DOCUMENT RESUME

ED 393 011 CE 071 241

AUTHOR Sabatini, John P.; And Others
TITLE Comparing Applied Literacy and Basic Skills Tests as Measures of Adult Literacy Performance.
INSTITUTION National Center on Adult Literacy, Philadelphia, PA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO NCAL-TR-95-03
PUB DATE Nov 95
CONTRACT R117Q00003
NOTE 46p.
AVAILABLE FROM National Center on Adult Literacy, University of Pennsylvania, Publications, 3910 Chestnut Street, Philadelphia, PA 19104-3111 (order no. TR95-03).
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adult Basic Education; *Adult Literacy; Adult Students; Comparative Analysis; Literacy Education; Predictive Measurement; Predictive Validity; *Student Evaluation; *Testing; *Test Selection; *Test Validity
IDENTIFIERS *Tests of Adult Basic Education; *Tests of Applied Literacy Skills

ABSTRACT

The Tests of Adult Basic Education (TABE) is a widely used multiple-choice test battery of basic skills in reading, language, and mathematics. The Test of Applied Literacy Skills (TALS) is an applied literacy battery consisting of document, prose, and quantitative literacy tests. The central issue in this study was the relationship of the TABE and TALS to each other as measures of change and as guides for instruction. All students for the project attended adult basic education (ABE) or General Educational Development (GED) classes at Rochambeau Schools in White Plains, New York, during the 1991-92 school year. The majority of the analyses were performed on the 168 students who completed all the initial TABE and TALS tests. Students were tested prior to placement into ABE/GED levels at the beginning of the 1992 school year, then again at mid-year and again at the end of the year. The main finding was that the TABE Mathematics Concepts and Applications Test scores were a stronger predictor of the TALS Document Test scores than the TABE Reading Comprehension scores. The pattern of results for predicting Prose Literacy scores mirrored those of Document Literacy, with Reading Comprehension taking on the role of strongest predictor. This result was interpreted as indicating a strong problem-solving component common to both tests that was not equivalent to basic skills reading measures. Some common sense starting points were proposed for deciding which tests to use for adult literacy assessment, including selecting tests that match the goals of instructional programs. (Appendixes include a list of 33 references and 11 data tables.) (YLB)

NCAL

NATIONAL CENTER ON ADULT LITERACY

COMPARING APPLIED LITERACY AND BASIC SKILLS TESTS AS MEASURES OF ADULT LITERACY PERFORMANCE

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

NCAL Technical Report TR95-03
November 1995

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

COMPARING APPLIED LITERACY AND BASIC SKILLS TESTS AS MEASURES OF ADULT LITERACY PERFORMANCE

John P. Sabatini
Richard L. Venezky
Page S. Bristow
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 TR95-03
NOVEMBER 1995

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. R117Q00003) 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 *(PB92-01, 7 pages)*

Oct 1992 *Life-Span and Life-Space Literacy: Research and Policy in National and International Perspectives*
Daniel A. Wagner *(OP92-01, 15 pages)*

Oct 1992 *Expanding Theories of Adult Literacy Participation*
Karen Reed Wikelund, Stephen Reder, & Sylvia Hart-Landsberg *(TR92-01, 30 pages)*

Oct 1992 *Invitations to Inquiry: Rethinking Staff Development in Adult Literacy Education*
Susan L. Lytle, Alisa Belzer, & Rebecca Reumann *(TR92-02, 44 pages)*

Dec 1992 *Developing the Professional Workforce for Adult Literacy Education*
Susan L. Lytle, Alisa Belzer, & Rebecca Reumann *(PB92-02, 11 pages)*

Jan 1993 *The Analysis of Adult Literacy: Problems in Factor Analysis Bib-Spiralled Item Administration*
David Kaplan *(OP93-01, 18 pages)*

Mar 1993 *The Impact of Workplace Literacy Programs: A New Model for Evaluation of Workplace Literacy Programs*
Larry Mikulecky & Paul Lloyd *(TR93-02, 180 pages)*

Mar 1993 *Literacy and Machines: An Overview of the Use of Technology in Adult Literacy Programs*
Terilyn C. Turner *(TR93-03, 86 pages)*

Jun 1993 *Myths and Misconceptions in Adult Literacy: A Research and Development Perspective*
Daniel A. Wagner *(PB93-01, 10 pages)*

Jun 1993 *Literacy and Development: Rationales, Assessment, and Innovation*
Daniel A. Wagner *(IP93-01, 50 pages)*

Jun 1993 *Early Childhood, Family, and Health Issues in Literacy: International Perspectives*
Laurel D. Puchner *(IP93-02, 45 pages)*

Sep 1993 *What Makes Workers Learn? The Role of Incentives in Workplace Education and Training*
Donald Hirsch & Daniel A. Wagner (Eds.) *(IP93-03, 243 pages)*

Sep 1993 *Prison Literacy: Implications for Program and Assessment Policy*
Anabel Newman, Warren Lewis, & Carolyn Beverstock *(TR93-01, 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. *(TR93-04, 150 pages)*

Sep 1993 *What Can Employers Assume About the Literacy Skills of GED Graduates?*
David Kaplan & Richard L. Venezky *(TR93-05, 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 *(TR93-07, 101 pages)*

Sep 1993 *When Less Is More: A Comparative Analysis for Placing Students in Adult Literacy Classes*
Richard L. Venezky, Page S. Bristow, & John P. Sabatini *(TR93-08, 46 pages)*

Sep 1993 *Metacognitive Aspects of Adult Literacy*
Scott G. Paris & Andrea Parecki *(TR93-09, 44 pages)*

Nov 1993 *Teamwork and Literacy: Learning From a Skills-Poor Position*
Sylvia Hart-Landsberg & Steve Reder *(TR93-06, 63 pages)*

Nov 1993 *Motivations for Learning: Voices of Women Welfare Reform Participants*
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 *Issues and Challenges in Adult Numeracy*
Iddo Gal *(TR93-15, 62 pages)*

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

Apr 1994 *Measuring Gain in Adult Literacy Programs*
Richard L. Venezky, Page S. Bristow, & John P. Sabatini *(TR94-12, 24 pages)*

Apr 1994 *Understanding Family Literacy: Conceptual Issues Facing the ...*
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, Ludo C. P. Scheffer, & Joel Hardman *(TR94-04, 84 pages)*

Jun 1994 *Literacy Transfer: A Review of the Literature*
Larry Mikulecky, Peggy Albers, & Michele Peers *(TR94-05, 21 pages)*

Jun 1994 *Instruction and Assessment for Limited-English-Proficient Adult Learners*
Ronald W. Solórzano *(TR94-06, 33 pages)*

Jun 1994 *Early Warning Signs of Functional Illiteracy: Predictors in Childhood and Adolescence*
Nazli Baydar, Jeanne Brooks-Gunn, & Frank F. Furstenberg *(OP94-01, 13 pages)*

Jul 1994 *Use It or Lose It? The Problem of Adult Literacy Skill Retention*
Daniel A. Wagner *(TR94-07, 27 pages)*

Jul 1994 *Technology: New Tools for Adult Literacy,* Videoconference Participant Materials
Joyce Harvey-Morgan, Christopher Hopey, & R. Karl Rethemeyer (Eds.)
(PR94-01, 58 pages)

Sep 1994 *Supply and Demand for Literacy Instruction in the United States*
Richard L. Venezky & Daniel A. Wagner *(TR94-10, 13 pages)*

Sep 1994 *The Professionalization of the Teacher in Adult Literacy Education*
Timothy Shanahan, Maureen Meehan, & Stephen Mogge *(TR94-11, 20 pages)*

NCAL publications to date (continued)

Sep 1994 *The Role of Literacy in the Wealth of Individuals and Nations*
Sue E. Berryman *(TR94-13, 15 pages)*

Oct 1994 *Abilities and Competencies in Adulthood: Life-Span Perspectives on Workplace Skills*
Jacqui Smith & Michael Marsiske *(TR94-12, 36 pages)*

Oct 1994 *Proceedings. Conference on Adult Mathematical Literacy*
Iddo Gal & Mary Jane Schmitt (Eds.) *(PR94-02, 130 pages)*

Nov 1994 *Literacy and Adults With Developmental Disabilities*
Karen A. Erickson, David A. Koppenhaver, & David E. Yoder *(TR94-15, 31 pages)*

Nov 1994 *Adult Basic Skills in OECD Countries: Policy Issues and a Research Agenda*
David Stern & Albert Tuijnman *(IP94-01, 12 pages)*

Dec 1994 *Who Counts in Adult Literacy Programs? A National Survey of Numeracy Education*
Iddo Gal & Alex Schuh *(TR94-09, 20 pages)*

Dec 1994 *Adult Numeracy Instruction: A New Approach,* Videoconference Participant Packet
Iddo Gal, Lynda Ginsburg, Ashley Stoudt, R. Karl Rethemeyer, & Caroline Brayer Ebby
(PR94-04, 58 pages)

Dec 1994 *Literacy and Welfare Reform: Are We Making the Connection?*
Elena Cohen, Susan Golonka, Rebecca Maynard, Theodora Ooms, & Todd Owen *(TR94-16, 47 pages)*

Jan 1995 *Self-Assessed Skill Needs and Job Performance*
Peter Cappelli & Nikolai Rogovsky *(TR94-08, 12 pages)*

Jan 1995 *Literacy and Voting Behavior: A Statistical Analysis Based on the 1985 Young Adult Literacy Survey*
David Kaplan & Richard L. Venezky *(TR94-14, 13 pages)*

Jan 1995 *Literacy and Older Adults in the United States*
Gail Weinstein-Shr *(TR94-17, 25 pages)*

Jan 1995 *Proxy Measurement of Adult Basic Skills: Lessons From Canada*
T. Scott Murray *(TR94-18, 18 pages)*

Jan 1995 *Using Large-Scale Assessment Results to Identify and Evaluate Generalizable Indicators of Literacy*
Irwin S. Kirsch & Ann Jungeblut *(TR94-19, 14 pages)*

Jan 1995 *Native Literacy and Language Roundtable Proceedings*
Joyce Harvey-Morgan (Ed.) *(PR94-03, 26 pages)*

Mar 1995 *The Military Experience and Workplace Literacy: A Review and Synthesis for Policy and Practice*
Thomas Sticht *(TR94-01, 78 pages)*

Apr 1995 *What Works? Literacy Training in the Workplace,* Videoconference Participant Materials
Joyce Harvey-Morgan (Ed.) *(PR95-01, 38 pages)*

May 1995 *Adult Literacy: The Next Generation*
An NCAL White Paper *(TR95-01, 29 pages)*

Nov 1995 *Making Sense of Technology Terminology for Adult Literacy: A Glossary and Annotated Bibliography*
Alycia Donohoe, Joseph Campbell, Camille Ciggs, R. Karl Rethemeyer, & Christopher Hopey *(PG95-01, 47 pages)*

Nov 1995 *Technology Planning for Adult Literacy*
Christopher E. Hopey & Joyce Harvey-Morgan *(PG95-02, 45 pages)*

Nov 1995 *Funding Technology in Adult Literacy*
Christopher E. Hopey & Joyce Harvey-Morgan *(PG95-03, 64 pages)*

Nov 1995 *Making the Right Choice: Evaluating Computer Software and Hardware for Adult Literacy Instruction*
Christopher E. Hopey, R. Karl Rethemeyer, & Jennifer A. Elmore *(PG95-04, 54 pages)*

Nov 1995 *Joining the On-Line Community: An Introduction for Adult Literacy*
R. Karl Rethemeyer *(PG95-05, 146 pages)*

NCAL publications to date (continued)

Nov 1995 *Comparing Applied Literacy and Basic Skills Tests as Measures of Adult Literacy Performance*
Richard L. Venezky, John P Sabatini, & Page S. Bristow *(TR95-03, 48 pages)*

Nov 1995 *Standards for Adult Literacy: Focal Points for Debate*
Regie Stites, Ellen Foley, & Daniel A. Wagner *(TR95-04, 34 pages)*

Nov 1995 *Adult Literacy in the United Kingdom: A History of Research and Practice*
Brian V. Street *(TR95-05, 54 pages)*

Nov 1995 *What Does "100% Juice" Mean? Exploring Adult Learners' Informal Knowledge of Percent*
Lynda Ginsburg, Iddo Gal, & Alex Schuh *(TR95-06, 44 pages)*

Nov 1995 *Learning to Read: Literacy Acquisition by Children and Adults*
Charles A. Perfetti & Maureen A. Marron *(TR95-07, 56 pages)*

Apr 1996 *The Infrastructure of Adult Literacy Education: Implications for Policy*
Hal Beder *(TR96-01, 32 pages)*

Apr 1996 *Evaluation of Workplace Literacy Programs: A Profile of Effective Instructional Practices*
Larry Mikulecky & Paul Lloyd *(TR96-03, 58 pages)*

Apr 1996 *A Review of Recent Workplace Literacy Programs and a Projection of Future Challenges*
Larry Mikulecky, Paul Lloyd, Lisa Horwitz, Sharon Masker, & Patti Siemantel
(TR96-04, 54 pages)

Information on ordering of NCAL publications may be addressed to Dissemination at NCAL.
Revised April 10, 1996

NCAL EDITORIAL REVIEW COMMITTEE (CONTINUED)

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, of the 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, and 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, White Plains Adult Literacy Program, White Plains Adult and Continuing Education Program 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 paper.

TABLE OF CONTENTS

COMPARING APPLIED LITERACY AND BASIC SKILLS TESTS AS MEASURES OF ADULT LITERACY PERFORMANCE

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

Abstract

The Tests of Adult Basic Education (TABE) is a widely used multiple-choice test battery of basic skills in reading, language, and mathematics. The Tests of Applied Literacy Skills (TALS) is an applied literacy battery consisting of document, prose, and quantitative literacy tests. The central issue in the present study is the relationship of the TABE and TALS tests to each other as measures of change and as guides for instruction. Students were tested prior to placement into ABE/GED levels at the beginning of the 1992 school year, then again at mid-year and again at the end of the year. The main finding was that the TABE Mathematics Concepts and Applications Test scores were a stronger predictor of the TALS Document Test scores than the TABE Reading Comprehension scores. The pattern of results for predicting Prose Literacy scores mirrored those of Document Literacy, with Reading Comprehension taking on the role of strongest predictor. This result was interpreted as indicating a strong problem-solving component common to both tests that was not equivalent to basic skills reading measures. The report concludes with some common sense starting points for deciding which tests to use for adult literacy assessment including selecting tests that match the goals of instructional programs.

INTRODUCTION

At the present time, two distinctly different types of standardized tests are available to adult literacy programs for measuring learner change: basic skills tests (e.g., reading, writing, numeracy) and applied literacy tests (sometimes referred to as functional literacy tests). The former are generally developed from factor structures representing the skills required for competent performance (e.g., vocabulary, literal comprehension, inferential comprehension, and critical analysis for reading). Although the particular skills may vary across tests, the basic procedures for defining skills are similar. In contrast, applied literacy tests are not defined by underlying skills but instead by document or text types and by the operations commonly performed on these documents or texts in applied settings.

A typical applied literacy test is based on the literacy demands of a set of specified domains (e.g., home, civic activities, shopping, work). Within each of these domains, typical documents will be identified along with the tasks that individuals must perform with them. Then, items are generated to simulate the documents and tasks, with document or task complexity varied to achieve a mix of item difficulties. Whatever skills might be required to perform the selected tasks is not a direct consideration in item selection, but these might be identified post hoc. The central issues in this report are whether the two types of tests are measuring the same or similar underlying abilities and whether they can be used interchangeably as measures of change. A set of secondary issues concerns the implications that instructors and students can draw from student performance on these two tests. In this study, we compare examples of these two approaches to adult literacy testing, using as exemplars the Tests of Adult Basic Education (TABE) (CTB/McGraw-Hill, 1987b) and the Tests of Applied Literacy Skills (TALS) (Simon & Schuster, 1990).

CONCEPTUAL OVERVIEW

NORM-REFERENCED ACHIEVEMENT TESTING IN ADULT EDUCATION

The nature of achievement test design is important because the choice of which test to administer drives instructional practices. Although the use of standardized achievement tests in schools is over 75 years old, the expanding use of norm-referenced tests can be linked to national attempts to monitor and control education. To take two well known examples, the use of norm-referenced tests expanded as a consequence of federal legislation in the 1950s and 1960s promoting math and science, and again in the 1980s as a consequence of "back to basics" educational reform initiatives (Calfee & Hiebert, 1991; Haney & Madaus, 1989; Madaus & Kelleghan, 1992).

Adult education 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 recruiting, 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 yet 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 states plans for programs to "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).

Although federal guidelines do not require basic skill tests, many state departments of education and adult programs use basic skills curriculum and instruction. The long tradition of basic skills instruction in schools makes it easier to plan and coordinate a diagnostic-prescriptive curriculum based on basic skills tests. Instructors and adult learners are familiar with their roles and responsibilities in basic skills-structured classrooms. With many adult programs using open-entry/open-exit policies, basic skills instruction provides a stable, individualized scope and sequence framework.

Norm-referenced, basic skills tests are themselves familiar cultural artifacts to most U.S. and world citizens. Educators, counselors, and administrators commonly discuss test performance in terms of percentile rankings, stanines, or grade-equivalent scores, rather than in absolute terms such as poor, fair, or good. Although often dreaded or feared, success on basic skills-style tests for the majority of adult students, both U.S. and foreign born, is associated with educational progress; at one time or another, basic skills tests were likely used to measure their children's or their own progress. Many want to prove to themselves and others that they can succeed on these tests. The institutional tradition and history of basic skills tests and instruction cannot be treated as trivial or inconsequential when considering new assessment and instructional approaches.

The most commonly used adult basic skills test is the TABE. In a recent nationwide survey of 427 ABE teachers, the TABE was mentioned by 59% as the only test regularly used or as one of several tests in regular use. The nearest competitor was the Adult Basic Learning Education (ABLE), mentioned by fewer than 5% of the respondents (Ehringhaus, 1991). The National Evaluation of Adult Education Programs reported that the TABE is used for initial diagnosis of learner needs by 68% of the 2,619 programs responding to their survey (Development Associates, 1992). New York City, as an example, requires reporting of TABE gain scores for all non-ESOL (English for Speakers of Other Languages) basic education students in Adult Literacy Initiative programs. Students must be tested within their first 12 hours of instruction and then retested after 50, 100, or 200 hours, according to the type of instruction they are receiving (tutorial, less than 9 hours/week, or greater than 9 hours/week, respectively) (Metis Associates, 1991).

Reliance on basic skills tests and instruction, however, is not universally accepted as the best approach to use in adult programs. Criticism of basic skills tests is widespread, though most research studies have concerned themselves with their use in primary and secondary schools (e.g., Archbald & Newmann, 1988). Basic skills tests are criticized for promoting basic skills instruction to the neglect of other instructional strategies and learning goals. The traditional conventions of the tests (including the multiple-choice format, long tests comprised of short items, the strict time limits, and the grade-equivalent scales) and the quality of passages have been questioned. Teachers have been

found to have difficulty in interpreting standardized test results accurately, thus limiting their use as guides to instruction (Calfee & Hiebert, 1991; Shepard, 1991).

APPLIED LITERACY TESTS AND THE TALS

Applied literacy tests are relatively new and unfamiliar, both to the general public and to educators, in comparison to the more traditional, basic skills tests. However, an applied literacy test was developed as early as 1937 by Guy Buswell at the University of Chicago and a steady stream of studies has been conducted in this genre since the 1970s (see Stedman & Kaestle, 1987 for a historical review). The California Assessment System (CASAS) (1989) is an example of one applied literacy test system that has been in use statewide since the early 1980s. The applied literacy approach underlies most recent national surveys of adults.

The definition of literacy governing the development of recent literacy surveys by the Educational Testing Service (ETS), including the Young Adult Survey, Department of Labor Survey, and the National Adult Literacy Survey (Kirsch & Jungeblut, 1986; Kirsch, Jungeblut, & Campbell, 1991, 1992; Kirsch, Jungeblut, Jenkins, & Kolstad, 1993) is: "Using printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential" (Kirsch, Jungeblut, & Campbell, 1991, p. 1-5). This definition emphasizes the application of literacy skills for specific purposes in specific contexts. Three distinct literacy constructs (prose, document, and quantitative) have been developed to describe the abilities underlying performance on different applied literacy item types, though the independence of these three scales is in question (Reder, in press).

The differences between applied literacy and basic skills approaches are immediately obvious from an examination of the two test types. The applied literacy items are constructed from representative literacy materials and common tasks from every day life, in contrast to the short, textbook passages or story problems of basic skills tests. The validity of reading comprehension and vocabulary as constructs of reading ability has the weight of history and tradition to argue for it, while the validity of prose, document, and quantitative ability as constructs of literacy is still open to debate. Nonetheless, applied literacy constructs have been defined and vigorously defended by their proponents (Guthrie, Britton, & Barker, 1991; Kirsch & Jungeblut, 1986; Kirsch & Mosenthal, 1990a). Their use in national assessments of adults such as the National Adult Literacy Survey (NALS) (Kirsch, Jungeblut, Jenkins, & Kolstad, 1993) has also bolstered their credibility. However, their widespread acceptance and adoption in adult education has been slow, in part no doubt because students and educators alike must learn to reconcile their traditional views of basic skills with the applied literacy approach.

It is important to remember that the psychometrics underlying the TALS is similar to that of the TABE; both use item response theory to select and calibrate items. The TALS, like the TABE, is a standardized, norm-referenced test battery. Both the TABE and TALS employ item response theory to locate items and individuals onto a single item response theory (IRT) scale. However, the TALS publishers have calibrated their IRT scale to the scale derived from the

Young Adults Literacy Survey.* They have not attempted to link the scale to elementary and secondary school norm-referenced tests; hence, there is no grade equivalency scale. Like the new literacy constructs of the TALS, the reliance on a scale linked to a national survey of young adults requires some getting-used-to by educators. An individual's scale score on the TALS represents the level of difficulty of printed materials and associated tasks that the adult is likely to be able to perform competently. The structure of the materials, the nature of the task, the familiarity with the content, and their interactions determine the difficulty of any test item. It is unclear, however, how to use these characteristics to determine the difficulty of an item, a priori, without further empirical validation. This scale ranks students; it does not imply instructional levels.

The TALS approach represents a slightly different way to interpret test results, but not a radically different way to construct tests. In contrast to the basic skills approach, the applied literacy approach describes the individuals or groups in terms of the types of literacy tasks they can perform, not by the underlying set of skills they possess. However, without a skill hierarchy (which in the basic skills approach roughly corresponds to an instructional program), using the TALS as a classroom tool for diagnosis and prescription of instruction is complicated, though some suggestions have been offered (Kirsch & Mosenthal, 1989, 1990b).

Criticism of applied literacy tests has taken two forms. The first argues that, in implementing the relativistic literacy definition given above, the resulting tests have been biased against ethnic and other demographic subgroups in the nation—groups that may seek self-improvement through education but have conflicting cultural values or personal goals (Gee, 1991; Levine, 1986). These critics argue that the outcome of the applied testing approach is to produce mainstream, middle-class citizens and productive workers to the exclusion of other educational goals. This line of criticism suggests that it is impossible to design (or agree upon) a set of test items that measure equivalent "applied" skills for various demographic and cultural groups across time and space (e.g., Levine, 1986).

A second line of criticism concerns the unconventional instructional theory underlying the design of test items. Unlike basic skills tests, item difficulty is manipulated independently of any hierarchy of knowledge and skills the individual may have achieved. The TALS designers argue that its prose, document, and quantitative literacy scales are independent, unidimensional traits of literacy. The cognitive subprocesses identified as underlying performance on these scales are locating, integrating, and generating information for prose; locating, cycling, and evaluating information for document literacy; and the four basic math operations and combinations of them for quantitative literacy. This framework provides sparse instructional or diagnostic guidance for the classroom teacher. Furthermore, in a recent reanalysis of the items on the three scales used in the YALS and NALS, Reder (in press) concluded that there is little empirical support for separate prose, document, and quantitative literacy scales.

How Well Do Psychometric Designed Tests Measure Different Psychological Constructs?

Although the technology for constructing, norming, establishing reliability, and scaling tests has become ever more sophisticated, some question whether psychometric test development techniques have kept pace with the theories of instruction and

* Although IRT scaling was used to select items and to scale them, norms reported for converting number correct to scale scores appear to be derived from a single-parameter scaling model such as the Rausch model.

psychology that should fundamentally undergird test design (Calfee & Hiebert, 1991; Nichols, 1994; Shepard, 1991; Snow & Lohman, 1989). Many of the assumptions underlying psychometric theory are being reevaluated (Linn, 1989; Messick, 1989, 1994; Moss, 1992, 1994). In recent years, norm-referenced, basic skills tests have been criticized because their structure continues to favor the measuring of discrete behavioral objectives over cognitive outcomes (Calfee & Hiebert, 1991; Snow & Lohman, 1989). Good reading comprehension, for example, is no longer considered to be the automatic consequence of acquiring a single hierarchy of subskills. For the most part, however, it is still easier to design a basic skills achievement test that assesses recall of factual knowledge and mastery of basic skills than it is to design one that assesses the efficiency of cognitive processes or problem-solving abilities.

The design of contemporary standardized, norm-referenced tests reflects the traditional purposes and contexts for which they are used. Norm-referenced tests are used in schools for placement, measuring student change, diagnosing individual differences, evaluating programs, and surveying abilities; often the same test is used simultaneously to achieve more than one of these purposes. By employing multiple-choice formats and computer scoring, today's tests are efficient, cost-effective tools for large-scale data collection, management, and analysis. They are also very efficient at ranking students against each other and norming populations. Only recently have performance assessments with constructed response items (i.e., all items that are not multiple choice) become available for many of these purposes.

The familiar external format of contemporary tests belies important internal, technical, and theoretical innovations. For example, the selection and calibration of items for many contemporary tests, including the most recent TABE revision, are based on applications of item response theory (IRT) (Lord, 1980), pioneered at ETS and used also in the construction of the TALS and Scholastic Aptitude Test. To apply item response theory to test construction, one must assume that a unidimensional underlying latent trait is being measured. For example, in the TABE, reading comprehension and vocabulary ability are assumed to be unidimensional, latent traits of reading ability, while in the TALS, prose, document, and quantitative ability are assumed to be unidimensional, latent traits of literacy.

These item selection techniques often increase a test's reliability by identifying troublesome items, a practice that at the same time virtually guarantees the appearance of unidimensionality. Another result is high intercorrelations between most psychometrically constructed subtests in a test battery, within and across domains (Calfee & Hiebert, 1991). When subtests are highly intercorrelated it is hard to tell whether it is because individuals have comparable levels of achievement across domains or whether it is an artifact of the test design. This criticism is true of both the TABE and the TALS, which use similar psychometric test design techniques. Item difficulty may depend as much on idiosyncratic features as on the labeled objective. Thus, in evaluating a test one must examine the theory underlying item design as well as its statistical properties. In a basic skills approach, the test designer has a theory of what items should be difficult based on a hierarchy of skills, while in the applied approach, face validity (i.e., authentic-looking materials and tasks) takes precedence over skill hierarchies. Both, however, control the difficulty of items by manipulating their length, complexity, and so forth.

Basic skills and applied literacy tests are clearly different on the surface level, and the surface level differences are themselves significant for educators and students. Some educators embrace the tradition, familiarity, and continuity of academic-style, basic skills tests, especially for students continuing into higher levels of formal education. Others look to the face validity of the relatively new-looking applied literacy tests to improve student motivation, especially for those with nonacademic goals. However, do basic skills and applied literacy tests differ on a deeper level? That is, are different underlying factors contributing to different performances on each? Or is the measurement of the deeper level ability differences constrained by the test design technology shared by both tests? If deeper differences turn out to be insignificant, that is, if performance on applied tests is reducible to basic skills ability, then surface level differences should be considered the defining characteristics of the tests.

CENTRAL ISSUES

The central issue in the present study is the relationship of the TABE and TALS tests to each other as measures of change and as guides for instruction. The TABE Vocabulary and Reading Comprehension Tests and the TALS Document and Prose Literacy Tests will be of primary concern throughout this report. However, as will become clear, the TABE Math Concepts and Applications and TALS Quantitative Literacy Tests will figure prominently in the interpretation of analysis results. Nonetheless, it is not the goal of this study to explain numeracy and we will limit our discussion to implications for measuring reading and document processing.

The reasons for highlighting the TABE reading tests are clear. They are the measures most often required by state and program administrators for reporting progress and most often used for initial placement. Also, their resemblance to primary and secondary school achievement tests makes them familiar to instructors and students. Finally, most ABE instruction is oriented towards improving primarily basic reading skills, secondarily writing and math skills, and on occasion life skills. The reasons for focusing on the TALS Document Test are that (a) processing and understanding documents in American society is as widespread a literacy practice as comprehending continuous prose and is of more practical importance than reading fiction, and (b) the literacy skills for processing document information are potentially the most different from the basic skills that underlie reading continuous prose or solving mathematics problems in the TABE. The locating, generating, and evaluating tasks embedded in the Prose Literacy Test also differ somewhat from the literal and inferential questions of the TABE Reading Comprehension Test, though not as dramatically as the Document items.

We will test the hypothesis that basic skills in reading underlie performance on applied literacy tasks. That is, a better reader should be better able to handle applied literacy tasks of everyday life. Put another way, basic skills should transfer to applied literacy tasks. This represents the rationale of both the TABE designers and the adult programs that focus on basic reading skills instructional approaches. Consequently, performance on the TABE Reading Tests (Vocabulary and Reading Comprehension) should be the best predictors of the applied literacy scores on the TALS. Furthermore, basic skills should be causally prior to applied literacy performance; that is, current basic skills scores should predict future levels of applied literacy ability.

The reverse hypothesis could also be true. Experience performing applied literacy tasks could incidentally increase an individual's basic reading skill ability or the capability to benefit from basic skills instruction. This is analogous to learning a foreign language by immersion into the community of a native speaking population, then later studying the grammar of the language. In such a scenario, applied literacy abilities would predict

future basic skill reading scores. Although this was not an original concern that motivated this study design, we will attempt to explore this possibility as well.

METHODS

STUDENTS

All 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 English as a second language (ESL), job skills, general continuing education, workplace literacy, neighborhood literacy, and family literacy in cooperation with White Plains elementary schools. Many of the students for this project were graduates of the school's ESOL programs. All of the students 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.

In this report, the majority of analyses were performed on the 168 students who completed all of the initial TABE and TALS tests (Starters). Initially, 213 students registered for classes, but only 168 completed the initial tests. Students remaining in the program were retested once at about the middle of the school year and then again near the end according to guidelines set by New York State (see Procedures section). Some analyses were performed on the 123 students remaining at the middle of the year (Midtermers). Some analyses were also performed on the 92 students who completed all of the TABE and TALS tests for the three testing periods (Persisters). The Starters 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 state election, almost none reported any health-related handicaps, one quarter 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.

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 five are certified teachers with a mean of 9 years and a range of 2 to 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 13 to 49 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; smaller amounts of time were spent on writing and 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.

Table 1 (see Appendix) 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 the Job Skills Education Program (JSEP) materials, whereas evening students had access to an optional, additional single night of JSEP instruction each week.

INSTRUMENTS

Tests of Adult Basic Education (TABE). The TABE is a battery of norm-referenced tests that require multiple-choice responses. 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. According to the publisher, the purpose of the battery is not to test specific life skills, but to test basic skills in the context of life skill tasks. The TABE Vocabulary Test measures mastery of synonyms, antonyms, homonyms, affixes, and words in context. The Reading Comprehension Test measures literal, inferential, and critical comprehension. The Mathematics Computation Test measures ability to do addition, subtraction, multiplication, and division. The Mathematics Concepts and Applications Test measures numerical concepts such as place-value, number sentences, and geometry, as well as the reasoning skills needed for practical problem solving.

Each test has four graduated but overlapping levels, (Easy, Medium, Difficult, and Advanced) with alternate forms available for each. Also available is a Locator Test for determining the appropriate level for full-scale testing. This Locator includes 25 multiple-choice vocabulary items and 25 multiple-choice arithmetic items and requires 37 minutes for administration. Test scores are converted by table lookup to scale scores that were derived from IRT scaling. Norming of the tests was done with about 6,300 examinees, divided among ABE enrollees, adult offenders, juvenile offenders, and vocational/technical school enrollees. Internal reliabilities of the separate TABE tests as measured by the Kuder-Richardon Formula (20) are mostly in the .80-.90 range (CTB/McGraw-Hill, 1987a). Validity, as measured through correlations with comparable GED tests, is moderate (.43-.64). However, this is probably not a fair assessment of validity due to differences in factor structures between the GED tests and the most closely matching TABE tests.

Tests of Applied Literacy Skills (TALS). The TALS is a battery of norm-referenced tests that use applied literacy tasks to measure an adult's ability to apply literacy skills in the contexts commonly encountered in everyday living. These instruments were developed from the experiences gained by ETS with the Young Adult Literacy Survey and the Department of Labor Literacy Survey (Kirsch & Jungeblut, 1986; Kirsch, Jungeblut, & Campbell, 1991). TALS items require short answer 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. According to the TALS publisher, the Document Literacy Test measures ability to identify and use information located in materials such as charts, indices, forms, and tables. The Quantitative Literacy Test requires performing arithmetic operations, alone or sequentially, using various printed materials. For example, respondents calculate wages, complete an order form, and determine the amount of interest from a loan advertisement. The Prose Literacy Test measures ability to read and interpret texts excerpted from newspaper articles, magazines, pamphlets, and books. Each test is divided into two sections that need to be scored separately. The two section scores then serve as indices to retrieve a single scale score from a publisher-supplied table. Scale scores were derived through IRT scaling from a norming study that involved 3,105 adults. Each test has two alternate forms and each test is to be administered in a 40-minute period. In this study, the TALS Document and Quantitative Tests were administered in Testings 1 and 2; all three TALS Tests were administered in Testing 3. TALS internal consistency reliabilities, as reported in the TALS technical manual, vary from .88 (Quantitative Literacy, Form B) to .92 (Prose and Document Literacy, both Form A). No external validity measures are reported.

PROCEDURES

Timing of Testing. The design called for three repeated measures of each student at regular intervals in the school year. The test 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. 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 2 (see Appendix).

Test 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 (E or D), based on their TABE Locator Test score. Students who scored less than 12 on the Locator Test were considered nonreaders and thus did not take the TABE (or TALS) battery. Students who received raw scores between 13 and 29 were given the E (Easy) level and students 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, 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 administer the oral reading tasks as described below.

In addition to the TABE and TALS, several other measures were administered to the students including a background questionnaire administered in the first test session, oral passage reading and word decoding tasks administered at test sessions one and three, and test evaluation forms administered after every test taken. Except for the background information, these measures were not central to the issues examined in this study and will not be discussed further.

Scoring. 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. Twenty percent of the TALS tests were rescored by another ETS-trained examiner; the interrater reliability was 99%.

RESULTS AND DISCUSSION

The results and discussion section will be organized as follows. The first three sections discuss descriptive statistics, statistical assumptions underlying regression analysis, and intercorrelations. The next four sections report on various regression models for predicting the TALS Document, Quantitative, and Prose Literacy Test scores, and the TABE Reading Comprehension Test scores.

In this study, the first question we ask is what is the total variance accounted for by the predictor variables and what are their unique contributions. To answer this question, we use standard multiple regression and report the adjusted squared multiple regression coefficient (R^2) and the squared semipartial correlation coefficient (sr^2), which can be interpreted as the unique variance accounted for by each of the variables in the model. The total variance (R^2) less the unique variance (sum of all sr^2s) leaves the variance shared by one or more of the predictor variables. Setwise regression will be used to answer questions regarding highly correlated and logically related pairs of variables (specifically, TABE Vocabulary and Reading Comprehension Tests; TABE Math Computation and Concepts and Applications Tests; TALS Document, Quantitative, and Prose Tests). For example, we may wish to know whether one variable of a pair is acting as a suppresser variable to another in a standard multiple regression model. Finally, we use hierarchical regression to answer questions concerning the additional unique variance contributed to an equation by predictor variables after the variance accounted for by the variables already in the equation.

DESCRIPTIVE STATISTICS

Descriptive statistics of groups were generated for the three testing periods. Table 3 (see Appendix) includes means, standard deviations, skewness, and kurtosis of the TABE and TALS batteries for three groups: the Initial Group at Session 1 (N=168), the Midtermer Subgroup at Session 2 (n=123), and the Persister Subgroup at Session 3 (n=91). We also include the means and standard deviations of the subsample of the Initial Group at Session 1 who were the Persister Group of Session 3. Finally, we include the

means and standard deviations from the norming samples as reported by the test publishers.

For the TABE, the group means for our sample were about a quarter of a standard deviation below the national sample means for ABE students for Vocabulary and Comprehension Tests. The Computation and Concepts and Applications Tests were about the same as the national means. For the TALS, the group mean was about a half of a standard deviation below the national mean for the Document and a quarter for the Quantitative. Better math versus reading performance on the TABE might be expected of enrollees in a basic skills program, particularly if a high percentage are nonnative speakers of English. Two thirds of the White Plains students reported that they were born outside the United States and its territories, and over half of those had been in the United States less than five years. Many were graduates of the English as a second language program.

Another interesting result to note is that the mean performance of the subsample of the 91 Persisters at Session 1 is lower than the mean for the Initial Group of 168. Thus, there was a tendency for higher ability students to leave the program prior to the final testing at Session 3. Some of the early leavers completed the GED test before the end of the program, though the vast majority of leavers were not in this category. We can only speculate that more able students are more likely to find a new job or move on to other opportunities prior to program completion. It should also be taken into account that most of those scoring at chance (i.e., equal to or less than the total number correct one would expect by chance if one guessed on each of the multiple-choice items of a subtest) on TABE subtests persisted in the program, accounting for much of the lower group means for the Persister Subgroup for Session 1.

STATISTICAL ASSUMPTIONS UNDERLYING REGRESSION ANALYSES

In a study of adult assessment, it is important to pay close attention to the number and type of exception students. The variance of adult performance is likely to be larger than an age cohort of high school students. Most test designers try to accommodate this wider range but it is impossible to make a test long enough to reliably test all student ability levels. Thus, a clustered group of extreme outliers can severely influence interpretation of group effects. Diagnostics were run to see if the data met regression analysis assumptions of normality, homoscedasticy, and collinearity, as well as to identify influential observations or outliers. Collinearity of the predictors was not a problem and will not be further discussed. Nonnormality and heteroscedasticity in the predictors will degrade the multivariate results, but the results may still be interpretable. On the other hand, nonnormality in the dependent variable is a serious violation of the assumptions.

Drawing all students from a single adult literacy site does not constitute a random sampling of the ABE population and unfortunately none of the test score distributions of the sample were normally distributed. Skewness and kurtosis are measures of the normality of a distribution. The closer to zero, the more normally distributed is the sample. The TALS Document Test scores over the three test sessions were not skewed, but were leptokurtic; that is, the distribution was too peaked or slender. The Quantitative and Prose Test scores were slightly bimodal, because of a short peak consisting of about 10 to 12 subjects who had near floor level scores. This may be the result of fatigue or

loss of motivation on the part of some students, since Document Literacy always preceded the Quantitative Literacy and Prose Literacy tests, although a similar fatigue effect is not visible in the longer TABE battery.

All of the TABE test distributions were skewed negatively and had positive kurtosis values. Negative skewness suggests a longer than expected tail at the low end, while positive kurtosis suggests a higher than expected peak in the distribution. Various transformations were performed on the variables. The square of TABE test scores improved the normality of the distributions somewhat. However, the subgroup of subjects who scored very low on at least one of the four TABE subtests continued to skew the distributions significantly. Also, visual examination of residual scatterplots of initial regression models showed some heteroscedasticity with the very low end TABE test scores tending to have smaller residuals than the rest of the scores. Many of the same subjects also were identified as multivariate outliers in the various regression models executed.

What to do with outliers is often a matter of judgment and compromise. We decided that the analysis results would be more interpretable if we eliminated all the low scorers in the TABE tests based on a single rule. Additional multivariate outliers would be eliminated based on the diagnostics for each model. This enabled us to work with consistent sample sets of the same students across the three test sessions, enhancing our ability to set up predictive models. The rule was to establish a cutoff based on two standard deviations from the sample mean for each subtest. A three standard deviations rule reduced some of the nonnormality, but not significantly. All subjects identified by this method were at the low end of the TABE scales; extreme high end scores were all within two standard deviations of the mean. According to the published norm tables for the TABE, the scores covered by this strategy account for less than 5% of the norming sample of ABE enrollees for each subtest. Table 4 (see Appendix) shows summary statistics for the reduced data sets for each session. Only reading comprehension at Session 2 continues to show relatively high skewness and kurtosis, but much reduced from the full sample group.

It should be noted that we have run the multivariate models under varying conditions including eliminating outliers based only on model diagnostics. The direction of effects and relative strengths of the predictor variables were not significantly different in any of these permutations than they are in the model results presented in this report. In general, the variance accounted for was stronger in the full sample models. To illustrate this point, we include the outliers in the initial model used to predict TALS Document Literacy scores for the purpose of letting the reader see how sensitive the model coefficients are to the presence of the outlier group. Our concern is that these results may be spuriously enhanced by the extreme values. Instead, the models reported in this paper are biased toward the middle ranges of ability as measured by the TABE tests. We prefer to make stronger claims about this narrower band of the target population and hope that others will be able to test the generality of our conclusions, rather than to speculate broadly about an extreme group not appropriately sampled in this study.

Before discussing the regression models, a few final words regarding the extreme scores and their effects on group means are in order. The outlier subgroup had TABE Locator Test scores that placed them in the TABE E level test. As mentioned, this outlier subgroup accounted for most of the nonnormality of the distribution of scale scores for the TABE, especially the negative skewness. This subgroup is highly influential when TABE group scores are aggregated, accounting for 10 to 15 point mean differences in the

first test session. Without this subgroup, the reading means approached the national sample means.

Because of regression to the mean, this subgroup is more likely by chance to score higher on retesting. Consequently, its members would potentially inflate mean gains disproportionally if they persisted in the program. The upwards effect on means occurs because so many extreme scores occurred at the bottom of the TABE reading scales, while there were relatively few extreme scores at the top to regress downward to balance the effect. In fact, the majority of these students did persist in the program until final testing. At that time, they did score higher as a group and were not identified as extreme by our two standard deviations rule nor in multivariate diagnostics. It is impossible to know statistically whether their gain scores represent the benefits of the program or the effects of regression to the mean.

CORRELATION ANALYSES

Intercorrelations among the tests are presented in Table 5 (see Appendix) for the three test session subgroups. There are moderate to high correlations across all the tests from a low of r=0.59 between the TABE Vocabulary and the TALS Document, to a high of r=0.81 between the TABE Vocabulary and Comprehension Tests. The highest correlations are between related test pairs: Reading Vocabulary-Comprehension, Mathematics Computation-Concepts/Applications, and Document-Quantitative Literacy. The correlations among tests generally increase slightly across sessions as fewer students are retained. In the reduced samples, the correlations were all slightly lower, but all of the interrelationships were nearly identical.

PREDICTING LITERACY PERFORMANCE

Predicting Document Literacy. We hypothesized that basic skills ability underlies applied literacy ability. To test this hypothesis, a multivariate regression model was constructed for Session 1 with TALS Document Literacy (DL1) scores as the dependent (criterion) variable and TABE Comprehension (RC1), Vocabulary (VC1), Computation (MC1), and Concepts/Applications (CN1) as predictor variables. Two cases were identified as multivariate outliers based on studentized residuals (p<.001) and were eliminated from the model. Regression coefficients (B), standardized coefficients (β), squared semipartial correlations (sr^2), and the adjusted multiple correlation for the Full Initial Group, the Reduced Initial Group, and the Persister Group subsample at Session 1 are reported in Table 6 (see Appendix).

The Initial Group model accounted for 59% of the variance in TALS DL1 scores with CN1 and RC1 significant predictors and MC1 approaching significance (F(4,167)=68.47, p<.0001). The Reduced Initial model accounted for 50% of the variance (F(4,145)=38.47, p<.0001). CN1 was the strongest predictor, with RC1 also a significant predictor. Further examinations of residuals and scatterplots and retests under slightly different conditions led to no significant changes to this model. Note that the relationships are similar to those in the Initial Group model, except that the addition of the outlier group makes the model fit stronger. As mentioned, the violation of model assumptions caused by the outlier group makes this result questionable. A similar model was

constructed for only the Persister Group subsample (n=89) for Session 1 with two outliers eliminated and similar results were also obtained; 66% of the variance was accounted for again by CN1 and RC1 ($F(4,84)=42.81$, $p<.0001$). The larger variances accounted for in the Persister Group subsample reflects the reduced variance available to analyze in the smaller sample size. However, observe that the pattern is similar.

In multiple regression models, correlations among the predictor variables reduce the unique variance reflected in the regression coefficients, because the overlap of variance among predictors is assigned only to the multiple correlation coefficient. Strongly correlated variables suppress the unique variance of their partners. In this study, VC-RC and MC-CN are strongly related pairs, the former a total reading measure, and the latter a total math ability measure. Univariate regression models were constructed for each of the TABE tests and the unique variance is reported in Table 6 in the column labeled R^2 (univariate). As the values indicate, CN1 (R^2 = .45) was the strongest predictor, but MC1 (R^2 = .29) was nearly as strong a predictor as RC1 (R^2 = .33), and even VC1 was a significant predictor (R^2 = .20). All univariate models were highly significant based on the F values ($p<.0001$).

To attenuate the amount of variance suppressed by related pairs of variables, a reduced model was computed with only the significant predictors from the original model. The results are presented in Table 6 as Model 2. The model accounts for 50% of the variance ($F(2,147)=76.40$, $p<.001$) with CN1 accounting for the greatest percentage of the unique variance (sr^2=0.17) and RC1 accounting for a lesser amount (sr^2=0.05). Thus, 28% of the variance in the model is overlap between RC1 and CN1. Put another way, CN1 alone would predict 45% of the variance of DL1, while RC1 contributes 5% unique additional explanatory power. Although 5% is statistically significant, its practical significance is limited. It also suggests that the higher correlation between the MC1 and CN1 resulted in the masking of the predictive power of the MC1 test which in the univariate case was similar to RC1. RC1 had a similar masking affect on VC1.

To test the robustness of this pattern of relations, similar models were constructed for Session 2 and Session 3 test scores predicting the corresponding Session 2 and Session 3 DL scores. The results are presented in Table 7 (see Appendix). In Session 2, 56% of the variance for the group was accounted for by CN2 and RC2 ($F(4,108)=37.15$, $p<.0001$). In Session 3, 70% of the variance for the group was accounted for by CN3 and MC3, with RC3 approaching significance ($F(4,79)=50.35$, $p<.0001$). Although this is a slightly different pattern than prior models, remember that RC and MC both account for nearly equal amounts of the variance as individual predictors. Consequently, it is consistent with the predictive relationship established. Note also that the math relationships to DL grow stronger over the year. Again, these patterns were similar with the low scorers, except that the amount of variance predicted was larger.

Predicting Future Document Literacy. Longitudinal data with multiple measurements on the same subjects permit us to test whether basic skill gains predict future applied abilities or vice versa. Since later events cannot cause earlier events, one can infer causal directions, or at least infer noncausal relationships. This basic logic only holds if one association is reliably greater than the reverse, that is, if prior TABE scores predict future TALS scores better than prior TALS scores predict future TABE scores. First we use TABE Session 1, 2, and 3 test scores to predict TALS Prose and Document literacy scores. Later we will see how well the TALS score predicts the TABE Reading Comprehension test score.

Multivariate models analogous to those described above were constructed for predicting Session 2 and 3 Document Literacy scores from Session 1 and 2 TABE

scores. The results are summarized in Table 8 (see Appendix). TABE scores at Session 1 predict 58% of the variance of Session 2 (DL2) scores ($F(4,110)=40.53$, $p<.0001$) and 58% of Session 3 (DL3) scores. ($F(4,76)=27.60$, $p<.0001$) Session 2 TABE scores predict 57% of Session 3 (DL3) scores($F(4,76)=24.01$, $p<.0001$). In each case, CN is the strongest predictor with RC accounting for a small, but significant percentage of the variance. The relationship does not change significantly over any of the three test sessions. The reduced models also show similar patterns.

One important effect to note is that TABE CN proves to be a stable predictor of TALS DL over all three sessions. In fact, the amount of unique variance contributed as a future predictor of DL by CN increases. That is, math concepts and applications may represent the most critical basic skills underlying applied document literacy, while reading comprehension skills are not as critical. This occurs even while instruction was focused primarily on basic skills in reading. One interpretation of this may be that, as ability in reading comprehension became more homogeneous over the year (because of instruction), it accounted for less unique variance, while individual differences related to math concepts and applications (i.e., problem-solving abilities) became a more significant predictor. Interestingly, DL group means showed statistically significant gains across all three test sessions, while RC and CN showed smaller significant gains from Session 1 to 2 and declines from Session 2 to 3. However, this variety of change patterns did not influence the stability of the longitudinal model predictors.

In summary, the TABE tests account for about 54% to 58% of the variance in the Document Literacy scores across all testing sessions. Although a significant chunk of the variance, this leaves about 40% to 50% of the variance unaccounted for by the basic skill measures. Most of the variance accounted for in the TALS Document Literacy test appears to be shared among all the TABE subtests. This shared variance might represent the fact that general achievement within an individual is usually correlated across a number of domains. However, the basic skills measures that contribute the largest amount of unique variance to the model are not the TABE Reading Tests, but rather the test of Math Concepts and Applications, which contributes about three times as much unique variance as Reading Comprehension in the initial model and up to eight times as much unique variance as a future predictor of Document Literacy scores.

At first glance we should be surprised that Concepts and Applications rather than (reading) Comprehension is the strongest predictor of Document Literacy. However, an inspection of the items for the two tests reveals a strong similarity. Both are problem oriented, requiring multistep operations for solutions. The TABE Comprehension Test, in contrast, involves traditional text comprehension tasks—main idea, author's intentions or purpose, inference of causes, and so forth. Document Literacy tasks seldom require the reading of an extended amount of text. All tasks require a search for specified information, followed by an operation on that information or the location of other information on the basis of some characteristic of the initially located item. Like the Concepts/Applications Test, the Document Literacy Test requires solution planning, monitoring of operations, testing for completion, and other problem-solving operations.

Predicting Prose Literacy. At this point, we have found little evidence that basic skills in reading significantly predict applied document literacy. This cannot be the ideal outcome for basic skills proponents. The impetus for basic skills instruction is the belief that basic skills can be easily transferred to other applied literacy tasks. However, even if reading does not transfer to locating, cycling, or evaluating information in a chart, map, or form, it should predict the ability to locate, integrate, and generate information in an applied prose task. To test this hypothesis, another set of regression models were run using Prose Literacy as the dependent variable and the four TABE tests as predictors. As mentioned earlier, we only had Prose Test scores for Session 3. This means that basic skills scores could be used to predict future Prose Literacy scores; however, we will not later be able to test the reverse hypothesis, that Prose Literacy predicts basic skills ability.

The regression results are presented in Table 10 (see Appendix). All of the models were significant predictors of Prose Literacy, accounting for 50% ($F(4,76)=20.93$, $p<.0001$), 49% ($F(4,78)=20.90$, $p<.0001$), and 53% ($F(4,74)=23.07$, $p<.0001$) of the variance over the three test sessions respectively. In general, the results mirror those for Document Literacy with RC contributing the largest amount of unique variance across the three sessions (sr^2=.12, .10, and 0.08 respectively), while CN contributed small, but significant additional unique variance in Sessions 2 and 3 (sr^2=0, .04, and .04 respectively). VC (R^2 = .32, .33, and .30 respectively) was about as strong a univariate predictor as CN (R^2 = .31, .31, and .39 respectively), but was masked in the multivariate models by RC; CN had a similar effect on MC.

It is interesting to note that the strength of the relationship to CN gained slightly over the year, while the strength of the other relationships remained constant. This trend is also apparent in the ratio of unique variance of RC to CN in the reduced models, which moves from about seven to one, to five to one, to three to one across the three sessions. Thus, the variance unique to reading comprehension shrank as the variance unique to concepts and applications grew. One interpretation of this may be that as ability in reading comprehension became more homogeneous over the year (because of instruction), it accounted for less unique variance in prose literacy, while individual differences related to math concepts and applications (i.e., problem-solving abilities) became a more significant predictor.

In summary, the pattern of regression results suggests that basic skills in reading comprehension are the strongest predictor of future applied prose literacy ability. However, it may be that the shared problem-solving ability that we have argued underlies document, quantitative, and math concepts and applications, also underlies the ability to locate, integrate, and evaluate prose materials. In other words, applied literacy in whatever form is a complex problem-solving domain, not just a basic skills reading measure.

Predicting Reading Comprehension. Perhaps we have been asking the question backwards. Adults with advanced problem-solving abilities as evidenced in the applied literacy tests may be better prepared to learn basic skills underlying reading comprehension. To test this hypothesis, multivariate regression models were constructed with TABE RC scores at Session 3 as the dependent (criterion) variable and TALS DL and QL and TABE CN scores as predictor variables for Sessions 1 and 2, and TALS DL, QL, PR and TABE CN for Session 3. TABE CN was included for comparisons, because of the strong relationship underlying it and the TALS tests. Reduced models with TALS DL and TABE CN as predictors are also presented, again because of the strong relationship underlying these two variables. Results are reported in Table 11 (see Appendix).

The models for predicting RC3 from Session 1 and Session 2 are both significant and similar. Session 1 scores predicted 36% of the variance ($F(3,78)=16.43$, $p<.0001$); Session 2 scores predicted 31% of the variance ($F(3,80)=13.57$, $p<.0001$). In both cases DL was the only significant predictor accounting for .07% and .08% unique variance respectively.

The models for predicting RC3 from Session 3 scores are more complex to interpret. Session 3 scores are highly significant predictors of RC3, accounting for 56% of the variance ($F(4,75)=25.67$, $p<.0001$). PR was a strong, significant predictor, accounting for 10% unique variance. CN was also significant, accounting for 4% unique variance. DL is no longer a significant predictor, but the negative coefficient suggests that the overlap of DL with CN is masking its predictive powers. The reduced model comparing DL and CN also was significant, accounting for 46% of the variance ($F(2,77)=31.63$, $p<.0001$); however, the relationship between CN and DL is reversed from Sessions 1 and 2. Now only CN is a significant predictor and accounts for 9% unique variance, while DL accounts for a nonsignificant 1%.

In summary, the results in Table 11 are equivocal. About half of the variance accounted for by the models in Sessions 1 and 2 for predicting RC3 scores is common to both DL and CN. As argued earlier, this may either be problem-solving ability or general achievement. Document Literacy scores are nonetheless better unique predictors of future Reading Comprehension scores than Math Concepts and Applications. When all the tests are taken at the end of the year, however, Math Concepts and Applications becomes a much stronger predictor both in overall variance accounted for (R^2 univ=43%) and in unique variance compared to DL. It is not as strong a predictor, however, as is Prose Literacy.

GENERAL DISCUSSION AND CONCLUSIONS

The central issue explored in this report is whether basic skills and applied literacy tests are measuring the same or similar underlying abilities. The results are complex, but we will try to present as unified a picture as possible, while acknowledging that some mysteries still remain. We also remind the reader that the data patterns found here need to be replicated in other studies before stronger claims can be justified. The main finding was that the TABE Mathematics Concepts and Applications Test was a stronger predictor of the TALS Document Test than TABE Reading Comprehension scores. This relationship was robust over time. Furthermore, Mathematics Concepts and Application scores at the beginning of the school year were also better predictors of future Document Literacy scores. On the surface, this appeared to be a surprising result because the content and structure of the Mathematics Concepts and Applications Test items appear different from the TALS items. The commonality, we argue, is in the problem-solving skills required to perform well on each test. The examinee who can set up a problem goal, reason through a sequence of steps, and set a strategy for maximizing performance is at an advantage.

The skills measured by the Concepts and Applications Test also seem to underlie the relationship between Document and Quantitative Literacy scores. As a predictor of Document Literacy, Quantitative Literacy adds only a small additional amount of unique variance in the presence of the Concepts and Applications tests. Concepts and Applications is actually a stronger predictor of Document Literacy than Quantitative Literacy, despite the differences in surface item characteristics and response format.

The low to moderate relationship of the Comprehension and Vocabulary tests as predictors of the Document Test suggest that the basic skills measured by these tests are not the same as the abilities underlying the TALS measures. This point is punctuated by the fact that the vast majority of the predictive power of the Reading Comprehension Test overlaps with the Concepts and Application Test, suggesting perhaps a general achievement factor. The TABE reading test scores are not interchangeable with the Document Literacy Test scores as measures of student progress. Thus, we have interpreted the shared relationship among the tests to be the result of two factors, correlations in general achievement across domains and individual differences in generalized problem-solving ability. This problem-solving ability may have developed as a result of exposure to mathematics in schools or to documents in society or to both. In any case, the relationship has more to do with the nature of the task or common experiences in both mathematics and document contexts, than with the nature of the document or print materials, since for the most part the Concepts and Applications diagrams and charts are much simpler and less diverse than those included in the Document Test.

The patterns of results for predicting Prose Literacy scores mirror those of Document Literacy, with Reading Comprehension taking on the role of strongest predictor, and Math Concepts and Applications as a smaller, but still significant predictor. However, unlike the Document scores, the relationship of Math Concepts and Applications grows stronger over the year, while the relationship to Reading Comprehension tended to diminish slightly. We tentatively put forward the hypothesis that the instructional focus on basic skills reading comprehension had the effect of diminishing the unique variance accounted for by this variable, while the individual differences in the Math Concepts and Applications variable remained constant. We suggest that the cognitive operations of locating, integrating, and evaluating in the applied prose literacy tasks have more in common with general problem solving than do the tasks associated with the reading comprehension items.

Finally, we looked at predictors of Reading Comprehension scores at Session 3 and found that Document Literacy was a stronger predictor of future scores than Math Concepts and Applications, though the relationship reversed when all tests were taken in the same session. Furthermore, the Prose Literacy Test was the most significant predictor of Reading Comprehension scores in Session 3. These results are consistent and mildly supportive of the hypothesis that ability in applied literacy document tasks is related to achievement in reading comprehension instruction. We attribute the strong relationship of Math Concepts and Applications at Session 3 to the similarity of the two TABE test types and general achievement correlations within an individual. On the other hand, as a predictor of future reading comprehension, general achievement and test format are less significant factors than prior exposure to solving problems in applied literacy settings, the ability measured by the TALS tests. Having said this, we acknowledge that other explanations could also account for the data pattern, and it is premature to draw strong conclusions.

Thus, we conclude that the TABE reading tests are not interchangeable with the Prose Literacy Test; however, they are more closely related to it than to the Document Literacy

Test. Common sense dictates that an instructor or program should be concerned with the relative importance of the tasks represented in each test type: word knowledge in the Vocabulary Test; literal and inferential comprehension in fiction and nonfiction reading passages in the Reading Comprehension Test; locating, integrating, and evaluating information in nonfiction prose materials in the Prose Literacy Test; or locating, cycling, and evaluating information in charts, tables, and maps in the Document Literacy Test.

Our results also suggest that at least the TALS Document and Prose Tests are not totally equivalent to each other, or to reading comprehension or other basic skills measures. Other recently reported research has suggested that the Document, Quantitative, and Prose scales may not be unique constructs (Reder, in press). In support of the Reder findings, we found that on the whole the intercorrelations among tests and variance accounted for by regression models in this study were large even for a relatively homogenous population of ABE students, suggesting that we are drawing distinctions based on relatively narrow bands of variance differences. One possibility for the distinctions among scales that we observed is that the TALS measures are better differentiated as constructs within an ABE population versus the nation of readers as a whole. That is, in subpopulations where general achievement and abilities are more homogeneous, individual differences in problem solving within prose, document, or quantitative material/task types becomes a more important factor.

The results reported here support some common sense starting points for deciding which test to use for adult literacy assessment. Based on the assumption that the test should match the instructional program, if reading comprehension strategies are taught and emphasized, then a test such as the TABE Comprehension Test may be appropriate. If vocabulary skills are emphasized, then a vocabulary test is appropriate. If interpreting and using everyday documents is a major objective, then the TALS Document Test is appropriate. In general, one can expect that students will be more motivated and perform better on a posttest measure when it covers materials that they have learned and practiced in class, rather than on a measure of some correlated skill that is supposed to explain their learning in another domain. For example, even though the TABE Concepts and Applications and TALS Document and Quantitative Tests may be measuring similar underlying abilities, they have very different face validity. Face validity is critical for students and teachers in setting goals and guiding instruction. The common sense approach is to start with the students' goals. Some students' goals may be to learn to apply literacy skills in the social settings necessary for improvement in occupational, civic, or family contexts. Others may want to learn mathematics in the more traditional approaches that are still taught in public schools. Learning in traditional instructional contexts may be important for students with aspirations to continue to higher levels of education.

Having said this, we have several general cautions. First, all of the tests favor students who are fast workers and clever problem solvers, even though these may not be the same traits that would be applied in many genuine literacy contexts. A slow, intense reader with a value of "getting it right the first time" may perseverate on a particular item for half the testing period and receive the same score as a fast, careless reader. The aged may also have scores that reflect inadequate time, not a lack of basic or applied skills. There is a preferred strategy for taking short-answer, timed tests and in an adult class there are likely to be widely divergent experiences regarding this dimension with students who

demonstrate similar ability in reading or document processing in other contexts. If tests are used to gather diagnostic information or if pre/post measures of gain are to be made, then adequate coaching of students should precede test taking to ensure valid measures. Where possible, probes should be made into problem areas to determine whether the performance on the test reflects the student's skill base or the nature of the assessment tool.

Finally, the quantitative precision of these instruments for individual and sometimes group scores may be overvalued. Both tests tend to be able to rank students relative to a national sample and each other, which may be sufficient for the purpose of placement into three or four levels of classes. However, what this ranking says about the abilities of the individuals is less clear. With the TABE, the four levels of tests, numerous table conversions, and grade equivalencies can all give an appearance of precision that is belied by large standard errors of measurement and occasional misplacement of students into testing levels. These tests should not be used to provide more than preliminary diagnostic hypotheses for teachers to begin to learn more about their students. Whenever attempting to measure gain, unless scrupulous care is taken to understand the threats to validity and reliability, any number of factors may result in spurious outcomes in individual or aggregate measures.

REFERENCES

Archbald, D., & Newmann, F. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school.* Reston, VA: National Association of Secondary School Principals.

Calfee, R., & Hiebert, E. (1991). Classroom assessment in reading. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research, volume II* (pp. 281-309). White Plains, NY: Longman.

CASAS. (1989). *CASAS adult life skills.* San Diego, CA: Author.

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.

Development Associates, Inc. (1992). *Executive summary of the first interim report from the National Evaluation of Adult Education Programs.* Arlington, VA: Author.

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

Gee, J. P. (1991). Discourse systems and aspirin bottles: On literacy. In C. Mitchell & K. Weiler (Eds.), *Rewriting literacy: Culture and the discourse of the other.* New York: Bergin & Garvey.

Guthrie, J. T., Britton, T., & Barker, K. G. (1991). Roles of document structure, cognitive strategy, and awareness in searching for information. *Reading Research Quarterly, 26,* 300-24.

Haney, W., & Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. *Phi Delta Kappan, 70,* 683-687.

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.

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.

Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey* (Report). Washington, DC: National Center on Educational Statistics.

Kirsch, I. S., & Mosenthal, P. B. (1989). Understanding documents: Building documents by combining simple lists. *Journal of Reading, 33,* 132-125

Kirsch, I. S., & Mosenthal, P. B. (1990a). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly, 25,* 5-30.

Kirsch, I. S., & Mosenthal, P. B. (1990b). Understanding documents: Mimetic documents: Diagrams. *Journal of Reading, 34,* 290-294.

Levine, K. (1986). *The social context of literacy.* London: Routledge & Kegan Paul.

Linn, R. L. (1989). Current perspectives and future directions. In R. L. Linn (Ed.). *Educational Measurement* (3rd ed., pp. 1-10). New York: Collier Macmillan.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Madaus, G. F., & Kelleghan, T. (1992). Curriculum evaluation and assessment. In P. W. Jackson (Ed.), *Handbook of research on curriculum* (pp. 119-156). New York: Macmillan.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*, 13-24.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-104). New York: Collier Macmillan.

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

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*, 5-12.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229-258.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*, 575-603.

Reder, S. (in press). *What does NALS measure? Issues of dimensionality and construct validity* (Technical Report). Philadelphia: University of Pennsylvania, National Center on Adult Literacy.

Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20*, 2-9.

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

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 263-332). New York: Collier Macmillan.

Stedman, L. C., & Kaestle, C. F. (1987). Literacy and reading performance in the United States, from 1880 to the present. *Reading Research Quarterly, 22*(1), 8-46

APPENDIX: TABLES

Table 1

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

	LEVEL														
	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

Testing Schedule, Dates and Numbers of Students Tested

	Session 1	Session 2	Session 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
Tests administered			
TABE Tests			
Locator	199	–	–
Vocabulary	185	136	101
RC.	185	136	101
MC.	185	136	101
Conc./CN.	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 3

Summary Statistics for TABE and TALS Tests for Three Test Sessions

	TABE				TALS		
Session 1							
	VC1	RC1	MC1	CN1	DL1	QL1	
Initial (*N*=168)							
M	699	716	761	715	246	266	
SD	79	61	70	61	49	55	
Skewness	-1.19	-1.74	-1.8	-1.02	0.04	-0.42	
Kurtosis	1.44	3.07	4.87	1.85	0.16	-0.5	
Persisters (*n*=91)							
M	687	703	748	709	239	260	
SD	82	65	80	64	53	57	
Session 2							
Midtermers (*n*=123)							
	VC2	RC2	MC2	CN2	DL2	QL2	
M	704	722	763	721	254	272	
SD	75	58	73	62	49	49	
Skewness	-1.08	-1.89	-2.06	-1.11	0.1	-0.08	
Kurtosis	1.34	3.8	5.49	2.25	0.04	-0.19	
Session 3							
Persisters (*n*=91)							
	VC3	RC3	MC3	CN3	DL3	QL3	PR3
M	697	705	765	717	260	248	252
SD	69	62	66	74	53	54	43
Skewness	-0.66	-1.28	-0.95	-1.17	-0.12	0.25	-0.55
Kurtosis	-0.05	1.59	1.2	2.07	-0.43	-0.24	-0.02
Mean and standard deviation of TABE and TALS norming samples							
	VC	RC	MC	CN	DL	QL	PR
M	719	733	758	712	283	284	292
SD	74	51	63	62	48	48	48

Note: TALS results scaled to link with the Young Adults Literacy Survey (Kirsch, Jungeblut, & Campbell, 1991); TABE means based on a sample of 1053 ABE enrollees (CTB/McGraw Hill, 1987b).

Table 4

Summary Statistics for Reduced TABE and TALS Tests for Three Test Sessions

	TABE				TALS		
Session 1							
Reduced Initial (n=152)							
	VC1	RC1	MC1	CN1	DL1	QL1	
M	715	730	775	728	254	275	
SD	58	38	46	45	43	48	
Skewness	-0.47	-0.95	-0.47	-0.18	0.39	-0.41	
Kurtosis	-0.61	0.94	0.74	0.04	0.25	-0.02	
Session 2							
Reduced Midterms (n=113)							
	VC2	RC2	MC2	CN2	DL2	QL2	
M	718	734	780	732	261	278	
SD	56	37	43	47	45	45	
Skewness	-0.32	-1.22	-0.45	-0.19	0.27	0.04	
Kurtosis	-0.75	1.37	-0.01	-0.61	0.29	-0.09	
Session 3							
Reduced Persisters (n=84)							
	VC3	RC3	MC3	CN3	DL3	QL3	PR3
M	705	716	778	730	267	255	256
SD	61	47	49	58	49	50	39
Skewness	-0.47	-0.79	-0.17	-0.47	-0.13	0.3	-0.55
Kurtosis	-0.72	0.28	-0.24	-0.46	-0.07	-0.01	0.29

Table 5

Correlation Matrices for Test Sessions 1, 2, and 3

Session 1 (n=168)

	DL1	QL1	VC1	RC1	MC1	CN1
DL1	1.00					
QL1	0.77	1.00				
VC1	0.59	0.65	1.00			
RC1	0.68	0.70	0.81	1.00		
MC1	0.67	0.72	0.66	0.68	1.00	
CN1	0.74	0.80	0.71	0.73	0.80	1.00

Session 2 (n=123)

	DL2	QL2	VC2	RC2	MC2	CN2
DL2	1.00					
QL2	0.75	1.00				
VC2	0.63	0.56	1.00			
RC2	0.67	0.63	0.83	1.00		
MC2	0.60	0.60	0.71	0.67	1.00	
CN2	0.72	0.72	0.67	0.68	0.87	1.00

Session 3 (n=91)

	DL3	QL3	VC3	RC3	MC3	CN3	PR3
DL3	1.00						
QL3	0.82	1.00					
VC3	0.61	0.66	1.00				
RC3	0.72	0.69	0.77	1.00			
MC3	0.81	0.78	0.71	0.76	1.00		
CN3	0.81	0.75	0.60	0.70	0.82	1.00	
PR3	0.75	0.70	0.69	0.78	0.64	0.68	1.00

Note: All correlations significant at p<.001.

Table 6

Multiple Regression: Predicting the TALS Document Literacy Test From the TABE Tests

Session 1								
	Initial (N=168)							
	B	β	sr^2					
Model 1								
INT	-241.02	0						
VC1	-0.05	-0.09	0					
RC1	0.25	0.31	0.03	**				
MC1	0.11	0.16	0.01	a				
CN1	0.36	0.44	0.06	***				
Adj. R^2	0.59							
	Reduced Initial (n=150)					**Reduced Initial (n=150)**		
	B	β	sr^2		R^2 (univ.)	B	β	sr^2
					Model 2			
INT	-352.10					-328.75		
VC1	-0.06	-0.09	0		0.20			
RC1	0.37	0.33	0.05	**	0.33	0.31	0.28	0.05 ***
MC1	0.07	0.07	0		0.29	0.48	0.51	0.17 ***
CN1	0.45	0.47	0.07	***	0.45			
Adj. R^2	0.50					0.50		
	Persisters (n=89)							
	B	β	sr^2		R^2 (univ.)			
INT	-240							
VC1	-0.03	-0.05	0		0.39			
RC1	0.21	0.27	0.02	*	0.50			
MC1	0.09	0.15	0.01		0.54			
CN1	0.40	0.51	0.06	***	0.63			
Adj. R^2	0.66							

Note: B=regression coefficient; β=standardized coefficient; sr^2=squared semipartial correlation; R^2(univ.)=the univariate correlation coefficient; adj. R^2=adjusted multiple correlation coefficient.

* $p<.05$.
** $p<.01$.
*** $p<.001$.
a $p<.10$.

Table 7

Multiple Regression: Predicting the TALS Document Literacy Test From the TABE Tests in Session 2 and Session 3

Session 2

Midtermers (n=113)

	B	β	sr^2 (uniq.)		R^2 (univ.)
INT	-364.98	0			
VC2	0	-0.01	0		0.23
RC2	0.31	0.25	0.03	**	0.34
MC2	-0.05	-0.05	0	a	0.32
CN2	0.61	0.63	0.12	***	0.53
Adj. R^2	0.56				

Session 3

Persisters (n=84)

	B	β	sr^2 (uniq.)		R^2 (univ.)
INT		0	.		
VC3	-0.10	-0.12	0.01		0.26
RC3	0.20	0.19	0.01	a	0.43
MC3	0.36	0.36	0.03	**	0.61
CN3	0.38	0.45	0.05	***	0.67
Adj. R^2	0.70				

Note: B=regression coefficient; β=standardized coefficient; sr^2=squared semiparitial correlation; R^2(univ.)=the univariate correlation coefficient; adj. R^2=adjusted multiple correlation coefficient.

* $p<.05$.
** $p<.01$.
*** $p<.001$.
a $p<.10$.

Table 8

Multiple Regression: Predicting the TALS Document Literacy Test at Session 2 and 3 From the TABE Tests at Sessions 1 and 2

Session 2 (DL2) from Session 1 TABE scores						Reduced Model (CN1 & RC1 only)			
(n=115)	B	β		sr^2	R^2	B	β		sr^2
INT	-351.86			(uniq.)	(univ.)	-337.24			
VC1	-0.09	-0.13		0.01	0.16				
RC1	0.29	0.26	**	0.03	0.29	0.21	0.19	**	0.03
MC1	0.03	0.03		0	0.33				
CN1	0.62	0.65	***	0.13	0.52	0.61	0.64	***	0.29
Adj. R^2	0.58								
Session 3 (DL3) from Session 1 TABE scores									
(n=81)	B	β		sr^2	R^2	B	β		sr^2
				(uniq.)	(univ.)				(uniq.)
INT	-372.01	0		.		-328.83			
VC1	-0.07	-0.08		0	0.22				
RC1	0.29	0.25		0.03	* 0.32	0.25	0.21		0.03 *
MC1	0.24	0.25		0.02	a 0.47				
CN1	0.41	0.43		0.04	** 0.54	0.58	0.61		0.25 ***
Adj. R^2	0.57								
Session 3 (DL3) from Test Session 2 TABE scores									
(n=81)	B	β		sr^2	R^2	B	β		sr^2
				(uniq.)	(univ.)				(uniq.)
INT	-356.36	0		.		-336.76			
VC2	-0.04	-0.05		0	0.24				
RC2	0.28	0.23		0.03	* 0.30	0.26	0.21		0.03 ***
MC2	0.07	0.07		0	0.37				
CN2	0.54	0.57		0.10	*** 0.52	0.57	0.60		0.24 ***
Adj. R^2	0.54								

Note: B=regression coefficient; β=standardized coeffiecient; sr^2=squared semipartial correlation; R^2(univ.)=the univariate multiple coefficient; adj. R^2=adjusted multiple correlation coefficient.

* p<.05.
** p<.01.
*** p<.001.
a p<.10.

Table 9

Multiple Regression: Predicting the TALS Document Literacy Test From the TABE Concepts and Applications Test and TALS Quantitative Literacy Test

Session 1					
	(*n*=150)				
	B	*β*	sr^2 (uniq.)		R^2 (univ.)
INT	-103.99	0			
CN1	0.36	0.38	0.06	***	0.45
QL1	0.34	0.38	0.05	***	0.45
Adj. R^2	0.51				

Note: *B*=regression coefficient; *β*=standardized coefficient; sr^2=squared semipartial correlation; R^2(univ.)=the univariate correlation coefficient; adj. R^2=adjusted multiple correlation coefficient.

* $p<.05$.
** $p<.01$.
*** $p<.001$.
a $p<.10$.

Table 10

Multiple Regression: Predicting the TALS Prose Literacy Test at Session 3 From the TABE Sessions 1, 2, and 3

						Reduced Model (CN1 & RC1 only)			
Session 1 (n=83)	B	β	sr^2 (uniq.)		R^2 (univ.)	B	β		sr^2
INT	-259.13	0				-253			
VC1	0.09	0.14	0		0.32				
RC1	0.47	0.49	0.12	***	0.48	0.54	0.56	***	0.20
MC1	0.05	0.07	0		0.25				
CN1	0.10	0.12	0		0.31	0.17	0.22	*	0.03
Adj. R^2	0.50					0.50			
Session 2 (n=83)	B	β	sr^2 (uniq.)		R^2 (univ.)	B	β		sr^2 (uniq.)
INT	-195.98					-254			
VC2	0.14	0.2	0.02		0.33				
RC2	0.44	0.45	0.1	***	0.45	0.51	0.53	***	0.18
MC2	-0.21	-0.23	0.02		0.18				
CN2	0.27	0.35	0.04	*	0.31	0.19	0.24	*	0.04
Adj. R^2	0.49					0.48			
Session 3 (n=79)	B	β	sr^2 (uniq.)		R^2 (univ.)	B	β		sr^2 (uniq.)
INT	-142.67					-157			
VC3	0.08	0.13	0		0.3				
RC3	0.34	0.44	0.08	***	0.49	0.4	0.51	***	0.15
MC3	-0.06	-0.09	0		0.28				
CN3	0.21	0.35	0.04	*	0.39	0.18	0.30	*	0.05
Adj. R^2	0.53					0.54			

Note: B=regression coefficient; β=standardized coefficient; sr^2=squared semipartial correlation; R^2(univ.)=the univariate correlation coefficient; adj. R^2=adjusted multiple correlation coefficient.

* p<.05.
** p<.01.
*** p<.001.
a p<.10.

Table 11

Multiple Regression: Predicting TABE Reading Comprehension From the TALS Tests

Session 1										
(*n*=82)										
	B	*β*	sr^2 (uniq.)		R^2 (univ.)		*B*	*β*		sr^2 (uniq.)
INT	482.54					INT	465.12			
DL1	0.31	0.37	0.07	**	0.32	DL1	0.35	0.41	***	0.12
QL1	0.09	0.10	0		0.27					
CN1	0.19	0.23	0.02	a	0.26	CN1	0.24	0.29	**	0.06
Adj. R^2	0.36						0.37			
Session 2										
(*n*=83)										
	B	*β*	sr^2 (uniq.)		R^2 (univ.)		*B*	*β*		sr^2 (uniq.)
INT	523.54					INT	524.84			
DL2	0.36	0.43	0.08	**	0.31	DL2	0.36	0.43	**	0.09
QL2	0	0	0		0.19					
CN2	0.15	0.19	0.01		0.24	CN2	0.15	0.18		0.02
Adj. R^2	0.31						0.32			
Session 3										
(*n*=80)										
	B	*β*	sr^2 (uniq.)		R^2 (univ.)		*B*	*β*		sr^2 (uniq.)
INT	375.94						385.74			
DL3	-0.08	-0.09	0		0.35	DL3	0.19	0.20		0.01
QL3	0.15	0.18	0.01		0.36					
CN3	0.26	0.34	0.04	*	0.43	CN3	0.39	0.50	***	0.09
PR3	0.53	0.43	0.1	***	0.48					
Adj. R^2	0.56						0.44			

Note: B=regression coefficient; *β*=standardized coefficient; sr^2=squared semipartial correlation: R^2(univ.)=the univariate correlation coefficient; adj. R^2=adjusted multiple correlation coefficient.

* $p<.05$.
** $p<.01$.
*** $p<.001$.
a $p<.10$.