DOCUMENT RESUME

ED 369 978 CE 066 455

AUTHOR Venezky, Richard L.; And Others
TITLE Measuring Gain in Adult Literacy Programs.
INSTITUTION National Center on Adult Literacy, Philadelphia, PA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO NCAL-TR93-12
PUB DATE Apr 94
CONTRACT R117Q0003
NOTE 52p.; A joint project of the New York State Education Department, the University of Delaware, and the Adult and Continuing Education Program of the White Plains, NY Public Schools.
AVAILABLE FROM National Center on Adult Literacy, Dissemination/Publications, 3910 Chestnut Street, Philadelphia, PA 19104-3111 (order no. TR93-12: $8 check or money order payable to "Kinko's Copy Center").
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Gains; Achievement Rating; Adult Basic Education; *Adult Literacy; *Adult Programs; Computer Simulation; Educational Research; *Literacy Education; Reading Achievement; *Student Evaluation

ABSTRACT

Problems in the measurement of gain in adult literacy programs were investigated through repeated testing of a group of students in adult basic education and General Educational Development classes and through computer simulations. Ninety-two students were tested at 3 different times over 7 months with a battery of norm-referenced reading and mathematics tests as well as with tests of reading rate and decoding developed especially for this study. Gain scores were found to vary across tests, with significant declines as well as gains. No significant differences in gains were found for amount of instructional time or for attendance rate, and a large amount of group heterogeneity was revealed through an analysis of growth patterns. Computer simulations for grade-equivalent stability showed that with populations smaller than 200, inconsistencies in grade-level intervals could account for a major portion of the yearly gain typically reported for adult literacy instruction. In contrast, simulations of regression to the mean caused by guessing on multiple-choice tests showed that this effect was relatively small. These results strongly supported the need to construct a multiple indicator system for evaluating adult literacy programs, a system that attends to the multiple goals of such programs and is free of elementary and secondary level conventions such as grade-equivalent scores. (Appended are 12 tables and 3 figures.) Contains 29 references.

NATIONAL CENTER ON ADULT LITERACY

MEASURING GAIN IN ADULT LITERACY PROGRAMS

Richard L. Venezky
Page S. Bristow
John P. Sabatini
University of Delaware

A Joint Project of the New York State Education Department, the University of Delaware, and the Adult and Continuing Education Program of the White Plains, NY Public Schools

NCAL TECHNICAL REPORT TR93-12
APRIL 1994

MEASURING GAIN IN ADULT LITERACY PROGRAMS

Richard L. Venezky
Page S. Bristow
John P. Sabatini
University of Delaware

A Joint Project of the New York State Education Department, the University of Delaware, and the Adult and Continuing Education Program of the White Plains, NY Public Schools

NCAL TECHNICAL REPORT TR93-12
APRIL 1994

This work was supported by funding from the National Center on Adult Literacy at the University of Pennsylvania, which is part of the Education Research and Development Center Program (Grant No. R117Q0003) as administered by the Office of Educational Research and Improvement, U.S. Department of Education, in cooperation with the Departments of Labor and Health and Human Services. The findings and opinions expressed here do not necessarily reflect the position or policies of the National Center on Adult Literacy, the Office of Educational Research and Improvement, or the U.S. Department of Education.

NATIONAL CENTER ON ADULT LITERACY
UNIVERSITY OF PENNSYLVANIA
3910 CHESTNUT STREET
PHILADELPHIA, PA 19104-3111
PHONE (215) 898-2100 FAX (215) 898-9804

The National Center on Adult Literacy (NCAL) was established in 1990 by the U.S. Department of Education, with co-funding from the Departments of Labor and Health and Human Services. The mission of NCAL addresses three primary challenges: (a) to enhance the knowledge base about adult literacy; (b) to improve the quality of research and development in the field; and (c) to ensure a strong, two-way relationship between research and practice. Through applied research and development and dissemination of the results to researchers, policymakers, and practitioners, NCAL seeks to improve the quality of adult literacy programs and services on a nationwide basis. NCAL serves as a major operating unit of the Literacy Research Center at the University of Pennsylvania.

NCAL publications to date include:

May 1992 *Matching Literacy Testing With Social Policy: What are the Alternatives?*
Richard L. Venezky (*PB92-1, 7 pages)*

Oct 1992 *Life-Span and Life-Space Literacy: Research and Policy in National and International Perspectives*
Daniel A. Wagner (*OP92-1, 15 pages)*

Oct 1992 *Expanding Theories of Adult Literacy Participation,*
Karen Reed Wikelund, Stephen Reder, Sylvia Hart-Landsberg (*TR92-1, 30 pages)*

Oct 1992 *Invitations to Inquiry: Rethinking Staff Development in Adult Literacy Education*
Susan L. Lytle, Alisa Belzer, Rebecca Reumann (*TR92-2, 44 pages)*

Dec 1992 *Developing the Professional Workforce for Adult Literacy Education*
Susan L. Lytle, Alisa Belzer, Rebecca Reumann (*PB92-2, 11 pages)*

Jan 1993 *The Impact of BIB-Spiralling Induced Missing Data Patterns on Goodness-of-Fit Tests in Factor Analysis*
David Kaplan (*OP93-1, 18 pages)*

Mar 1993 *The Impact of Workplace Literacy Programs: A New Model for Evaluation of Workplace Literacy Programs*
Larry Mikulecky, Paul Lloyd (*TR93-2, 180 pages)*

Mar 1993 *Literacy and Machines: An Overview of the Use of Technology in Adult Literacy Programs*
Terilyn C. Turner (*TR93-3, 86 pages)*

Jun 1993 *Myths and Misconceptions in Adult Literacy: A Research and Development Perspective*
Daniel A. Wagner (*PB93-1, 10 pages)*

Jun 1993 *Literacy and Development: Rationales, Assessment, and Innovation*
Daniel A. Wagner (*IP93-1, 50 pages)*

Jun 1993 *Early Childhood, Family, and Health Issues in Literacy: International Perspectives*
Laurel D. Puchner (*IP93-2, 45 pages)*

Sep 1993 *What Makes Worker Learn? The Role of Incentives in Workplace Education and Training*
Donald Hirsch, Daniel A. Wagner, ed. (*IP93-3, 243 pages)*

Sep 1993 *Prison Literacy: Implications for Program and Assessment Policy*
Anabel Newman, Warren Lewis, Carolyn Beverstock (*TR93-1, 219 pages)*

Sep 1993 *Management Information Systems in Adult Education: Perspectives from the States and from Local Programs*
Mark A. Kutner, Lenore Webb, Rebecca Herman, Pelavin Associates, Inc. (*TR93-4, 150 pages)*

Sep 1993 *What Can Employers Assume about the Literacy Skills of GED Graduates?*
David Kaplan, Richard L. Venezky (*TR93-5, 45 pages)*

NCAL publications to date (continued)

Sep 1993 *Should Reading-Disabled Adults Be Distinguished From Other Adults Seeking Literacy Instruction? A Review of Theory and Research*
Anne E. Fowler, Hollis S. Scarborough *(TR93-7, 101 pages)*

Sep 1993 *When Less Is More: A Comparative Analysis forPlacing Students in Adult Literacy Classes*
Richard L. Venezky, Page S. Bristow, John P. Sabatini *(TR93-8, 46 pages)*

Sep 1993 *Metacognitive Aspects of Adult Literacy*
Scott G. Paris, Andrea Parecki *(TR93-9, 44 pages)*

Nov 1993 *Teamwork and Literacy: Learning from a Skills-Poor Position*
Sylvia Hart-Landsberg, Steve Reder *(TR93-6, 63 pages)*

Nov 1993 *Motivations for Learning: Voices of Women Welfare Reform Participant*
Karen Wikelund *(TR93-10, 54 pages)*

Nov 1993 *Initiating Practitioner Inquiry: Adult Literacy Teachers, Tutors, and Administrators Research Their Practice*
Susan L. Lytle, Alisa Belzer, Rebecca Reumann *(TR93-11, 69 pages)*

Nov 1993 *Coalition Building for Adult Literacy: Historical and Organizational Perspectives*
Anabel P. Newman, Bernadette Lehman *(TR93-13, 68 pages)*

Nov 1993 *Effective Service Delivery in Adult Literacy Programs: A Policy Review and Recommendations*
Judith Ann Koloski *(TR93-14, 46 pages)*

Dec 1993 *Adult Literacy Training and the Integration of Human Services*
Elizabeth R. Reisner *(TR93-16, 30 pages)*

Dec 1993 *Issues and Challenges in Adult Numeracy*
Iddo Gal *(TR93-15, 62 pages)*

Apr 1994 *Measuring Gain in Adult Literacy Programs*
Richard L. Venezky, Page S. Bristow, John P. Sabatini *(TR93-12, 24 pages)*

Apr 1994 *Understanding Family Literacy: Conceptual Issues Facing the Field*
Vivian L. Gadsden *(TR94-02, 32 pages)*

Apr 1994 *Children, Parents, and Families: An Annotated Bibliography on Literacy Development In and Out of Program Settings*
Vivian L. Gadsden *(TR94-04, 84 pages)*

Information on ordering of NCAL publications may be addressed to Dissemination at NCAL.

Revised May 20, 1994

ACKNOWLEDGMENTS

Many people contributed to the success of this project, including Garrett W. Murphy, Director, Division of Continuing Education Planning and Development; Cynthia T. Laks, Chief, Bureau of Continuing Education Program Development; Camille Fareri, Associate, Bureau of Continuing Education Program Development, New York State Education Department; George K. Tregaskis, Consultant, D. Keil Associates; Andrew L. Morzello, Director of White Plains Adult and Continuing Education; Ann Serrao, Supervisor of Instruction, White Plains Adult Literacy Program; M. Howard Davis, Supervisor of Guidance Services, White Plains Adult Literacy Program; the teachers, counselors and students of the White Plains Adult and Continuing Education Program; a cadre of student workers at the University of Delaware, particularly Alex Hamilton, Mike Beall, Jenny Stanberry, Nina Patti, and Paige Kelty; and Mary Seifert of Yardley, Pennsylvania.

However, to Ann Serrao, Supervisor of Instruction at the Rochambeau School, we owe more than can be expressed here for her continuing support and for her frequent efforts to secure and clarify data, locate assistants, arrange for individuals to be tested, and to keep accurate records on all aspects of the data collection. We are also indebted to Bob Calfee, Irwin Kirsch, Andy Kolstad, and Dan Wagner for comments on an earlier version of this report.

TABLE OF CONTENTS

MEASURING GAIN IN ADULT LITERACY PROGRAMS

Richard L. Venezky
Page S. Bristow
John P. Sabatini
University of Delaware

Abstract

Problems in the measurement of gain in adult literacy programs were investigated through repeated testing of a group of students in ABE and GED classes and through computer simulations. Ninety-two students were tested at three different times over seven months with a battery of norm-referenced reading and mathematics tests as well as with tests of reading rate and decoding developed especially for this study. Gain scores were found to vary across tests, with significant declines as well as gains. No significant differences in gains were found for amount of instructional time or for attendance rate, and a large amount of group heterogeneity was revealed through an analysis of growth patterns. Computer simulations for grade-equivalent stability showed that with populations smaller than 200, inconsistencies in grade level intervals can account for a major proportion of the yearly gain typically reported for adult literacy instruction. In contrast, simulations of regression to the mean caused by guessing on multiple-choice tests showed that this effect was relatively small. These results strongly support the need to construct a multiple indicator system for evaluating adult literacy programs; a system that attends to the multiple goals of such programs and is free of elementary and secondary level conventions such as grade-equivalent scores.

INTRODUCTION

Programs that receive federal funding are required to measure and report certain factors related to program quality and learner achievements. The most recent codification of these requirements is in amendments to the Adult Education Act, contained in the National Literacy Act of 1991 (P.L. 102-73). Section 331 (a) (2) of these amendments requires state agencies to develop and implement "indicators of program quality," attending at a minimum to recruitment, retention, and literacy skill improvement. Another section states that assistance to programs should be based in part on learning gains made by educationally disadvantaged adults, and another section requires, as part of an application for federal assistance, a description of "how the applicant will measure and report progress" on meeting recruitment, retention, and educational achievement goals. These provisions are in addition to the requirement in the original act that state plans for programs must describe how the sponsoring organization will "gather and analyze data (including standardized test data) to determine the extent to which the adult programs are achieving the goals set forth in the plan..." (Section 352).

For a number of years the New York State Education Department has required that the Tests of Adult Basic Education (TABE) (CTB/McGraw-Hill, 1987b) be given to students in state-funded literacy programs on entry and on exit, and that the programs report mean learner gains in grade equivalents. The TABE, which is described more fully below, is a battery of tests for vocabulary, reading comprehension, language, spelling, and mathematics abilities. All of these tests were constructed to measure basic skills using skill models for the areas involved. With the publication of several new approaches to assessing adult literacy and in recognition of a growing discomfort within the adult literacy community over basic skills assessment, the New York State Education Department sought to explore alternatives to the TABE. The present study resulted from discussions with the New York State Education Department and the White Plains Adult and Continuing Education Program at the Rochambeau School in White Plains, New York, where the testing was done. This report presents findings on one aspect of this project, the measurement of learner gain. Other components of the larger project are reported in Venezky, Bristow, and Sabatini (1993) and Sabatini, Venezky, and Bristow (1994).

Measurement of learning is one of the most complex issues facing the study of schooling (Bryk & Raudenbush, 1987; Harris, 1963; Willett, 1988). Although claims are often made that 100-150 hours of instruction are required to achieve a grade level of progress in adult literacy programs (cf. Mikulecky, 1987; Sticht, 1982), few studies have reported data to support this claim. In fact, where claims such as this appear to be supported, contradictions exist. For example, in a longitudinal analysis of New York Literacy Assistance Center data, Metis Associates (1991) reported that 2,055 students who remained in city literacy programs for two years achieved an average of 9.9 months progress in reading in their first year, but only 3.5 months of progress in their second year, with an average of 182 contact hours per year. An inspection of the mean gains for different entry level scores showed a decline of mean gain with increasing entry level score, indicating that

increased test-taking ability might account for a significant portion of the gain.[1]

Another problem with most of the gain scores reported in evaluation studies (e.g., Metis Associates, 1991) is that they are based upon grade equivalents, which are not equal interval scales (see below). Although all test scales present problems of interpretation, grade equivalents are especially problematic when applied to adults because they are dependent upon the content and pace of the elementary/secondary curriculum. Among the many irregularities that contribute to the difficulties in interpreting grade equivalents are differential growth rates across subject areas, a flattening of the age-achievement growth curve at the higher grade levels, and the assumption of no loss or gain in ability over the summer (Reynolds, 1981; Thorndike & Hagen, 1977).

Lack of an equal interval scale may not be a problem when sample sizes are large; however, it is a potential problem when sample sizes are small or when other (equal interval) scales are used concurrently. Under either of these conditions, means and differences of grade-equivalent scores could be highly misleading. For example, an item response theory (IRT) scale score change of three points in one interval of the TABE Level D Comprehension scale (757-760), corresponding to a change from 31 to 32 items correct, is equated to a change of 0.4 grade equivalents. At a slightly higher point on the scale, however, the same three-point scale score change (771-774) corresponds to a 1.6 grade-equivalent difference. Thus, equal changes in performance at different intervals on the scale can result in vastly different grade-equivalent changes. The result of aggregating grade equivalents and computing averages from them can lead, therefore, to substantially different gains or losses from what is implied by the scale score differences.

Grade equivalents applied to adults are problematic for other reasons as well. First, many adult educators consider grade-equivalent scales to be misleading in that they are based on how particular skills develop in children and may not necessarily represent similar development in adults. To tell an adult that he or she reads or writes at a fourth-grade level may be more demeaning than it is informative. Even if an adult scores the same on a particular reading test as the average fourth grader, the strengths and weakness of the adult in reading will probably not be the same as those of the school level student.

Another problem with grade-equivalent scores is that they are a reflection of how literacy is taught in the school curriculum, which is vastly different from what ABE programs teach. The former traditionally stresses narrative fiction—plot, character, author's purpose, and so forth—almost to the exclusion of the functional literacy skills that are the core of most adult programs (Venezky, 1982).[2] Charts, TV schedules, prescription labels, job application forms, and the like are infrequently encountered in elementary school basal readers. When they do appear, they receive abbreviated attention and are usually surrounded by a *cordon sanitaire* to isolate them from the more honored selections from the canon of award-winning children's fiction. The levels at which functional literacy skills develop in the K-12 continuum may be as dependent upon mathematics, social studies, and science instruction as they are on reading instruction.

Finally, even if grade equivalents were meaningful for adult needs, they would remain for many adult students an unwelcome reminder of an unpleasant and unsuccessful schooling experience. With grade equivalents rests also an inference of childlike abilities, baggage that adult educators would prefer not to carry in striving to assist those whose self-images related to education are often weak and whose attitudes towards formal education may be far from positive.

A further measurement problem is that ABE and mathematics classes teach far more than reading comprehension, yet most reporting of student progress is for reading alone. Typical ABE classes stress reading, writing, mathematics, and life skills. Some students may progress in life skills or writing, for example, but this will usually not be reflected in the scores reported to higher administrative levels. For GED classes, a similar situation exists, in that GED classes focus on the five subtests of the GED examination: mathematics, science, social studies, language arts, and writing. Even though progress toward these goals can be, and usually is, measured by the official GED practice tests, some state and local reporting requirements still mandate basic skills reading tests for measuring learner gain.

Whatever primacy reading may have in adult education, it is not the sole instructional goal nor is it an adequate proxy for the other skills of interest. Although tests of basic skills may show moderate to high correlations across tests due to test-taking ability, general intelligence, and skill relationships (e.g., word problems in mathematics), there is no evidence to suggest a similar correlation for gain scores in basic skills, which are determined more by the content actually taught and practiced.

Even the measurement of reading poses a question in that this entity can be assessed by either basic skills tests or by functional literacy tests. In a separate study, it was demonstrated that these two approaches, at least as exemplified by popular standardized instruments, are far from equivalent (Sabatini, Venezky, & Bristow, 1994). In a multiple regression of TABE and TALS scores (and scores on oral reading and decoding tests), the best predictor of functional literacy, using the TALS Document scale as a dependent measure, was not Reading Comprehension nor Vocabulary, but Mathematical Concepts and Applications. The TALS requires problem-solving ability for its open-response items, including solution planning, multistep operations, and the like; this ability is captured best by a test that requires many of the same skills to be used in the solution of word problems.

All of these issues are further confounded by the practice of reporting mean gain scores for classes or programs based upon pretests and posttests without attention to individual growth, group heterogeneity, or student attrition. Testing at only two points in time (i.e., two-wave designs) necessarily gives the impression of linear growth, when, in fact, many other growth curves are possible and probable. With the extremely divergent mixtures of individuals whose abilities are aggregated when adult literacy programs report gain scores, important differences in growth patterns are disguised (Willett, 1988). Unfortunately, without unobtrusive data collection techniques (e.g., computer-assisted instruction with instruction-based assessment), repeated measurement is limited to relatively long intervals. Standardized achievement tests, for example, could not be given every month

due to their negative influence on program retention and practice effects that would confound measurement.

Finally, attrition rates are relatively high in adult literacy programs, averaging from 25% to 60% of the initial enrollment. Under these conditions, regressions to the mean could occur, thus biasing pretest/posttest differences, if: (a) test items are susceptible to guessing (i.e., multiple-choice items) and (b) the cohort of students who exit early does not represent a random distribution across pretest performance. For example, if the students who remained to the end of the program scored lower, on the average, than those who left early, the expected mean performance of those who remained would be higher on a retest than on their pretest, even one administered immediately after the pretest.

Given the concerns expressed above, the research reported here focused on four specific issues: (a) the variability of learning across subject areas, (b) differences between basic skills and functional literacy gains (or losses) as a result of instruction, (c) the reliability of aggregated grade-equivalent scores, and (d) the size of regression to the mean effects for multiple-choice tests.

A. METHODS

1. OVERVIEW

The data reported here were gathered during the 1991-92 school year from adult literacy classes at the Rochambeau School in White Plains, New York. Because extensive batteries of tests were to be used in an instructional setting, a number of constraints were accepted by all parties involved in this study. Testing had to have, at most, a small impact on the students involved and could not become a barrier to their retention in the program. Similarly, testing had to be minimally intrusive on the classrooms and instructors and could not impose additional burdens on the staff without appropriate and agreed upon compensation. Participation by students in the experiment had to be voluntary, with observance of all customary and required human subjects' safeguards.

These were the most obvious constraints. Nevertheless, other constraints were present that limited the range of comparisons possible. One serious limitation of this study was a constraint that every study of this nature faces; adults come voluntarily to literacy programs and many leave prior to the final testing date, due either to early completion (e.g., GED) or to other reasons. Of the 213 individuals who entered the instructional program in the fall of 1991 seeking assistance with literacy skills, 168 completed all of the initial TALS and TABE tests. Of these 168, 123 completed the second testing of the TABE and TALS, and 92 retook the complete set of tests after 360 hours of instruction (day students) or 120 hours (evening students).[3] It should be noted, however, that according to the New York State Education Department, the retention rate of the White

Plains program is above average for the state of New York adult programs, and New York state programs typically exceed national adult education averages for retention.

2. SUBJECTS

All subjects (hereafter called *students*) for this project attended ABE or GED classes at the Rochambeau School in White Plains, New York, during the 1991-92 school year. This school is the site of the White Plains Adult and Continuing Education Program and is used exclusively for that purpose. Besides the ABE and GED classes, the school offers an extensive number of programs, including ESL, job skills, general continuing education, workplace literacy, neighborhood literacy, and family literacy in cooperation with White Plains elementary schools. Many of the students were graduates of the school's ESL programs. All of them attended ABE 1, ABE 2, ABE 3, or GED classes voluntarily, either during the day or in the evening. No survey data are available on the students' reasons for enrollment, but the program staff believe that improvement of job potential was the most common motivating force.

Three groupings of students will appear throughout this report with others mentioned occasionally for special reasons. The three main groupings are: (a) the 213 students who registered for classes and for whom some background information is available (enterers), (b) the 168 students who completed all of the initial TABE and TALS tests (starters), and (c) the 92 students who completed all of the TABE and TALS tests for the three testing periods (persisters). The students were predominantly foreign born, non-Caucasian, low income, and either not married or separated from their spouses. There were slightly more males (53%) than females (47%), and 60% were in the age range of 26-50 years. Few voted during the past five years in a national or s , election, almost none reported any health-related handicaps, one quarte read a newspaper daily, and nearly three quarters considered themselves sufficiently literate to handle the reading demands of home, work, and family. Most also claimed to have relatively extensive literacy practices, as evidenced by self-reports of newspaper, magazine, book, and other types of reading.

3. INSTRUCTION

The ABE/GED staff is composed of five teachers, one of whom teaches both day and evening classes. Of the other four teachers, two teach day classes and two teach evening classes. Three counselors support the ABE/GED programs as well as the ESL programs at the school. The day teachers work full-time, are members of the local teachers union, and receive benefits. The evening teachers work part-time and are paid on an hourly basis without benefits. All 5 are certified teachers with a mean of 9 years and a range of 2-22 years of experience teaching adults.

Separate classes were held, day and evening, for ABE 1, ABE 2, ABE 3, and GED. Class sizes ranged from 25-32 students, with average attendance in the 16-25 range. Teachers described their classroom instruction as varied and flexible. The majority of class time was spent on instructional and practice activities to improve reading, followed by writing and then mathematics activities. A small amount of time was also spent on life skills.

Basic skills were emphasized, particularly in the ABE classes. Instructional groupings varied from one-on-one (or two), to small groups, to large groups, with some use of peer tutoring.

Students were assigned to classes on the basis of TABE Total Reading scores. However, slightly different methods were used for those entering the program this year compared to those who were continuing from the previous year. (See the Appendix to Venezky, Bristow, & Sabatini, 1993 for the complete assignment algorithm.)

Table 1 shows the distribution of students by ABE/GED level and by day or evening sessions. Day classes met for 20 hours of instruction each week, while evening classes met for 6 hours of instruction each week. In addition, day ABE 2, ABE 3 and GED classes had access for two class hours each day to a computer laboratory using Job Skills Education Program (JSEP) materials, while evening students had access to an additional, optional, single night of JSEP instruction each week.

4. INSTRUMENTS

A. TESTS OF APPLIED LITERACY SKILLS (TALS)

The TALS is a battery of norm-referenced tests that use functional literacy tasks to measure an adult's ability to apply literacy skills in contexts commonly encountered in everyday living (Kirsch, Jungeblut, & Campbell, 1991). These instruments were developed from the experiences gained by ETS with the Young Adult and Department of Labor Literacy Surveys (Educational Testing Service, 1992; Kirsch & Jungeblut, 1986). TALS items require short answers and other constructed responses as opposed to multiple-choice responses. The TALS battery is composed of three tests: Document Literacy, Prose Literacy, and Quantitative Literacy. In this study, the TALS Document and Quantitative Tests (Form A) were administered in testings 1 and 2; all three TALS tests (Form B) were administered in testing 3.

B. TESTS OF ADULT BASIC EDUCATION (TABE)

The TABE is a battery of norm-referenced tests that require multiple-choice responses and is the most frequently used commercial test in adult literacy programs (Development Associates, 1992; Ehringhaus, 1991). The tests administered in this study were the Vocabulary, Reading Comprehension, Mathematics Computation, and Mathematics Concepts and Applications tests, all of which were administered at each testing period. (Form 5 was administered in testings 1 and 2; Form 6 in testing 3.) Each test has four graduated but overlapping levels (Easy, Medium, Difficult, Advanced) with alternate forms available for each. Also available is a Locator Test for determining the appropriate level for full-scale testing. This Locator Test includes 25 multiple-choice vocabulary items and 25 multiple-choice arithmetic items and requires 37 minutes for administration.

C. ORAL READING TASKS

i. Passages

Although adults engage in silent reading far more often than oral reading, studies of low-literacy adults have shown that oral reading ability is a consistent indicator of reading comprehension (Bristow & Leslie, 1988). Change in oral reading rate might, therefore, be an adequate indicator of reading improvement. To test this possibility, four passages for oral reading were selected from a variety of instructional materials commonly used in adult basic education and GED programs. Each passage was selected from expository materials of ascending difficulty and was minimally adapted for length. The resulting passages varied from 188 to 328 words. The topics (sleepwalkers, lightning, plastic trash, and fever) were selected because of their high familiarity among virtually all adult populations. Two comprehension questions were prepared for each passage. One question required factual recall and the other required an inference based upon textual material. The questions were administered solely to focus the adults' attention on comprehension as they read orally.

ii. Decoding

The decoding tasks consisted of seven lists of six pseudowords each, designed to be of increasing difficulty. List one consisted of three-letter CVC pseudowords; list two consisted of four-letter pseudowords, with consonant clusters and short vowels. List three included four- and five-letter pseudowords with variant consonant pronunciations, digraphs, long vowels and silent "e"s. List four contained pseudowords with vowel combinations (digraphs). List five contained some multisyllabic pseudowords and more complex vowel and consonant combinations. List six contained two- and three-syllable pseudowords, while list seven contained four- and five-syllable pseudowords composed of high frequency syllables. A summary of these lists with sample test items is shown in Table 2. Internal consistency of the entire test, as measured by Kuder-Richardson Formula 20 (KR-20), varied from 69% (first testing period) to 81.2% (third testing period).

D. BACKGROUND QUESTIONNAIRE

The background questionnaire, adapted from the questionnaire developed and used by ETS in the National Adult Literacy Survey (NALS), is composed of six sections: general and language background, educational background and experiences, political and social participation, labor force participation, literacy activities, and demographic information. The questionnaire was administered individually in an interview format and required 15-20 minutes for completion.

5. OTHER DATA COLLECTION

Additional data were collected through interviews with students, teachers, and administrators; class visits; inspection of student work; and attendance records. Except for the attendance data, which were entered into various analyses described below, these data were used to characterize the instructional programs.

6. PROCEDURES

A. TIMING OF TESTING

The TABE and TALS batteries were administered at the beginning of instruction, after 60 (evening) or 120 (day) hours of instruction, and after 120 (evening) or 360 (day) hours of instruction. (For the TALS, only the Document and Quantitative tests were administered during the first and second testing periods; at the final testing, all three TALS tests were included.) The oral reading tasks were administered only at the initial and final testing. Although 213 students (enterers) were initially enrolled and, therefore, eligible for testing, only 168 (starters) completed all of the initial TALS and TABE tests. Of these, 145 were enrolled when the second testing occurred, and of these, 123 completed the second round of tests. In the third testing, 92 (persisters) of the 101 students still enrolled completed all of the tests. Testing for the day students occurred in September, late October (after 120 hours of instruction), and February (after 360 hours of instruction). The evening students were tested in September, December (after 60 hours of instruction), and March (after 120 hours of instruction). The complete testing schedule, including dates, tests administered, and numbers of students tested, is shown in Table 3.

B. TABE/TALS ADMINISTRATION

For each testing period, students were randomly assigned to take either the TABE or the TALS on day 1; the remaining tests were given on the next class day. Each set of tests was administered in a single sitting; group administration in classrooms utilized the publisher's standardized instructions, including time limits. During the first testing period, students were placed into one of two levels of the TABE (Easy or Difficult) based on their TABE Locator Test score. Those who scored less than 12 on the Locator Test were considered nonreaders and thus did not take the TABE (or TALS) battery. Those who received raw scores between 13 and 29 were given the E (Easy) level, and subjects who scored above 39 were given the D (Difficult) level. Students whose scores were between 30 and 39 were randomly assigned to either the D or the E levels. Normally these students would have been placed in Level M (Medium), but since the tests overlap considerably in difficulty levels, little loss in precision was projected. Once assigned to a level, a student was tested at that level for all three testings. All test administrators attended a three-hour training session that prepared them to use the TABE and TALS standardized administration procedures and to adm er the oral reading tasks as described below.

C. ORAL READING TASKS

Subjects were individually tested on the oral reading tasks in a half-hour session that was audiotaped. Twelve examiners participated in the testing; of the twelve, six were ABE/GED teachers, two were guidance counselors, and four were college students who had prior experience working with the adult population being tested. The oral reading tasks included two sections: decoding and oral reading of passages. For the decoding section, students were told that they would be shown lists of made-up words that they were to read aloud. Prior to reading the lists, they read a sample to give them practice with made-up (pseudo)words. Once they appeared to understand

the task, the lists were administered. The made-up words were presented on cards that displayed six words each. Each student read the seven lists in order with no interruptions unless he or she began to have difficulty. If a student was obviously struggling, the examiner asked, "Would you like to stop?" The student made the decision to stop or continue.

Regardless of whether the student completed the decoding lists, he or she was asked to read the oral reading passages. Students read orally as many of the four selections as they could and answered comprehension questions after each. The passages were ordered according to difficulty from easiest to hardest. Students were told that the examiner could not help them if they had difficulty and that after reading they would be expected to answer questions without looking back at the passage. Comprehension questions were included to assure that they were focusing on comprehension as a goal in reading. The instructions given here encourage use of what Carver (1990) calls the learning process. (Carver found that oral reading rate for an individual varies with the reading task. Skimming and scanning, for example, can be done significantly faster than *rauding*, Carver's term for reading comprehension of relatively easy materials. Learning and memorizing are the slowest forms of reading, according to Carver.)

Students were told prior to reading that they could stop at any point if the reading became too difficult for them. During the first round of testing, 47 students exercised this option before completing all four passages. Data on passages begun but not completed were excluded from the study. After reading each passage, students answered two comprehension questions and made a self-assessment of the difficulty of the passage. Students were asked to rate the passage as either easy, hard, or just about right for them.

D. BACKGROUND QUESTIONNAIRE

Each background questionnaire was administered in interview format by one of eight examiners who had attended a three-hour training session on administration procedures. For all questions with multiple choice responses, hand cards were presented which listed the potential responses. The choices were read to the student from the card, after which he or she selected one of the alternatives (or supplied a new one). Students were notified at the beginning of the interview that they could refuse to answer any question.

7. COMPUTER SIMULATIONS OF GRADE-EQUIVALENT SCORES AND REGRESSION TO THE MEAN

The issue examined in the grade-equivalent simulations was the variation in mean grade equivalents for groups of students who had identical mean scale scores. Because the particular relationship of the grade-equivalent scale to the IRT-obtained scale, two groups of students could have identical scale score means but different grade-equivalent means. If the variability of grade-equivalent means was large for identical scale score means, then aggregate grade equivalents could be potentially misleading. This variability is referred to here as stability (or instability). Tests of grade-equivalent stability and regression to the mean were carried out with computer intensive methods (Noreen, 1989). To test the stability of grade-equivalent scores relative to IRT scale scores, populations of different sizes and scoring characteristics were simulated. In all of these simulations, a mean and a standard deviation

for a normal scale score distribution were selected for a population of a specified size. Then, scores for members of the population were drawn randomly from this distribution.

For each scale score drawn, the corresponding grade-equivalent score was located from the conversion tables provided by the publisher or determined through interpolation of the data given in these tables (CTB/McGraw-Hill, 1987c). Once all scale scores were drawn for a specified population and their corresponding grade equivalents determined, the mean of the grade equivalents was computed. This process was done 1,000 times to give a distribution of 1,000 pairs of means, from which the standard deviation of the grade-equivalent means was computed.

To maintain a fixed mean for scale scores, each scale score drawn was paired with the scale score that would yield the desired group mean for that sampling. Thus, if the desired mean was 667, and 650 was drawn, then 684 was also included. Through this procedure, N/2 random scores were drawn to create a total of N scores.

To test the effect of the variance of the scale score distribution on grade-equivalent stability, standard deviations of 30, 40, 50, 60, and 70 were tested for scale score means of 650, 667, 689, 702, and 750, using a population size of 100. This size was selected because it closely approximated the sample size for persisters in this study. The standard deviation range selected was based upon the TABE's norming standard deviation for ABE students (50.7) (CTB/McGraw-Hill, 1987a, p. 49). The scale scores were selected to cover a range from five items correct above the chance level to about five items below a perfect score. (Samples drawn randomly from distributions with means near the scale boundaries and with large standard deviations tend to become highly skewed; therefore, scores near the boundaries of the scale were avoided.)

To test the influence of sample size on grade-equivalent stability, five different population sizes (10, 50, 100, 200, 500) were tested at each of five different scale score means (650, 667, 689, 702, 750), with the scale score standard deviation again set at 50.

To quantify regression to the mean, it was assumed that each observed score (S_O) was a linear combination of independent components, consisting of a true score (S_t), a guessing score (g), and a small random disturbance (d). S_t represents the number of items on a test for which a specified individual has the ability to respond correctly; g is the number of remaining items on the test (Total - S_t) for which an individual would select the correct answers by chance. With four alternatives for each TABE item, the probability of selecting a correct alternative by chance was 25%. The disturbance (d) was an amount that, in theory, should be added or subtracted from each score to account for above or below normal test alertness, motivation, distractibility, and so forth. In these simulations, however, the disturbance factor was assumed to be small and was, therefore, ignored.

In each simulation, the maximum regression possible was tested. First, raw scores for 200 subjects were generated, using a normal distribution of scores; then, the 100 highest scoring subjects were removed. These latter

subjects represented students who left a hypothetical program after pretesting but before posttesting. (With these students present for the posttesting, any tendency for abnormally low pretest scores to regress upward would, in theory, be counterbalanced by abnormally high pretest scores regressing downward.) Next, new observed scores were generated for the remaining students, holding their true scores constant. (That is, only the guessing component, g, was recomputed.) This second set of simulated scores represented what might have been a retesting of the lowest 100 scoring subjects within a short time period after the first testing. The difference in group means for the two "testings," using only the lowest scoring 100 students, was taken to represent a regression effect. As with the grade-equivalent simulations, 1,000 difference scores were generated by this method for each statistic reported.

8. SCORING

A. ORT DECODING AND READING PASSAGES

The ORT Decoding lists were scored by a linguistics doctoral student trained by the author who designed the decoding task. A random sample of decoding protocols was rescored by the author to verify that scoring procedures were being followed systematically. A number of problems were uncovered in the scoring procedures, the most serious of which relates to the evaluation of non-native English pronunciations. In many cases, the distinction between incorrect decoding and a non-native pronunciation of correct decoding was difficult to make. Because of the high percentage of non-native English speakers in the main sample for this study, this finding led to a cautious interpretation of decoding scores.

The ORT reading passages were timed from the audiotape using a stopwatch to determine minutes and seconds for the reading of each passage. All passages were retimed to verify the accuracy of the timing procedure. Agreement between the two timings was over 90%. Disagreements of more than three seconds were resolved by a third timing. Since correlations of reading rates across the passages for the first and third testing periods ranged from 0.85-0.98, rates on a single passage (Lightning) were used for the analyses described below.

B. TABE/TALS

The TABE tests were scored twice, initially by test examiners and later by project personnel at the University of Delaware. Discrepancies were resolved by a third scoring. For the Locator Test, scoring errors made by the initial scorers totaled 11.8% for the vocabulary section and 11.3% for the mathematics section. Seventy-three percent of these errors were within two items of the correct score. The TALS tests were scored by an ETS-trained scorer, utilizing the standardized scoring criteria (Kirsch et al., 1991). Twenty percent of the TALS tests were rescored by another ETS-trained examiner; the interrater reliability was 99%.

B. RESULTS

1. GAIN SCORES

Means and standard deviations for the various tests across the three testing periods are shown in Tables 4-6. Mean scores at testing 1 increase monotonically from ABE 1 through ABE 3 for both day and evening classes. The differences between ABE 3 and GED, however, are less distinct, especially for the day classes. This pattern, with minor changes, holds for the other two testing periods. Mean gain scores were examined for each of the eight groupings (four class levels by two times of offering). These data are displayed in Table 7 for changes from testing 1 to testing 3, the two testing periods for which the oral tests were administered along with the TABE and TALS. In general, these data do not show any relationship between size of gain and entry level ability, as is often found in adult literacy evaluations (e.g., Metis Associates, 1991). The only testing familiarity pattern occurs for TABE Comprehension in the day classes, where the mean gains from lowest to highest levels of instruction were 41.6 (ABE 1), 5.0 (ABE 2), 0.2 (ABE 3), and -5.8 (GED), respectively.

All classes showed gains for Document Literacy, Mathematics Computations, and reading rate. All classes except the evening ABE 1 class showed losses for the TALS Quantitative Literacy Test, and all except one (evening ABE 3) showed losses for decoding. For the remaining tests, more gains were made than losses.

Because of the relatively small sample sizes within each class, the four day classes were collapsed into a single group as were the four evening classes, and the resulting aggregated gain scores were compared. These results are shown in Table 8 where some unexpected results can be observed. For example, day students, who averaged more than three times as many contact hours as the evening students, did not consistently outgain the evening students. In fact, the evening students outgained the day students for TABE Vocabulary and Mathematics Computation and for the two TALS tests. However, multiple analyses of covariance (MANCOVA), with testing 1 scores as covariates, showed no significant differences between day and evening students for any of the tests ($p \leq .05$).[4]

When the student scores are summed across all instructional categories (with unequal amounts of instruction between day and evening classes), t-tests showed significant positive changes in performance from testing 1 to testing 3 for TALS Document Literacy, TABE Mathematics Computation, and oral reading rate (see Table 9). In contrast, TALS Quantitative Literacy and decoding showed significant declines. None of these gains or losses reached one half of a standard deviation, and those for decoding and oral reading rate were particularly small. From testing 1 to testing 2, all mean TABE and TALS scores for the 92 persisters increased significantly. However, from testing 2 to testing 3, the only significant increase was for TALS Document Literacy; mean scores for TALS Quantitative Literacy and TABE Comprehension declined significantly. TABE Concepts and Applications declined for this interval and TABE Computations increased,

but neither of these changes was significant. (TABE Vocabulary declined an insignificant 0.4.)

Intercorrelations among gain scores were uniformly low, ranging from -0.21 (Mathematics Computation-Quantitative Literacy) to 0.37 (Concepts and Applications-Mathematics Computation). Gain scores also did not correlate significantly with attendance (based upon a subsample of 48 day and evening students). The attendance data were tested for both total sessions attended and for percentage of total possible sessions attended: neither correlated significantly with any gain score.

Group heterogeneity was investigated by summarizing for TALS Document Literacy (the test that showed the most consistent results) the different growth patterns across the three testing periods. These are shown in Table 10, where it can be seen that only 30.4% of the students gained consistently across the three testing periods. About 26% gained from testing 1 to testing 2 and then declined from testing 2 to testing 3. In contrast, about 20% declined from testing 1 to testing 2 and then gained from testing 2 to testing 3. About 70% of the students gained from testing 1 to testing 2, whereas only about 57% gained from testing 2 to testing 3.

A second approach to measuring progress is shown in Table 11, where the number of individuals who, in theory, were ready to progress to the next level of instruction as measured by the TABE Total Reading score or the TABE Total Mathematics score is shown. The same procedures used for placement at the beginning of instruction at White Plains were applied to testing 1 and testing 3 scores to identify those ready to progress to the next instructional level. For ABE classes, an individual was ready to move to the next level if his or her score was above the score range defined for the class in which the individual was currently placed. For example, the score range for ABE 2 (Total Reading) is 682-724. Therefore, anyone who scores above 724 is ready to move up to a higher level class. The two GED classes were omitted from these analyses because no higher level class is available for their graduates.

As Table 11 shows, 14 of 64 students in ABE 1 through ABE 3 classes were ready by the third testing period to move up to a higher level class. Unfortunately, as Table 11 also shows, exactly the same number of students were ready to move to a higher level class after testing 1, but due to either placement error or teacher judgment were retained at a lower level. Thus, the net gain based on Total Reading score was zero. For Total Mathematics, the net gain was 7 (of 64) or 11%. Therefore, by the promotion criterion, only a small gain over pretest performance occurred, and this was restricted to mathematics.

If the various tests were highly unreliable or if the testing conditions varied dramatically from testing to testing, true gains in reading and mathematics skills would be difficult to detect. Furthermore, correlations of the same or related tests over the different testing periods would tend to be low. However, within the same test, correlations over testing periods were high: TABE Comprehension at testing 1, for example, correlated 0.86 with TABE Comprehension at testing 2 and 0.89 at testing 3 for the 92-subject sample (persisters). Similarly, TALS Document Literacy at testing 1 correlated 0.85 with itself at testing 2 and 0.82 at testing 3 for the same

sample. These correlations are remarkably high, indicating that the ordering of the students changed little from testing 1 to testing 3.

The small decrease (6.2%) in decoding ability does not appear to be meaningful, given the problems uncovered in scoring an oral pronunciation test for students whose native language was not English. Oral reading rates, which improved significantly from testing 1 to testing 3, nevertheless, remained quite low (mean rate at testing 3 was 111 words per minute). Furthermore, an increase of only seven words per minute after six to eight months of instruction is not substantial, representing only about a 7% increase in performance. Although holistic scoring might reveal qualitative improvements not detected by rate measures, the low gain scores strongly suggest that the students involved did not make educationally meaningful gains in reading rate.

2. GRADE-EQUIVALENT STABILITY

Results of the grade-equivalent stability simulations are shown in Figures 1 and 2. The graph of grade-equivalent standard deviation versus scale score for various population sizes (Figure 1) shows that for population sizes of 200 or more, the grade-equivalent variability is small. (The scale score standard deviation was set at 50 for the data shown in Figure 1.) However, for smaller population sizes, considerable fluctuation occurs. Since the mean grade-equivalent gain for adults in the second and third years of New York City ABE and GED programs was reported by Metis Associates (1991) to be just above 0.3 years per year of study, a grade-equivalent standard deviation of even 0.1 could result in a high proportion of reported gain (or loss) being due to the peculiarities of the grade-equivalent scale rather than to true improvement in performance. Figure 2 shows that as the variability of scale scores within a group with a specified mean increases, the variability of the corresponding grade-equivalent scores also increases. (The population size was fixed at 100.)

To extend these results to practical situations, imagine that 100 different adult literacy programs had enrollments of 50 students each and each program reported identical mean pretest scale scores of 702 in Reading Comprehension. If these programs reported mean grade equivalents instead of mean scale scores, their scores would most likely no longer be equivalent; instead, they would probably be normally distributed with a mean of 5.4 years and a standard deviation of 0.14 years. With this distribution, 32% of the reported scores would either be larger than 5.14 or smaller than 4.86. This means that if no change in mean test performance occurs for any of these groups from pretest to posttest, many would still report, by grade equivalents, relatively large gains or losses.

Table 12 illustrates how two different groups of 10 students could have identical mean scale scores but different mean grade equivalents. This example could also be viewed as a case in which identical pretest and posttest scale means resulted in a 0.55 mean gain in grade equivalents.

3. Regression to the Mean

In contrast to the grade-equivalent simulations, the regression to the mean simulations demonstrated that only a small proportion of the change in performance might be attributed to this effect. Figure 3 shows the relationship between mean true score, that is, the number of items (out of 40) that a hypothetical subject would know the correct answers for, and shift in mean total score due to attrition of the top half of the population. In these simulations, a total score for each of 200 subjects is generated as a linear combination of a true score and a guessing score. The true score is drawn from a normal distribution of a specified mean with a standard deviation of seven. The guessing score is generated as the sum of correct guesses on all of the items not known, where the probability of guessing correctly on each item is 0.25.

Then, under the assumption of attrition of the top scoring 100 subjects, a mean total score is generated for the remaining 100 subjects. Finally, the guessing score for each of these remaining subjects is redrawn, and the new mean for the group of 100 subjects calculated. After 1,000 cycles of this procedure, the mean difference between first and second total scores for the lowest scoring 100 subjects is calculated. This is plotted in Figure 3 for 8 different true score means between 15 and 33. Given that these are worst case scenarios, wherein it is assumed that the 100 highest scoring subjects leave before posttesting, the sizes of the differences are relatively small, ranging from 0.67 (true score=15) to 0.302 (true score=33) items. Under more normal conditions, wherein the early leavers are distributed across the pretest score range, but with a higher mean, these differences would be greatly reduced.

Under the worst possible scenario of complete guessing on all items of a multiple-choice test, and with removal of the highest scoring half of the population, the increase in mean score after recomputation of random guessing scores was only 1.5 items. This represents the maximal regression to the mean that could be obtained and requires that none of the subjects knows the correct answers for any of the items, that is, that each guesses on all items. These simulations clearly demonstrate that regression to the mean for a typical multiple-choice test is a marginal threat to score reliability.

Conclusions

Of the four issues presented at the beginning of this study, three were shown to be major problems for adult literacy assessment, and one was not. Grade equivalents were found to be unstable relative to scale scores for sample sizes smaller than 200, but regression to the mean resulting from guessing on multiple-choice items was shown to be small and generally insignificant.

Gains made by students in the ABE and GED classes varied by subject and by test. Significant positive changes in functional literacy, as measured

by the TALS Document Literacy Test, occurred, but basic reading skills, as measured by the TABE Vocabulary and Reading Comprehension Tests, did not show a corresponding change. Mathematics Computation ability improved significantly, but Quantitative Literacy ability decreased significantly. Mathematics Concepts and Applications ability did not change at all. Similarly, oral reading rate increased significantly and decoding performance declined significantly, but changes were small.

Given the irregularity of these patterns, it is difficult to attribute them to a single factor. One possibility could have been a decrease in test-taking motivation at testing 3, since four of the six measures taken at testings 2 and 3 either declined or showed no effective change. However, two increased, one significantly, and the correlations of scores from testing 1 to testing 3 and from testing 2 to testing 3 were high, indicating that students scored similarly across tests. Supporting this conclusion are the intertest correlations and regression analyses, which are similar for the first and third testing periods. Another possibility is differences in difficulty between alternate forms of the tests involved. (One form of each test was used for testings 1 and 2; the alternate form was used for testing 3.) However, both test publishers claim that all pairs of alternate forms have comparable difficulty levels.

The differences found in gains across testing periods and testing instruments illustrate the dangers in basing program evaluation on any single measure. If a total TABE Reading score were reported for this academic year, the result would have been no significant change in performance. If, on the other hand, TALS Document Literacy scores were reported, a significant increase of about 0.4 standard deviations would have been submitted; if the TALS Quantitative Literacy scores were reported, the result would have been a significant decline in ability, equal to about a quarter of a standard deviation.

Group heterogeneity, as marked by different growth patterns on TALS Document Literacy, was extremely large. About 90% of the growth patterns fell into four different categories: continually increasing (31.5%), increase + decrease (25%), decrease + increase (20.7%), and increase + level (13%). Although this is a subject of continuing investigation, the characteristics that distinguish individuals in these various pattern groups have not been determined. Level performance from one testing period to the next could result from the student focusing on different skills than the ones tested, or on a consolidation of learning that is not revealed by the test. However, there is no tenable explanation for the large number of performance declines, outside of measurement error, which probably does not account for more than a small percentage of the total.

If these are true declines, then skill retention is clearly an issue to be studied (cf. Wagner, in press). One hypothesis is that skills are often not sufficiently practiced to become automatic. Therefore, if testing occurs soon after skill acquisition, the probability of satisfactory performance on that skill is high; on the other hand, if testing is offset from the completion of instruction, significantly lower performance will most likely be observed. Although it might be expected, based upon other evaluation studies, that the

highest gains would be found among those with the lowest entry-level scores, no such pattern was found here.

Another puzzling outcome was that neither attendance rate nor hours of instruction was a good predictor of gain for any of the measures. It was expected that day students, who attended classes for 20 hours per week, would outperform evening students, who only attended for 6 hours per week. Why no clear advantage for the day students was found remains to be resolved. One possible explanation is that the evening students were primarily those who were already working and, therefore, their presence indicated a particularly strong desire for improvement. In contrast, the day students may have been less sure of the value of further education and, therefore, less willing to invest the extra time and energy needed outside of class for academic advancement.

Given how small the gains were, a simpler explanation is that whatever is being acquired is not revealed by the measures selected so that differences between day and evening students are not being adequately assessed. This may be true and has been offered in similar situations as an explanation for low or insignificant gains (e.g., Fingeret & Danin, 1991). However, interviews with teachers and observation in classes revealed that instructional emphasis was placed on basic skills similar to those assessed by the TABE. It was expected that the TABE Vocabulary Test would show improved ability for almost any approach to teaching reading so long as the students received sufficient exposure to printed words. This is not to argue that nothing was acquired by the students; some of the tests did show significant improvements. In addition, life skills and writing may have improved but were not assessed, even though instructional time was devoted to them.

Low validity or reliability of the measures used in this study could also account for some of the results obtained. These issues are examined more fully in another part of the larger study mentioned earlier (Sabatini, Venezky, & Bristow, 1994). In general, both the TALS and the TABE had moderate to high content validity for adults, but the TABE Reading Comprehension Test was based on an outdated model of reading skill. The TALS tests were found to be reliable as measured by rank correlations between reported and obtained item difficulties and by Student-Problem (S-P) score table analysis (Harnisch & Linn, 1981; McArthur, 1987). In contrast, the TABE rank correlations were quite low and the S-P plots more irregular. These findings on the TABE are alarming, given the widespread use of this test for individual and program assessment. Based on the results reported here, use of the TABE for measuring either individual or program progress cannot be recommended.

Even with these findings on reliability, the possibility of highly limited true gains needs to be considered. Many of the persisters were students who failed to acquire basic skills in elementary and secondary schooling. It is not known how many were diagnosed during schooling as having a learning disability, but it is suspected that many were. Perhaps with more intensive instruction or with more functionally situated methods, they would make better gains. Nevertheless, it is possible, but not proven here, that these students have learning difficulties that cannot be easily overcome by the methods commonly used in ABE and GED instruction.

Whatever the explanation for the limited gains reported here, mean gain scores, no matter which scale is used, give an uninformative and inadequate view of program performance. A low to moderate mean gain score could result from a small number of students making large increases in performance while all the others made no gains, or by a large number of students making small but insignificant gains. Which of these occurs is important to know if program effectiveness is an issue. Furthermore, the impact of early leavers on gain scores needs to be examined. In spite of the tendency to view these students as dropouts, the truth is that many leave because of successful completion of their goals. This is evident with students who pass the GED; these students should not under any circumstances be counted as dropouts, nor should they be assessed with instruments other than the GED Practice Tests or the GED itself.

Comparative information on student performance, based upon national standards, is important for policy makers and could be useful to both instructors and students. Without such data, the ability of programs to prepare adults for work, citizenship, and home management may be difficult to evaluate. It is critical to develop evaluation policies that are based upon valid and reliable indicators for adults and that attend to what programs actually teach. Furthermore, the reporting of these data must allow a more complete portrait than what is possible from a single aggregate score.

These results narrow the search for a sound assessment policy in many instances and provide at least tentative answers in others. First, all of the customary caveats concerning the external validity of the sample are offered. The majority of the data derives from those who remained in the instructional programs. Very little is known about the background characteristics and literacy skills of the 45 students who registered but did not complete the first round of tests. A little more is known about those who completed the first round but not the second, and even more is known about those who completed the second round but not the third. Most striking about those who terminated early is that they scored higher on every measure at the first testing than those who were available and were tested at the third round of tests.

Why should the leavers be superior in ability to those who remained? One possibility is that the leavers, on average entering at a higher achievement level, reached their goals more rapidly. Another possibility is that the leavers were more highly motivated to achieve and, therefore, moved on more quickly into jobs, technical training courses, and the like, or passed the GED tests prior to the end of instruction. Of the 76 students who were tested at testing 1 but not at testing 3, we know that 13 left because they passed the GED prior to testing 3. Others may also have passed the GED at sites other than White Plains and, therefore, not been listed as such in the program's attendance logs.

When students leave ABE programs before the formal end of the programs, their departures tend to be viewed as losses for both the students, who may have failed to achieve their goals, and for the programs, which must report lower retention. But under certain circumstances, early withdrawal can be positive. For example, a student may come to an ABE program to learn specific skills, such as fractions and decimals. Once these

abilities are acquired, the student may move on to other interests. Although program directors may realize that a student's chances for gainful employment are limited without at least a GED, the program has, nevertheless, fulfilled a particular need. For GED students, the situation can be much clearer. Students in GED classes often take the GED Practice Tests and on the basis of these results decide whether or not to attempt the GED Tests. If they take the GED and pass, they leave the program as graduates, having succeeded fully. The students who remain at the end of an academic year are the ones who have failed to pass the GED. If all of the students who depart early do so because they pass the GED, then the persisters are, in effect, the students of most concern and the ones we should expect to show the lowest test scores for basic and applied skills.

One could speculate from students' vision screening data (Bristow, 1992) that some of the persisters have physical barriers to learning, thus requiring more time to reach their goals than those who left early. It is not known, however, if those who persisted had higher incidences of vision problems than those who left because there is no data at all on the vision abilities of the leavers. It is also not known how persistence in adult literacy programs relates to progress.

This study concludes that a single reading score is inadequate for measuring student progress in adult literacy programs and that grade equivalents are unreliable for estimating gain over time when fewer than 200 scores are aggregated. On the other hand, regression to the mean proved to be only a marginal threat to the reliability of mean test scores with multiple-choice tests. Most striking, however, was the unreliability found for the most commonly used adult literacy test (TABE). In contrast, the functional literacy test examined (TALS) proved to be reliable. However, the change in scores for these two test batteries was highly dissimilar, thus demonstrating the nonequivalence of basic and functional skills, at least as measured by these tests.

These findings strongly support the need to construct a multiple indicator system for evaluating adult literacy programs, a system that attends to the multiple goals of adult literacy classes and that is free of elementary- and secondary-level conventions such as grade-equivalent scores.

ENDNOTES

[1] The argument is that the lowest scoring students generally have the poorest test-taking skills and, therefore, do much worse than expected for their skill levels when encountering a test after a long absence from schooling or testing. However, after a few weeks in adult literacy instruction, they often acquire better test-taking strategies and more confidence in their own abilities and, therefore, score closer to their true ability levels.

[2] More recently, school-based reading programs have begun shifting to a more balanced approach to reading; however, basal readers continue to stress award-winning fiction over all other forms of writing.

[3] In addition, 175 students completed background questionnaires; however, certain types of background information (e.g., sex) were available from the program records for all 213 students who entered the program in the fall. Of the 92 students who completed all of the full-scale TABE and TALS tests, 90 also completed the TABE Locator Test. (The remaining two were continuing students for whom Locator scores were not available.)

[4] Although students were not assigned randomly to day and evening classes, the assumption in this analysis is that whatever led to the particular distribution observed was similar in effect to a random distribution.

REFERENCES

Bristow, P. S. (1992, Fall). Vision screening: A must for adult education programs. *NCAL Connections, 1*, 6.

Bristow, P. S., & Leslie, L. (1988). Indicators of reading difficulty: Discrimination between instructional- and frustration-range performance of functionally illiterate adults. *Reading Research Quarterly, 23*(2), 200-218.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*, 147-158.

Carver, R. P. (1990). *Reading rate: A review of research and theory*. San Diego, CA: Academic Press.

CTB/McGraw-Hill. (1987a). *Technical Report: Tests of Adult Basic Education (TABE) Forms 5 and 6*. Monterey, CA: Author.

CTB/McGraw-Hill. (1987b). *Tests of Adult Basic Education* (TABE). Monterey, CA: Author.

CTB/McGraw-Hill. (1987c). *Examiner's Manual Tests of Adult Basic Education (TABE) Forms 5 and 6*. Monterey, CA: Author.

Development Associates. (1992). *Executive summary of the first interim report from the national evaluation of adult education programs*. Arlington, VA: Author.

Educational Testing Service. (1992). *Beyond the school doors: The literacy needs of job seekers served by the U.S. Department of Labor*. Princeton, NJ: Author.

Ehringhaus, C. (1991). Teachers' perceptions of testing in adult basic education. *Adult Basic Education, 1*(3), 138-154.

Fingeret, H. A., & Danin, S. T. (1991, January). *"They really put a hurtin' on my brain:" Learning in Literacy Volunteers of New York City*. Durham, NC: Literacy South.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*(3), 133-146.

Harris, C. W. (Ed.). (1963). *Problems in measuring change*. Madison, WI: University of Wisconsin Press.

Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults* (Report No. 16-PL-01). Princeton, NJ: National Assessment of Educational Progress.

Kirsch, I. S., Jungeblut, A., & Campbell, A. (1991). *ETS Tests of Applied Literacy Skills: Administration and scoring manual*. New York: Simon & Schuster.

McArthur, D. L. (1987). Analysis of patterns: The S-P technique. In D. L. McArthur (Ed.), *Alternative approaches to the assessment of achievement* (pp. 79-98). Boston: Kluwer.

Metis Associates. (1991). *Analysis of New York City's 1989-1990 adult literacy data base*. New York: Literacy Assistance Center.

Mikulecky, L. (1987). The status of literacy in our society. In J. Readence & S. Baldwin (Eds.), *Research in literacy: Merging Perspectives: Thirty-sixth yearbook of the National Reading Conference* (pp. 211-235). Rochester, NY: National Reading Conference.

Noreen, E. W. (1989). *Computer intensive methods for testing hypotheses: An introduction*. New York: Wiley.

Reynolds, C. R. (1981). The fallacy of "two years below grade level for age" as a diagnostic criterion for reading disorders. *Journal of School Psychology, 19*, 350-358.

Sabatini, J. P., Venezky, R. L., & Bristow, P. S. (in press). *A comparison of the TABE and TALS as measures of adult literacy abilities* (Technical Report). Philadelphia, PA: University of Pennsylvania, National Center on Adult Literacy.

Simon & Schuster. (1990). *ETS Tests of Applied Literacy Skills.* Englewood Cliffs, NJ: Author.

Sticht, T. G. (1982). *Basic skills in defense.* Alexandria, VA: Human Resources Research Organization.

Sticht, T. G. (1990). *Testing and assessment in Adult Basic Education and English as a second language programs.* San Diego, CA: Applied Behavioral & Cognitive Sciences, Inc.

Thorndike, R. L., & Hagen, E. P. (1977). *Measurement and evaluation in psychology and education* (4th ed.). New York: Wiley.

Venezky, R. L. (1982). The origins of the present-day chasm between literacy needs and school literacy instruction. *Visible Language, 16*, 113-127.

Venezky, R. L., Bristow, P. S., & Sabatini, J. P. (1993). *When less is more: Methods for placing students in adult literacy classes* (Technical Report TR93-8). Philadelphia, PA: University of Pennsylvania, National Center on Adult Literacy.

Wagner, D. A. (in press). *Use it or lose it? What we don't know about adult literacy skill retention* (Technical Report). Philadelphia, PA: University of Pennsylvania, National Center on Adult Literacy.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education, 15*, 345-422.

APPENDIX A: TABLES

Table 1

Distribution of Students by Level and by Sessions for Initial Sample (Init.), Testing 1 (T1), and Testing 3 (T3)

	Levels														
	ABE 1			ABE 2			ABE 3			GED			All		
Session	Init.	T1	T3	Init.	T1	T3	Init.	T1	T3	Init.	T1	T3	Init.	T1	T3
Day	24	17	9	20	16	11	27	23	13	31	24	11	102	80	44
Evening	29	20	15	20	17	9	13	12	7	49	39	17	111	88	48
Combined	53	37	24	40	33	20	40	35	20	80	63	28	213	168	92

Note: With the exception of the evening GED group, only one class was offered per level and session. For the evening GED group, two classes were initially offered.

Table 2

Decoding Test

List	Contents	Example
1	Simple CVCs	vun
2	CVC with consonant clusters	hent
3	CVC & CVCe with digraph consonants	shafe
4	CVC with digraph vowels	spawk
5	1 & 2 syllables with common prefixes and endings	refarbed
6	2 & 3 syllables with common prefixes and endings	impentive
7	4 & 5 syllables with common prefixes and endings	disfactible

(All lists have six items.)

Table 3

Testing Schedule: Dates and Numbers Tested

	Testing 1	Testing 2	Testing 3
Dates Testing Began:			
Day Students	9/5/91	10/29/91	2/12/92
Evening Students	9/23/91	12/16/91	3/30/92
Test Administered:			
TABE Tests			
Locator	199	—	—
Vocabulary	185	136	101
Comprehension	185	136	101
Math Comp.	185	136	101
Con. & Apps.	184	136	101
TALS Tests			
Document	201	145	98
Quantitative	199	145	98
Prose	—	—	98
Background Questionnaire	175	—	—
Oral Reading Tasks			
Decoding	189	—	101
Oral Reading	185	—	101
Vision Screening	34	—	—

Table 4

Mean (Standard Deviation) Scores for ABE and GED Levels by Session, Testing 1

Day Session (n = 43)

	Level			
Test	ABE 1 (n=8)	ABE 2 (n=11)	ABE 3 (n=13)	GED (n=11)
TABE				
Comprehension	604.1 (71.1)	719.0 (22.4)	742.5 (23.0)	744.8 (22.8)
Vocabulary	609.1 (76.6)	677.5 (39.2)	722.2 (47.4)	735.7 (38.4)
Computation	692.9 (84.4)	740.1 (56.3)	795.5 (53.1)	779.1 (42.4)
Concepts & Applications	653.8 (58.7)	697.5 (55.2)	755.3 (47.3)	724.6 (39.6)
Locator Vocabulary[a]	7.6 (3.6)	12.9 (3.7)	18.7 (3.1)	19.6 (4.2)
Locator Math[a]	11.1 (4.9)	14.7 (5.6)	18.7 (5.6)	17.4 (4.2)
TALS				
Document	193.8 (38.5)	257.3 (50.8)	278.5 (47.4)	239.1 (29.8)
Quantitative	206.3 (41.0)	261.8 (51.2)	283.1 (50.6)	252.7 (42.7)
Oral Reading				
Rate[b]	70.1 (29.9)	105.9 (21.5)	118.6 (26.6)	115.1 (25.6)
Decoding[c]	16.4 (9.7)	20.0 (3.9)	29.8 (6.6)	24.6 (5.3)

(table continued on next page)

Table 4 (continued)

Mean (Standard Deviation) Scores for ABE and GED Levels by Session, Testing 1

Evening Session (n=47)

	Level			
Test	ABE 1 (n=14)	ABE 2 (n=9)	ABE 3 (n=7)	GED (n=17)
TABE				
Comprehension	628.5 (55.3)	703.1 (30.3)	727.3 (17.4)	749.2 (11.0)
Vocabulary	596.9 (97.7)	676.6 (39.2)	733.1 (40.6)	749.7 (29.4)
Computation	663.7 (92.7)	745.1 (31.0)	774.3 (23.2)	804.9 (27.7)
Concepts & Applications	632.9 (56.7)	709.7 (32.0)	734.4 (20.0)	760.6 (36.4)
Locator Voc[d]	11.7 (5.1)	13.6 (5.5)	17.9 (3.1)	17.8 (3.8)
Locator Math[d]	8.8 (4.5)	13.6 (4.6)	16.3 (2.9)	21.1 (2.9)
TALS				
Document	178.6 (34.6)	223.3 (31.6)	234.3 (21.5)	282.4 (41.8)
Quantitative	202.9 (39.5)	256.7 (47.4)	305.7 (26.4)	308.2 (35.9)
Oral Reading				
Rate[e]	61.5 (33.9)	107.1 (22.6)	123.3 (19.3)	125.7 (29.5)
Decoding[f]	13.8 (7.0)	19.2 (8.8)	22.9 (3.5)	22.8 (7.5)

a n= 7, 11, 13, 11, for ABE 1,2,3 and GED respectively.
b n= 7, 11, 12, 10, for ABE 1,2,3 and GED respectively.
c n= 8, 11, 12, 10, for ABE 1,2,3 and GED respectively.
d n= 13, 9, 7, 17, for ABE 1,2,3 and GED respectively.
e n= 12, 9, 7, 16, for ABE 1,2,3 and GED respectively.
f n= 14, 9, 7, 17, for ABE 1,2,3 and GED respectively.

Note: Oral reading rate is reported in words per minute; decoding scores are reported in total correct out of 42. All other scores are scale scores.

Table 5

Mean (Standard Deviation) Scores for ABE and GED Levels by Session, Testing 2

Day Session (n= 43)

	Level			
Test	ABE 1 (n=8)	ABE 2 (n=11)	ABE 3 (n=13)	GED (n=11)
TABE				
Comprehension	656.4 (57.5)	729.7 (19.9)	751.0 (19.8)	746.1 (25.6)
Vocabulary	640.9 (60.0)	689.5 (37.2)	733.8 (40.3)	740.7 (42.9)
Computation	719.0 (69.6)	769.0 (42.7)	798.5 (46.6)	788.7 (34.7)
Concepts & Applications	674.3 (65.5)	719.6 (42.4)	756.5 (44.4)	730.5 (48.7)
TALS				
Document	208.8 (31.8)	256.4 (36.1)	290.0 (51.8)	247.3 (53.7)
Quantitative	222.5 (36.9)	268.2 (41.4)	295.4 (53.5)	264.5 (41.6)

(table continued on next page)

Table 5 (continued)

Mean (Standard Deviation) Scores for ABE and GED Levels by Session, Testing 2

Evening Session (n=47)

	Level			
	ABE 1	ABE 2	ABE 3	GED
Test	(n=14)	(n=9)	(n=7)	(n=17)
TABE				
Comprehension	637.2 (90.3)	712.2 (36.6)	742.7 (17.3)	754.8 (15.7)
Vocabulary	598.4 (80.2)	690.9 (22.6)	736.1 (45.5)	759.9 (48.2)
Computation	665.9 (99.1)	755.1 (32.7)	788.9 (21.7)	808.7 (26.2)
Concepts & Applications	644.4 (68.3)	717.0 (37.9)	741.9 (31.7)	768.8 (37.1)
TALS				
Document	200.7 (33.4)	234.4 (42.2)	257.1 (30.4)	290.6 (41.2)
Quantitative	225.0 (42.7)	268.9 (40.1)	284.3 (28.8)	310.0 (35.7)

Table 6

Mean (Standard Deviation) Scores for ABE and GED Levels by Session, Testing 3

Day Session (n = 43)

	Level			
Test	ABE 1 (n=8)	ABE 2 (n=11)	ABE 3 (n=13)	GED (n=11)
TABE				
Comprehension	645.8 (32.2)	724.0 (35.0)	742.7 (25.9)	739.0 (29.0)
Vocabulary	632.8 (45.0)	670.7 (53.8)	731.3 (42.6)	740.5 (35.4)
Computation	714.4 (62.9)	762.5 (46.6)	800.6 (52.3)	790.0 (43.8)
Concepts & Applications	664.3 (50.7)	732.8 (59.7)	741.5 (94.4)	729.5 (46.8)
TALS				
Document	213.8 (32.0)	265.5 (51.6)	295.4 (51.3)	260.0 (33.5)
Quantitative	185.0 (27.3)	231.8 (37.1)	275.4 (56.7)	241.8 (36.3)
Oral Reading				
Rate[a]	84.4 (38.1)	112.3 (23.6)	129.1 (32.0)	119.7 (25.3)
Decoding[b]	15.3 (9.3)	16.9 (6.9)	27.8 (6.9)	23.4 (5.1)

(table continued on next page)

Table 6 (continued)

Mean (Standard Deviation) Scores for ABE and GED Levels by Session, Testing 3

Evening Session (n=47)

	Level			
Test	ABE 1 (n=14)	ABE 2 (n=9)	ABE 3 (n=7)	GED (n=17)
TABE				
Comprehension	617.2 (69.9)	703.6 (44.5)	723.4 (17.6)	749.4 (12.2)
Vocabulary	619.4 (69.9)	681.9 (45.8)	729.9 (38.0)	756.5 (26.6)
Computation	686.7 (58.4)	756.1 (22.0)	785.6 (26.6)	818.9 (27.9)
Concepts & Applications	631.2 (68.8)	722.8 (55.0)	742.6 (32.9)	769.6 (29.5)
TALS				
Document	203.6 (34.1)	248.9 (48.6)	292.9 (37.7)	292.9 (34.6)
Quantitative	211.4 (39.0)	240.0 (26.9)	268.6 (14.6)	304.7 (42.9)
Oral Reading				
Rate[c]	69.9 (30.4)	110.9 (29.9)	133.0 (21.8)	130.6 (25.5)
Decoding[d]	12.5 (8.8)	17.0 (9.8)	23.6 (4.2)	21.6 (7.7)

a n= 8, 10, 13, 11, for ABE 1,2,3 and GED respectively.
b n= 8, 10, 13, 11, for ABE 1,2,3 and GED respectively.
c n= 12, 9, 7, 15, for ABE 1,2,3 and GED respectively.
d n= 14, 9, 7, 16, for ABE 1,2,3 and GED respectively.

Table 7

Mean Gain (Standard Deviation) for ABE and GED Levels by Session, Testing 1 to Testing 3

Day Session (n = 43)

	Level			
Test	ABE 1 (n=8)	ABE 2 (n=11)	ABE 3 (n=13)	GED (n=11)
TABE				
Comprehension	41.6 (48.2)	5.0 (24.2)	0.2 (17.2)	-5.8 (14.6)
Vocabulary	23.6 (62.7)	-6.8 (45.5)	9.1 (36.0)	4.7 (24.9)
Computation	21.5 (37.4)	22.4 (28.9)	5.2 (22.5)	10.9 (27.9)
Concepts & Applications	10.5 (32.7)	35.3 (41.9)	-13.8 (67.9)	4.8 (31.8)
TALS				
Document	20.0 (18.5)	8.2 (31.9)	16.9 (23.2)	20.9 (19.7)
Quantitative	-21.3 (27.0)	-30.0 (21.9)	-7.7 (24.5)	-10.9 (29.5)
Oral Reading				
Rate[a]	11.5 (13.8)	7.9 (7.2)	11.5 (14.5)	2.5 (11.2)
Decoding[b]	-1.1 (2.1)	-3.1 (5.7)	-1.7 (4.2)	-0.6 (3.1)

(table continued on next page)

Table 7 (continued)

Mean Gain (Standard Deviation) for ABE and GED Levels by Session, Testing 1 to Testing 3

Evening Session (n=47)

	Level			
Test	ABE 1 (n=14)	ABE 2 (n=9)	ABE 3 (n=7)	GED (n=17)
TABE				
Comprehension	-11.3 (38.3)	0.4 (31.4)	-3.9 (11.3)	0.1 (6.9)
Vocabulary	23.4 (71.1)	5.3 (28.6)	-3.3 (26.4)	6.8 (21.4)
Computation	23.0 (43.1)	11.0 (14.5)	11.3 (16.8)	13.9 (16.4)
Concepts & Applications	-1.7 (45.5)	13.1 (30.0)	8.1 (22.7)	8.9 (18.0)
TALS				
Document	25.0 (20.7)	25.6 (34.0)	58.6 (39.8)	282.4 (41.8)
Quantitative	8.6 (21.4)	-16.7 (41.2)	-37.1 (20.6)	308.2 (35.9)
Oral Reading				
Rate[c]	3.7 (10.0)	3.8 (14.9)	9.7 (14.3)	10.0 (5.7)
Decoding[d]	-1.3 (4.7)	-2.2 (5.2)	0.7 (4.4)	-0.8 (2.4)

a n= 7, 11, 12, 10, for ABE 1,2,3 and GED respectively.
b n= 8, 11, 12, 10, for ABE 1,2,3 and GED respectively.
c n= 12, 9, 7, 16, for ABE 1,2,3 and GED respectively.
d n= 14, 9, 7, 17, for ABE 1,2,3 and GED respectively.

Table 8

Mean Gain (Standard Deviation) by Session, Testing 1 to Testing 3

Test	Mean Gain		
	Day Session	Evening Session	t
TABE			
Comprehension	7.6 (20.8)	-3.8 (25.1)	1.904
Vocabulary	6.6 (42.1)	9.9 (43.7)	-0.368
Computation	14.1 (28.5)	15.7 (26.8)	-0.276
Concepts & Applications	8.0 (50.0)	6.4 (30.9)	0.181
TALS			
Document	16.3 (23.9)	24.9 (36.5)	-1.336
Quantitative	-16.7 (26.4)	-7.4 (36.1)	-1.395
Oral Reading Rate	8.3 (12.1)[a]	6.9 (10.9)[b]	0.524
Decoding	-1.7 (4.0)[c]	-1.0 (4.1)[d]	-0.745

Note: No t-values are significant at $p < 0.05$. TABE Comprehension approaches significance with a $p = 0.06$. For the day session $n=43$, and for the evening session $n=47$.

[a] $n = 39$
[b] $n = 41$
[c] $n = 40$
[d] $n = 46$

Table 9

Mean Gain (Standard Deviation) for Testing Periods 1 - 3, All Subjects (N=92)

	Mean Gain		
Test	1-2	2-3	1-3
TALS			
Document Literacy	11.0 (29.0) *	9.6 (30.1) *	20.5 (30.9) *
Quantitative Literacy	9.1 (25.5) *	-21.3 (29.8) *	-12.2 (32.1) *
TABE			
Comprehension	13.0 (35.3) *	-10.6 (33.4) *	2.4 (29.7)
Vocabulary	10.2 (36.9) *	- 0.4 (41.0)	9.9 (44.6)
Mathematics Computation	11.7 (28.8) *	4.4 (31.4)	16.1 (28.8) *
Mathematics Concepts & Applications	10.0 (33.9) *	-2.7 (44.5)	7.3 (40.7)
Oral Reading			
Rate	—	—	7.5 (11.4) *
Decoding	—	—	-1.3 (4.0) *

* $p \leq 0.01$

Note: Oral reading and decoding were administered only at testing periods 1 and 3. Changes for TABE and TALS tests are in scale score points; for decoding, in raw score points; and for Oral Reading, in words per minute.

Table 10

Growth Patterns, Testings 1, 2, and 3 for TALS Document Literacy (N=92)

Testing Interval			
1-2	2-3	Number	Percentage
Increase	Increase	28	30.4
Increase	Level	12	13.0
Increase	Decrease	24	26.1
Decrease	Increase	18	19.6
Decrease	Level	0	0
Decrease	Decrease	2	2.2
Level	Increase	4	4.3
Level	Level	2	2.2
Level	Decrease	2	2.2

Table 11

Numbers of Students Who Qualified for a Higher Instructional Level (N= 92)

Total Reading

Class	n	Qualifying Range	Number Qualified for Promotion at Testing 1	Number Qualified for Promotion at Testing 3
Day				
ABE 1	9	< 682	1	2
ABE 2	11	682-724	3	3
ABE 3	13	725-749	4	4
Evening				
ABE 1	15	< 682	5	3
ABE 2	9	682-724	0	1
ABE 3	7	725-749	1	1
Total			14	14

(table continued on next page)

Table 11 (continued)

Numbers of Students That Qualified for a Higher Instructional Level (N= 92)

Total Math

Class	n	Qualifying Range	Number Qualified for Promotion at Testing 1	Number Qualified for Promotion at Testing 3
Day				
ABE 1	9	< 685	4	5
ABE 2	11	685-729	5	7
ABE 3	13	730-763	7	10
Evening				
ABE 1	15	< 685	6	4
ABE 2	9	685-729	4	4
ABE 3	7	730-763	2	5
Total			28	35

Table 12

Calculation of Different Mean Grade Equivalents (GE) for Populations with Identical Mean Scale Scores

Comprehension, Level M (Form 5)

Group 1 (n=10)			Group 2 (n=10)		
Scale	Raw	GE	Scale	Raw	GE
685	20	4.4	560	8	2.1
603	11	2.4	582	10	2.3
765	38	9.9	603	11	2.4
650	15	3.1	617	12	2.5
721	29	5.8	629	13	2.7
582	10	2.3	758	37	8.8
776	39	10.9	765	38	9.9
659	16	3.3	776	39	10.9
725	30	6.0	821	40	10.9
685	20	4.4	717	28	5.5
Mean Scale:		685	Mean Scale:		685
Mean GE:		5.25	Mean GE:		5.80

APPENDIX B: FIGURES

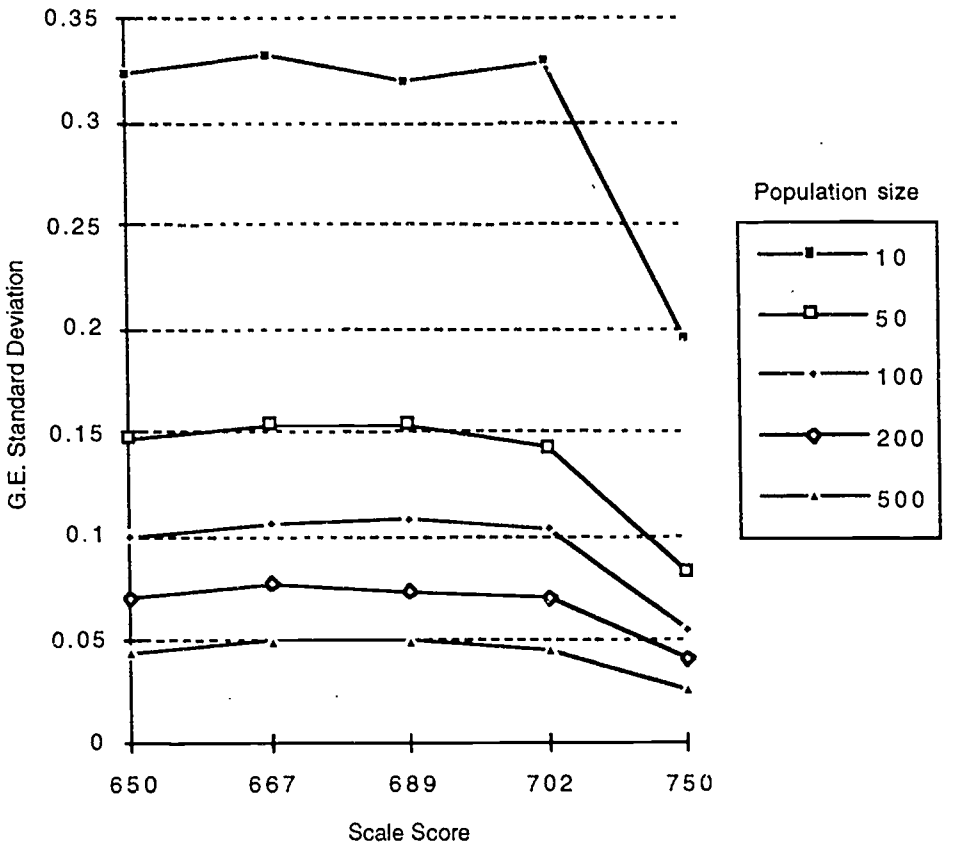

Figure 1. Grand Equivalent Variability for Different Population Sizes and Scale Scores (Scale Score s.d.=50)

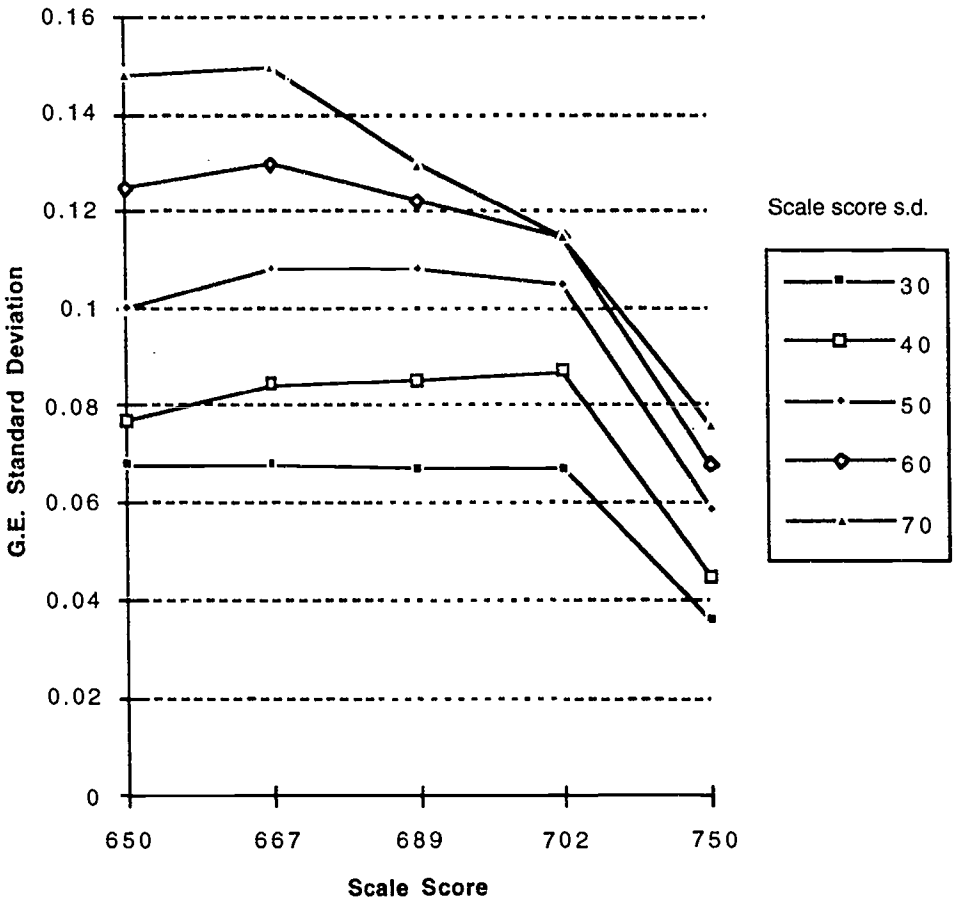

Figure 2. Grade Equivalent Variability for Different Scale Score Means and Standard Deviations (N=100)

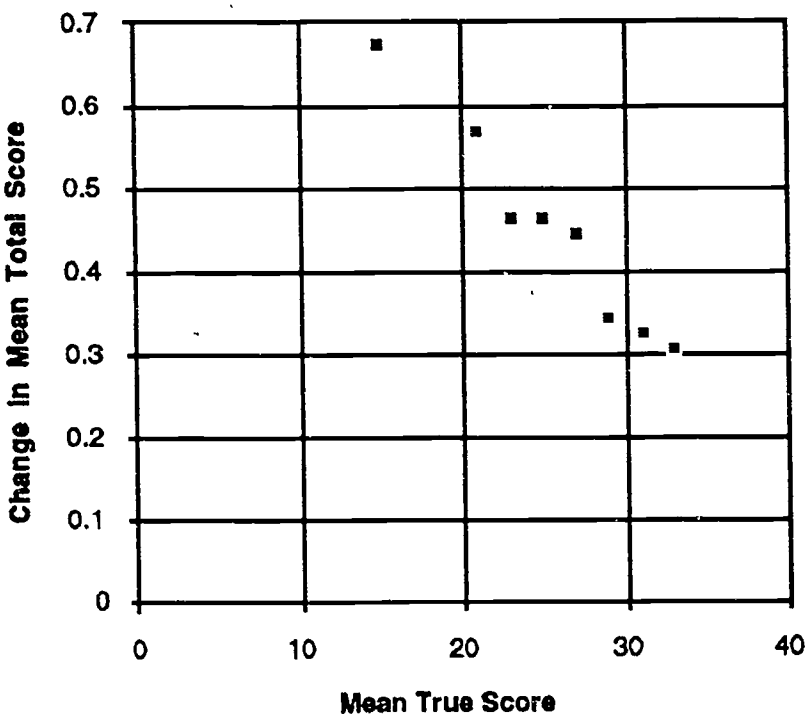

Figure 3. Mean Change in Total Score Due to Attrition of the Top One-half of a Population for Different True Scores (N=200)