DOCUMENT RESUME

ED 220 811 CS 006 800

AUTHOR Valentine, Thomas
TITLE An Examination of the Validity of Two Item Classification Schemes for the Reading Comprehension Subtest of the Tests of Adult Basic Education.
PUB DATE Oct 82
NOTE 100p.; M.Ed. Thesis, Rutgers the State University of New Jersey.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Adult Basic Education; *Classification; Evaluation Methods; *Item Analysis; Reading Comprehension; Reading Diagnosis; *Reading Tests; *Validity
IDENTIFIERS *McGraw Hill Test of Adult Basic Education

ABSTRACT

A study was conducted to determine the validity of two separate item classification schemes for the reading comprehension subtest of the McGraw-Hill Test of Adult Basic Education, Level D, Form 4. The first of the two schemes was developed by the publisher and presented as a means of classifying errors for the purpose of diagnosing individual learning needs. The second was developed by the researcher through a subjective examination of the 45 subtest items. Data for the study consisted of test results from 242 adult students enrolled in an urban career assessment program. In the first stage of the study, a factor analysis was used to determine whether the subtest was truly multidimensional, since this is a prerequisite for meaningful item classification schemes. The results showed that the subtest was, in fact, multidimensional. The second stage of the study examined the discriminant validity of the classification schemes by determining the number of items within each subskill category that were more highly correlated with the corrected total score of the subskill category of which they were a part than with the total scores for each of the remaining subskill categories. Results indicated that the researcher-developed classification scheme evidenced a 9% increase in the number of items manifesting discriminant validity. The findings suggested that neither classification scheme offered sufficient evidence of discriminant validity to establish its practical utility. (Author/FL)

AN EXAMINATION OF THE VALIDITY OF TWO ITEM CLASSIFICATION SCHEMES FOR THE READING COMPREHENSION SUBTEST OF THE TESTS OF ADULT BASIC EDUCATION

AN ABSTRACT OF A THESIS

SUBMITTED TO THE FACULTY

OF THE GRADUATE SCHOOL OF EDUCATION

OF

RUTGERS

THE STATE UNIVERSITY OF NEW JERSEY

BY

THOMAS VALENTINE

IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE

OF

MASTER OF EDUCATION

COMMITTEE CHAIRPERSON: Gordon A. Larson, Ed.D.

NEW BRUNSWICK, NEW JERSEY OCTOBER, 1982

AN EXAMINATION OF THE VALIDITY OF TWO ITEM CLASSIFICATION SCHEMES FOR THE READING COMPREHENSION SUBTEST OF THE TESTS OF ADULT BASIC EDUCATION

A THESIS

SUBMITTED TO THE FACULTY

OF THE GRADUATE SCHOOL OF EDUCATION

OF

RUTGERS

THE STATE UNIVERSITY OF NEW JERSEY

BY

THOMAS VALENTINE

IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE

OF

MASTER OF EDUCATION

NEW BRUNSWICK, NEW JERSEY OCTOBER, 1982

APPROVED: Gordon A. Larson

Gordon G. Darkenwald

Martin Kling

DEAN: Irene Athey

ACKNOWLEDGEMENTS

First and foremost, I would like to extend my sincere gratitude to the members of my committee, who offered just the right blend of acumen and patience throughout the execution of this, at times, troublesome study. They gave me the freedom to direct my own learning, to stumble when I insisted upon it, but never to fall.

In addition, I would like to offer my sincere thanks:

- to Jeff Smith, who knows not only how to juggle numbers, but how to toss them gently to a nervous novice.
- to Steve Traylor of the New Brunswick Career Preparation Center, who supplied the needed data without hesitation.
- to my dear friends, Ralph and Woodrow, who nudged me along the way, and who can well appreciate a study of this nature.
- to the adult learners of New Brunswick, who have motivated my research, improved my teaching, and enriched my life.

I dedicate this thesis to Deborah, the finest wife a man could want.

TABLE OF CONTENTS

TABLE OF CONTENTS -- Continued

LIST OF TABLES

LIST OF TABLES -- Continued

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Background of the Problem

Program administrators in adult education are, of necessity, pragmatic. Administrative decisions are typically made in a context of restricted, and often tenuous, budgets, and of time constraints dictated by a limited staff and by adult students with limited amounts of time available for educational endeavors. It is not surprising, then, that in many cases, practical adequacy must take precedence over impractical perfection, and the "best" administrative decision is superseded by a more expedient alternative.

In selecting assessment materials, adult education administrators look for tests that will provide the maximum amount of information on large numbers of participants for the minimum investment of hours and dollars. McGraw Hill's Test of Adult Basic Education (TABE), a standardized, silent test of reading, mathematics, and language proficiency, is a very economical test. Examiners require no special training, the entire battery can be administered in approximately three hours, hand-scoring is rapid and

easy, and the testing materials themselves are comparatively inexpensive. Partly for these reasons, and partly due to an absence of serious competition in the field, the TABE has virtually cornered the market in basic skills testing in adult education in New Jersey. Unfortunately, however, many program administrators have come to regard this single test as the answer to all of their assessment needs.

Each year, thousands of adults participating in basic, secondary, and vocational education programs are required to take the TABE. For many federally-funded vocational training programs, rigid entrance criteria are set in the form of grade equivalent scores in reading and mathematics, as measured by the TABE. Alternate forms of the TABE are used in adult basic education as pre- and post- tests to measure academic growth resulting from instruction. In adult secondary education, TABE scores are widely used to indicate an individual's readiness to take the General Educational Development (GED) Test, or to certify that an individual has attained a prerequisite level of basic skills proficiency for the award of a high school diploma.

Obviously, it is impossible for any one test to adequately accomplish all of the functions for which the TABE is employed. It could, perhaps, be argued that the use of the TABE as a screening instrument for vocational training programs is justified, since the comparison of individual performance on a series of common tasks is the proper

function of a norm-referenced test; such an argument, however, implies tacit acceptance of the validity of the largely school-like test content in assessing aptitude for vocational training, as well as the acceptance of a norming procedure based entirely on a sample of elementary and secondary school students -- the TABE never was normed on an adult population. The TABE's other uses, however, as a measure of individual growth and as an instrument used to certify exit competencies, are, at best, highly suspect.

This study, however, was solely concerned with still another ancillary use to which the TABE is put in adult education, namely as a diagnostic instrument. Although the test authors carefully avoid the term "diagnosis" in all of their published materials, the scoring instructions for the TABE provide for an "Analysis of Learning Difficulties." This analysis, which simply consists of classifying observed errors according to subskill breakdown within each of the subtests, allegedly makes it "possible to plan instruction that will focus on the student's particular needs" (Examiner's Manual, 1976, p. 37).

Despite the existence of tests specifically designed for the diagnosis of learning needs, the fact remains that the TABE Analysis of Learning Difficulties is widely used in the field, and often serves as the only attempt at formal diagnosis for the purpose of planning individualized instruction.

The practical efficacy of the Analysis of Learning Difficulty hinges on the validity of the item classification scheme presented by the test publisher in the examiner's manual. Each of the 45 items in the reading comprehension subtest is assigned to one of four "content categories" (Examiner's Manual, p. 32), namely Using Reference Skills, Recalling Facts, Understanding Main Ideas, and Making Inference.

Nowhere, in either the Examiner's Manual or in the Technical Report (1978) for the TABE, are these content categories, or more appropriately, subskill categories, defined. No evidence is offered concerning the reliability or validity of these measures. Since test consumers are provided with no statistical or substantive rationale for the subskill categories, users of the TABE Analysis of Learning Difficulties are simply submitting to the publisher's assertion that such categories exist as viable and integral constructs.

Statement of the Problem

It was the purpose of this study to examine the validity of two item classification schemes for the Reading Comprehension subtest of the TABE, Level D, Form 4. The first of the item classification schemes is presented by the publisher in the Examiner's Manual. The second was developed through a subjective analysis of test items during the early stages of this study. It was hoped that an

examination of the validity of the two item classification schemes would accomplish two major goals, namely:

- To operationally define reading comprehension, as measured by the TABE.
- To determine the extent to which each of the item classification schemes is of practical value in the diagnosis of individual learning needs.

The major research questions addressed by this study are:

1. Is the reading comprehension subtest unidimensional or multidimensional?
2. To what extent are the subskill categories presented in each of the item classification schemes internally consistent?
3. To what extent are the subskill categories presented in each of the item classification schemes hierarchical?
4. To what extent do the subskill categories presented in each of the item classification schemes show evidence of divergent validity?
5. Do the item classification schemes differ in terms of internal consistency and discriminant validity?

Overview of the Study

In order to accomplish the goals of the study, and in

order to answer the research questions stated above, four major steps were undertaken. The first step employed a subjective analysis; the three remaining steps represent a number of empirical analyses based upon test results from a sample of 242 adult examinees.

The first step consisted of the development of an alternative item classification scheme, derived inductively from a subjective examination of the test items and without reference to the original item classification scheme presented by the publishers. The impetus for the development of an alternative scheme grew out of this writer's dissatisfaction with the extremely general nature of some of the publisher's subskill categories, most notably the category entitled "Making Inferences." This broad category includes 22 of the 45 test items and appears to measure abilities ranging from simple paraphrasing to what one reviewer has termed "exercises in logical reasoning which are quasi-mathematical" (Donlon, in Buros, 1978, p. 117).

The development of the alternate item classification scheme represents a markedly different approach than that presumably used by the test authors in developing the original scheme. The latter employed a prescriptive, deductive approach, in that, presumably, they determined the subskill categories to be assessed, and attempted to write items to fit those categories. The alternative item classification scheme utilized a descriptive, inductive

approach; the existing test items were scrutinized, and, based on a subjective appraisal of common skill demands of the items, descriptive subskill categories were derived and refined.

The second step of the study consisted of a factor analysis of the inter-item correlation matrix calculated from the test results. The basic intent of the factor analysis was to determine the extent to which the TABE reading comprehension subtest is, in fact, multidimensional, as opposed to unidimensional. According to Henrysson (1971), "If only one large general factor is found, except for some other very small loadings, the test is measuring only one main dimension" (p. 154). Had the results indicated that the subtest was unidimensional, the use of either item classification scheme would have been, at best, problematical, since the subtest would have appeared to be measuring a single construct.

A secondary purpose for conducting a factor analysis was to determine whether the results would lend support to one of the two item classification schemes. Finally, there was the possibility that, if a very strong factor structure emerged in the analysis, a third item classification scheme would be indicated, different from both the original and the proposed alternative.

In the third step of the study, the nature of each subskill category, as defined by both the original item

classification scheme and the proposed alternative, was examined. The internal consistency of each subskill category was determined, and an intercorrelation matrix was computed for each of the item classification schemes. The mean p value for items in each category was calculated, in order to determine the extent to which the subskills represent levels of ascending difficulty.

The fourth, and final, step of the study attempted to determine the discriminant validity of the subskill categories, as defined by both the original item classification scheme and the proposed alternative. Using a method adapted from Hunt (1957), it was determined which of the items within each subskill category correlated significantly more highly with the corrected total score of its own subskill category than with the total score of each of the other subskill categories. Stated more concretely, if an item purports to measure skill A, it should correlate more highly with the total score of other items measuring skill A than with the total scores of items measuring skills B, C, or D.

Importance of the Study

There is little uniformity in the labels which are assigned to reading tests, subtests, and subskill categories within subtests. Lennon (1962) reports that an inspection of existing test catalogs would lead one to believe that tests can measure as many as 80 distinct reading

skills and abilities. Such alarming overspecificity strongly suggests that skill labels are more likely to reflect the skills which publishers would like to see their tests measure than the separate, identifiable skills which the tests actually measure.

Kerfoot (1968) asserts that the assessment of reading comprehension ability is encumbered by a "problem of inconsistency in both theoretical base and descriptive terminology" (p. 42). As a solution to this problem, he calls for the development of operational definitions of reading comprehension based on specific reading tasks.

Carefully developed item classification schemes have the potential to clarify, in effect, to define operationally, the type of reading comprehension measured by a given test. The labels used in any item classification scheme, of course, will still be subjective and somewhat arbitrary, but if it can be empirically demonstrated that test items cluster according to a known pattern, the item clusters themselves can serve as the best definitions of component subskills, names and labels not withstanding.

This study represents an attempt to determine the validity of two such item classification schemes, and thus, in an indirect way, to operationally define reading comprehension as it is measured by the TABE. If such definition is always desirable, it is more so in the case of the TABE; if a single test is going to continue to be used to decide

the future of thousands of adults, the least that can be done is to determine what the test is truly measuring.

Limitations of the Study

The molecular nature of this study precludes an investigation into certain broader issues which are fundamental to the testing of reading comprehension in adult education. In essence, this study is restricted to the validity of two item classification schemes as applied to one subtest of one form of one level of one test. If the research method employed here has some general applicability, the findings are clearly not generalizable beyond the performance of similar populations on the Reading Comprehension subtest of the TABE, Level D, Form 4.

The fact that the research focuses on the diagnostic component of a norm-referenced survey test should not be read as an endorsement of the use of norm-referenced tests for diagnosis. For sound diagnostic testing, test construction must focus on the adequate coverage of defined content. For a norm-referenced survey test, the goal of test construction is basically that of sorting a group of individuals based on their performance on a series of common tasks. Fisher (1978) maintains that "the manner in which norm-referenced tests are constructed virtually precludes content validity" (p. 1). The fact remains, however, that the TABE is being widely used for diagnostic

purposes, and it seemed worthwhile to examine it in hopes of defining, and perhaps improving it, rather than simply to denounce it.

Furthermore, this study was undertaken with the known restriction that the subtest content could not be altered, and that any suggested refinements would be restricted to the proposal of an alternative item classification scheme to facilitate the interpretation of test results. Nowhere are revisions suggested, particularly in terms of the deletion, refinement, or addition of test items.

Factor analysis, as it is used in this study, does not represent an attempt to construct a generalizable theory of reading. Rather, it was used quite simply to determine whether the subtest is unidimensional or multidimensional, and if the latter, whether the discovered factor structure was parsimonious enough to indicate meaningful subskill components.

In summation, this study is simply an attempt to determine the validity of two interpretive frameworks, or item classification schemes, for the results of a test that, for good or bad, is widely used for diagnosis in adult basic skills testing. Ultimately, the quality of the information one gets from a test is a function of the quality and type of test one administers. A comprehensive diagnostic instrument would yield better quality information, but since the use of the TABE continues, an attempt to

improve its diagnostic efficacy seems desirable.

Definition of Terms

Convergent valitity: the property of empirical measurement possessed by an item or test when it correlates at a high level with other items or tests purporting to measure the same trait, construct or skill (Farr, 1968).

Discriminant validity: the property of empirical measurement possessed by an item or test when it correlates at a lower level with items or tests purporting to measure different traits, constructs, or skills than with items or tests purporting to measure the same trait, construct, or skill (Farr, 1967). Also called divergent or differential validity.

Item classification scheme: a system for the grouping of test items according to the common skills they purport to measure.

Subskill: normally, a component of a very general trait, construct or skill, such as intelligence or reading ability. In this paper, the term is used in reference to alleged components of reading comprehension, which, in less molecular studies, is itself regarded as a subskill of reading.

CHAPTER II

REVIEW OF THE LITERATURE

This chapter consists of two major sections. The first section focuses on some of the issues involved in the definition, isolation, and measurement of subskills in reading comprehension. The second reviews certain statistical and psychometric attempts to isolate and measure those subskills, namely those studies employing factor analysis and those employing two types of correlational analyses in an attempt to establish the discriminant validity of subskill measures.

Section One: Issues in the Measurement of Reading Comprehension Subskills

Reading tests traditionally, and almost universally, have been directed at the measurement of three broad aspects of reading: vocabulary, comprehension, and speed (Traxler, 1958). These three logically derived components of the reading process, while broad enough to foster general, if tenuous, agreement among test constructors working in a field remarkable for a lack of agreement, are simply too broad to produce the kind of explicit information on individual examinees which is immediately applicable to instructional settings.

The demand for more specific information on reading ability has led test constructors to devise subtests and part scores within subtests which purport to measure a myriad of alleged reading skills. Traxler (1951) reviewed 28 published reading tests which purported to measure 49 distinct aspects of reading; Lennon (1962) surveyed existing test catalogs and found that this "list may be extended, if not ad infinitum, at least ad some seventy or eighty alleged reading skills and abilities" (p. 327). Unfortunately, this variety is more a function of idiosyncratic labelling on the part of test constructors than it is a credit to the precision of measurement in the reading field, and scholars have warned against the uncritical acceptance of the subskill categories proposed by test constructors (Farr, 1968, 1969; Lennon, 1962), since "In every instance, this division is arbitrary, since there is almost no research evidence supporting it" (Farr, 1969, p. 33).

The simple assertions of test constructors and publishers can hardly be regarded as evidence that subskills exist as valid, measurable constructs. Lennon (1962) sums up the major issues in the testing of subskills as follows:

> It is one thing -- and a necessary thing -- to make a careful analysis of reading ability, to spell out its various supposed components in detail, and to prepare extensive lists or charts of the specific skills or abilities to

> serve as statements of desired goals or outcomes of the reading program. It is quite another thing to demonstrate that these manifold skills or abilities do, in fact, exist as differentiable characteristics of students; and still a third thing to build tests which are in truth measures of one or another of these skills, and not of some more general, pervasive reading ability. (pp. 327-328)

The definition, validation, and measurement of reading comprehension subskills has generated a good deal of scholarly debate. On the one hand, many writers have conducted subjective "armchair analyses" of reading comprehension, carefully delineating subskills with little or no reference to experimental or statistical evidence, and with little or no attempt at validation (Auerbach, 1971; Davis, 1972). This broad category of writers is quite varied, and ranges from serious scholars employing careful, if subjective, logic to expedient test constructors and publishers who, with little regard for theory (Kingston, 1960), base their subskills on a priori goals of assessment and on appeals to content validity. On the other hand, certain researchers have attempted to discover and isolate distinct subskills through a variety of statistical and psychometric techniques. (See reviews in: Davis, 1972; Farr, 1969; Lennon, 1962.) Research of this type, which flourished from the early 1940's to the early 1970's, endeavored to infer, through factor analysis, or to validate, through various correlational procedures, the subskills of reading comprehension as manifested in test

data.

It is not surprising that virtually all empirical attempts to identify subskills of reading comprehension have focused on the product, in terms of test results, rather than on the process of reading comprehension. Traxler (1957) points out:

> Since, except for a superficial estimate of speed, no aspect of silent reading can be measured without interrupting the process, we customarily resort to a kind of addendum to the reading process itself. We ask a series of questions when the reading is finished and hope that the answers to those will indicate the quality of the comprehension which took place while the reading was being done. (p. 2)

Such a condition undoubtedly limits the utility of these analyses for the building of a generalizable theory of reading; as Farr (1968) points out, caution must be exercised in distinguishing between "the makeup of reading ability versus the validity of sub-skills of reading ability as measured by most standardized tests" (p. 189). Raygor (1966), in criticizing the research of Holmes and Singer (1964), points to a limitation common to all empirical studies based on test data: the validity and reliability of the findings are dependent upon the validity and reliability of the tests used to gather data (Farr, 1969).

Validity is a perennial problem in standardized tests of reading comprehension. Test publishers typically advance claims of content validity, an amorphous concept,

based on appearances rather than evidence, and one which Messick (1980) has termed "not validity at all" (p. 1015). When empirical evidence is supplied, it almost always takes the form of correlation coefficients, showing that a test or subtest has a substantial positive relationship with other tests purporting to measure the same skill or ability. Such validity coefficients are comparatively easy to obtain, considering that most measures of academic abilities tend to correlate substantially. In the case of tests of reading ability measured by multiple choice questions, the much harder task is to establish that a subtest or subscore correlates at a lower level with subtests or subscores measuring allegedly different constructs (Farr, 1968).

Messick (1980) argues:

> Construct validation entails both confirmatory and disconfirmatory strategies, one to provide convergent evidence that the measure in question is coherently related to other measures of the same construct as well as to other variables that it should relate to on theoretical grounds, and the other to provide discriminant evidence that the measure is not related ur[illegible]v to exemplars of other distinct cons[illegible] .. (p. 1016)

Discriminant validity, while always desirable in terms of defining the construct being measured, is essential if a subtest or part score is to have any practical diagnostic value (Farr, 1968).

The extent to which one might reasonably expect subskills of reading comprehension to exhibit discriminant

validity, however, is largely determined by the pattern of relationships existing among the subskills. Chapman (1969) has proposed three hypothetical matrices of intercorrelations among unspecified reading comprehension subskills which are illustrated in Table 1. Few scholars would subscribe to the first matrix, which postulates that comprehension subskills are learned and used independently; if this were the case, however, evidence of discriminant validity would be easily obtained providing that measurement was precise enough to capture the different subskills. The second matrix, which postulates that reading comprehension is unidimensional, would suggest that the subskills are totally without discriminant validity. The third matrix postulates that subskills are separate but correlated, and that the more basic skills are included, at least in part, in the higher order skills. Such hierarchical patterning, though supported by common logic and empirical evidence (Davis, 1972), necessarily confounds any demonstration of discriminant validity.

Section Two: Psychometric and Statistical Evidence of Reading Comprehension Subskills

In 1941, Traxler (quoted in Lennon, 1962) expressed the hope that the controversies surrounding the isolation and measurement of reading comprehension subskills might find a "mathematical resolution . . . by means of a thorough-going factor analysis of the abilities which

TABLE 1

CHAPMAN'S HYPOTHETICAL MATRICES
OF INTERCORRELATIONS*

Uncorrelated, or Isolated, Skills Theory

Test	A	B	C	D
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1

Global-Skill Theory

Test	A	B	C	D
A	1.00	.90	.90	.90
B	.90	1.00	.90	.90
C	.90	.90	1.00	.90
D	.90	.90	.90	1.00

Hierarchical Skills Theory

Test	A	B	C	D	E
A	1.00	.7I	.58	.62	.44
B	.71	1.00	.81	.71	.63
C	.58	.81	1.00	.87	.78
D	.62	.71	.87	1.00	.89
E	.44	.63	.78	.89	1.00

*Reproduced from Davis, 1972, p. 671.

enter into silent reading" (pp. 327-328). The much-reviewed factor analytic studies of reading which were conducted in the ensuing two decades (see detailed reviews in: Davis, 1972; Farr, 1969; and Lennon, 1962), however, were characterized by inconsistent or inconclusive findings. According to Farr (1969),

> The studies showed only limited agreement as to the number of factors: some named only one factor (Conant, 1942, for instance) while others (such as Davis, 1941) found six. That there should be such disparity is not surprising: factor analysis studies are dependent on both the data collected and the manner in which it is collected. The same tests were not used in each study and those which were used measured a wide array of elements ranging from personality factors, social studies and science achievement, and intelligence to reading, as defined by as many publishers and researchers as tests that were used. Given this situation, it is hardly surprising that the factors thought to comprise reading lack consistency from study to study. (p. 3)

Lennon (1962), in reviewing these same studies, concluded from various findings:

> that we may recognize and hope to measure reliably the following components of reading ability: (1) a general verbal factor, (2) comprehension of explicitly stated material, (3) comprehension of implicit or latent meaning, and (4) an element which might be termed 'appreciation'. (p. 334)

Lennon's conclusions, however, are somewhat puzzling in light of the disparate nature of the findings of these studies, and have been termed by Farr (1969) as "perhaps an over-simplification" (p. 3).

All of the factor analytic studies mentioned above differ, both in method and goal, from a more recent study by Powers and Gallas (1980). The earlier studies attempted to build generalizable theories of reading by factor analyzing the intercorrelation matrix of aggregate scores on selected tests. Powers and Gallas employed more molecular data to ask a considerably more modest research question: What is the factor structure of the Comprehensive Tests of Basic Skills Reading Comprehension Test, Level 2, Form S, when administered to sixth and seventh grade Title I students? Despite seemingly adequate sample sizes (220 sixth graders; 361 seventh graders; 45 test items), the findings indicate remarkedly different factor structures for the two groups. The authors offer no suggestion as to why two such similar populations should produce factor structures with virtually no overlap, nor do they attempt to label the many factors (19 for the sixth graders; 17 for the seventh graders) with eigenvalues greater than one. They do conclude, however, that the test is multidimensional, and that, due to the lack of coherence among test items, care should be exercised in interpreting test results for Title I students.

Generally speaking, factor analysis has not succeeded in clarifying the confusion in the field regarding the nature of reading comprehension subskills. For the most part, findings have required so much in the way of.

subjective interpretation that they are often open to debate. Several studies have been reinterpreted by subsequent researchers, and radically different conclusions drawn; see, for example, Thurstone's (1946) and Thorndike's (1973 - 1974) reanalyses of the Davis (1941) data. When factor analysis has been used to generate theories of reading -- and this seems to have been its primary application in the field -- the results are ambivalent and the factors few. It is not at all unusual to infer a single factor called "general comprehension" or "general reading ability" (Artley, 1944; Conant, 1942; Harris, 1947; Stoker & Kropp, 1960; Thurstone, 1946; Traxler, 1941). When factor analysis was used by Powers and Gallas (1980) to decompose a single test, the results were so fractionated as to defy meaningful interpretation, particularly with relation to comprehension subskills.

Hunt (1957) examined the discriminant validity of the six subskill categories suggested by Davis's (1941) original findings. Twenty-one judges classified each of 204 multiple-choice items in terms of the following categories: word knowledge, reasoning ability, literal meanings, inference, organization, and literary devices and techniques. Each of the subskill categories contained 34 items, and each item "received sufficient agreement from the consultants" (p. 163) in order to be included in one of the categories. Hunt's sample consisted of 370 college students enrolled in

a reading improvement course. Hunt approximated the correlation coefficients between each of the 204 items and total scores for each of the six categories. These estimated coefficients were then corrected for self-correlation and for attenuation resulting from the unreliability of the total scores. Finally, the mean of these twice - corrected coefficients was obtained between each of the 34 items and the category in which it was included, and between each of the 34 items and the five categories in which it was not included.

When these means were compared, it was discovered that only vocabulary items correlated significantly more highly with the subskill category of which they were a part than with the remaining categories. Hunt concludes:

> For practical purposes, then, each group of items based on the reading passages discriminates equally for all of the five criterion skill measures. The different item groups are apparently not measuring differences in performance on the part of the examinees. (p. 169)

Davis (1972) points out that, whereas Hunt's findings may be regarded as evidence that the items in each category measured much the same general ability, it is important to note that "tiny components of variance unique to certain types of items might be lost in approximation procedures used in item analysis and in the corrections for self-correlation and for attenuation" (pp. 664-665).

In a study attempting to investigate both the convergent

and the discriminant validity of subtests of several upper level reading tests, Farr (1968) employed a procedure suggested by Campbell and Fiske (1959). This procedure, commonly called the multitrait-multimethod model, consists of calculating a common correlation matrix for a number of tests, each having several subtests in common with the other tests. For example, if Tests A, B, and C each contain subtests purporting to measure subskills d, e, and f, a 9 x 9 matrix would be constructed. The diagonals of the matrix would contain standard validity coefficients; skill d, as measured by each of the three tests, should evidence substantial positive correlations. The off-diagonal triangles contain correlation coefficients between supposedly different skills; if these are substantially lower than the values in the diagonals, this can be taken as evidence of discriminant validity. Moreover, if the relative magnitude of the coefficients in each of the triangles give evidence of a common pattern, conclusions may be drawn regarding the relationships of the various subskills.

Farr (1968) constructed multitrait - multimethod correlation matrices for three secondary reading tests (administered to 67 ninth graders) and for three college level reading tests (administered to 91 undergraduate education majors). The subtests in question were vocabulary, comprehension, and speed. Farr's findings were somewhat

inconclusive:

> This lack of discriminant validity is most apparent in the study with the ninth grade students. With the college population, some of the subtests did seem to give evidence of discriminant validity. (p. 190)

Farr argues, however, more for the value of the procedure than for the importance of his findings:

> One of the most important findings of the preceding studies is the recognition of the value of the Campbell-Fiske model for investigating the construct validity of the sub-tests of reading test batteries. The general confusion in investigating construct validity, usually by the use of factor analysis, is exemplified by the Lennon (1962) article. (p. 189)

The two methods (Farr, 1968; Hunt, 1957) discussed above for determining the discriminant validity of subtests of reading have a common limitation, in that they can only be used for validating existing subskills (as represented by subtests and part scores), and are useless for the purposes of test construction or for discovering skill clusters among existing items. Moreover, the multitrait-multimethod model, which appears to be more a method of data presentation than an actual empirical procedure, is further limited by the fact that it can only supply relative information about the validity of existing measures of subskills. A major shortcoming of the Hunt method, as operationalized above, is the use of approximated correlations, and the means of those approximated correlations; with the improved computer capabilities

of the past 25 years, such approximations are no longer necessary.

CHAPTER III

PROCEDURES

In an attempt to determine the validity of two item classification schemes for the TABE Reading Comprehension subtest, this study employed a sequence of disparate procedures. In order to facilitate the description of these procedures, this chapter is organized in seven separate sections.

Section One describes, in narrative form, the method that was used to develop the alternate item classification scheme. Section Two describes the data base for the empirical procedures described in the remaining five sections. Sections Three through Seven describe the statistical procedures used to answer the five research questions stated in Chapter I. For the convenience of the reader, these questions will be restated as the appropriate sections are described.

Section One: The Development of an Alternative Item Classification Scheme

The 45 test items were subjectively examined by this investigator without reference to the original item classification scheme. No a priori decisions were made regarding either the nature of the subskill categories to

be derived, or the number of categories to be created. Each item was examined individually, and a brief notation was made describing the skill demands of the items. After each of the items was described, the notations were examined to determine which of the items appeared to be tapping common skills. This originally resulted in 11 "clusters" of items, with the largest cluster containing eight items and the smallest containing only one. The items in similar clusters were re-analyzed, and the descriptive notations for several clusters were broadened to subsume the smaller clusters, which were deemed impractical due to their overspecificity. Ultimately, six subskill categories were isolated and defined.

Section Two: Collection and Preparation of the Data

Data were gathered at a large, urban career counseling and assessment program in central New Jersey between May and October of 1980. Although no demographic information is available on the specific sample of examinees whose test results comprised the data for this study, program regulations required that all participants were out-of-school adults and currently unemployed. Program staff provided the additional information that the participants were largely minority, and that educational level was heterogeneous.

All program participants were required to take the TABE

as part of the assessment component of the program. During the period in which data were collected for this study, a total of 298 participants took the TABE, Level D, Form 4. This entire population was used for this study.

Of the 298 total population, 242 remained after the data were "cleaned." Data cleaning consisted of deleting the test results of any examinee who fell into one or both of the following categories:

-Non-completers. Any test missing a response to any item was deleted; because of the timed nature of the test, most missing responses occurred at the end of the test.

-Those scoring at or below the chance level.

The responses for the surviving 242 cases were coded dichotomously (correct-incorrect) and keypunched.

Section Three: Principal Axes Factor Analysis

A factor analysis of the inter-item correlation matrix was conducted to answer the following research question:

Is the reading comprehension subtest unidimensional or multidimensional?

The data were factor analyzed using SPSS, subprogram FACTOR, method PA2; method PA2 is an exploratory principal axes factoring method which replaces the main diagonal of the correlation matrix with estimates of communalities, and then improves these estimates through an iteration procedure (Kim, 1975).

The initial factor solution was then rotated to an oblique solution. Korth (1975) points out that the use of an oblique transformation is "based on the belief that it is better to let the data set the solution, rather than imposing the possibly artificial restriction of orthogonality" (p. 166-167). Korth further maintains that the oblique transformation is especially appropriate when there are theoretical reasons to believe that the factors may reveal a hierarchical structure.

Section Four: Calculation of Reliability Coefficients

A reliability coefficient, using Kuder-Richardson Formula 20 (KR20), was calculated for the total subtest and for each of the subskill categories specified by the two item classification schemes. This was done in order to address the following research question:

> To what extent are the subskill categories presented in each of the item classification schemes internally consistent?

The KR20 coefficients are based on the obtained inter-item correlation matrix, and thus are sound indicators of internal consistency.

The number of items in a test, or in this case, in a subskill category, greatly effect the magnitude of the obtained KR20 coefficients. In order to compare directly the internal consistency of subskill categories having

different numbers of items, an adjustment is needed. Consequently, the Spearman-Brown Prophecy Formula was used to calculate what the coefficients would be if each of the subskill categories contained 25 items of a quality and nature identical to the items they now contain. This hypothetical increase in the number of items to an arbitrary constant was proposed only for heuristic purposes, i.e., to make possible direct comparisons among categories, and should not be read as a recommendation to alter the test.

Section Five: Calculation of Intercorrelations and p Values

Separate intercorrelation matrices for each of the item classification schemes were computed. This was done as the first step to answering the following research question:

> To what extent are the subskill categories presented in each of the item classification schemes hierarchical?

The resulting matrices were then subjectively compared to the three hypothetical matrices of reading comprehension skills proposed by Chapman (1969) and presented in Chapter II of this report. Ideally, such a comparison would suggest that the subskill categories are either hierarchical, orthogonal, or measuring a single, global construct.

In addition, the mean p value for items in each category was computed, in order to determine the extent

to which the subskills represent levels of ascending difficulty.

Section Six: Determination of the Discriminant Validity of the Subskill Categories

A three-step procedure was used for each of the test items as it is classified according to each of the item classification schemes, in order to answer the following research question:

> To what extent do the subskill categories presented in each of the item classification schemes show evidence of discriminant validity?

Each of the three steps was repeated for each of the 45 test items to determine the discriminant validity of the original item classification scheme; the entire process was then repeated to assess the validity of the alternative item classification scheme.

Since the procedure is rather complicated, this description will be illustrated with an example from the actual analysis, as it was performed on one item from the Reference Skills subskill category from the original item classification scheme. According to this scheme, Reference Skills are measured by Items 1 through 6; the other subskill categories in the scheme are Recall, Main Idea, and Inference.

In the first step, each item was correlated with the

corrected total score of its own subskill category. The resulting point biserial correlation coefficient will be termed the corrected item-total correlation. In the example, Item 1 was correlated with the sum of items 2 through 6.

In step two, each item was correlated with the total scores of the subskills categories of which it is not a part. In this example, Item 1 was correlated with Recall, then with Main Idea, and finally with Inference; this step yields three correlation coefficients.

Step three consisted of a series of one-tail t tests ($p < .05$), testing the hypothesis that a given item will correlate significantly more highly with its own subskill category than with each of the other categories. Because of the highly dependent nature of the obtained correlation coefficients, the following formula, developed by Hotelling (in Walker and Lev, 1953, p. 257) was used to compute the t statistic:

$$t = (r_{xz} - r_{yz}) \frac{(N - 3)\ (1 + r_{xy})}{2(1 - r^2_{xy} - r^2_{xz} - r^2_{yz} + 2r_{xy}r_{xz}r_{yz})}$$

in which z = the item score

x = the corrected total for its own subtest

y = the total score on another subtest

In the example, then, three t tests would have to be performed to determine whether Item 1 correlated more highly with itw own corrected subtest total than with Recall, Main Idea, and Inference, respectively.

In order to complete step three, 135 t tests were conducted for the original item classification scheme (i.e., 45 items x 3 t tests), and 225 t tests were conducted for the alternative item classification scheme, which included six subskill categories (i.e., 45 items x 5 t tests).

Despite the fact that these statistical tests were performed on individual items, they were, in fact, testing the validity of the subskill categories by determining how well the items in a given category "fit" together. In order to make the results of the 360 t tests interpretable, and in order to facilitate comparisons between the two item classification schemes, each item was assigned one of the following labels, depending on the results of the t tests conducted according to each scheme:

valid: an item is considered valid if it correlates significantly more highly with the corrected-total of its own subtest than with each of the other subtests.

indeterminate: an item is considered to possess indeterminate validity if it correlates significantly more highly with the corrected total of its own subtest than with at least one, but not all, of the other subtests.

invalid: an item is considered invalid if it does not correlate significantly more highly with the corrected-total of its own subtest than with any one of the other subtests.

Section Seven: Comparison of the Two Item Classification Schemes

This last section represents an attempt to synthesize some of the preceding findings in order to answer the following research question:

Do the item classification schemes differ in terms of internal consistency and discriminant validity?

No statistical tests were performed, but the two schemes were compared in terms of:

- the extent to which they "fit" the inferred factor structure
- calculated KR20 coefficients
- reliability coefficients after adjustment with the Spearman-Brown Prophecy Formula
- the extent to which they matched Chapman's proposed subskill matrices
- the extent to which the presented subskills indicate ascending levels of difficulty
- the number of invalid, indeterminate, and invalid items.

CHAPTER IV

FINDINGS AND DISCUSSION

In order to facilitate the description of the findings and the accompanying discussion, this chapter will be divided into seven sections roughly parallelling the seven sections presented in Chapter III. For the convenience of the reader, the research questions will be restated as the appropriate sections are presented.

Section One: The Alternative Item Classification Scheme

The alternative item classification scheme contained six subskill categories. The names of the categories, and the reading demands imposed by test items in each category, appear below.

1. Reference. Items in this category require the examinee to locate explicitly stated information in tabular and non-textual arrays.
2. Literal. Items in this category require the examinee to locate explicitly stated information in the stimulus passages. In all cases, the vocabulary and syntax of the test item closely resemble the vocabulary and syntax of the stimulus passage.

3. Paraphrase. Items in this category require the examinee to locate explicitly stated information in the stimulus passages. In all cases, the vocabulary and/or the syntax of the test item is markedly different from that of the stimulus passage.

4. Vocabulary in Context. Items in this category require the examinee to derive the meaning of a word or phrase from information supplied in the stimulus passage.

5. Reasoning. Items in this category require the examinee to engage in various types of reasoning based on information supplied in the stimulus passages and non-textual arrays. Eight of the 13 items in this category require the examinee to engage in symbolic reasoning tasks based on presented grids, and on described grids which the examinee must construct according to specifications provided in the stimulus.

6. Synthesis. Items in this category require the examinee to recognize the main idea/central thought of passages or specific paragraphs.

The publisher's original item classification scheme presented only four subskill categories, and provided neither definition nor rationale for the presented categories. The original categories are: Using Reference

Skills (hereafter, Reference), Recalling Facts (hereafter, Recall), Understanding Main Ideas (hereafter, Main Idea), and Making Inferences (hereafter, Inference).

With the exception of the Reference categories, the two item classification schemes are substantially different. Figure 1 depicts the intersection of the two schemes, while Table 12 in Appendix A presents the classification of the specific items according to the two schemes.

Section Two: Description of the Data Base

The data for all empirical analyses consisted of the completed subtests of 242 adult examinees, as described in the preceding chapter. Summary statistics for the distribution of obtained total subtest scores are presented in Table 13 in Appendix B. Although all measures of central tendency were toward the higher end of the scale, the scores evidenced substantial variance overall.

Such was not the case for some of the subskill categories. (See Table 2.) The Reference categories in both schemes evidenced extremely restricted variance, and the majority of the remaining categories displayed at least moderate restriction. The restricted variances of the subskill categories undoubtedly has a suppressing, but not specifiable, impact on many of the later correlational analyses of this study.

Section Three: Results of the Factor Analysis

This segment of the study was conducted in order to

Original Subskill Categories

Alternative Subskill Categories	Reference	Recall	Inference	Main Idea	Row Total
Reference	6		1		7
Literal		6			6
Paraphrase		4	4	1	9
Vocabulary in Context			5		5
Reasoning		1	11	1	13
Synthesis			1	4	6
Column Total	6	11	22	6	

Figure 1. Crosstabulation of item frequency for two item classification schemes.

TABLE 2

DESCRIPTIVE STATISTICS FOR DISTRIBUTIONS OF SUBSKILL CATEGORY RAW SCORES FOR SAMPLE POPULATION (n=242)

Category	Number of Items	Mean	Standard Deviation	Range	Maximum	Minimum
Original Scheme						
Reference	6	5.3	1.0	6	6	0
Recall	11	8.6	1.9	9	11	2
Inference	22	14.4	4.0	18	22	4
Main Idea	6	4.1	1.2	6	6	0
Alternative Scheme						
Reference	7	6.3	1.0	6	7	1
Literal	6	4.9	1.2	5	6	1
Paraphrase	9	7.1	1.9	6	9	3
Vocabulary in Context	5	4.0	1.2	5	5	0
Reasoning	13	6.7	2.7	13	13	0
Synthesis	5	3.5	1.2	5	5	0

answer the following research question:

Is the reading comprehension subtest unidimensional or multidimensional?

In addition, the results of the exploratory factor analysis were examined to see if they would lend support to one or the other of the item classification schemes, or if they could be used as a foundation for the creation of a third and different scheme.

During the course of the factor analysis, three different factor solutions were considered. The first, which involved retaining any factors with eigenvalues greater than 1.0 (Kim and Mueller, 1978), retained 16 factors (see Table 14 in Appendix C). The second, which was attempted in light of the preponderance of small factors resulting from the original analysis, used a scree test in order to determine whether the number of factors might reasonably be reduced. A scree test simply consists of a subjective examination of the plotted eigenvalues in order to determine the point at which the eigenvalues "begin to level off forming a straight line with an almost horizontal slope" (Kim and Mueller, 1978, p. 44). The plotted eigenvalues for the 16 factors indicated a marked levelling after the fourth factor (see Figure 2); consequently, the second solution retained four factors. The third, and final, solution which was considered was a one factor solution, since, if the test was, in fact, unidimensional,

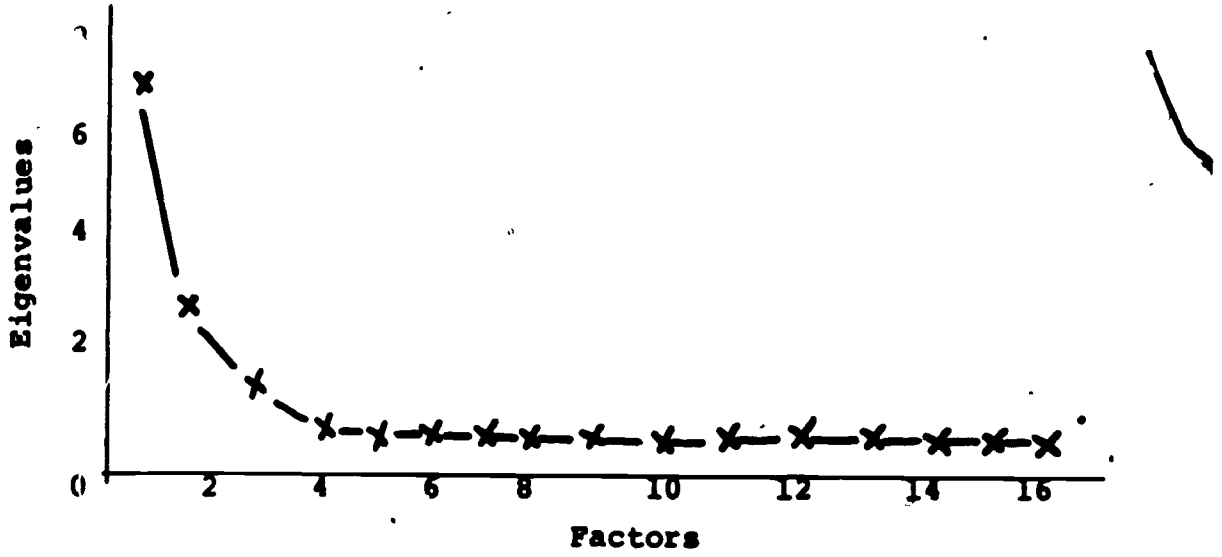

Figure 2. Eigenvalues for the first 16 factors of factor analysis of the principal axes TABE Reading Comprehension Subtest (n=242)

a single large factor should account for a good deal of the variance (Henryssón, 1971).

Table 3 allows for a comparison of the three solutions. None of the solutions is particularly good, in that anywhere from 24 percent to 31 percent of the items do not load on any factor. The fact that Solution III, the one factor solution, accounts for only 15.1 percent of the total variance suggests that the subtest is not testing a single, undifferentiated construct.

Additional evidence for the multidimensionality of the test is suggested by the very low intercorrelations of the 16 factors in Solution I. When oblique rotation is used, as it was in the current analysis, there is no imposition of orthogonality on the derived factors. As Kim and Mueller (1978) point out, "if one finds an orthogonal structure when oblique rotations are applied . . . then one can claim that the underlying structure is orthogonal" (p. 78). The mean intercorrelation coefficient for the 120 entries in the 16 x 16 factor correlation matrix is .03, and the standard deviation is .12. This suggests that there are 16 virtually uncorrelated factors, or dimensions, at work in the subtest.

Interpretation of Solution I was problematical, since nothing approaching simple structure was evident in the factor matrix. Loadings were generally low, and it proved necessary to arbitrarily define an unusually low cutoff

TABLE 3

THREE FACTOR SOLUTIONS FOR THE TABE READING COMPREHENSION SUBTEST WHEN ADMINISTERED TO 242 ADULT EXAMINEES

Solution Number	Number of Factors	Percent Common Factor Variance	Number of Items Loading* on One Factor	Number of Items Loading on Two or More Factors	Number of Items Not Loading on any Factor
I	16	61.4	31	3	11
II	4	28.0	32	1	12
III	1	15.1	31	0	14

*Only factor loadings ≥.30.

point in deciding which items loaded "substantially" on a given factor. Table 15 in Appendix C depicts all factor loadings greater than or equal to .30 for the model. Because of the generally weak loadings, because only a few items loaded on each of the 16 factors, and because the items that did load on any one of the factors did not seem to be tapping a common skill not being measured by non-loading items on the test, no substantive interpretation of the factors was warranted. Furthermore, it was concluded that the factor analysis did not suggest the formulation of a third and different item classification scheme.

The final goal of the factor analysis was to determine whether the results would lend support to either of the item classification schemes. As depicted in Figures 3 and 4, items in each of the subskill categories were, in every case, spread out over a number of factors. It was therefore concluded that the factor analysis did not support either scheme.

Section Four: Findings Relating to Reliability and Internal Consistency

This segment of the study focuses on the following research question:

> To what extent are the subskill categories presented in each of the item classification schemes internally consistent?

To answer this question, and to determine the practical

Factors

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Reference		3			1	1			1							
Recall	2					1	1	1	1				1	1	1	1
Inference	3		1	2	1			2	1	2	1	1		1		
Main Idea					1		1		1		1	1	1			

Figure 3. Crosstabulation of item frequencies: factors by subskill categories from original item classification scheme

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Reference		3			1	1			1							
Literal	1						1	1	1							
Paraphrase	1				1	1			1		1	1		1	1	1
Vocabulary in Context	2		1													
Reasoning	1			2				2		2	1		1	1		
Synthesis					1		1		1			1				

Figure 4. Crosstabulation of item frequencies: factors by subskill categories from alternative item classification scheme

utility of the categories in diagnosing learning difficulties, KR20 coefficients were calculated for the total subtest and for each of the subskill categories.

The KR20 coefficient for the total subtest score, though only of passing interest for this study, was only marginally adequate at .86. The KR20 coefficients for the subskills, not surprisingly, did not fare as well (see Table 4).

Three interrelated factors which directly affect the reliability of a given test are test length, the level of ability of the group tested, and the range of ability in the group (Thorndike and Hagen, 1977). All three of these factors had a negative impact on the reliabilities of the subskill categories. First, as was shown in an earlier section (see again Table 2), the variance of scores in the subskill categories was quite narrow. Second, the level of ability, as evidenced by the means in Table 4 and by the p values of the items (see next section), was at the high end of the scale in each of the subskill categories. Both of these factors, that is, a narrow range and a high level of ability, tend to restrict the total variance of each of the categories, which in turn restricts the inter-item correlations, which, ultimately, attentuate the KR20 coefficients. Third, and most obvious, the number of items in each of the subskill categories is exceedingly small.

TABLE 4

RELIABILITY DATA FOR TOTAL SUBTEST AND FOR SUBSKILL CATEGORIES (n=242)

Measure	Number of Items	Mean Inter-Item Correlation	KR 20 Coefficient	Standard Error of Measurement
Original Scheme				
Reference	6	.19	.50	.7
Recall	11	.11	.55	1.2
Inference	22	.13	.78	1.9
Main Idea	6	.07	.30	1.0
Alternative Scheme				
Reference	7	.16	.51	.7
Literal	6	.12	.44	.9
Paraphrase	9	.16	.64	1.1
Vocabulary in Context	5	.17	.51	.8
Reasoning	13	.12	.64	1.6
Synthesis	5	.10	.38	.9
TOTAL SUBTEST	45	.12	.86	2.6

The low KR20 coefficients presented in Table 8 are, in and of themselves, sufficient evidence of the psychometric inadequacy of the subskill categories for the diagnosis of individual learning needs. Reliability is a prerequisite of validity, since a "test must measure something before it can measure what we want it to measure" (Thorndike and Hagen, 1977, p. 87).

In order to facilitate comparisons among subskill categories containing different numbers of items, and in order to determine the extent to which the inadequacy of the observed KR20 coefficients is a function of test length (as opposed to quality), the Spearman-Brown Prophecy Formula was used to predict the hypothetical reliability of each subskill category had it contained 25 test items. The formula assumes that the quality of the test items and the nature of the examinees remains constant.

The results appear in Table 5. Comparisons are somewhat problematical without a reference point, since there is no theoretically defined "good" reliability coefficient for a 25 item test. To provide such a reference point, a hypothetical reliability coefficient for the total subtest, if reduced to 25 items, was computed. For the original item classification scheme, two of the four subskill categories surpass the hypothetical reliability coefficient for the total test, and two fall short. For the alternative

TABLE 5

HYPOTHETICAL RELIABILITY COEFFICIENTS FOR SUBSKILL CATEGORIES AND TOTAL SUBTEST CONTAINING A UNIFORM NUMBER OF ITEMS

Measure	Number of Item	Reliability Coefficients
Original Item Classification Scheme		
Reference	25	.81
Recall	25	.74
Inference	25	.80
Main Idea	25	.64
Alternative Item Classification Scheme		
Reference	25	.79
Literal	25	.77
Paraphrase	25	.83
Vocabulary in Context	25	.84
Reasoning	25	.77
Synthesis	25	.75
TOTAL SUBTEST	25	.77

item classification scheme, three of the six subskill categories surpass the hypothetical reliability coefficient for the total subtest, two match it, and one falls slightly short. The results of this analysis, though somewhat tenuous, suggest a superiority of the alternative item classification scheme.

(Identical results could have been obtained by simply examining the inter-item correlations, on which the calculation of the reliability coefficients are based. The hypothetical reliability coefficients were used only because they are in a more familiar metric, and thus more readily understandable.)

Section Five: Intercorrelations and p Values

Intercorrelation matrices for the subskill categories in each of the item classification schemes were calculated in an attempt to answer the following research question:

> To what extent are the subskill categories presented in each of the item classification schemes hierarchical?

The matrices are presented in Tables 6 and 7.

The intercorrelation coefficients presented in Tables 6 and 7 are, with few exceptions, relatively uniform. All of the coefficients for the alternative scheme (Table 7), and five out of six for the original scheme (Table 6), fall somewhere between .31 and .59. Consequently, neither matrix appears to fit any of the patterns set out in Chapman's (1969)

TABLE 6

INTERCORRELATION COEFFICIENTS FOR SUBSKILL CATEGORIES FROM ORIGINAL ITEM CLASSIFICATION SCHEMES (n=242)

Category	1	2	3	4
1. Reference	1.00	.45	.45	.31
2. Recall		1.00	.72	.49
3. Inference			1.00	.56
4. Main Idea				1.00

TABLE 7

INTERCORRELATION COEFFICIENTS FOR SUBSKILL FROM ALTERNATIVE ITEM CLASSIFICATION SCHEME (n=242)

Category	1	2	3	4	5	6
1. Reference	1.00	.41	.45	.37	.34	.33
2. Literal		1.00	.59	.58	.49	.43
3. Paraphrase			1.00	.56	.57	.50
4. Vocabulary in Context				1.00	.45	.37
5. Reasoning					1.00	.41
6. Synthesis						1.00

three hypothetical matrices (see again Table 1).

In addition, the coefficients are suspiciously low when one considers that the subskill scores represent measures obtained, in many cases, from the reading of common stimulus passages. The low intercorrelations can, to a large extent, be explained by the low reliabilities of the subskill measures themselves, since the magnitude of a correlation coefficient is always attenuated by reliabilities less than 1.0.

Standard formulas to correct for attenuation, however, would require that an individual examinee's errors on the two measures being correlated are random and orthogonal (Walker and Lev, 1953; Carmines and Zeller, 1979). This is clearly not the case in regard to the subskill categories; in many cases a single stimulus passage on the subtest is followed by test items from several subskill categories. Walker and Lev (1953) point out that if the assumption of random and orthogonal errors is violated, the formula to correct for attenuation will overestimate the corrected reliability coefficient. Attempts to apply the correction formula to this data resulted in inflated, and therefore invalid, coefficients (see Tables 8 and 9).

In summation, the intended comparison of the observed intercorrelation matrices to Chapman's hypothetical matrices was frustrated by the low reliabilities of the subskill categories. Little confidence could be placed

TABLE 8

INTERCORRELATIONS COEFFICIENTS* FOR SUBSKILL CATEGORIES FROM THE ORIGINAL ITEM CLASSIFICATION SCHEME CORRECTED FOR ATTENUATION

Category	1	2	3	4
1. Reference		.86	.73	.80
2. Recall			1.10	1.21
3. Inference				1.16
4. Main Idea				

*Coefficients > 1.0 are the result of overestimation due to non-random error.

TABLE 8

INTERCORRELATIONS COEFFICIENTS* FOR SUBSKILL CATEGORIES FROM THE ORIGINAL ITEM CLASSIFICATION SCHEME CORRECTED FOR ATTENUATION

Category	1	2	3	4
1. Reference		.86	.73	.80
2. Recall			1.10	1.21
3. Inference				1.16
4. Main Idea				

*Coefficients > 1.0 are the result of overestimation due to non-random error.

TABLE 9

INTERCORRELATION COEFFICIENTS* FOR SUBSKILL CATEGORIES FROM THE ALTERNATIVE ITEM CLASSIFICATION SCHEME CORRECTED FOR ATTENUATION

Category	1	2	3	4	5	6
Reference		.87	.79	.73	.60	.75
Literal			1.11	1.22	.92	1.05
Paraphrase				.98	.89	1.01
Vocabulary in Context					.79	.84
Reasoning						.83
Synthesis						

*Coefficients > 1 0 are the result of overestimation due to non-random error.

in the attenuated, uncorrected coefficients; still less could be placed in the inflated, corrected coefficients. As a result, no conclusions were drawn regarding the possible hierarchical nature of the subskill categories.

In the next phase of the analysis, item difficulty p values were calculated for the items in each subskill category. The mean p values were generally high (see Table 10). In the case of the Reference categories in both of the item classification schemes, the mean p values were so high that it is doubtful that these categories contribute much useful information about individual performance difference of examinees. The high p values obtained suggest that, in many cases, the items on the subtest were simply too easy for the sample population.

The mean p values suggest that the subskill categories represent levels of ascending difficulty. In many cases, however, the differences between the means of adjacent categories was small enough to be of no substantive significance.

Section Six: Discriminant Validity of the Subskill Categories

A three-step procedure, described in Chapter III, was used to answer the following research question:

To what extent do the subskill categories presented in each of the item classification schemes show evidence of discriminant validity?

TABLE 10

SUMMARY STATISTICS FOR ITEM DIFFICULTY (p) VALUES
BY SUBSKILL CATEGORY AND FOR TOTAL
SUBTEST

Measure	Mean	SD	Minimum	Maximum
Original Scheme				
Reference	.89	.13	.64	.99
Recall	.78	.13	.47	.95
Main Idea	.68	.18	.50	.94
Inference	.66	.20	.32	.96
Alternative Scheme				
Reference	.90	.12	.64	.99
Literal	.82	.10	.65	.95
Vocabulary in Context	.79	.11	.66	.96
Paraphrase	.78	.09	.64	.93
Synthesis	.71	.17	.54	.94
Reasoning	.51	.15	.32	.79
TOTAL SUBTEST	.72	.19	.32	.99

TABLE 10

SUMMARY STATISTICS FOR ITEM DIFFICULTY (p) VALUES BY SUBSKILL CATEGORY AND FOR TOTAL SUBTEST

Measure	Mean	SD	Minimum	Maximum
Original Scheme				
Reference	.89	.13	.64	.99
Recall	.78	.13	.47	.95
Main Idea	.68	.18	.50	.94
Inference	.66	.20	.32	.96
Alternative Scheme				
Reference	.90	.12	.64	.99
Literal	.82	.10	.65	.95
Vocabulary in Context	.79	.11	.66	.96
Paraphrase	.78	.09	.64	.93
Synthesis	.71	.17	.54	.94
Reasoning	.51	.15	.32	.79
TOTAL SUBTEST	.72	.19	.32	.99

A subskill category was considered to manifest discriminant validity to the extent that items in the category correlated more highly with the corrected-total of that category than with the other categories in the subtest. An item was considered valid if it diverged from all of the other subskill categories in the subtest, and invalid if it failed to diverge from even one category. If an item fell somewhere in between, it was labelled indeterminate.

In general, subskill categories in both item classification schemes offered little evidence of discriminant validity (see Table 11). In each of the categories, at least 20 percent of the test items were invalid, and in the Main Idea category, all of the items were invalid.

Comparisons between the two item classification schemes is confounded by the fact that they contain different numbers of categories. For the alternative item classification scheme, which contains more categories, it is at once harder for an item to be valid (since it has to diverge from five distinct categories) and to be invalid (since there is a greater chance that it will diverge from at least one). Consequently, for the total subtest, one would expect to find more indeterminate items in the alternative item classification scheme than in the original scheme.

As shown in Table 11, this was not the case. For both schemes, 51 percent of the items fell into the

TABLE 11

NUMBER AND PERCENTAGE OF VALID, INDETERMINATE, AND INVALID ITEMS FOR EACH SUBSKILL CATEGORY AND FOR THE TOTAL SUBTEST, ACCORDING TO TWO ITEM CLASSIFICATION SCHEMES

Measure	Valid		Indeterminate		Invalid	
	n	(%)	n	(%)	n	(%)
Original Scheme						
Reference	1	(17)	3	(50)	2	(33)
Recall	0	(0)	3	(27)	8	(73)
Inference	0	(0)	17	(77)	5	(23)
Main Idea	0	(0)	0	(0)	6	(100)
Total Subtest	1	(2)	23	(51)	21	(47)
Alternative Scheme						
Reference*	2	(29)	3	(43)	2	(29)
Literal	0	(0)	1	(17)	5	(83)
Paraphrase	0	(0)	7	(78)	2	(22)
Vocabulary in Context	2	(0)	3	(78)	2	(22)
Reasoning	3	(23)	6	(46)	4	(31)
Synthesis	0	(0)	2	(40)	3	(60)
Total Subtest	5	(11)	23	(51)	17	(38)

*Rounding error caused percentages in this category to total to 101.

indeterminate category. The alternative scheme contained 9 percent fewer invalid terms, but it cannot be determined from the data whether this was due to an improvement in discriminant validity of the subskill categories or simply to an increase in the number of categories. Significant, however, is the fact that the alternative scheme contains 9 percent more valid items, since this can only be attributed to an improvement in discriminant validity of the scheme, or, more accurately, of two of the subskill categories of the scheme.

Despite the improvement in the alternative scheme, however, neither scheme demonstrated a sufficient number of valid items to establish the discriminant validity of its subskill clusters.

Section Seven: Comparison of the Two Item Classification Schemes

This last section reports no new analyses. Instead it represents an attempt to synthesize the findings of Sections Three through Six, above, in order to answer the following research question:

> Do the item classification schemes differ in terms of internal consistency and discriminant validity?

It should be noted that this section is restricted to a comparison of the two schemes. The more important issue, namely, whether either one of them has any practical value,

will be discussed in Chapter V.

First, although the factor structure seemed to suggest multidimensionality, neither of the schemes was supported by the results. Items from the subskill categories in both schemes loaded on several factors.

Second, the KR20 coefficients for subskill categories in both schemes were low. Generally speaking, neither scheme was markedly better than the other. The mean KR20 coefficient for the original scheme was .53, and for the alternative scheme it was .52.

Third, the results of using the Spearman-Brown Prophecy formula to raise the number of items in each category to an arbitrary constant suggested that, in terms of internal consistency and hypothetical reliability coefficients, the alternative item classification scheme was somewhat superior. Four out of six categories (67%) in the alternative scheme obtained hypothetical reliability coefficients at or above the level of the total subtest, while only two of the four categories (50%) in the original scheme obtained such coefficients. These results were confirmed, in fact determined, by the inter-item correlations within the subskill categories.

Fourth, because the intercorrelations of the subskill categories in each scheme were severely attenuated by low reliabilities, no conclusions were drawn regarding the possibility of hierarchical structure. Both schemes

demonstrated subskill categories of ascending mean item difficulty. The range of mean p values for the alternative scheme (from .51 to .90) was greater than that for the original scheme (from .66 to .89).

Fifth, the number of items demonstrating discriminant validity was 9% greater for the alternative scheme than for the original scheme. The number of invalid items was 9% less for the alternative scheme than for the original scheme, but this finding could have resulted from the increase in the number of subskill categories in the alternative scheme.

In summation, the alternative item classification scheme is slightly, but probably not substantially, better than the original item classification scheme in terms of internal consistency (as evidenced by the hypothetical reliability coefficients produced by application of the Spearman-Brown Prophecy Formula, and by the inter-item correlations) and in terms of discriminant validity (as evidenced by the number of valid items).

CHAPTER V

CONCLUSIONS AND IMPLICATIONS FOR PRACTICE

This chapter will present the conclusions drawn from findings of the study and the implications of the findings for testing practices in adult education.

Conclusions

By means of a series of interrelated analyses, the study attempted to accomplish two major goals. The first goal was to define operationally reading comprehension as it is measured by the Reading Comprehension subtest of the TABE, Level D, Form 4. The second goal was to determine the practical utility of two separate item classification schemes for that subtest in the diagnosis of the individual learning needs of adult examinees.

The study has the potential to define operationally reading comprehension, as measured by the subtest in one of three ways: through exploratory factor analysis, through the validation of the original item classification scheme, or through the validation of the alternative item classification scheme. Had the factor analysis resulted in a strong, interpretable factor structure, the study would have produced an empirically based definition of reading

comprehension, consisting of labelled factors, the items loading on each factor, and the relative importance of each factor, as indicated by the percent of variance explained. Had either of the item classification schemes been validated, reading comprehension would have been defined in terms of the distinct subskill categories according to which test items were classified. The results of the analyses, however, do not allow for any of these approaches to operational definition.

The factor analysis resulted in 16 factors which together accounted for 61.4 percent of the variance. Interpretation of these factors was confounded by weak loadings, by the fact that several items loaded on more than one factor, and by the fact that 14 of the 45 test items failed to load substantially on any of the factors. When the goal is to define and explain a 45-item test, a 16-factor solution which lacks simple structure and which accounts for only 61.4 percent of the variance is the antithesis of parsimony. More and better information about the subtest could be obtained by simply and directly examining the test items themselves. The results of the factor analysis did suggest that reading comprehension, as measured by the subtest, is multidimensional. That multidimensionality, however, is of a highly fractionated, as opposed to a neatly compartmentalized, natur . The

dimensions were not reflected in either of the item classification schemes examined in this study, and it is unlikely that the dimensions could be reflected by any reliable item classification scheme.

Neither of the item classification schemes examined in this study operationally define the test. The results of the analyses indicate that the subskill categories are seriously lacking in both internal consistency, in terms of inter-item correlations, and in discriminant validity, in terms of the number of items which correlate more highly with the category of which they are a part than with other categories in the schemes. The findings suggest that the subskill categories are little more than labels, in that there is no empirical evidence that they represent viable, separable constructs.

These same findings argue against the practical utility of the two item classification schemes for the differential diagnosis of learning needs. The fact that the alternative scheme performed slightly better than the original scheme is hardly encouraging, since this is simply a case of comparing the bad with the worse. Even without examining the inter-item correlations and the number of valid versus invalid items resulting from the discriminant validity analyses, one would have to conclude that the subskill categories have little practical value based only on the low KR20 coefficients obtained, and the resultant high

standard errors of measurement. There are simply too few items in each of the categories to obtain reliable measurement, and without reliability, there can be no validity.

Implications for Practice

Teachers in adult education use the TABE Analysis of Learning Difficulties to plan instruction out of a desire to meet the immediate learning needs of individual adult learners in the most efficient way possible. The findings of this study suggest, however, that the Analysis of Learning Difficulties for the Reading Comprehension Subtest is based on an item classification scheme comprised of subskill categories which are characterized by low reliability and questionable validity. To plan instruction on the basis of such categories could easily result in the type of misdirected and inefficient instruction that the teachers are trying to avoid.

Succinctly stated, the Analysis of Learning Difficulties for this subtest has little practical value for diagnosis. Whenever possible, programs should abandon the use of the Analysis of Learning Difficulties and should substitute diagnostic instruments of demonstrated reliability and validity. Those programs that continue to use it should be extremely wary of placing too much confidence in the results. In all cases, individual instruction plans based on the Analysis of Learning Difficulties should be regarded

as tenuous, and should be revised or refined based on teacher observations and informal performance measures.

BIBLIOGRAPHY

Artley, A. S. A study of certain relationships existing between general comprehension and reading comprehension in a specific subject matter area. Journal of Educational Research, 1944, 37, 464-73.

Auerbach, I. An analysis of reading comprehension tests: Final Report. Project No. 0-A-074. Cambridge, Mass.: Harvard University Graduate School of Education, 1971.

Buros, O. K. The eighth mental measurements yearbook. Highland Park, N.J.: Gryphon Press, 1978.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait - multimethod matrix. Psychological Bulletin, 1959, 6., 81-105.

Carmines, E. G., & Zeller, R. A. Reliability and validity assessment. Sage University Paper Series on Quantitative Applications in the Social Sciences, 1979, 17 (series no. 07-017.) Beverly Hills: Sage Publications.

Chapman, C. A. An analysis of three theories of the relationships among reading-comprehension skills. Symposium on psycholinguistics and reading presented at the meeting of the International Reading Association, Kansas City, May, 1969 (Cited in Davis, 1972.)

Conant, M. M. The construction of a diagnostic reading test. N.Y.: Teachers College Press, Columbia University, 1942.

Davis, F. B. Fundamental factors of comprehension in reading. Unpublished doctoral dissertation, Harvard University, 1941.

Davis, F. B. Psychometric research on comprehension in reading. Reading Research Quarterly, 1972, 7, 628-678.

Examiner's Manual, Level-D. Tests of Adult Basic Education. Monterey, Ca.: CTB/McGraw-Hill, 1976.

Farr, R. The convergent and discriminant validity of several upper level reading tests. In G. B. Schick & M. M. May (Eds.), Multidisciplinary aspects of college-adult reading. The seventeenth yearbook of the National Reading Conference. Milwaukee: The National Reading Conference. Milwaukee: The National Reading Conference, Inc., 1968, 181-191.

Farr, R. Reading: What can be measured? Newark, Del.: International Reading Association, 1969.

Fisher, D. Literacy: Meeting the challenge: Assessment of reading competencies. Paper presented at the National Right to Read Conference, Washington, D.C., May, 1978.

Harris, C. W. Measurement of comprehension of literature: II. studies of measures of comprehension. School Review, 1948, 56, 332-42.

Henrysson, S. Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Holmes, J. A., & Sioger, H. Theoretical models and trends toward more basic research in reading. Review of Educational Research, 1964, 34, 127-55.

Hunt, L. C. Can we measure specific factors associated with reading comprehension? Journal of Educational Research, 1957, 51, 161-171.

Kerfoot, J. F. Problems and research considerations in reading comprehension. In Mildred A. Dawson (Ed.), Developing comprehension including critical reading. Newark, Delaware: International Reading Association, 1968, 38-44.

Kim, J. Factor analysis. In Norman H. Nie (Ed.), SPSS: Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill, 1975, 468-514.

Kim, J., & Mueller, C. W. Factor analysis: Statistical methods and practical issues. Sage University Paper Series on Quantitative Applications in the Social Sciences, 1978, 14 (series no. 07-014.) Beverly Hills: Sage Publications.

Kingston, A. J. The measurement of reading comprehension. In O. Causey & E. Bliesmer (Eds.), Research and evaluation in college reading. The ninth yearbook of the National Reading Conference for College and Adults. Fort Worth: Texas Christian University Press, 1960.

Korth, B. Exploratory factor analysis. In D. J. Amick & H. J. Walberg (Eds.), Introductory multivariate analysis for educational, psychological, and social research. Berkeley: McCutchan, 1975, 113-169.

Lennon, R. T. What can be measured? The Reading Teacher, 1962, 15, 326-37.

Messick, S. Test validity and the ethics of assessment. American Psychologist, 1980, 35, 1012-1027.

Powers, S., & Gallas, E. J. The factorial structure and item composition of Comprehensive Tests of Basic Skills, reading comprehension test, when administered to ESEA Title I sixth and seventh grade students. Paper presented at the sixty-fourth annual meeting of the American Educational Research Association, Boston, April, 1980. (ERIC Document Reproduction Service No. ED 194 527)

Raygor, A. L. Problems in the susstrata-factor theory. Reading Research Quarterly, 1966, I (3), 147-50.

Stoker, H. W., & Kropp, R. P. The predictive validities and factorial context of the Florida state wide ninth-grade testing program battery. Florida Journal of Educational Research, 1960, I, 105-14.

Technical Report. Test of Adult Basic Education. Monterey, Ca.: CTB/McGraw-Hill, 1978.

Thorndike, R. L. Reading as reasoning. Reading Research Quarterly, 1973-1974, 9 (2), 135-147.

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

Thurstone, L. L. Note on a reanalysis of Davis' reading tests. Psychometrika, 1946, 11, 185-188.

Traxler, A. E. A study of the Van Wagenen-Dvorak Diagnostic Examination of Silent Reading Abilities. Educational Records Bulletin, 1941, 31, 33-41.

Traxler, A. E. Critical survey of tests for identifying difficulties in interpreting what is read. In William S. Gray (Ed.), Promoting growth toward maturity in interpreting what is read (supplementary educational monographs no. 74). Chicago: University of Chicago Press, 1951.

Traxler, A. E. Values and limitations of standardized reading tests. In Arthur E. Traxler (Ed.), Evaluation of reading (supplementary educational monographs no. 88). Chicago: University of Chicago Press, 1958.

Walker, H. M., & Lev, J. Statistical inference. New York: Holt, Rinehart and Winston, 1953.

APPENDIX A

TWO ITEM CLASSIFICATION SCHEMES FOR THE TABE READING COMPREHENSION SUBTEST, LEVEL D, FORM 4

TABLE 12

TWO ITEM CLASSIFICATION SCHEMES FOR THE TABE READING COMPREHENSION SUBTEST, LEVEL D, FORM 4

Item	Original	Alternative
1	Reference	Reference
2	Reference	Reference
3	Reference	Reference
4	Reference	Reference
5	Reference	Reference
6	Reference	Reference
7	Main Idea	Synthesis
8	Inference	Vocabulary in Context
9	Recall	Paraphrase
10	Recall	Paraphrase
11	Inference	Reasoning
12	Recall	Paraphrase
13	Inference	Reasoning
14	Main Idea	Synthesis
15	Inference	Vocabulary in Context
16	Retall	Literal
17	Inference	Vocabulary in Context
18	Recall	Literal
19	Recall	Literal
20	Inference	Paraphrase

TABLE 12 -- CONTINUED

Item	Original	Alternative
21	Main Idea	Synthesis
22	Inference	Paraphrase
23	Inference	Paraphrase
24	Inference	Paraphrase
25	Inference	Reasoning
26	Main Idea	Synthesis
27	Inference	Vocabulary in Context
28	Recall	Paraphrase
29	Inference	Vocabulary in Context
30	Recall	Literal
31	Recall	Literal
32	Recall	Literal
33	Recall	Reasoning
34	Main Idea	Paraphrase
35	Inference	Snythesis
36	Main Idea	Reasoning
37	Inference	Reference
38	Inference	Reasoning
39	Inference	Reasoning
40	Inference	Reasoning
41	Inference	Reasoning
42	Inference	Reasoning

TABLE 12 -- CONTINUED

Item	Original	Alternative
43	Inference	Reasoning
44	Inference	Reasoning
45	Inference	Reasoning

APPENDIX B

DISTRIBUTION OF SCORES FOR SAMPLE POPULATION

TABLE 13

DESCRIPTIVE STATISTICS FOR DISTRIBUTION OF TOTAL RAW SCORES FOR SAMPLE POPULATION (n=242)

Mean	32.4
Standard Deviation	6.8
Median	32.7
Mode	32.0
Skewness	-0.3
Range	31.0
Maximum	45.0
Minimum	14.0

APPENDIX C

RESULTS OF EXPLORATORY FACTOR ANALYSIS

TABLE 14

RESULTS* OF PRINCIPAL AXIS FACTOR ANALYSIS OF THE TABE READING COMPREHENSION SUBTEST WHEN ADMINISTERED TO 242 ADULT EXAMINEES

Factor	Eigenvalues	Percent of Variance	Cumulative Percent of Variance
1	6.80	15.1	15.1
2	2.21	4.9	20.0
3	1.95	4.3	24.3
4	1.64	3.6	28.0
5	1.56	3.5	31.4
6	1.49	3.3	34.8
7	1.45	3.2	38.0
8	1.36	3.0	41.0
9	1.30	2.9	43.9
10	1.26	2.8	46.7
11	1.21	2.7	49.4
12	1.18	2.6	52.0
13	1.12	2.5	54.5
14	1.07	2.4	56.9
15	1.03	2.3	59.2
16	1.01	2.3	61.4

*Only factors with eigenvalues $\geq$ 1.00 are presented.

TABLE 15

FACTOR LOADINGS* FOR OBLIQUELY ROTATED SOLUTION OF A PRINCIPAL AXIS FACTOR ANALYSIS OF THE TABE READING COMPREHENSION SUBTEST WHEN ADMINISTERED TO 242 ADULT EXAMINEES

Factor	Item	Loading
Factor 1	9	.33
	11	.33
	15	.41
	16	.33
	27	.40
Factor 2	1	.71
	2	.33
	5	.68
Factor 3	8	.72
Factor 4	43	.56
	44	.35
Factor 5	6	.42
	14	.51
	22	.38
Factor 6	3	.79
	10	.35
Factor 7	29	.32
	30	.61
Factor 8	31	.38
	40	.65
	41	.54
Factor 9	6	.44
	7	.45
	19	.59
	20	.35
Factor 10	42	.55
	44	.40

TABLE 15 -- CONTINUED

Factor	Item	Loading
Factor 11	25	.69
	34	.37
Factor 12	24	.30
	26	.40
Factor 13	33	.51
	36	.46
Factor 14	9	.32
	13	.61
Factor 15	12	.53
Factor 16	28	.54

*Only factor loadings $\geq$ .30 are presented.