by the Office of Educational Research and Improvement. The five-year mission of the National Center on Effective Secondary Schools is to learn how secondary schools can improve the achievement of all students, especially those who are disadvantaged and less successful students. The particular objective of the At-Risk Program Area is to understand how at-risk students are affected by special programs, efforts at schoolwide improvement, and districts' policies regarding admission and attendance.

At-risk students are defined by Wehlage (1986b) as ..."those youth who have serious personal and/or academic problems that are likely to lead to dropping out" (p. 1). These students now compose more than 25% of the American secondary school population

Data Gathering in the At-Risk Project

A constellation of methods of data gathering has been selected to help assess the impact of various intervention programs on at-risk secondary school students. Data will be collected in six ways: (1) direct program observation, (2) interviews of students, teachers, administrators, and other persons knowledgeable about the programs, (3) a survey of adolescent personal and social orientations via an instrument developed at the Wisconsin Center for Education Research, (4) measures of schools' effectiveness in reducing truancy, the dropout rate, and disciplinary problems with students, (5) a standardized test in reading and (6) samples of students' writing.

This constellation of methods of data gathering is significant in order

to avoid what Kilpatrick (1979) terms the illusion of curriculum evaluation. He argues the "curriculum" does not exist nor can it be measured in a global sense, but must be viewed and evaluated in terms of its work or products, its situation-specific nature. Rather than attempt to "evaluate a curriculum" and compare the effectiveness of different curriculums, Kilpatrick suggests gathering evidence of the instructional products of a curriculum. He asserts this procedure requires "...the skills of the reporter, the historian, the anthropologist, and the critic" (p. 169).

The title of one chapter of a recent text (1984) concerning testing is "The Public Stake in Proper Test Use." The text, an outcome of a 1983 Educational Testing Service Conference called the Uses and Misuses of Tests, explores the appropriate role of testing in American education. John Casteen, Secretary of Education for the State of Virginia, writing about "proper test use" (p. 11), states that good tests "verify or validate what people do," they "sustain learning and document competence." The major question which this paper must address is the proper role of tests in research concerning at-risk secondary school students. Among the questions one must ask are: What results should evaluation of these students' work offer to the students, their teachers, researchers? Why should one course of data gathering or method of evaluation be pursued rather than another? What are the general dilemmas involved in testing and the particular problems associated with testing educationally at-risk youths? How does one best gather a project's work and capture its working methods, plans, procedures, and internal methods of evaluation without placing a burden on program participants?

Testing

The purpose of this section of the paper is to provide readers with background in current thought concerning the evaluation of students' achievement through standardized tests.

A test may be defined as a collection of pieces of information about student achievement (Tyler & White, 1979) or a systematic and deliberate way of sampling a student's behavior or thinking (Stodolsky, 1975). Among the purposes a testing program may serve are the following (Haney, 1985): the evaluation of student progress, provision of diagnostic information about individuals for instructional purposes, informing student grouping and placement decisions, identification of curricular goals for emphasis, evaluation of instructional programs, provision of information for planning, and provision of accountability to school boards and communities. Some critics (Hein, 1975) of American schooling and its practices charge that testing serves a major sorting function in schools and tends to support the existing social and political system. Kilpatrick (1979), writing in a volume published by the International Commission on Mathematical Instruction, argues that evaluation is composed of three processes: an interactive process involving knowledge, values, and beliefs; a psychological process affecting how people view themselves; and a socio-political process which leads to decisions concerning who will be educated, how they will be tracked, and what will be taught. Further, testing can be used to assist researchers and practitioners in questioning and experimenting with the existing structures of schooling and the assumptions upon which they are built.

Decisions concerning issues such as educational equality and teacher,

school, and program effectiveness are increasingly affected through the use of such standardized achievement test data. As this practice has increased, Tyler and other researchers are questioning its value (Airasian & Madaus, 1983; Haertel & Calfee, 1983; Tyler, 1974).

Ralph W. Tyler (1974) explains the use of achievement tests for the study of educational effectiveness.

> Since 1925 the accepted design for studying educational effectiveness has been, at some initial point, to give an achievement test called a pretest, and at a final or later point, a comparable posttest. The gain in the mean score was taken as a measure of the educational effectiveness of the program, method, or instructional material. In case a pretest could not be given, the final scores could be compared with test norms or with the scores obtained from comparable groups not following the program method or material under study (p. 143).

Airasian and Madaus (1983, p. 105) argue there are five basic problems with this use of achievement tests to detect performance differences between schools or programs. They question the use of a test intended to assess at one level, the individual, for assessing the performance of another level, the school or program. Standardized achievement tests are designed to assess the skills and make decisions concerning individual students. Yet, such tests are some times used to make decisions about the entire group of those students. Airasian and Madaus point out that the performance of an individual student may differ greatly from the mean performance of the group of students.

They argue that standardized achievement tests do not reflect the specific content and objectives of a program as the test items are reduced to those representing the lowest common denominator, or are not part of the content taught in a curriculum.

A third problem cited is the low correlation between the methods of instruction and learning and that of measurement. That is, the multiple choice format of tests requires students to select the best answer from provided alternatives, whereas, class time teacher requests require students to supply or construct their own response.

In addition, the authors question the use of a total or single score on an achievement test as the dependent variable in studies of school or program effectiveness as such scores mask, rather than highlight, the differences between schools or programs. Differences between schools or programs, they argue, are more likely to be at the specific skill level than at the total score level.

Finally, they write that influences concerning differential programs or instructional effectiveness are best derived when the processes underlying test performance are closely linked to instruction. This issue becomes the focal point of their argument against the use of standardized achievement tests for providing differential data on teachers, schools, or programs.

Tyler and White (1979) summarize the recommendations of the 1978 National Institute of Education sponsored conference on Research on Testing.

> Educational tests are now predominantly used for four purposes: accountability, selection, evaluation, and classroom guidance. Problems in each area can be identified in the context of contemporary criticisms of tests: Selection procedures are not completely fair to minority students. The use of tests for accountability is imperfect, and now new tests for accountability are urged. The tests are not a positive force in classroom teaching and, in some regards, are perceived as inhibiting and constrictive. Finally, the tests are not broad enough in scope to allow for fully satisfactory evaluation of educational programs (p. 11).

Eisner (1985) supports this viewpoint, acknowledges the aura of precision associated with numbers derived from testing and argues that

numbers are but one limited kind of reporting device.

Romberg (1986) writes that standardized tests do reasonably well what they were designed to do, rank order respondents in regard to a particular type of mental ability or skills or show the respondent rank in a group, and are easy to develop, administer, purchase, and comprehend. The greatest problem he notes is their use to address problems for which they were not intended. For example, it is claimed that standardized tests are used by elected officials and educational administrators to compare the quality of teachers' instruction, schools, and school districts. Romberg underscores the misleading nature of such comparisons.

Although standardized tests are designed to rank order respondents on some measure, the purposes such ranking serves and the damage done to those in the bottom of the rank are subjects of concern. In a text titled Barriers to Excellence: Our Children At-Risk (1985) The National Coalition of Advocates for students reports findings of its study of "at-risk" children. Funded by numerous corporate sponsors and philanthropic foundations, their Board of Inquiry Project held hearings in ten United States cities to hear testimony concerning the strengths and weaknesses of the educational system, particularly as it relates to differential treatment of children by race, class, sex, language, and handicap.

Among the concerns voiced in the text is the narrowing of school curriculum to a focus on teaching to the standardized tests. Increased pressure is placed on teachers to raise students' scores. Students of lower income, minority, and handicapped status too often find their scores at the bottom of the test's bell curve where their weaknesses, rather than their strengths, are highlighted.

In addition to problems with the general notion of using tests to measure the effectiveness of educational innovations, there exist problems of variable student test performance.

Students' Test Performance

A number of factors affect a student's test performance, therefore caution is advised when drawing conclusions from any single student test score. General factors affecting test performance include (Harris & Sipay, 1985) the "test-wiseness" of the student, the lack of passage dependence of a comprehension task, the task demands placed upon students in the testing situation, students' response accuracy, test anxiety and by such a collection of unpredictable causes as tiredness, guessing, inattentiveness, and lack of motivation.

Students' "test-wiseness" may affect their performance in any one testing situation. Harris (1985) has defined test-wiseness as "...the ability to use test taking skills to their fullest in order to obtain the highest score possible" (p. 169). Because students differ widely in their test-taking ability, Daley (1977) encourages teachers to prepare students for test-taking situations. This preparation includes explanations of why the tests are given, how results are used, the sorts of thinking skills required, and the role of the teacher during the test. These preparations are designed to develop test-wiseness in all students.

Another factor which may determine test outcome is the passage dependence of a comprehension task. Students may be able to answer some reading comprehension test items without reading the passage. Such questions are said to lack passage dependence (Harris, 1985). This may be attributed to students' prior experiences, the presence of the answer in

previous questions concerning the passage, or the distinctive appearance of the correct answers, which may be longer and more precise than the other choices.

Daley (1977) summarizes research concerning students' attitudes towards testing. She describes differences found by Neulinger in secondary school students' attitudes towards testing based upon social background and personality characteristics.

> The student in the lower socioeconomic and less educated domain saw the test as identifying him or her--but not as a member of an elite. The identification was the equivalent of being degraded The school, which is supposed to upgrade his or her abilities (as students see it), condemns the student before he or she gets a chance. The test excludes him or her from places of higher learning (p. 59).

The consequences of frequent negative feedback concerning test performance for at-risk students is clear: further alienation from school.

Daley (1977) also discusses the negative effects of low self-concept, anxiety, and low motivation on students' test scores. Further, she deplores the poor testing environment; most often students sit in large, poorly lit rooms with infrequent breaks between test subsections.

Further, the task demands placed upon a student may vary from test to test and from testing situation to classroom pursuit of a similar task. Eisner (1985) argues that a major problem with achievement tests is related to the context in which they are given. Students in such testing situations know they are expected to do their best work in a form and context they know are artificial.

In addition, students' response accuracy must be taken into account when using test scores for evaluation as well as diagnostic purposes. For example, two students may achieve the same score, but one works slowly and

answers few problems while the other works rapidly and guesses on many items. The way in which the student approaches the test is not described by test data.

Sarason (1983) speculates that students may spend more time worrying how their work will be evaluated than they do absorbed in academic pursuits. He suggests that teachers increase their feedback to students regarding skill acquisition and skill strengthening. Sarason also suggests that test anxiety may be reduced when students receive information concerning the evaluative tasks they are presented as well as the strengths and weaknesses they as individuals bring to the task.

Finally, the illness, lethargy, or lack of motivation or attention felt by students in a given testing situation will potentially affect their score. There are numerous factors which affect student test performance and consequently, success in school. Current testing practices demand the attention of teachers and researchers.

Future Evaluation Techniques

Concern for improved measures of student performance and program effectiveness has existed for many years. While researchers continue to critique current practices, they also look forward to a new era where testing and instruction are used to create a different learning environment.

Participants at the 1978 National Institute of Education sponsored conference on Research on Testing developed a vision of the future where continual collection of test data would inform the instructional process. Conference Chairpersons Ralph Tyler and Sheldon White summarized this vision (1979, p.25):

> In this vision of the future, school tests as we know them would cease to exist. The intrusive, specialized,

> institutionalized activity called testing would be absorbed into a new kind of learning and testing environment. Computers could accept inputs from students and teachers on an almost continual basis, extracted from the rich tapestry of ongoing learning activities. Instructional systems would accumulate an educational portfolio for each student, including a wide range of interrelated performance and situational descriptions. One would be as unlikely to cease all instructional activities to test a student as one is to stop conversing with a child in order to test his or her linguistic competence. Instead, testing would be a continuously collected data base. Some of these aggregation would have an immediate impact on ongoing learning activities; others would be remote from the moment of data collection.

Despite the criticism of researchers concerning the appropriateness of using standardized achievement tests and the promise of a future where teaching and testing are intertwined, the dilemma of gathering data concerning special programs for at-risk secondary school students remains.

Parents, teachers, and members of the community continue to seek the sort of information provided by standardized testing. Until educators, parents, policymakers, and the public understand what tests can and cannot do, data from such measures will continue to be sought as indicators of how well students, teachers, programs, and schools are performing (Tyler & White, (p. 18).

Haertel & Calfee (1983), writing in the Journal of Educational Measurement, describe the current role of test experts and test users (p. 130).

> For the time being, test experts and test users in the schools can strive to become more aware of the limitations of objectives stated in behavioral terms, can scrutinize more closely the processes and knowledge provided by tests, and can work to create a stronger demand for better measures of achievement.

Given the limitations and cautions described above, we have chosen to

test students' academic achievement in one area, skill in communication. One component of the assessment of students' communications skills is the assessment of their reading skills; a second is the examination of students' skills of written expression. Authors of several recent reports concerning the status of the American educational enterprise emphasize the importance of communication skills (Boyer, 1983; Goodlad, 1984; Sizer, 1984).

Dr. Ernest Boyer, President of the Carnegie Foundation for the Advancement of Teaching has expressed this concern for the development of communication skills. Dr. Boyer views the mastery of English as the primary, most essential goal of education and the teaching of clear written expression as the central objective of the secondary school. He writes (1983, p. 85):

> The first curriculum priority is language. Our use of complex symbols separates human beings from all other forms of life. Language provides the connecting tissue that binds society together, allowing us to express feelings and ideas, and powerfully influence the attitudes of others. It is the most essential tool for learning. We recommend that high schools help all students develop the capacity to think critically and communicate effectively through the written and spoken word.

Development and assessment at-risk students' skills of communication is a concern of teachers and researchers. The assessment of at-risk students' reading skills is one component of evaluation to be conducted as part of this study.

<u>Testing Students' Reading Skills</u>

Walter Hill (1970) describes the major functions of "regular school achievement assessment" in reading as follows:

> The major functions served are: to use on-going testing program results as first screening of reading difficulty cases, to compare individual and group reading progress with other skills progress, to provide gross measurements

of achievement-capacity differentials (p. 149).

An Oregon State Department of Education document (1977) also describes the purpose of norm-referenced reading tests:

> Usually norm-referenced (standardized) reading tests are used when the purpose of evaluation is to compare the achievement of students in one program, class, district, or state with another group of students (p. 23).

Although widely used at both the elementary and secondary school levels, reading achievement tests have received continued criticism.

Collins and Haviland, researchers at the Center for the Study of Reading, criticize most reading tests as tests of students' background knowledge (1979). This issue has been discussed earlier as part of the problem of passage dependence. Therefore, they correlate highly with IQ tests, which also measure background knowledge.

Collins and Haviland warn that higher level reading skills are so entwined with background knowledge that the following skills are among those difficult to measure:

> ...the ability to understand the conventions of punctuation and paragraphing, the ability to find specific information in a text, the ability to recognize and recover from wrong hypotheses about the text, and the ability to recognize and use high level text structure" (p.140).

Petrosko (1977), summarizing a study of 352 standardized tests of reading comprehension and 373 standardized vocabulary measures, also cautions test users (p. 25):

> But researchers should be cognizant of special problems with language oriented tests. Probablythe most significant of these is the passage dependence of Reading Comprehension tests. Tuinman (1973-74) found that some items in Reading Comprehension tests are not dependent on the passage of prose they follow. Such items are answered correctly at a higher than chance rate by subjects who do not read the passage with which the items are ostensibly linked. Needless to say, this weakness in measurement

needs to be noted by a prospective test user.

Collins and Haviland suggest the following remedies for the problem of passage dependence: (a) design tests around experiences and motivations common to all children taking the test, (b) construct tests tailored to each student's needs, or (c) attempt to distinguish between children's two major strategies for decoding unknown words (either sounding out the unknown or guessing at the unknown based on its initial letters, context, and background knowledge) and teach them how to alternate these strategies for optimal success (p. 140, 141).

In addition to problems of passage dependence, reading tests have been criticized for their division of the reading task into numerous discrete skills. Debate concerning what is measured on tests of reading has raged for many decades (Lennon, 1970). Perusal of catalogs of reading tests yields the names of as many as seventy or eighty subtests allegedly measuring reading skills and abilities. Yet, one should not assume that reading is composed of an aggregate of these subsets. Lennon, writing in an issue of The Reading Teacher published twenty-five years ago, discusses the issue of analysis of reading ability.

> It is one thing--and a necessary thing--to make a careful analysis of reading ability, to spell out its various supposed components in detail, and to prepare extensive lists or charts of the specific skills or abilities to serve as statements of desired goals or outcomes of the reading program. It is quite another thing to demonstrate that these manifold skills or abilities do, in fact, exist as differentiable characteristics of students; and still a third thing to build tests which are in truth measures of one or another of these skills, and not of some more general, persuasive reading ability (p. 20).

Despite such criticisms and recommendations, most current reading tests remain a series of increasingly difficult passages followed by comprehension

questions. Correct answers are, therefore, frequently a function of a student's background knowledge. It appears that such tests are inherently biased towards those in upper income brackets as a greater number and variety of life experiences builds a greater background knowledge of the world.

Johnston (1984) captures the reasons for this uneasy match between the goals and critique of reading researchers and the tests used to assess student skills. Group administration and ease of testing and scoring have contributed to the current state of testing in reading.

> The major approach to the assessment of reading that is currently in use (largely group silent reading tests) seems to have been the result of an ideological thrust that favored ease of use over all else. The reasoning is probably best captured by Anderson and Dearborn (1952):
>
> If the reader will now ask himself this double-headed question as to (1) just how he is going to find out how much and how well the individual pupils in a class understood or comprehend what they have read silently and (2) just how he is going to make it easy for the teacher or tester to score the findings in terms of age and grade norms, he will come to understand why the tests of silent reading are as and what they are, and why they have so many shortcomings and limitations (p. 301).

In other words, ease of group administration and scoring have remained a powerful incentive for maintenance of these tests.

General problems associated with achievement testing, the composition of reading tests, and the special nature of the school population to be tested combined to make selection of an appropriate reading test a difficult task. Although the students enrolled in the programs to be studied are generally labelled "at-risk," they do not share a uniform set of academic skills. While many of the students lack the academic skills necessary for success in secondary school, some students enrolled in these programs do

achieve at grade level competency and could be college bound. Selection of a reading test which accurately measures the reading skills of such a broad population is difficult.

Criteria for Selection of Reading Test

The following general criteria shaped the selection of an appropriate reading test to administer to at-risk secondary school students in the study:

1. The test may be administered to an individual or a group. Students absent on the date of testing or those who enter a program following the testing date require individual administration of the test.

2. Multiple forms of the test are available. Pretest and posttest forms of the test are planned for use in the fall and spring of two successive years.

3. Time for test administration is brief. Length of testing time is considered significant for two reasons. First, testing time competes with valuable instructional time. Second, educationally at-risk secondary school students are resistant to monitoring of their academic achievement through testing. Lengthy periods of testing do not enhance student motivation or the school environment.

4. Test results are easy to interpret. Students, parents, teachers, and researchers should be able to understand test results.

5. Procedures for test administration, scoring, and interpretation are standardized.

6. The test provides information useful for decision-making at the school level as well as aids in decision-making concerning program effectiveness.

7. Out of grade level testing is possible. Students in the at-risk population vary in their level of academic achievement. Students termed educationally at-risk include both low and high achieving youths. Therefore, students may need to be tested on criteria other than that of the grade level at which one generally takes a particular test.

8. The test provides information concerning student reading comprehension which may be used for varying purposes. Comparisons of student achievement across time and between programs is anticipated.

9. The test controls as much as possible for passage dependence and bias.

The search for an appropriate reading test was conducted via nomination of tests by experts in the areas of reading and testing and measurement, a search of the literature concerning reading tests, and interviews of persons involved in the administration of local, state, and federal programs for educationally at-risk youth. Among the tests considered appropriate for use were the National Assessment of Educational Progress, Test of General Educational Development, the Nelson-Denny Reading Test, the Iowa Test of Educational Development, the California Achievement Test, the Comprehensive Test of Basic Skills and the Degrees of Reading Power Test.

Tests were rejected as appropriate for use on varying grounds. Tests with numerous subtests required for the assessment of students' skills in a particular area, as well as tests with large total administration time were rejected. For example, the Iowa Test of Educational Development was not chosen for use as it has three reading comprehension subtests: social studies, reading, natural sciences reading, and literary materials. Total time for administration of these tests was too lengthy for the purposes and

constraints of this study. Other grounds for eliminating a test form consideration included: a test structure in which a separate form of the test was to be used at each grade level, complicated and difficult to comprehend reporting procedures, and (in the case of NAEP tests) the availability of the most current materials.

Degrees of Reading Power Test

The standardized reading test chosen for administration to secondary school students in the at-risk program study is the Degrees of Reading Power. The test will be administered to students in the fall and spring of 1986-87 and 1987-88.

The Degrees of Reading Power (DRP) has been described by reviewer Michael Kibby (1981) as:

> A term used to describe a specific reading test, a general methodology of testing reading comprehension, and a method of calibrating the difficulty of reading material (p. 416).

The DRP tests for grades three through twelve were developed as a response to the New York State Education Department's call for a reading measure which would define the most difficult materials students could read. Developed by The College Board in concert with the New York State Education Department, the Carnegie Corporation, and Touchstone Applied Science Associates, Inc., it is based upon a Rasch latent trait theory model using the ability of the student test-taker and the difficulty of the test items.

Koslin, Koslin, and Zeno (1979) discuss the problems of current norm-referenced and criterion-referenced tests in reading which led to their research and development of the Degrees of Reading Power test.

Regarding norm-referenced tests:

> Thus, for several reasons, norm-referenced tests represent an unsatisfactory approach to the measurement of effective-

> ness. Clearly they are unsatisfactory as effectiveness measures because they do not directly measure what students have learned in terms of some sought-after (socially validated and valued) outcome. Since norm-referenced achievement tests are designed to discriminate among individuals, they cannot be expected to be sensitive measures of educational outcomes. In addition, it is not always clear what norm-referenced data are actually measuring. Certainly no definition of what is meant by reading with comprehension is stated in conjunction with operational procedures for measurement (p. 313).

Regarding criterion-referenced tests:

> In short, building a model of reading comprehension-a prerequisite to using conventional criterion-referenced tests as effectiveness measures-would require work in at least four areas: (1) identifying and validating all the reading subskills; (2) operationally defining what is meant by reading comprehension; (3) showing that reading with comprehension does not occur in the absence of a subskill and (4) establishing the relationships between subskills and comprehension so that the proper weights and aggregation rules could be assigned (at different stages of reading development (p. 315).

They argue the reading comprehension ability that an adult or secondary school graduate must attain is operationally defined by the "average" measured readability characteristic of materials those persons are expected to read. Therefore, standards of performance and incremental steps toward that level of performance can be developed. Further, they assert that reading comprehension can be defined in terms of an individual's ability to process information in a text, correctly selecting from several alternatives that which fits the accepted meaning of that text.

The authors of the DRP believe this model of prose difficulty makes possible a quantitative definition of reading tasks and a definition of ability in terms of the difficulty of materials which can be read. The independent, instructional, and frustration levels of the reader and the reading level of texts are defined in DRP units.

In a paper produced by the Division of Educational Testing of the New York State Department of Education (1979), the following uses of DRP scores were suggested:

> ...to decide if a student is able to read material required in a course, to match reading materials to a student's ability level, to group students of comparable reading ability, and to measure individual and group gain in reading ability (p. 11).

Scores are reported in DRP units ranging in value from 15 to 99 on a scale derived from the Bormuth Mean Cloze Readability Formula. This formula bases its readability calculations upon four linguistic variables: number of letters, number of words, number of words on the Dale Long List from the Dale-Chall formula and number of sentences. Three reading levels are indicated for students based upon their cumulative DRP units: the independent level, that level of text which can be read by students with pleasure or can correctly answer 90% of the comprehension questions; the instructional level, that level of text which can be read with instructional help or can correctly answer 75% of the comprehension questions; and the frustration level, that level of text which the student is unlikely to comprehend or can answer only 50% of the comprehension questions.

Fitzgerald (1979) calls this instrument a form of meaning-based testing which may influence the way reading is taught.

> In sum, the introduction of a comprehension measure with a meaning-base should influence the curriculum offerings. A language centered, active processing mode of instruction would appear more responsive to the task defined by the DRP. Much new information has been generated over the past decade concerning our understanding of the comprehension process, but much still remains unclear. What is clear is that the instruction within the classroom has not kept pace with this new information. The introduction of a new testing instrument may help to reduce this gap (p. 8).

DRP tests are either hand scored or mailed to The College Board for

scoring and reporting services. These reports can include frequency, cumulative frequency, percentage, and cumulative percentage distributions for intervals of 5 DRP units. National percentile ranges, normal curve equivalents, and local norms may also be requested and provided (The College Board, 1985). Individual report forms concerning student performance are exceptionally clear and helpful in understanding a student's reading levels.

Each of the five forms (for grades 3-5, 5-7, 7-9, 9-12, and 12-14) of the DRP test are available in two series, PA and PB, and contain a series of non-technical, nonfiction, 325-word passages arranged in order of difficulty. Each multi-paragraph passage contains seven test items, one deletion per two to seven sentences in the passage. Students select the replacement word from five choices which appear to the right of the sentence in which the deletion occurs.

The Division of Educational Testing of the New York State Education Department and Touchstone Applied Science Associates also report the average readability (in DRP units) of across the curriculum materials for grades four through twelve in an annual Readability Report (Kibby, p. 418, 419). Thus, teachers can easily match their students' reading strengths to an appropriate instructional level and textbook. A microcomputer software package called MicRA--DRP is also available to determine the DRP units of other reading materials.

A New York State Education Department document (1979) explains the manner in which a student's score in DRP units is used to select appropriate instructional texts.

> For example, assume that a particular student answered 48-items correctly on the DRP Field Trial Form A. By referring to Figure 1 it can be seen that this student has a .90 (independent level) likelihood of success in

> reading materials with a DRP unit value of 39. Materials in which selections are at a level around 39 DRP Units can therefore be used profitably by the student for independent reading. At the instructional level the `lower band' and `upper band' denote a range of difficulty. Tests that have a majority of selections within the 47-55 range can therefore be used instructionally with this same student, with a .75 likelihood of success. The further the materials fall toward the upper bound of his or her instructional range (55), the more assistance will be required for the student to use these materials profitably (p.11).

These tests do not have test-taking time limits, but The College Board (1985) suggests "sufficient time" be allotted for test-taking and that one class period is adequate.

Kibby (1985) notes that the initial studies establishing the validity and reliability of the DRP were conducted with 5,000 fourth and sixth graders in New York State. Kibby points out that 47% of the students in the initial study were non-white and 40% were from homes of lower socioeconomic status, and that communities with less than a 20% minority population were excluded.

While he points out a number of potential problems, such as that mentioned above, he concludes that:

> I am greatly impressed by the DRP methodology. It has a sound rationale, is reasonably well constructed, provides information that has significance in the world of textbooks and the classroom, and has potential for further development and additional uses (p. 427).

The level of the test selected for administration to students in the at-risk study is the PA-4 and PB-4 grade 7-9 form. This level of the test provides a broad band of material for the test-taker, beginning with very simple passages and ranging upward to sophisticated reading material. It is designed to capture the broad spectrum of reading skill which students in the study possess. Although another form of the test for grades 9-12 is

available, the grade 7-9 form appears more responsive to the varied reading abilities of students in the at-risk study. The grade 7-9 form of the DRP provides adequate samples of reading passages at both ends of the skill spectrum, low and high. Therefore, it appears the more appropriate choice for testing the reading skills of at-risk students.

The Degrees of Reading Power test responds well to the criteria for test selection explored earlier in this paper. The Degrees of Reading Power is a criterion-referenced measure which may be administered to an individual or group. There are multiple forms of the test available, test administration time is brief, and test results are standardized and easy to interpret. Out of grade level testing is possible and the test controls as much as possible for paragraph dependence and bias. It also provides information concerning reading comprehension which may be used for varying purposes, at the classroom, school, and district level.

Researchers in the at-risk project realize that the administration of a single test cannot yield all of the data needed to properly evaluate any reading program.

Farr (1970) emphasizes the importance of a comprehensive effort of evaluation of a reading program. He stresses that while group administration of a standardized test is frequently viewed as a program of evaluation, other valuable components are necessary.

> In addition, continuous evaluation of teaching procedures, instructional materials, curriculum oranization, and the objectives of the program must be planned as an integral part of the total evaluation program (p. v).

This effort to evaluate students' reading skills is viewed as a single measure of students' ability to comprehend reading material similar to that they encounter during the school day. The use of a single reading test is

not a comprehensive effort of reading program evaluation, but a part of a broader effort to assess program impact on educationally at-risk secondary school youths.

Mathematics Testing

The original plan for this project included the selection of both a standardized reading test and a standardized mathematics test for administration to at-risk secondary school students. However, research into the problem of providing an adequate fit between the goals of data gathering, the varied skills of the at-risk secondary school students, the varied goals of their teachers, and the current state of research and testing in mathematics did not yield an appropriate testing instrument.

Such a mathematics test would need to fulfill the following requirements:

a. Provide scores to compare students in varied programs of educational intervention. Most of these programs provide students with assistance in improving mathematics skills; however, different programs address the problem with different objectives and materials.

b. Provide a means of testing secondary students whose level of skill and needs for instruction range from basic arithmetic to geometry and algebra. At-risk students are not all low academic achievers. Youths enrolled in special programs for at-risk secondary school students vary in the reasons for their placement. Enrollment in a special program may result from truancy and other disciplinary problems, low achievement and course failure, drug or alcohol dependency, or pregnancy (Wehlage, 1986).

c. Teachers in special programs for at-risk students also have varied needs for information concerning their students' performance in mathematics.

Teachers in programs for at-risk students already provide testing based upon individual program goals. In some cases, teachers must administer state mandated competency or achievement tests as well as tests designed to inform their particular mathematics curriculum. A third level of testing administered by researchers is neither desirable in terms of time spent in testing nor valuable to inform comparisons of program effectiveness (Romberg, 1986; Taylor, 1979).

Researchers in mathematics, testing and measurement criticize the use of standardized tests on numerous grounds.

Romberg, in a paper prepared for the 1986 National Conference on the Influence of Testing in Mathematics Education, argues for new assessment procedures in mathematics. He describes and critiques the use of norm-referenced standardized tests, profile achievement tests, and objective or criterion-referenced tests.

Romberg discusses four features of norm-referenced standardized tests. First, he notes these tests are designed to rank individuals on a single, uni-dimensional trait and cautions that the derived score resulting from such testing is not a direct measure of that trait.

> It is as if one were measuring the Houston Rocket's basketball star Ralph Sampson's height but not reporting that he is 7'4. Rather what is reported is that he is at the 99th percentile for American men. Furthermore, for standardized tests there is no theoretically single trait (like height) that is being assessed (p. 15).

He further explains that individuals' high or low scores from such tests are simply outcomes of the comparison of individual scores with a norm population. Therefore, no assignment of "good" or "bad" referents should be made to any individual's score in relation to the trait.

Romberg's third point concerning norm-referenced tests is the

assumption of item equivalence. Items on these tests can neither be assumed equivalent to one another nor representative of well-defined domain.

Last, he reminds us that the validity of such tests is predictive validity; that is, the tests predict an individual's performance on some future task. An example of predictive validity is that SAT test results are reasonable predictions of a person's future college performance.

Romberg contrasts the norm-referenced standardized test with profile achievement tests in which mathematics content topics are crossed with hypothesized cognitive levels in grade level matrices. Profile tests do not assume one underlying trait, but are based on the premise that mathematics instruction at any grade level concerns numerous topics. Profile tests are designed to sample the performance of a group as opposed to an individual and their validity is based upon content or curriculum validity.

While useful for providing data about groups, profile tests are not useful for providing information concerning individuals as each student takes only a sample of items. Profile tests are more expensive to develop, harder to administer and score, and difficult to interpret. Romberg also questions the theoretical assumptions upon which they are based.

Objective or criterion-referenced tests are given to individuals at the end of a unit of instruction. Satisfactory performance levels are generally pre-specified via percentile of correct answers. Although Romberg believes these tests can be useful (to ascertain whether a concept or skill has been learned), and are easily scored and interpreted, he finds three weaknesses in their use. They are expensive to develop; aggregation across objectives is not reasonable; and the objectives are assumed independent rather than interdependent. Further, test items which measure higher level problem

solving are difficult to develop for such tests.

Romberg summarized his concerns regarding norm-referenced standardized tests, profile achievement tests, and objective or criterion-referenced tests.

> The main point to be made is that while these tests have been useful for some purposes and undoubtedly will continue to be used, they are products of an earlier era in educational thought. Like the Model T Ford assembly line, objective tests were considered as an example of the application of modern scientific techniques in the 1920's. Today we ought to be able to do something better (p. 27).

Critics of standardized mathematics tests discourage their use and value for numerous reasons, including measurement of a specific objective or instructional aim, provision of information concerning the efficacy of an entire program diagnosis of an individual's strengths and progress toward educational goals, or program evaluation.

Taylor (1979) discusses a National Council of Supervisors of Mathematics Position Paper on Basic Mathematical Skills concerning the effectiveness of standardized testing. The National Council of Supervisors of Mathematics report emphasized that educators and the public have accepted and placed too much faith in standardized tests. The council acknowledged that standardized tests do provide comparative data which allow the rank ordering of individuals, schools, and districts, but have the following limitations: they are not necessarily developed to measure specific instructional objectives and measure only a sample of any program's content. The National Council of Supervisors of Mathematics report concluded that because standardized tests do not provide enough information about how much mathematics a student knows, they are insufficient devices for reporting individual growth in mathematics skill.

Taylor (1979) presents further support for the view that norm-referenced standardized tests are inappropriate for both individual student assessment and program evaluation. He quotes the 1975 report of the National Committee on Mathematical Education (appointed by the Conference Board of the Mathematical Sciences in 1974). This report indicates particular concern for the lack of diagnostic power of standardized tests. The National Committee on Mathematical Education reported concern that standardized test scores represent average performance in a number of cognitive levels in a range of content areas. Therefore, an individual student's relative strengths and weaknesses are not determined. The standardized test, then, does not aid the evaluation of students' progress toward their educational goals.

Another concern echoed throughout the literature on standardized testing is that what can be tested will be tested, that those skills or pieces of knowledge which are testable via easily administered and scored multiple-choice items are those which receive our teaching as well as our testing attention (Eisner, 1985; Frederiksen, 1984; Taylor, 1979; Kilpatrick, 1979). Testing in mathematics, then, tends to focus on low level skills. Taylor (p. 106) outlines this problem:

> In basic skills testing there is a danger of focusing on isolated low-level skills and neglecting to determine if these skills can be effectively combined to solve problems.

Frederiksen (1984) does not disparage the concern for improvement in basic skills, but questions the reliance on objective tests for evidence of improvement. Further, he fears reliance on such tests does not serve higher order thinking and problem-solving skills.

> Improvement in basic skills is of course much to be

> desired, and the use of tests to achieve that outcome is not to be condemned. My concern, however, is that reliance on objective tests to provide evidence of improvement may have contributed to a bias in education that decreases effort to teach other important abilities that are difficult to measure in the multiple choice tests. A recent NAEP report suggests that there is such bias (p. 195).

Frederiksen reports on National Assessment of Educational Progress data which support his concerns regarding testing. NAEP data show that test items in mathematics, reading, writing, and science indicate performance in "basic skills" is not declining. Rather, Frederiksen writes, performance on items reflective of more complex cognitive processes is declining. For example, he reports 1982 NAEP data concerning mathematics performance shows 90% of seventeen year olds could perform simple addition and subtraction, but performance on problems regarding mathematical principles fell from 62% to 58% and on problem solving from 33% to 29%. Frederiksen is concerned with the focus on the low level, testable concepts which abound in mathematics tests.

While critics of standardized mathematics tests differ in the focus of their concerns, the consensus of expert opinion appears to be a call for new ways to assess students' skills in mathematics. This seems particularly important for at-risk secondary school students who often come to the testing situation with low motivation and a history of low test scores. Their teachers also question the utility of placing students in such a situation. All too often such tests, while mandated by local, state, or federal authorities, do not inform the curriculum. Rather, they provide further experiences in failure for many students.

While mathematics testing will not be conducted by researchers in the at-risk project, data concerning students' mathematics achievement will be

gathered at each of the study sites. This material will include scores of tests administered by local school personnel, interviews of teachers and students concerning mathematics instruction and achievement, and observation of mathematics instruction. The focus of this inquiry is the local program in mathematics instruction and its impact about student achievement.

Collection of Writing Samples

Clear and effective writing is an important communication skill. The paucity of writing conducted in American secondary schools is a continuing concern of researchers, teachers, employers, and the general public (Applebee, 1981). Such concern is evident in the formation of the Commission on Writing by the Council for Basic Education and the subsequent publication of its findings in Empty Pages: A Search for Writing Competence in School and Society (1979). The commissioners initiate their discussion of teaching, reading, and writing with twelve assumptions concerning the role of communication in society.

1. The life of any culture rests on that rock-bottom device of social bonding, language.

2. Therefore, the teaching and learning of the language should have as their ultimate goal (in addition to more immediate aims) the continued health and improvement of the culture.

3. One way to achieve this healthy state, as well as to effect the improvement, is to liberate the intelligence of citizens by insuring that they have the ability to read and write.

4. The liberation of the intelligence should not be confused with `socialization,' `acculturation,' self-expression,' or the `search for identity.' The teaching of language, and notably of writing, should not be subordinated to purely private purposes, let alone fleeting trends or fashions. It should be anchored in the best means of expression so far attained by our culture.

5. English teachers are primarily the best means we have of transmitting language skills. They are not, or should not be, primarily entertainers, welfare workers, group therapists, priest-parson-rabbi surrogates, librarians,, or sitters. Even if not primarily so, all teachers (and even administrators) are or should be teachers of English and therefore, to some degree of writing.

6. The job (insofar as they do not do these things for themselves) is to teach the students to talk, think, read, and write in the language known as Standard English. Oral Standard English and written Standard English may differ, but the differences between them are less marked than those distinguishing the accepted language from ethnic, dialectical, jargon, or vogue English.

7. Writing is inseparable from thinking, reading, speaking, listening, and studying. Though it has its own norms and uses its own pedagogy, it is part of a circle of connected activities.

8. Since reading and writing are intimately connected, learning to write depends on exposure to high-quality reading material.

9. Teachers themselves must have learned and must continue to practice writing. This obligation rests on them as it does upon the student.

10. Achieving some competence in writing is both the right and responsibility of all members of a democracy. We cannot afford to reproduce in the domain of literacy the Two Nations--the Poor and the Rich--identified by Disraeli in the domain of property.

11. Clear and effective writing is not simply a skill or a socioeconomic advantage. Because it expresses the integrity (or dishonesty) of an intellectual process, it is a moral activity.

12. Finally, we believe that every normal Jane and Johnny can, if properly taught, learn how to write clearly, competently, and correctly (p. 2, 3).

The authors of the National Assessment of Educational Progress report titled "Writing Objectives: 1983-84 Assessment" argue that writing contributes to students' subject knowledge and self knowledge.

> Students need to understand that writing, like talking, composes and expresses our thoughts while providing a record of our thinking that can be reflected on, developed and changed. The act of writing can help students review and refine the ideas presented in textbooks and class discussion. Even more important, students can use exploratory writing to test their understandings of new concepts and principles and to participate in new ways of thinking. In this way, writing makes learning participatory, rather than passive.
>
> Writing is also a powerful tool for self discovery and self expression. Through letters, diaries and free-writing exercises, people can clarify for themselves and others what they think and believe. Writing enables them to express their emotions in a concrete form and then stand back, as a more detached observer might, to grasp more fully what they feel and why. Thus, writing offers a special opportunity to focus, analyze and understand our thoughts and feelings (p. 6, 7).

Such statements concerning the importance of writing highlight the need for information concerning the written communication skills of at-risk secondary school students.

Effective skills of written communication contribute to success in school, employment, and general life tasks. Therefore, the at-risk project plans a program of data collection (in the fall and spring of 1986-87 and 1987-88) concerning students' writing skills. Writing samples will be collected from all students enrolled in one of the at-risk project study sites. The National Assessment of Educational Progress design for the collection of data concerning students' writing skills is the model for this effort.

Although the NAEP studies gathered data (in the school years 1974, 1979, and 1984) on three types of writing tasks: informative, persuasive, and imaginative, the at-risk project will collect data only on students' abilities to conduct persuasive writing. A focus on one type of writing is a function of both time and ease of administration. Teachers in these

programs already relinquish valuable time to local and state testing and are requested to administer a reading test as well as collect writing samples for the at-risk project. The varied size, organization, and schedules of the nine project sites make the collection of data concerning three writing types a difficult, if not impossible task. A persuasive writing task is selected as a focus because of its real life utility. Applying for a job, arguing for lower taxes, better law enforcement, or increased day-care facilities, protesting job discrimination or requesting an adjustment on a miscalculated bill, all require skills of written persuasion.

The 1986 NAEP document describing decade long trends in writing defines the persuasive task as an attempt to "...bring about some action or change...its aim is to influence others" (p. 19).

Two kinds of scoring will be conducted on these papers. Primary trait and holistic scoring of each paper will be conducted once per year by members of the Wisconsin Writing Project on the campus of the University of Wisconsin-Madison. The NAEP document (1986) "NAEP-Writing Trends Across the Decade, 1974-1984" describes these scoring methods (as used in the assessment).

> One set of analyses is based on primary trait scoring and focuses on the writers' effectiveness in accomplishing the particular task that was set; it is sensitive to the writers' understanding of audience, as well as to the inclusion of specific features necessary to accomplish the purpose of each informative, persuasive, or imaginative writing task. The other set of analyses is based on general impression or holistic scoring and focuses on the writers' overall fluency in responding to each particular writing task; it is sensitive to a range of different skills, including organization, quality of content, grammar and usage, spelling, punctuation, and choice of words to express an idea.

The primary trait scores for the persuasive writing task (in the NAEP studies) were determined via assignment of a paper to one of the following

four categories:

a. **unsatisfactory** - the writer fails to take a stand or does not support their statements;

b. **minimal** - the writer clearly takes a stand and supports their statements with one or more appropriate reasons consistent with a viewpoint, or embeds an important supporting statement within a number of less important reasons;

c. **adequate** - the writer takes a clear stand and supports it with a brief argument or several interrelated reasons;

d. **elaborated** - papers in which the writer pursues an extended argument or a list of interrelated reasons to support their argument.

Wisconsin Writing Project evaluators will also use these categories. Each paper will also be given a holistic rating. The method of evaluation called "holistic scoring" is described in a recent text (1981) concerning The Evaluation of Composition Instruction.

> In the evaluation of student writing, the `holistic assessment method' yields an overall judgment of quality without recommendations for improvement, or indeed without any identification of the exact strengths or weaknesses of the performance are. Since a holistic assessment can be achieved fairly reliably and rather quickly (and hence inexpensively), it is quite widely used and is very valuable in determining whether there has been an improvement in the overall quality of composition in a particular class or school. Such data are exactly what one needs for evaluation of programs for improving student writing (p. 9).

Teachers who have received training in these scoring techniques as part of their participation at the Wisconsin Writing Project, one site of the National Writing Project, will meet in the spring of 1987 and 1988 to score papers from the fall and spring samples of each year.

Portfolios of writing will also be collected at each program site.

These will include samples of students' written work completed as a naturally occurring part of their curriculum.

In addition to the collection and scoring of persuasive writing samples, researchers visiting each site will attempt to gather data concerning the following questions:

1. How often are students in special programs asked to write paragraph length or longer products?

2. How much time is spent on any individual writing assignment?

3. Are prewriting or planning activities incorporated into assignments?

4. Are opportunities for revision offered to students?

5. Do teachers' evaluations of papers include feedback other than letter or number grades?

6. For example, are comments and/or suggestions for improvement included as part of the evaluation of written work?

7. Do teachers hold small group or individual conferences with students concerning their writing?

This combination of methods of data gathering is designed to inform researchers and teachers concerning at-risk program impact on students' writing skills.

Conclusion

Data collected from scoring of the Degrees of Reading Power test and evaluation of writing samples, in conjunction with direct program observation, interviews, and surveys, will be used in three ways. First, information gathered from scoring of tests and writing samples will help inform program designers, teachers, and administrators concerning the

effects of special programs on at-risk students. Second, the strengths and weaknesses of individual students and the range of skills of students within a program will be highlighted. Further, this information will allow researchers to compare and contrast the elements of various special programs under study.

The administration of the Degrees of Reading Power test and the collection and scoring of writing samples are designed as part of a multi-dimensional effort to refine effective secondary programs for educationally at-risk youth.

Wehlage (1986b) states that educators and policymakers have developed three major responses to at-risk students: early identification, special programming in high school, and systemic changes in the process of schooling. Wehlage argues that whatever course of action or combination of actions are attempted to remediate the dropout problem, researchers need to identify, describe, and evaluate special program components. The at-risk project attempts to meet this challenge through multiple efforts at data collection. Increasing numbers of youths are dropping out of American secondary schools. Wehlage writes (p. 4):

> To avert this human tragedy, schools will need to devise new strategies and responses that are constructive with respect to this group of students. Both specific strategies that are designed for students with special problems as well as more general and systemic changes to make high schools more effective and engaging with at-risk students will be required.

Through multiple strategies of information gathering, researchers hope to identify, describe, evaluate, and inform others of the most effective tools to stem the tide of secondary school dropouts.

References

Airasian, P. W., & Madaus, G. F. (1983, Summer). Linking testing and instruction: Policy issues. Journal of Educational Measurement, 20, 103-118.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1986). Writing trends across the decade, 1974-84. Washington, DC: National Assessment of Educational Progress.

Berlak, H. (1986). Testing in a democracy. Educational Leadership, 2, 16-17.

Boyer, E. L. (1983). High school: A report on secondary education in America. New York: Harper and Row Publishers.

Brunig, R. (1985). Review of degrees of reading power test. In J. Mitchell (Ed.), The Ninth Mental Measurements Yearbook, I, 443-444. Lincoln, NE: Buros Institute of Mental Measurement.

Casteen, J. T. (1984). The public stake in proper test use. In C. W. Daves (Ed.), The uses and misuses of tests, (pp.). San Francisco: Jossey-Bass Publishers.

Collins, A., & Haviland, S. (1975). Children's reading problems. In R. U. Taylor, & S. H. White (Eds.), Testing, teaching, and learning (pp. 136-145). Washington, DC: National Institute of Education.

Curtis, M. E., & Glaser, R. (1983, Summer). Reading theory and the assessment of reading achievement. Journal of Educational Measurement, 20, 133-147.

Daley, T. T. (1976). The student and testing. In Testing and the public interest: Proceedings of the 1976 ETS invitational conference, (pp. 55-64). Princeton, NJ: Educational Testing Service.

Davis, B. G., Scriven, M., & Thomas, S. (1981). The evaluation of composition instruction. Inverness, CA: Edgepress.

Eisner, E. M. (1985). The art of educational evaluation. London: The Falmer Press.

Fadiman, C., & Howard, J. (1979). Empty pages: A search for writing competence in school and society. Belmont, CA: Signature Books.

Farr, R. (1970). Preface. In R. Farr (Ed.), Measurement and evaluation of reading (pp. v-vi). New York: Harcourt, Brace, & World, Inc.

Fitzgerald, T. P. (1979, April). Meaning-based testing and curriculum implications. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39(3), 193-202.

Goodlad, J. I. (1984). A place called school: Prospects for the future. New York: McGraw-Hill Book Company.

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

Haney, W. (1986). Making testing more educational. Educational Leadership, 43(2), 4-15.

Haney, W. (1984). Testing reasoning and reasoning about testing. Review of Educational Research, 4, 597-654.

Harris, A. J., & Sipay, E. R. (1985). How to increase reading ability: A guide to developmental and remedial methods. New York: London.

Hein, G. E. Standardized testing: Reform is not enough. (1975). In M. D. Cohen (Ed.), Testing and evaluating new views (pp. 27-30). Washington, DC: Association for Childhood Education International.

Hill, W. (1970). Evaluating secondary reading. In R. Farr (Ed.), Measurement and evaluation of reading (pp. 126-153). New York: Harcourt, Brace, & World, Inc.

Johnston, P. (1984). Assessment in reading. In P. D. Pearson (Ed.), Handbook of reading research (pp. 147-182). New York: Longman Books.

Kibbey, M. (1981). Test review: The degrees of reading power. Journal of Reading, 24(5), 416-427.

Kilpatrick, J. (1979). Methods and results of evaluation with respect to mathematics education. In New trends mathematics teaching: Vol. 4. Paris: UNESCO.

Koslin, B. L., Koslin, S., & Zeno, S. (1975). Towards an effectiveness measure in reading. In R. W. Taylor and S. H. White (Eds.), Testing, teaching, and learning (pp. 311-334). Washington, DC: National Institute of Education.

Lennon, R. T. (1970). What can be measured? In R. Farr (Ed.), Measurement and evaluation of reading (pp. 18-34). New York: Harcourt, Brace, & World, Inc.

National Coalition of Advocates for Students. (1985). Barriers to excellence: Our children at-risk. Boston, MA: Author.

New York State Education Department (Division of Educational Testing). (1979). Degrees of reading power: Description of a new kind of reading test and its related technology (ERIC Document Reproduction Service No. ED 170 712).

Oregon State Department of Education. (1977). Reading in the secondary school (ERIC Document Reproduction Service No. ED 145 366).

Petrosko, J. M., & Hufano, L. (1975). An assessment of the quality of high school mathematics tests. Paper presented at the meeting of the National Council on Measurement in Education, Washington, DC.

Romberg, T. A. (1986, June). Measures of mathematical achievement: Problems and influences. Paper presented at the National Conference on the Influences of Testing on Mathematics Education, Los Angles, CA.

Sarason, I. G. (1983). Understanding and modifying test anxiety. In S. B. Andersen, & J. S. Helmick (Eds.), On educational testing (pp. 133-149). San Francisco: Jossey-Bass Publishers.

Sizer, T. R. (1984). Horace's compromise: The dilemma of the American high school. Boston: Houghton Mifflin Co.

Stodolsky, S. S. (1975). What tests do and don't do. In M. D. Cohen (Ed.), Testing and evaluation: New views (pp. 13-17). Washington, DC: Association for Childhood Education International.

Taylor, R. (1975). Mathematics testing: A view from the schools. In R. W. Taylor, & S. H. White (Eds.), Testing, teaching and learning (pp. 98-111). Washington, DC: National Institute of Education.

Tyler, R. W. (1974). The use of tests in measuring the effectiveness of educational programs, methods, and instructional materials. In R. W. Tyler, & R. M. Wolf (Eds.), Crucial issues in testing (pp. 143-155). Berkeley, CA: McCutchan Publishing Corporation.

Tyler, R. W., & White, S. (1975). Chairmen's report. In R. W. Tyler & S. H. White (Eds.), Testing, teaching, and learning (pp. 1-6). Washington, DC: National Institute of Education.

Wehlage, G. G. (1983). The marginal high school student: Defining the problem and searching for policy. Children and Youth Services Review, 5(4), 321-342.

Wehlage, G. G. (1986). At-risk students and the need for high school reform. Madison, WI: National Center on Effective Secondary Schools.

Wehlage, G. G. (1986). Programs for at-risk students: A research agenda. Unpublished manuscript, University of Wisconsin, National Center on Effective Secondary Schools.