A CONSTRUCT VALIDATION OF ADAPTIVE ACHIEVEMENT TESTING

Isaac I. Bejar
and
David J. Weiss

RESEARCH REPORT 78-4
NOVEMBER 1978

PSYCHOMETRIC METHODS PROGRAM
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MN 55455

This research was supported by funds from the Army Research Institute, Air Force Human Resources Laboratory, Defense Advanced Research Projects Agency, Navy Personnel Research and Development Center, and Office of Naval Research, and monitored by the Office of Naval Research.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
Research Report 78-4

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
A Construct Validation of Adaptive Achievement Testing

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Isaac I. Bejar and David J. Weiss

8. CONTRACT OR GRANT NUMBER(s)
N00014-76-C-0627

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Psychology
University of Minnesota
Minneapolis, MN 55455

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
P.E.: 61153N PROJ.:RR042-04
T.A.: RR042-04-01
W.U.: NR150-389

11. CONTROLLING OFFICE NAME AND ADDRESS
Personnel and Training Research Programs
Office of Naval Research
Arlington, VA 22217

12. REPORT DATE
November 1978

13. NUMBER OF PAGES
28

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES
This research was supported by funds from the Army Research Institute, Air Force Human Resources Laboratory, Defense Advanced Research Projects Agency, Navy Personnel Research and Development Center, and Office of Naval Research, and monitored by the Office of Naval Research.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
testing
achievement testing
computerized testing
adaptive testing
sequential testing
branched testing
individualized testing
tailored testing
programmed testing
response-contingent testing
automated testing
item characteristic curve theory

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
The construct validities of conventional paper-and-pencil and adaptive achievement tests were compared using data from two independent groups of 269 and 230 college students. Two adaptive achievement tests were computer administered to each group using the stradaptive testing strategy; each group also completed two conventional classroom paper-and-pencil achievement tests. All achievement tests were drawn from the same pool of achievement test items on which item characteristic curve (ICC) parameters had been determined. Students were also administered two stradaptive vocabulary tests. All tests

were scored by maximum likelihood estimation using the three-parameter logistic model. A nomological net was specified describing the relationships of the achievement tests to the achievement constructs and their relationships with the vocabulary construct and the vocabulary tests. The parameters of the net were estimated by fitting the observed intercorrelations among the test scores to the nomological net, using the methodology of linear structural equations. Maximum likelihood estimates of the parameters of the nomological net indicated essentially equal validities for the classroom and adaptive tests in four comparisons. However, the validity of the adaptive tests was effectively higher than that of the classroom tests, since equal validities were achieved with from 25% to 31% fewer items. The data also permitted an analysis of the effects of verbal ability on achievement test performance, separately for the conventional and adaptive tests. The results from a confirmatory maximum likelihood factor analysis showed a larger influence of verbal ability on achievement test performance at the first administration of the adaptive test. This result was attributed to a necessity to learn how to use the computer equipment with verbal instructions, which may have further reduced the validity of the adaptive tests. Combined with the facts that the adaptive tests were obtained under volunteer conditions while the classroom tests were obtained under "motivated" grading conditions, the results of this study indicate that computer-administered adaptive tests can provide more valid measurement of achievement than conventional paper-and-pencil tests.

CONTENTS

Acknowledgements

Data utilized in this report were obtained from volunteer students in General Biology, Biology 1-011, at the University of Minnesota, during fall quarter 1976 and winter quarter 1977; appreciation is extended to these students for their participation in this research. The cooperation of Kathy Swart and Norman Kerr of the General Biology staff in providing access to the students, as well as their encouragement and important contributions of time and ideas to this research, are also deeply appreciated.

Technical Editor: Barbara Leslie Camm

A Construct Validation of Adaptive Achievement Testing

In the last decade there has been an increasing amount of research on adaptive or tailored ability testing (Weiss, 1976). In general, this research has shown that adapting ability tests to the individual is beneficial in terms of (1) reducing test anxiety and increasing test-taking motivation (Betz & Weiss, 1976) and (2) providing measurement of higher precision (McBride & Weiss, 1976; Vale, 1975). More recently, interest has extended to achievement testing as well (Bejar, Weiss, & Gialluca, 1977; Bejar, Weiss, & Kingsbury, 1977; Brown & Weiss, 1977; Reckase, 1977). The question of the validity of adaptive testing has not yet been investigated, however, either in the ability testing domain or in the adaptive measurement of achievement.

A few studies have examined the "fidelity" of adaptive testing strategies, where fidelity is defined as the correlation between true ability level and ability level estimated by an adaptive testing procedure (e.g., McBride & Weiss, 1976; Urry, 1976; Vale & Weiss, 1975). Of necessity, however, these studies were computer simulation studies in which "true" ability was known and testees were simulated by a mathematical model. Other studies which have examined the validity of adaptive tests (e.g., Linn, Rock, & Cleary, 1969) were real-data simulation studies in which responses to adaptive tests were simulated from the responses of students to conventional paper-and-pencil tests. Thus, no live-testing studies have been reported in which tests were administered adaptively and in which the comparative validity of conventional paper-and-pencil testing and adaptive strategies was examined.

Narrowly defined, validity consists of ascertaining how well an individual's performance on a criterion of interest can be forecasted from knowledge of his/her test performance on the test being validated (e.g., Cronbach, 1971). The usual procedure for this kind of validation consists of assessing the relationship, or correlation, of the scores on the criterion with the scores on the test being validated.

When the interest is in comparing the validities of two or more testing procedures, this approach to validation could give misleading results, since a test consists of several components, each of which can determine to some extent a testee's performance on the test. As a result, the correlation between scores on tests administered in different ways may be partially determined by the components shared by the testing procedure being validated and the criterion (Bejar, 1977). Thus, if the correlation between scores from Testing Procedure A and Criterion C is higher than the correlation between scores from Testing Procedure B and Criterion C, this may not necessarily be evidence that Testing Procedure A is more valid than Testing Procedure B. The apparent difference in validity could be due simply to the fact that Test A and Criterion C were measured under similar conditions and thus had more method variance in common. For example, both the test and the criterion performance might be measured under conditions which were arbitrarily high speeded, and the resulting correlation would reflect this common speededness.

A broader and more appealing view of the validation process is construct validation (Campbell & Fiske, 1959; Cronbach & Meehl, 1955). In this context the question is not how well some criterion is predicted; rather, the goal is identification of the constructs that account for test performance. This is done by postulating a nomological net--a theory describing the laws and hypotheses that relate observables to observables, observables to constructs, and constructs to constructs. The validation process then consists of ascertaining whether the data support the theoretical hypotheses in the nomological net. If the data are in accord with the hypothesis, the problem becomes one of estimating the strength of the relationship between the different components of the net. The practical problem of assessing the relative validity of two testing procedures becomes one of determining how well each measures the construct it is supposed to measure. This can be approached by assessing the relationship between the observed scores derived from each of the testing procedures and the constructs that the testing procedures are designed to measure.

Purpose

The purpose of this study was to assess the relative construct validities of two testing procedures for measuring achievement--a conventional paper-and-pencil test and a computer-administered adaptive test. A nomological net was specified and fitted to the intercorrelations among four measures of achievement and to measures of verbal ability. A secondary purpose of the study was to estimate the relationships of verbal ability to achievement test performance.

Method

Data for this study were obtained from students enrolled in a large introductory biology course at the University of Minnesota during the fall and winter quarters of the 1976-1977 school year. The analysis was based on volunteers for which the following six scores were available:

1. Classroom biology achievement test, first midquarter (MQ1C)
2. Classroom biology achievement test, second midquarter (MQ2C)
3. Adaptive biology achievement test, first midquarter (MQ1A)
4. Adaptive biology achievement test, second midquarter (MQ2A)
5. Adaptive vocabulary test at first midquarter (VOC1)
6. Adaptive vocabulary test at second midquarter (VOC2)

The classroom midquarter tests, MQ1C and MQ2C, were the tests normally given in the course for grading purposes. Data on both the adaptive achievement and vocabulary tests were collected from students who volunteered to participate in the research in exchange for extra points toward their final course grade.

Subjects

Data were available on students from two academic quarters. During the fall quarter, 394 students had volunteered to take an adaptive midquarter test based on the material from the first classroom biology midquarter test and 386 volunteered for the adaptive midquarter test based on the material from the second classroom biology midquarter test. However, only 269 students participated at both occasions; data analysis for fall quarter data was based on this group. For winter quarter, 317 students volunteered to participate in the first adaptive midquarter test administration and 349 volunteered to participate in the second; data analysis for winter quarter data was based on the 230 students who participated in both adaptive midquarter tests.

Procedure

At both the first and second adaptive test administrations, the volunteer students were first given the adaptive multiple-choice verbal ability test (VOC1, VOC2) followed by the adaptive multiple-choice biology test (MQ1A, MQ2A) based on the content covered in the classroom biology midquarter tests. The adaptive tests were administered by means of cathode ray terminals (CRTs) connected to a Hewlett-Packard real-time computer system. Instructional screens explaining the operation of the equipment were presented prior to testing (DeWitt & Weiss, 1974). A proctor was present in the testing room at all times to assist students with the equipment. Each test item was presented separately at the rate of 960 characters per second on the CRT screen. Students responded by pressing the key corresponding to the chosen alternative. During the fall quarter administration, feedback was provided after each response (i.e., each student was informed whether or not he/she had answered each test item correctly); if an incorrect answer was given, the student was told which answer was correct. During the winter quarter administration, immediate feedback was not provided. There were no time limits imposed on the tests. At the completion of testing, students received a printed report which listed questions answered incorrectly and provided the correct answers.

The classroom biology achievement test data (MQ1C, MQ2C) were obtained from course instructors.

Achievement Tests

Item pool. The development of the item pools used in this study has been described by Bejar, Weiss, and Kingsbury (1977). Briefly, the answer sheets for two classroom biology midquarter tests from two previous academic quarters were used as raw data for obtaining the item parameters--discrimination (a), difficulty (b), and guessing (c)--of the logistic item characteristic curve (Birnbaum, 1968) for each item. For the fall quarter administration, 114 items covering the content of the first midquarter were available; the pool covering the content of the second midquarter contained 112 items. For the winter administration, 44 items were added to the first midquarter pool and 49 were added to the second midquarter pool; thus, there were a total of 158 items in the first midquarter item pool and a total of 161 in the second midquarter pool. Both the adaptive and classroom achievement tests were constructed from the same item pool.

Adaptive achievement tests. The adaptive achievement tests were administered by the stradaptive strategy (Weiss, 1973). The entry point was selected based on student-reported GPA. At the beginning of the adaptive testing session, students were asked to state their grade point average (GPA) by selecting one of nine equally spaced intervals from 2.00 to 4.00 (DeWitt & Weiss, 1974, p. 49). For example, students reporting GPAs in the lowest interval began testing in the least difficult stratum, whereas students choosing the highest GPA interval began in the most difficult stratum.

The branching strategy used in the stradaptive test was the standard "up-one/down-one" procedure. That is, if an item was answered incorrectly or with a "?," the next unadministered item from the next easier stratum was administered; if an item was answered correctly, the next unadministered item from the next more difficult stratum was administered.

A variable criterion was used to terminate testing on the stradaptive test. After a student answered five items in a stratum, if he/she answered 20% or fewer correctly, testing was terminated. If testing was not terminated by this criterion after 50 items had been administered, no further items were administered.

To construct item pools which could be used for administration of stradaptive tests, each of the two pools (Midquarters 1 and 2) was structured by forming nine strata of increasing difficulty. Mean stratum difficulties were chosen so that there would be approximately the same number of items per stratum. Within each stratum the items were ordered in terms of their discriminations unless this resulted in items covering the same content area appearing consecutively. Appendix Tables A and B show the item difficulties and discriminations for items in the nine strata into which the first and second midquarter item pools were structured. Table 1 summarizes that information by showing the mean and standard deviations of the discrimination (*a*), difficulty (*b*), and guessing (*c*) parameter estimates for the fall and winter item pools. For both the first and second midquarter tests, the mean discriminations, difficulties, and "guessing" parameters were essentially identical for the two quarters.

Table 1
Mean and Standard Deviation of Item Parameter Estimates of the Fall and Winter Item Pools for the First and Second Adaptive Achievement Midquarter Tests (MQ1A and MQ2A)

Test	Number of Items	*a* Discrimination Mean	*a* Discrimination *S.D.*	*b* Difficulty Mean	*b* Difficulty *S.D.*	*c* "Guessing" Mean	*c* "Guessing" *S.D.*
MQ1A							
Fall	114	1.21	.46	.19	1.21	.27	.08
Winter	158	1.20	.44	.16	1.19	.27	.08
MQ2A							
Fall	112	1.20	.41	.16	1.16	.28	.09
Winter	161	1.20	.39	.11	1.16	.28	.08

Classroom achievement tests. The classroom biology midquarter test each quarter included 55 items which the course staff selected by a combination of pedagogical criteria and procedures from classical test theory. Their aim in constructing these tests was to produce a "good" test for purposes of course grading. Students were instructed to answer 50 items of their choice. For purposes of this research, however, the classroom achievement tests were shorter than 50 items, since item parameter estimates were not available for some of the items. The item parameter estimates for the items in MQ1C and MQ2C for the fall administration are in Appendix Table C; those for the winter administration are in Appendix Table D.

Table 2 shows the means and standard deviations of estimates of the three item parameters for MQ1C and MQ2C for the fall and winter administrations. Constrasting these figures to those in Table 1, it is evident that the items for MQ1C were, on the average, less discriminating than those in the adaptive test pool; the items in MQ1C were also less discriminating than those in the adaptive test pool, but the differences between the two pools were smaller.

Table 2
Mean and Standard Deviation of Item Parameter Estimates for the First and Second Classroom Achievement Midquarter Tests (MQ1C and MQ2C) in the Fall and Winter Administration

	Number	*a*		*b*		*c*	
	of	Discrimination		Difficulty		"Guessing"	
Test	Items	Mean	*S.D.*	Mean	*S.D.*	Mean	*S.D.*
MQ1C							
Fall	39	1.09	.27	.11	1.14	.29	.06
Winter	45	1.09	.31	.08	1.33	.25	.09
MQ2C							
Fall	41	1.17	.44	.07	1.20	.28	.07
Winter	44	1.14	.40	-.06	1.29	.25	.08

Adaptive Vocabulary Tests

The adaptive vocabulary test was also administered by the stradaptive strategy. The same entry point and termination rule used in the biology achievement test were used for the vocabulary test, except that the maximum number of items in the vocabulary test was set at 40.

The development of the vocabulary item pool has been described by McBride and Weiss (1974); the procedures for estimating the item parameters used for the vocabulary tests are described in Prestwood and Weiss (1977). For the fall administration, the same pool consisting of 321 items was used for the first and second midquarters. During winter quarter, however, the pool was split into two comparable halves consisting of 160 and 161 items each, used for the first and second midquarter administrations, respectively. Appendix Table E provides the item parameters for the stradaptive vocabulary tests.

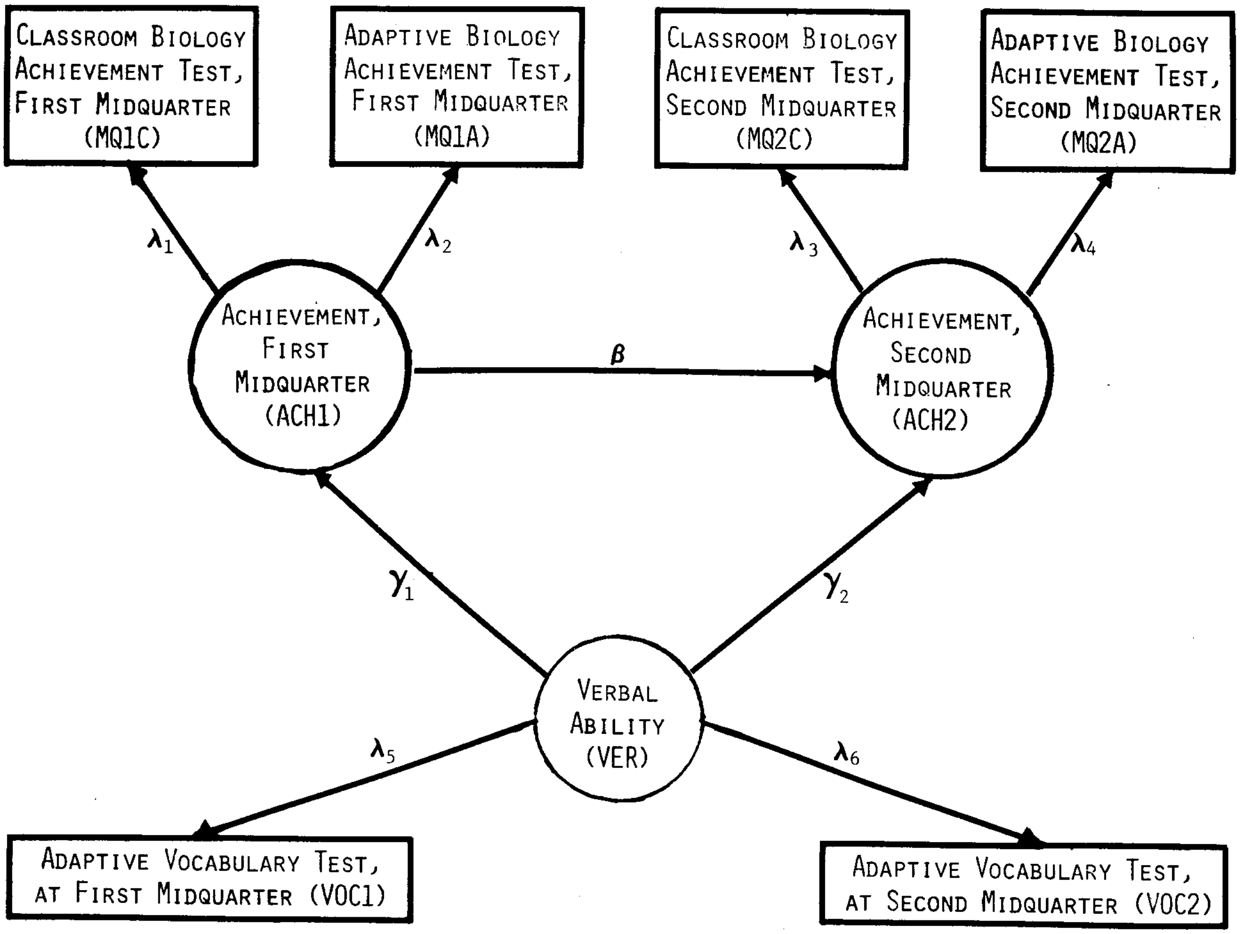
Scoring

All tests were scored by maximum likelihood estimation, specifying Birnbaum's (1968) three-parameter logistic model as the response model. The item parameter estimates were edited by the scoring program so that the maximum value of the discrimination parameter (*a*) was set to 2.5, the maximum absolute value of the difficulty parameter (*b*) was set to 3.00, and the maximum value of the guessing parameter (*c*) was set to .35. In estimating achievement scores, omitted items were ignored in the computations. The convergence criterion was set to .0001, and a maximum of 50 iterations was allowed in the maximum likelihood scoring.

Nomological Net

The nomological net investigated consisted of three constructs, each measured twice (see Figure 1)--achievement at the first midquarter (ACH1), achievement at the second midquarter (ACH2), and verbal ability (VER). ACH1 and ACH2 were each measured once by the classroom biology achievement midquarter tests (MQ1C, MQ2C) and once by the adaptive biology achievement midquarter tests (MQ1A, MQ2A). VER was also measured twice--once during the administration of MQ1A and once during the administration of MQ2A. The arrows connecting the constructs and the constructs with their observable measures symbolize the parameters of the nomological net to be estimated. Thus, Figure 1 postulates that verbal ability (VER) influenced achievement at the first midquarter (ACH1) and achievement at the second midquarter (ACH2). Achievement at the second midquarter (ACH2) in turn was hypothesized to be influenced both by achievement at the first midquarter (ACH1) and by verbal ability (VER).

Figure 1
Nomological Net for Construct Validation
of Classroom and Adaptive Achievement Tests


Classroom Biology Achievement Test, First Midquarter (MQ1C)
Adaptive Biology Achievement Test, First Midquarter (MQ1A)
Classroom Biology Achievement Test, Second Midquarter (MQ2C)
Adaptive Biology Achievement Test, Second Midquarter (MQ2A)
λ_1
λ_2
λ_3
λ_4
Achievement, First Midquarter (ACH1)
β
Achievement, Second Midquarter (ACH2)
γ_1
γ_2
Verbal Ability (VER)
λ_5
λ_6
Adaptive Vocabulary Test, at First Midquarter (VOC1)
Adaptive Vocabulary Test, at Second Midquarter (VOC2)

For construct validation comparisons of the adaptive and conventional paper-and-pencil achievement tests, the parameters of interest were those that estimated the relationships between the observables and their corresponding constructs

(λ_1 through λ_4). These parameters may be referred to as the validities of the observable achievement scores. Thus, in the context of Figure 1 the major purpose of this study was to compare the validities for the adaptive achievement tests (λ_3 and λ_4) with the validities for the conventional classroom paper-and-pencil achievement tests (λ_1 and λ_2) in two independent sets of data.

The nomological net in Figure 1 also focuses on the effects of verbal ability on biology achievement at both midquarters (γ_1 and γ_2) and on the dependence of achievement at the second midquarter on achievement at the first midquarter (β). This part of the model is relevant from a substantive point of view because it indicates the degree to which assimilation of instruction is dependent on verbal ability. From a psychometric point of view, however, the effects of verbal ability on achievement test performance are equally important, since individual differences in verbal ability could possibly affect the validity of the achievement scores, particularly when the method of administration was different in the two testing procedures (i.e., the adaptive test was computer administered and the classroom test was paper-and-pencil). Thus, a second objective of this investigation was to assess the influence of verbal ability on test performance under the two modes of administration.

Data Analysis Methodology

Estimating the parameters of the nomological net. Traditionally, construct validation hypotheses have been partially investigated by factor analytic techniques. However, in recent years the methodology of linear structural equations (Goldberger & Duncan, 1973) has been applied to these kinds of questions (e.g., Schmitt, 1978) as a result of computational developments due primarily to Jöreskog (e.g., Jöreskog & van Thillo, 1972). Structural equations methodology is a more general analytic technique than factor analysis, but it is very much related to it. In general, a structural equations model consists of three parts. One of these parts models the interrelationships among the endogenous or dependent variables. The second part models the interrelationships among the exogenous or independent variables. The modeling of both sets of variables is by means of factor analytic models; that is, it is assumed that the interrelationships within the dependent and independent variable sets can be accounted for by a factor analytic model. Finally, the third part of the structural equations model connects the constructs or factors derived separately from the dependent and independent variables.

The application of this methodology to a nomological net such as that shown in Figure 1 has been discussed by Jöreskog and Sörbom (1976); the following discussion utilizes their notation. To construct the model, the nomological net can be separated into the three parts indicated above.
The first part, the factor model for the dependent variables in the nomological net of Figure 1, is seen in Equation 1:

$$\begin{pmatrix} \text{MQ1C} \\ \text{MQ2C} \\ \text{MQ1A} \\ \text{MQ2A} \end{pmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ \lambda_2 & 0 \\ 0 & \lambda_3 \\ 0 & \lambda_4 \end{bmatrix} \begin{pmatrix} \text{ACH1} \\ \text{ACH2} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix} \qquad [1]$$

This is simply an orthogonal two-factor model for the four biology achievement scores (MQ1C, MQ2C, MQ1A, MQ2A). The two factors postulated were achievement in biology at the first midquarter (ACH1) and at the second midquarter (ACH2). The ε_i's are the unique components associated with each observable measure.

For estimation purposes λ_1 and λ_3 were set in the estimation program to 1.0, while λ_2 and λ_4 were free to take on any values. The loadings of MQ1C and MQ1A (λ_1 and λ_3) were fixed at 1.0 in order to make the model identified, that is, to insure the uniqueness of each parameter estimate. The uniqueness variances, $\sigma^2_{\varepsilon_i}$, were also estimated by the program.

The second part of the model describing the structures of the independent variables is given by Equation 2:

$$\begin{pmatrix} \text{VOC1} \\ \text{VOC2} \end{pmatrix} = \begin{pmatrix} \lambda_5 \\ \lambda_6 \end{pmatrix} (\text{VER}) + \begin{pmatrix} \varepsilon_5 \\ \varepsilon_6 \end{pmatrix} . \qquad [2]$$

Equation 2 indicates that performance on the vocabulary tests is accounted for by the single construct, verbal ability (VER). For purposes of estimation, λ_5 was set to 1.0 in order to make the model identified. Thus, the parameters to be estimated were λ_6 and $\sigma^2_{\varepsilon_5}$ and $\sigma^2_{\varepsilon_6}$.

Finally, the third part of the model relates the two achievement constructs of biology (ACH1, ACH2) and verbal ability (VER). This relationship was postulated to be

$$\begin{bmatrix} 1 & 0 \\ \beta & 1 \end{bmatrix} \begin{matrix} \text{ACH1} \\ \text{ACH2} \end{matrix} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} (\text{VER}) + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} . \qquad [3]$$

The parameters to be estimated in this part of the model were β, which indicates the strength of the relationship between achievement at two points in time; γ_1 and γ_2, which indicate the strength of the relationship of verbal ability with achievement at the first midquarter and at the second midquarter; and finally, the variances of the residuals, ζ_1 and ζ_2.

Expanding on Equation 3, ACH1 and ACH2 can be expressed as

$$\text{ACH1} = \gamma_1 \text{ VER} + \zeta_1 \qquad [4]$$

$$\text{ACH2} = (\gamma_2 - \beta\gamma_1) \text{ VER} + (\zeta_2 - \beta\zeta_1) . \qquad [5]$$

ACH1 is the sum of two effects, verbal ability (VER) and ζ_1, a residual component. ACH2, on the other hand, is a function of verbal ability, achievement at the first midquarter, and a residual $(\zeta_2 - \beta\zeta_1)$. Note that if β=0--that is, ACH1 has no effect on ACH2--Equation 5 reduces to

$$ACH2 = \gamma_2 VER + \zeta_2 \quad . \qquad [6]$$

It was assumed that the expected value of ACH1, ACH2, and VER was zero. The expected value of the ε_i's was also zero. The ε_i's were assumed to be uncorrelated and independent among and between themselves and uncorrelated with ACH1, ACH2, and VER. The residuals (i.e., ζ_1, ζ_2) were also assumed to be uncorrelated and to have a mean of zero. In addition to these assumptions, it was assumed that the joint distribution of the observed variables was multivariate normal and that the sample size was large; therefore, maximum likelihood estimates of the parameters in Equations 1 through 3 could be obtained by using the program LISREL (Jöreskog & van Thillo, 1972).

Estimating the influence of verbal ability on test performance. The nomological net described in Figure 1, which was postulated to account for achievement in biology, did not allow the estimation of the effect of verbal ability on achievement test performance. The role of this type of method variance analysis in the validation process has been recognized since Campbell and Fiske's formalization of the multitrait-multimethod matrix (1959). However, precise methods for estimating the proportion of method variance did not become available until the development of maximum likelihood factor analysis (Boruch & Wolins, 1970; Jöreskog, 1974). This methodology was used to estimate the effects of verbal ability on achievement test performance.

An orthogonal factor model was postulated to account for the interrelationships among the six observed scores. The pattern matrix associated with the proposed model is shown in Table 3. An "X" indicates that the variable was permitted to load on the factor. A "0" indicates that a variable was not permitted to load on the factor. Setting certain loadings to zero permits the definition of "clean" factors, while at the same time it introduces restrictions in the estimation procedure which are necessary to insure that the model as a whole is identified.

Table 3
Factor Model Postulated to Account for Variation Among Six Observed Scores

	Factor			
Variable	I	II	III	IV
MQ1A	X	X	X	0
MQ2A	X	X	0	X
MQ1C	X	X	X	0
MQ2C	X	X	0	X
VOC1	X	0	0	0
VOC2	X	0	0	0

Note. An "X" means the corresponding parameter is "free" to take any value. A "0" indicates the parameter is "fixed" to take the value 0.

The model in Table 3 allows the identification of four influences on the observed scores or sources of variance. The first source of variance may be called a verbal ability factor, since it was the only factor on which the verbal scores (VOC1 and VOC2) were allowed to load. The loadings of the four achievement scores on this factor indicate the effect of verbal ability on achievement test performance. The second factor may be called an achievement factor because only the four achievement scores (MQ1A, MQ2A, MQ1C, MQ2C) were allowed to load on it. The third and fourth factors are "occasions" factors because they capture the unique variability associated with the first and second midquarter tests, respectively, MQ1A, MQ1C and MQ2A, MQ2C.

Models such as that shown in Table 3 can only be estimated with factor analysis programs which permit restricted solutions. A number of such programs exist. The program ACOVS (Jöreskog, Gruvaeus, & van Thillo, 1970) was used in these analyses. This program obtains maximum likelihood estimates of each of the loadings under the usual stochastic assumptions of factor analysis. If the sample size is large and the data are multivariately distributed, the measure of fit computed by the programs is distributed as a χ^2 variable with known degrees of freedom.

Data Analysis

Subject pool. During the fall and winter administrations, 269 and 213 students, respectively, had completed all six tests. However, data for some students were eliminated from all analyses for one of two reasons: (1) If the scoring procedure failed to converge on any one of the six scores, that student was eliminated from the analyses; (2) If a student's maximum likelihood score on the adaptive test was too "discrepant" from the classroom test maximum likelihood score, the student was eliminated from the analyses. Specifically, the difference in each student's maximum likelihood scores--MQ1C-MQ1A and MQ2C-MQ2A--was computed. If the absolute value of either score difference was 2.00 or larger, the student was excluded. Invariably, the difference was positive for the students eliminated, which indicated that the student performed on the adaptive achievement test two units below his/her classroom achievement test performance. The rationale for excluding such students was that they probably were not "doing their best" taking the adaptive achievement test, since it was a volunteer situation. After excluding students for either of these two reasons, there were 213 and 187 students, respectively, who had taken all six tests during fall and winter administrations. The analyses and results that follow are based on these students only.

Distributional analysis. An assumption needed to obtain maximum likelihood estimates of the parameters by fitting the structural model to a correlation matrix is that the distribution of the scores be multivariate normal. Although some procedures for testing multivariate normality exist (e.g., Andrews, Gnanadesikian, & Warner, 1973), they are not easily implemented. For that reason, the univariate normality of each score was investigated instead. If the multivariate distribution of a set of scores is normal, it would follow that the component scores are each also normally distributed. However, demonstrating that each score is normally distributed does not guarantee that the joint distribution of all scores will be multivariate normal.

The univariate normality of each of the scores was tested by means of the Kolmogorov-Smirnov statistic (see, e.g., Lindgren, 1968). According to this test, if the observed cumulative frequency exceeds the theoretically expected frequency by a certain amount, the hypothesis that the distribution is normally distributed is rejected. The statistic is

$$D = \text{MAX} \mid F_0(x) - F_E(x) \mid \qquad [7]$$

where

D is the absolute value of the maximum discrepancy,
$F_0(x)$ is the observed cumulative frequency of x, and
$F_E(x)$ is the expected cumulative frequency of x.

The null hypothesis was tested at the .05 level for each variable in each quarter.

Results

Distributional Analysis

The results of application of the Kolgomorov-Smirnov test to the maximum likelihood scores on each of the six tests is shown for the fall, winter, and combined data in Table 4. All six scores were judged normally distributed in each quarter and in the combined data. As can be seen, the probability of the null hypothesis was high in every instance, with a minimum value of p=.17 for the VOC2 data in the fall and combined groups. Thus, the results lend support to the assumption that the joint distribution of observable scores may be multivariately normally distributed.

Table 4
Results of the Kolgomorov-Smirnov Test of Normality for Fall, Winter, and Combined Groups

	Test					
Group and Statistic	MQ1C	MQ2C	MQ1A	MQ2A	VOC1	VOC2
Fall (N=213)						
Maximum Discrepancy	-.04	-.03	.05	-.04	-.05	-.07
Probability	.94	.99	.74	.81	.62	.17
Winter (N=187)						
Maximum Discrepancy	-.06	.04	.05	-.04	-.05	-.07
Probability	.56	.98	.63	.83	.71	.30
Combined Groups (N=400)						
Maximum Discrepancy	-.04	-.03	-.04	-.04	-.04	-.06
Probability	.63	.92	.59	.55	.59	.17

Test Score Intercorrelations

Estimates of the parameters in Figure 1 were obtained by fitting the model to a correlation matrix. Thus, the first step toward that goal was the computation of the intercorrelations among the six maximum likelihood scores. These intercorrelations, along with the means and standard deviations of each score, are shown in Table 5 for the fall and winter data separately and combined. In general, the variabilities of the classroom achievement test scores (MQ1C and MQ2C) were higher than the variabilities of the corresponding adaptive achievement test scores (MQ1A, MQ2A). This suggests that the volunteers were more homogeneous with respect to achievement than was the class as a whole. Another contrast seen in Table 5 is that the mean achievement scores on the classroom tests were higher than the corresponding means for the adaptive tests. Since the adaptive achievement test was taken anywhere between one day and three weeks after the classroom achievement test, this may indicate that some forgetting took place. An alternative explanation for the lower means on the adaptive achievement tests is that the students were less motivated to perform to their full capabilities on the adaptive test; scores on the adaptive achievement test did not count toward their course grades, while their grades were based on scores from the classroom tests.

Table 5
Means and Standard Deviations and Intercorrelation Matrices of Six Scores for Fall, Winter, and Combined Data

Group and Test	Mean	*S.D.*	Test				
			MQ1C	MQ2C	MQ1A	MQ2A	VOC1
Fall (*N*=213)							
MQ1C	.551	1.028					
MQ2C	.434	.898	.699				
MQ1A	.024	.883	.741	.665			
MQ2A	-.048	.874	.665	.748	.692		
VOC1	-.454	.966	.230	.239	.335	.246	
VOC2	-.329	.967	.274	.277	.375	.278	.890
Winter (*N*=187)							
MQ1C	.529	.975					
MQ2C	.438	.904	.619				
MQ1A	-.120	.915	.782	.586			
MQ2A	.014	.815	.619	.768	.629		
VOC1	-.473	.983	.387	.408	.376	.378	
VOC2	-.418	1.052	.371	.349	.346	.331	.851
Combined (*N*=400)							
MQ1C	.541	1.000					
MQ2C	.436	.900	.662				
MQ1A	-.043	.900	.758	.625			
MQ2A	-.019	.847	.644	.756	.657		
VOC1	-.463	.973	.302	.319	.354	.305	
VOC2	-.371	1.001	.320	.311	.362	.300	.870

As expected, the intercorrelation matrices show that the achievement test scores were more highly correlated among themselves than they were with the vocabulary scores. Within the achievement data, the highest correlations in all three matrices were between tests taken on the same material (i.e., MQ1A and MQ1C, and MQ2A and MQ2C).

Nomological Net Analysis

Validity of classroom and adaptive achievement tests. The results of fitting the validity model to the fall data are shown in Table 6. The χ^2 reported at the bottom is a measure of the overall fit of the model to the data.

Table 6

Standardized Maximum Likelihood Parameter Estimates of Achievement Model Fitted to Fall Data (N=213) and Winter Data (N=187)

Parameter	Description	Estimate Fall	Estimate Winter
λ_1	Validity of classroom biology achievement test, first midquarter (MQ1C)	.853	.893
λ_2	Validity of classroom biology achievement test, second midquarter (MQ2C)	.866	.868
λ_3	Validity of adaptive biology achievement test, first midquarter (MQ1A)	.869	.876
λ_4	Validity of adaptive biology achievement test, second midquarter (MQ2A)	.864	.884
λ_5	Validity of adaptive vocabulary test at first midquarter (VOC1)	.890	.972
λ_6	Validity of adaptive vocabulary test at second midquarter (VOC2)	.999	.876
β	Regression of achievement at second midquarter (ACH2) on achievement at first midquarter (ACH1)	.925	.734
γ_1	Regression of achievement at first midquarter (ACH1) on verbal ability (VER)	.380	.447
γ_2	Regression of achievement at second midquarter (ACH2) on verbal ability (VER)	-.031	.123
ζ_1	Variance of residuals for achievement, first midquarter (ACH1)	.855	.800
ζ_2	Variance of residuals for achievement, second midquarter (ACH2)	.165	.361
ϕ	Variance of verbal ability (VER)	1.000	1.000
$\sigma^2_{\varepsilon 1}$	Error variance for MQ1C	.521	.451
$\sigma^2_{\varepsilon 2}$	Error variance for MQ2C	.496	.482
$\sigma^2_{\varepsilon 3}$	Error variance for MQ1A	.500	.496
$\sigma^2_{\varepsilon 4}$	Error variance for MQ2A	.504	.467
$\sigma^2_{\varepsilon 5}$	Error variance for VOC1	.456	.236
$\sigma^2_{\varepsilon 6}$	Error variance for VOC2	.000	.483
χ^2		8.39	4.04
df		6	6
p		.21	.67

A better fit of the model to the data is indicated by higher values of p, the probability of the observed χ^2 value. In this particular case, the probability was .21, indicating an adequate fit of the model to the fall data. This may be considered evidence in favor of the validity of the nomological net postulated earlier. However, to determine whether the adaptive or classroom tests were more valid measures of achievement requires examination of the values of the parameter estimates.

The first four lines of Table 6 show the standardized loadings (validities) of the four achievement measures on their respective constructs. For the first midquarter in the fall group, the coefficient (λ_1) was .853 for the classroom achievement test (MQ1C); for the adaptive achievement test (MQ1A), the coefficient (λ_3) was .869. The corresponding data for the second midquarter in the fall group (λ_2 and λ_4) were .866 for MQ2C and .864 for MQ2A.

The last column of Table 6 shows the results for winter data. Again, the fit statistic at the bottom of the table indicated that the nomological net postulated for these data was a reasonable summary of the intercorrelations among the six scores. Moreover, for the winter data the fit was better than for the fall data (p=.67 vs. .21).

The validity coefficients for the four biology achievement tests (λ_1 through λ_4) indicated that for the winter data the first classroom midquarter test (MQ1C) was slightly more valid than the corresponding adaptive midquarter test (λ_1=.893 vs. λ_3=.876). This was a reversal of the findings with fall data where the adaptive midquarter test was found to be more valid (λ_1=.853 vs. λ_3=.869). However, for winter data, the second adaptive midquarter test was more valid than the classroom counterpart (λ_4=.884 vs. λ_2=.868), whereas for fall both testing procedures were found to be about equally valid (λ_4=.864 vs. λ_2=.866).

Table 7 summarizes the construct validity correlations in Table 6 and provides information on the average numbers of items in the classroom and adaptive biology achievement tests. As Table 7 shows, both testing procedures achieved essentially equal validities in both quarters. However, in both cases the adaptive achievement tests achieved essentially the same level of validity with considerably fewer items, on the average. For the fall data, the average length of the first adaptive achievement midquarter test was 24.1 items, while that of the first classroom achievement midquarter test was 35 items; the difference of 11 items represents a reduction of 31% in the length relative to the classroom achievement test with a slight increase in validity. For the other three tests, reductions due to adaptive achievement testing were 27% and 25% for both winter tests, again with essentially no differences in validities.

Thus, the adaptive achievement test was effectively more valid, since it required fewer items to yield scores as valid as the classroom achievement test. However, it may be noted that the adaptive achievement tests were drawn from item pools with a higher mean discrimination than the items in the

Table 7
Construct Validity Correlations (*r*) for
Classroom and Adaptive Biology Achievement Tests

Group and Test	Classroom Average No. Items	Classroom *r*	Adaptive Average No. Items	Adaptive *r*
Fall Quarter (*N*=213)				
First Midquarter	35.0	.853	24.1	.869
Second Midquarter	37.0	.866	27.2	.864
Winter Quarter (*N*=187)				
First Midquarter	40.4	.893	30.4	.876
Second Midquarter	40.2	.868	30.0	.884

classroom achievement tests. This, however, seems to be an inherent advantage of the adaptive achievement testing procedure and not an unfair one. Additional research comparing adaptive and conventional paper-and-pencil achievement tests will be necessary to determine whether the effectively higher validity of adaptive tests was due to the higher average item discriminations or to the process of adapting the test to each student.

Other parameters of the nomological net. Table 6 also shows the estimated regression (β) of achievement at the second midquarter (ACH2) on achievement at the first midquarter (ACH1), for both the fall and winter data. For the fall data this coefficient was very high (.925), suggesting that subsequent achievement was largely determined by previous achievement. For the winter data the regression coefficient was β=.734, which suggested a decrease in the influence of ACH1 on ACH2; but since these are standardized estimates, that conclusion may not be completely justified because of possibly different variabilities in achievement between the two quarters.

The regression coefficients (γ_1, γ_2) of the achievement constructs (ACH1 and ACH2) on verbal ability for both fall and winter data are also shown in Table 6. For the fall data, achievement at the first midquarter (ACH1) seemed to be more influenced by verbal ability (γ_1=.380) than achievement at the second midquarter (γ_2=-.031). Since the regression of ACH2 on VER is a partial regression weight, the fact that it was close to zero indicates that verbal ability did not influence achievement at the second midquarter beyond the influence that it exerted through ACH1. The amount of achievement variance that remained unexplained after taking into consideration verbal ability is indicated by the residual variances of ACH1 and ACH2, ζ_1 and ζ_2. Since the solution was standardized, these data can be interpreted directly as proportions of variance. Thus, for ACH1 most of the variance (85%) remained unexplained in this model. The other 15% was explained, in this case, by verbal ability. By contrast, for ACH2, the proportion left unexplained was only 17%, i.e., verbal ability and achievement at the first midquarter accounted for 83% of the variance.

As was true of the fall data, in the winter data verbal ability had a moderate, but somewhat larger, influence on achievement at the first midquarter.

This was reflected in the residual variance of ACH1 (ζ_1), which was 80% as compared with 85% for the fall data. Thus, verbal ability accounted for 5% more variance of ACH1 in the winter data than in the fall data. There was, on the other hand, an increase in the winter data in the proportion of the unexplained variance of ACH2 (ζ_2). In the fall data that proportion was 17%; in the winter data it was 36%.

Effect of Verbal Ability on Achievement Test Performance

Table 8 shows the maximum likelihood estimates of the factor pattern matrix associated with the four-factor model postulated to account for the intercorrelations among the six tests for the fall and winter data combined. The χ^2 statistic of 6.15 with 1 degree of freedom (p=.013) suggests that the fit was statistically not very good. However, the residual correlation matrix (i.e., the difference between the observed correlation matrix and the reproduced correlation matrix computed using the solution in Table 8) was nearly zero with the largest residual correlation being -.014, which suggests an adequate fit of the data to the model.

Table 8
Maximum Likelihood Solution for Four-Factor Model for Fall and Winter Data Combined (N=400)

	Factor			
	I	II	III	IV
Test	Ability	Achievement	Occasion 1	Occasion 2
MQ1C	.334	.750	.329	.0
MQ2C	.337	.726	.0	.334
MQ1A	.384	.703	.334	.0
MQ2A	.324	.739	.0	.331
VOC1	.928	.0	.0	.0
VOC2	.937	.0	.0	.0

Note. χ^2=6.15; df=1; p=.013

The variance component estimates derived from the solution in Table 8 are shown in Table 9. These were obtained by squaring the corresponding loadings. The first row of Table 9 shows the proportion of performance variance in each test accounted for by verbal ability. For the two classroom achievement midquarter tests (MQ1C and MQ2C), the proportion was .11. For the first adaptive achievement midquarter (MQ1A), that proportion was .15; and for the second adaptive achievement midquarter (MQ2A), it was .10.

The second row of Table 9 shows the proportion of variance due to achievement in biology. For the first and second classroom achievement midquarter tests (MQ1C and MQ2C) and the second adaptive achievement midquarter (MQ2A), between 53% and 55% of the variance was due to biology achievement. For the first adaptive achievement midquarter, the corresponding percentage was 49%. The next two rows show the proportion of occasion-specific variance associated with the four achievement tests. In all cases, that proportion was .11. Finally, the last row shows the proportion of variance unaccounted for in each test, which was essentially constant for each of the achievement tests.

Table 9
Variance Components for Fall and Winter Data Combined (*N*=400)

	Test					
Source of Variance	MQ1C	MQ2C	MQ1A	MQ2A	VOC1	VOC2
Verbal Ability	.11	.11	.15	.10	.86	.88
Achievement	.55	.53	.49	.55	.00	.00
Occasion 1	.11	.00	.11	.00	.00	.00
Occasion 2	.00	.11	.00	.11	.00	.00
Residual	.23	.25	.25	.24	.14	.12
Total	1.00	1.00	1.00	1.00	1.00	1.00

Discussion and Conclusions

The focus of the study was to assess the validities of two testing procedures--conventional paper-and-pencil and adaptive--in the context of a meaningful nomological net or model of the achievement process. This model was illustrated in Figure 1 and found to fit data from two academic quarters very well. In the context of this model, validity was indexed by the loading of observed performance on the corresponding achievement construct. It was found that out of four comparisons, the adaptive procedure was somewhat more valid in two cases, equally valid in one, and somewhat less valid in another. However, in all instances, the adaptive procedure was at least 25% shorter on the average than the conventional paper-and-pencil testing procedure. Thus, in a practical sense, the adaptive testing procedure was considerably more valid in all instances.

While these results demonstrate the increased efficiency of adaptive testing in practical situations, the results also raise questions of a theoretical nature. Previous results reported by Bejar, Weiss, and Gialluca (1977) indicated that the adaptive test provided higher levels of information than did the conventional paper-and-pencil test, even though the adaptive test was shorter on the average. The substantial differences in information in favor of the adaptive testing procedure would lead to the expectation that the scores from the adaptive testing procedure would likewise be substantially more valid while at the same time reducing test length. However, this expectation was not totally fulfilled. This might have resulted from the presence of situational factors during the administration of the adaptive test which were not present during the classroom paper-and-pencil administration.

One such factor was identified in the present study--namely, the larger influence of verbal ability in the first adaptive test administration. The results from the confirmatory factor analysis helped in understanding the findings from the nomological net analysis with respect to the validity of adaptive and conventional paper-and-pencil achievement testing scores by corroborating the differential influence of verbal ability on test performance. The data showed that performance on the first adaptive achievement midquarter test (MQ1A) was more dependent on verbal ability than was performance on the other achievement tests. This may have been due to the fact that learning to properly operate the testing equipment was dependent to some extent on verbal ability. By contrast, the occasion-specific influence on each of the achievement tests was the same. This suggests that the increased influence of verbal ability on the first adaptive achievement midquarter test reduced the role that achievement could otherwise have played. As a result, the validity of the first adaptive achievement midquarter test reported earlier was probably underestimated.

The net result of this situational difference between the adaptive and conventional classroom paper-and-pencil test administrations may have been to introduce a bias into the achievement estimates and corresponding information. That is, the item characteristic curves (ICC) derived from conventional classroom paper-and-pencil administration of a test (which was used to parameterize the items used in the adaptive administration) may not have been an accurate portrayal of the relationship between performance on a test item and achievement the first time the item was administered by computer.

This situational vulnerability of the ICC model may be surprising in view of the "invariant" nature of ICC models. However, the invariance property of ICC models pertains to populations responding to a test under similar circumstances. There is nothing in the theory to suggest that the model is situationally invariant. Whether this is the case or not is a matter of empirical test. In the present study, not only was the medium of administration different but so were the motives for taking the test. That is, the adaptive test data were obtained on volunteers, while the classroom test data were used for grading purposes. In view of these differences, the expectation that the adaptive procedure would be substantially more valid may have been unrealistic.

It is clear from this discussion that further validation studies of adaptive testing should be careful to equate as much as possible the conditions of administration. Specifically, the appropriateness of ICCs derived under circumstances different from those surrounding adaptive testing should be carefully evaluated.

The focus of this investigation has been on the psychometric properties of adaptive and conventional paper-and-pencil testing; however, because of the construct validation approach, the results presented here seem to have relevance to a larger question--namely, the identification of some of the components underlying competence and achievement (see Glaser, 1976). Historically, construct validation has played a minor role in the achievement testing field. One reason for this is that users of achievement tests, as well as some psychometricians (e.g., Shoemaker, 1975) are primarily concerned with content and predictive validity. Their orientation is behavioristic; the question they ask is, what can this individual do? Tests which address this question are called criterion-referenced tests (Glaser, 1963; Glaser & Nitko, 1971; see Hambleton, Swaminathan, Algina, & Coulson, 1978, for a recent review); however, Messick (1975) argues persuasively that tests must also be construct referenced. That is, to fully understand test scores, the processes, attributes, and traits determining test performance must be understood.

Since verbal ability is an indicator of information-processing efficiency in short-term memory (Hunt, Lunneborg, & Lewis, 1975; Glaser, 1976), the results of this study give an indication of the influence this cognitive mechanism has on achievement, at least within this course. Knowledge of the cognitive mechanism underlying achievement would seem to be a prerequisite to adaptive instruction. For instance, Glaser (1976) has suggested,

> They [tests] will have to assess performance attainments and capabilities that can be matched to available educational options in more detailed ways than can be carried out with currently used testing and assessment procedures. (Glaser, 1976, p. 21)

The role of achievement testing in this broader context is to provide information relevant to instructional decisions about an individual in an instructional course. The results of the present study have demonstrated that adaptive testing can fulfill that assignment more efficiently than conventional paper-and-pencil testing.

References

Andrews, P. F., Gnanadesikian, R., & Warner, J. L. Methods of assessing multivariate normality. In P. R. Krishnaiah (Ed.), Multivariate analysis III. New York: Academic Press, 1973, pp. 95-116.

Bejar, I. I. Applications of adaptive testing in measuring achievement and performance. In D. J. Weiss (Ed.), Applications of computerized adaptive testing (Research Report 77-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1977. (NTIS No. AD A038114)

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. An information comparison of conventional and adaptive tests in the measurement of classroom achievement (Research Report 77-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977. (NTIS No. AD A047495)

Bejar, I. I., Weiss, D. J., & Kingsbury, G. G. Calibration of an item pool for the adaptive measurement of achievement (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1977. (NTIS No. AD A044828)

Betz, N. E., & Weiss, D. J. Psychological effects of immediate knowledge of results and adaptive ability testing (Research Report 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1976. (NTIS No. AD A027170)

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addision-Wesley, 1968.

Boruch, R. F., & Wolins L. A procedure for estimation of trait, method, and error variance attributable to a measure. Educational and Psychological Measurement, 1970, 30, 547-574.

Brown, J. M., & Weiss, D. J. An adaptive testing strategy for achievement test batteries (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977. (NTIS No. AD A046062)

Campbell, D. T., & Fiske, D. W. Convergent and divergent validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education, 1971.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

DeWitt, L. J., & Weiss, D. J. *A computer software system for adaptive ability measurement* (Research Report 74-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, January 1974. (NTIS No. AD 773961)

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 1963, *18*, 519-521.

Glaser R. Components of a psychology of instruction. *Review of Educational Research*, 1976, *46*, 1-24.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education, 1971.

Goldberger, A. S., & Duncan, A. S. *Structural equation models in the social sciences*. New York: Seminar Press, 1973.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1978, *48*, 1-48.

Hunt, E. G., Lunneborg, C. E., & Lewis, J. What does it mean to be high verbal? *Cognitive Psychology*, 1975, *7*, 194-227.

Jöreskog, K. G. Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2). San Francisco: W. H. Freeman & Co., 1974.

Jöreskog, K. G., Gruvaeus, G. T., & van Thillo, M. *ACOVS--A general computer program for the analysis of covariance structures* (Research Bulletin 70-15). Princeton, NJ: Educational Testing Service, 1970.

Jöreskog, K. G., & Sorbom, D. Statistical models and methods for test-retest situations. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in psychological and educational measurement*. New York: John Wiley & Sons, 1976, pp. 135-158.

Jöreskog, K. G., & van Thillo, M. *LISREL--A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables* (Research Bulletin 72-56). Princeton, NJ: Educational Testing Service, 1972.

Lindgren, B. W. *Statistical theory* (2nd ed.). New York: Macmillan Publishing Co., Inc., 1968.

Linn, R. L., Rock, D. A., & Cleary, T. A. The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, 1969, *29*, 129-146.

McBride, J. R., & Weiss, D. J. *A word knowledge item pool for adaptive ability measurement* (Research Report 74-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1974. (NTIS No. AD 781894).

McBride, J. R., & Weiss, D. J. *Some properties of a Bayesian adaptive ability testing strategy* (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (NTIS No. AD A022964).

Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 1975, *30*, 955-966.

Prestwood, J. S., & Weiss, D. J. *Accuracy of perceived test-item difficulty* (Research Report 77-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, May 1977. (NTIS No. AD A041084).

Reckase, M. D. *Computerized achievement testing using the simple logistic model*. Paper presented at the 1977 Annual Meeting of the American Educational Research Association, New York, NY, 1977.

Schmitt, N. Path analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, 1978, *2*, 157-173.

Shoemaker, D. M. Toward a framework for achievement testing. *Review of Educational Research*, 1975, *45*, 127-148.

Urry, V. W. A five-year quest: Is computerized adaptive testing feasible? In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing*. Washington, DC: U.S. Civil Service Commission, 1976, pp. 97-102.

Vale, C. D. Strategies of branching through an item pool. In D. J. Weiss (Ed.), *Computerized adaptive trait measurement: Problems and prospects* (Research Report 75-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1975. (NTIS No. AD A018675)

Vale, C. D., & Weiss, D. J. *A simulation study of stradaptive ability testing* (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1975. (NTIS No. AD A020961)

Weiss, D. J. *The stratified adaptive computerized ability test* (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973. (NTIS No. AD 758376)

Weiss, D. J. *Computerized ability testing, 1972-1975* (Final Report). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, April 1976. (NTIS No. AD A024516)

APPENDIX: SUPPLEMENTARY TABLES

Table A

Item Number, Discrimination (*a*), Difficulty (*b*), and Guessing (*c*) Parameters for Items in the Midquarter 1 Stradaptive Item Pool

Item	*a*	*b*	*c*
Stratum 9			
(15 items)			
3209	2.50	2.29	.29
3417	2.50	3.00	.35
3033	1.54	2.44	.35
3440*	1.52	2.00	.30
3251	2.50	2.39	.35
3406	1.31	2.48	.35
3045	1.02	2.48	.27
3242	.94	2.40	.35
3407	1.02	2.41	.29
3263*	.99	2.29	.35
3241	.91	2.09	.17
3414	.88	2.29	.32
3402	.83	2.44	.35
3247	.82	2.42	.35
3228	.67	2.49	.31
Mean (F)	1.34	2.43	.32
Mean*(W)	1.33	2.39	.32
Stratum 8			
(20 items)			
3409	2.50	1.28	.00
3234	2.50	1.73	.00
3018	.89	1.25	.35
3204	1.14	1.66	.35
3422	1.47	1.50	.35
3411	1.36	1.23	.35
3250	.91	1.94	.29
3206	.74	1.51	.21
3410	1.30	1.34	.31
3429	1.25	1.24	.28
3419	1.23	1.48	.25
3421	1.17	1.15	.35
3436*	1.12	1.59	.35
3271*	.95	1.32	.30
3061*	.95	1.57	.30
3427	.92	1.51	.26
3449*	.91	1.26	.14
3063*	.91	1.51	.35
3074*	.84	1.79	.35
3420	.68	1.62	.35
Mean (F)	1.29	1.46	.26
Mean*(W)	1.19	1.47	.27
Stratum 7			
(20 items)			
3408	2.50	1.05	.31
3437	1.95	.66	.28
3258	1.24	.81	.35
3432	1.72	.67	.35
3048	1.35	.66	.33
3413	1.40	.76	.35
3448*	1.40	.73	.30
3439*	1.36	.64	.32
3219	1.23	.62	.21
3072*	1.02	.65	.32
3277*	1.00	1.04	.35
3035	.90	.68	.28
3433	1.35	.86	.30
3447*	1.18	.93	.32
3064*	.94	.86	.24
3230	.90	.87	.35
3444*	.88	.78	.35
3012	.75	.80	.35
3260	.71	.84	.28
3056*	.71	.89	.26
Mean (F)	1.28	.78	.31
Mean*(W)	1.22	.79	.31
Stratum 6			
(19 items)			
3047	1.66	.44	.29
3079*	1.61	.27	.35
3213	.93	.52	.35
3041	1.51	.23	.35
3062*	1.47	.43	.30
3405	1.40	.55	.32
3445*	1.19	.44	.34
3218	.82	.58	.12
3019	1.31	.29	.29
3207	.70	.46	.28
3431	.70	.28	.34
3000	1.24	.52	.35
3046	1.18	.24	.22
3042	1.15	.37	.27
3050	1.13	.35	.18
3066	1.05	.53	.31
3034	1.01	.37	.28
3262	.81	.47	.35
3438	.70	.21	.27
Mean (F)	1.13	.40	,28
Mean*(W)	1.14	.40	.29
Stratum 5			
(15 items)			
3282*	2.06	-.02	.35
3220	1.79	-.03	.26
3005	1.43	.11	.35
3425	1.36	.17	.23
3053	1.12	.12	.00
3214	1.12	.03	.23
3412	1.12	.19	.35
3051	1.29	.21	.28
3279*	.99	.01	.28
3403	.99	.18	.19
3069*	.88	-.01	.35
3211	.88	.01	.13
3002	.82	.13	.14
3426	.68	.07	.22
3423	.66	.16	.27
Mean (F)	1.11	.11	.22
Mean*(W)	1.15	.09	.24
Stratum 4			
(13 items)			
3256	2.31	-.33	.26
3430	1.15	-.30	.29
3031	1.47	-.33	.35
3254	3.38	-.17	.22
3237	1.54	-.37	.18
3404	.65	-.29	.35
3244	1.35	-.44	.23
3058*	1.05	-.43	.35
3240	.98	-.28	.15
3268*	.97	-.28	.18
3208	.76	-.16	.12
3006	.77	-.37	.33
3259	.69	-.41	.20
Mean (F)	1.27	-.31	.25
Mean*(W)	1.23	-.32	.25
Stratum 3			
(19 items)			
3021	1.96	-.49	.21
3217	1.06	-.48	.14
3052	1.71	-.93	.00
3055*	1.71	-.65	.24
Stratum 3, con t.			
3215	1.59	-.82	.23
3011	1.32	-.86	.20
3435*	.83	-.61	.35
3216	1.27	-.62	.18
3054*	1.29	-.93	.31
3221	1.25	-.52	.17
3049	1.15	-.71	.18
3255	1.14	-.72	.26
3067*	1.07	-.76	.21
3246	1.10	-.72	.28
3022	1.01	-.48	.30
3272*	1.06	-.81	.35
3017	.99	-.58	.16
3076*	.94	-.73	.21
3224	.80	-.50	.37
Mean (F)	1.26	-.65	.20
Mean*(W)	1.22	-,68	.22
Stratum 2			
(20 items)			
3023	2.40	-1.15	.35
3202	1.81	-.99	.21
3415	.85	-.96	.35
3245	1.34	-.96	.21
3236	1.26	-1.20	.33
3020	1.23	-1.28	.17
3028	1.12	-1.26	.35
3226	1.09	-.98	.20
3210	1.04	-1.22	.35
3239	1.04	-1.13	.21
3013	1.00	-.97	.35
3267*	1.02	-1.22	.23
3257	.98	-1.02	.25
3070*	.95	-1.28	.22
3036	.92	-1.18	.16
3014	.86	-1.24	.14
3060*	.86	-1.31	.29
3274*	.85	-1.05	.26
3238	.82	-1.06	.21
3032	.77	-1.06	.27
Mean (F)	1.16	-1.10	.26
Mean*(W)	1.11	-1.13	.26
Stratum 1			
(17 items)			
3077*	2.50	-1.39	.20
3027	1.67	-1.38	.35
3443*	1.07	-1.64	.35
3249	.91	-1.69	.17
3428	.90	-1.56	.35
3073*	1.43	-1.57	.31
3205	1.25	-1.53	.19
3078*	1.24	-1.65	.35
3057*	1.20	-1.35	.26
3065*	1.17	-1.66	.35
3235	1.15	-1.40	.28
3029	1.13	-1.50	.28
3201	1.07	-1.34	.23
3008	.96	-1.75	.18
3252	.79	-1.77	.35
3003	.96	-1.76	.34
3044	.87	-1.42	.15
Mean (F)	1.06	-1.55	.26
Mean*(W)	1.19	-1.55	.28

Note. Items with asterisks are those which were added to the pool Winter quarter. All other items were in the pool both Fall and Winter quarters.

Table B

Item Number, Discrimination (*a*), Difficulty (*b*), and Guessing (*c*) Parameters for Items in the Midquarter 2 Stradaptive Item Pool

Item	*a*	*b*	*c*
Stratum 9			
(18 items)			
3831	2.50	1.96	.06
3690	2.50	2.36	.24
3833*	2.50	2.85	.35
3904	2.45	1.48	.28
3805	2.50	2.38	.35
3698	2.11	2.82	.35
3901	1.55	2.62	.35
3835*	1.21	2.28	.35
3620	2.04	2.97	.35
3697	1.56	3.00	.35
3810	.92	2.20	.27
3664	1.11	1.60	.35
3625	.98	1.66	.35
3622	.95	2.53	.35
3841*	.87	2.13	.35
3651	.95	2.30	.35
3728*	.91	2.55	.35
3712*	.75	1.64	.30
Mean (F)	1.70	2.31	.31
Mean*(W)	1.58	2.30	.32
Stratum 8			
(18 items)			
3615	1.69	1.17	.29
3916	1.39	1.14	.35
3673	1.51	1.11	.31
3804	.95	1.42	.35
3733*	1.24	1.40	.35
3719*	1.18	1.08	.31
3921*	.91	1.23	.29
3827	.87	1.35	.35
3716*	1.14	1.14	.27
3642	1.11	1.11	.24
3902	.73	1.49	.29
3627	1.03	1.07	.35
3681	1.03	1.54	.35
3676	.89	1.51	.25
3644	.88	1.25	.35
3717*	.83	1.25	.35
3670	.80	1.11	.35
3647	.79	1.14	.35
Mean (F)	1.05	1.26	.32
Mean*(W)	1.05	1.25	.32
Stratum 7			
(15 items)			
3743*	2.14	.68	.32
3661	1.90	.68	.32
3674	1.72	.63	.26
3909	1.34	.77	.35
3662	1.54	.93	.27
3654	1.51	.84	.21
3669	1.45	.70	.32
3623	1.42	.74	.31
3912	.95	.70	.19
3734*	.89	.96	.35
3700	.84	.85	.30
3659	1.37	.67	.29
3635	1.17	.66	.35
3612	1.12	.75	.35
3616	.86	.62	.25
Mean (F)	1.32	.73	.29
Mean*(W)	1.35	.75	.30

Item	*a*	*b*	*c*
Stratum 6			
(20 items)			
3707*	1.75	.55	.31
3746*	1.59	.43	.30
3806	1.57	.48	.35
3925*	1.14	.48	.35
3658	1.24	.32	.35
3905	.98	.35	.20
3738*	1.34	.40	.35
3605	1.22	.57	.34
3815	.95	.58	.35
3611	1.22	.39	.32
3675	1.21	.40	.28
3820	.92	.38	.12
3665	1.19	.54	.22
3709*	1.19	.30	.35
3724*	1.14	.37	.30
3819	.76	.53	.35
3918*	.66	.35	.23
3614	.79	.46	.35
3923*	.63	.38	.31
3626	.65	.52	.25
Mean (F)	1.06	.46	.29
Mean*(W)	1.11	.44	.30
Stratum 5			
(15 items)			
3742*	1.89	.27	.35
3745*	1.58	-.07	.20
3720*	1.45	.26	.29
3607	1.38	.09	.35
3811	1.15	.22	.35
3908	1.15	.07	.31
3649	1.32	.11	.22
3632	1.23	.27	.35
3718*	1.22	.16	.33
3629	1.11	-.03	.35
3732*	.96	-.01	.35
3633	.94	-.08	.35
3609	.78	.18	.35
3730*	.75	.01	.10
3618	.64	-.05	.00
Mean (F)	1.08	.09	.29
Mean*(W)	1.17	.09	.28
Stratum 4			
(19 items)			
3744*	1.94	-.35	.30
3708*	1.62	-.20	.16
3631	1.53	-.18	.35
3814	1.26	-.32	.35
3903	1.21	-.43	.31
3671	1.51	-.14	.26
3701	.82	-.15	.35
3643	1.40	-.50	.25
3914	.98	-.39	.16
3693	1.13	-.24	.24
3725*	1.09	-.52	.24
3710*	1.02	-.33	.30
3653	.83	-.51	.33
3660	.78	-.39	.14
3922*	.64	-.26	.30
3606	.71	-.22	.14
3663	.69	-.17	.33
3696	.68	-.35	.00
3656	.63	-.31	.34
Mean(F)	1.01	-.31	.25
Mean*(W)	1.08	-.31	.26

Item	*a*	*b*	*c*
Stratum 3			
(17 items)			
3634	1.79	-.58	.30
3739*	1.68	-.61	.35
3809	1.27	-.61	.35
3924*	1.13	-.79	.18
3672	1.57	-.80	.15
3737*	1.41	-.66	.34
3915	1.08	-.61	.16
3640	1.43	-.69	.35
3906	.87	-.66	.14
3812	.82	-.63	.13
3682	1.33	-.72	.34
3637	1.29	-.73	.28
3636	1.24	-.63	.27
3641	1.20	-.65	.22
3711*	1.05	-.56	.35
3608	1.04	-.78	.16
3705*	.87	-.58	.14
Mean (F)	1.24	-.67	.24
Mean*(W)	1.25	-.66	.25
Stratum 2			
(20 items)			
3735*	1.63	-.94	.35
3648	1.59	-.96	.33
3807	1.52	-1.10	.17
3907	1.43	-1.08	.35
3704*	1.39	-1.13	.23
3655	1.37	-.90	.35
3813	1.20	-.97	.17
3919*	1.30	-.98	.21
3680	1.33	-1.01	.16
3808	.99	-1.00	.30
3686	1.26	-.88	.29
3721*	1.23	-1.20	.22
3821	.90	-.92	.35
3679	1.21	-.94	.17
3685	1.19	-1.01	.16
3668	.97	-.87	.14
3684	.86	-.85	.14
3703*	.83	-1.16	.21
3617	.79	-1.11	.14
3713*	.75	-1.18	.33
Mean (F)	1.19	-.97	.23
Mean*(W)	1.19	-1.01	.24
Stratum 1			
(19 items)			
3741*	1.63	-1.56	.35
3910	1.58	-1.59	.21
3692	1.53	-1.28	.35
3825	1.09	-1.38	.34
3639	1.47	-1.80	.35
3638	1.35	-1.54	.21
3913	1.31	-1.31	.19
3837*	1.09	-1.59	.25
3715*	1.16	-1.63	.26
3920*	1.12	-1.34	.23
3842*	1.01	-1.55	.35
3695	1.09	-1.73	.22
3731*	1.05	-1.67	.35
3832	.99	-1.74	.32
3838*	.99	-1.68	.35
3613	.86	-1.74	.33
3683	.85	-1.31	.14
3657	.81	-1.74	.35
3610	.80	-1.33	.14
Mean (F)	1.14	-1.54	.26
Mean*(W)	1.15	-1.55	.28

Note. Items with asterisks are those which were added to the pool Winter quarter. All other items were in the pool both Fall and Winter quarters.

Table C
Item Discrimination (*a*), Difficulty (*b*), and Guessing (*c*)
Parameters for Classroom Tests MQ1C and MQ2C in Fall Quarter

MQ1C				MQ2C			
Item No.	*a*	*b*	*c*	Item No.	*a*	*b*	*c*
3060	.86	-1.31	.29	3922	.64	-.26	.30
3067	1.07	-.76	.21	3904	2.45	1.58	.28
3065	1.17	-1.66	.35	3918	.66	.35	.23
3056	.71	.89	.26	3921	.91	1.23	.29
3063	.91	1.51	.35	3919	1.30	-.98	.21
3073	1.43	-1.57	.31	3920	1.12	-1.34	.23
3058	1.05	-.43	.35	3923	.63	.38	.31
3274	.85	-1.05	.26	3924	1.13	-.79	.18
3271	.95	1.32	.30	3801	.80	-.17	.35
3055	1.71	-.65	.24	3841	.87	2.13	.35
3072	1.02	.65	.32	3838	.99	-1.68	.35
3057	1.20	-1.35	.26	3833	2.50	2.85	.35
3064	.94	.86	.24	3837	1.09	-1.59	.25
3069	.88	-.01	.35	3835	1.21	2.28	.35
3054	1.29	-.93	.31	3641	1.20	-.65	.22
3066	1.05	.53	.31	3708	1.62	-.20	.16
3268	.97	-.28	.18	3718	1.22	.16	.33
3267	1.02	-1.22	.23	3728	.91	2.55	.35
3272	1.06	-.81	.35	3665	1.19	.54	.22
3070	.95	-1.28	.22	3730	.75	.01	.10
3008	.96	-1.75	.18	3719	1.18	1.08	.31
3019	1.31	.29	.29	3705	.87	-.58	.14
3062	1.47	.43	.30	3713	.75	-1.18	.33
3061	.95	1.57	.30	3703	.83	-1.16	.21
3262	.81	.47	.35	3709	1.19	.30	.35
3263	.99	2.29	.35	3707	1.75	.55	.31
3447	1.18	.93	.32	3721	1.23	-1.20	.22
3443	1.07	-1.64	.35	3717	.83	1.25	.35
3438	.70	.21	.27	3715	1.16	-1.63	.26
3448	1.40	.73	.30	3716	1.14	1.14	.27
3435	.83	-.61	.35	3720	1.45	.26	.29
3439	1.36	.64	.32	3744	1.94	-.35	.30
3436	1.12	1.59	.35	3745	1.58	-.07	.20
3449	.91	1.26	.14	3746	1.59	.43	.30
3440	1.52	2.00	.30	3711	1.05	-.56	.35
3437	1.95	.66	.28	3710	1.02	-.33	.30
3427	.92	1.51	.26	3724	1.14	.37	.30
3445	1.19	.44	.34	3725	1.09	-.52	.24
3444	.88	.78	.35	3731	1.05	-1.67	.35
				3712	.75	1.64	.30
				3704	1.39	-1.13	.23
Mean	1.09	.11	.29	Mean	1.17	.07	.28

Table D
Item Discrimination (*a*), Difficulty (*b*), and Guessing (*c*) Parameters for Classroom Tests MQ1C and MQ2C in Winter Quarter

MQ1C				MQ2C			
Item No.	*a*	*b*	*c*	Item No.	*a*	*b*	*c*
3287	.85	-1.28	.13	3750	.93	-1.79	.34
3292	.68	1.39	.35	3926	.93	-1.56	.16
3219	1.23	.62	.21	3845	1.71	.26	.29
3290	1.16	-.57	.20	3763	1.23	1.95	.28
3214	1.12	.03	.23	3762	1.97	-1.56	.17
3268	.97	-.28	.18	3772	.74	-.84	.35
3289	1.14	-1.45	.35	3759	.99	-.14	.21
3293	.96	-1.30	.14	3768	1.11	-1.55	.17
3291	.65	.52	.35	3756	1.10	-.21	.28
3249	.91	-1.69	.17	3749	1.05	-1.77	.22
3083	1.05	-.90	.13	3757	1.18	-1.60	.18
3090	1.48	-1.65	.18	3755	1.03	-.12	.16
3054	1.29	-.93	.31	3747	1.11	-1.69	.18
3084	1.22	-1.06	.15	3753	.91	-.55	.17
3092	.98	-.65	.15	3654	1.51	.84	.21
3082	1.05	2.27	.35	3673	1.51	1.11	.31
3011	1.32	-.86	.20	3716	1.14	1.14	.27
3095	.79	-1.20	.12	3700	.84	.85	.30
3085	1.16	-1.81	.35	3773	1.69	1.62	.27
3423	.66	.16	.27	3748	.85	1.31	.35
3453	1.19	.48	.22	3766	1.12	1.41	.35
3456	1.03	2.71	.35	3760	1.28	-1.58	.18
3454	1.10	2.66	.35	3758	.89	-1.45	.15
3460	1.99	1.59	.34	3703	.83	-1.16	.21
3452	.75	1.98	.31	3853	1.05	.12	.17
3406	1.31	2.48	.35	3854	1.03	-.19	.31
3461	.94	1.51	.35	3852	.69	-1.78	.35
3457	.90	1.87	.28	3850	.89	1.83	.35
3459	.84	-.29	.26	3851	.76	.18	.23
3407	1.02	2.41	.29	3752	1.24	-.50	.19
3458	1.46	-1.10	.15	3769	1.15	-.39	.16
3432	1.72	.67	.35	3751	.80	1.91	.35
3455	.96	-.61	.31	3770	2.50	1.73	.00
3420	.68	1.62	.35	3622	.95	2.53	.35
3433	1.35	.86	.30	3761	.84	1.27	.32
3412	1.12	.19	.35	3767	1.02	-.04	.30
3462	1.31	-1.03	.17	3930	1.21	-.44	.35
3285	.79	-.60	.11	3904	2.45	1.58	.28
3294	.76	-.68	.19	3918	.66	.35	.23
3041	1.51	.23	.35	3903	1.21	-.43	.31
3091	1.64	.58	.30	3928	1.00	.65	.35
3089	.92	-.37	.30	3929	.96	-1.76	.22
3093	.75	-.94	.11	3813	1.20	-.97	.17
3096	1.48	-1.48	.16	3927	1.01	-1.34	.16
3086	.74	-.67	.35				
Mean	1.09	.08	.25	Mean	1.14	-.06	.25

Table E

Item Discrimination (*a*) and Difficulty (*b*) Parameter Estimates for Vocabulary Items by Stratum and Midquarter Subpool (*c*=.20 for All Items)

Midquarter 1			Midquarter 2			Midquarter 1			Midquarter 2		
Item No.	*a*	*b*	Item No.	*a*	*b*	Item No.	*a*	*b*	Item No.	*a*	*b*
Stratum 1 (18 items)			Stratum 1 (18 items)			Stratum 3 (continued)			Stratum 3 (continued)		
7	3.00	-2.32	71	3.00	-2.32	173	.88	-1.06	86	.89	-1.19
102	3.00	-2.36	64	3.00	-2.36	204	.88	-.74	637	.88	-1.02
28	3.00	-2.63	25	3.00	-2.63	285	.84	-1.02	227	.81	-1.25
42	3.00	-2.63	14	2.21	-2.46	640	.78	-1.06	584	.76	-.68
24	1.75	-2.37	11	1.75	-2.58	185	.68	-.68	189	.76	-1.19
9	1.45	-2.24	70	1.29	-2.24	232	.67	-1.25	322	.67	-1.09
99	1.24	-2.67	206	1.11	-2.19	112	.61	-.78	235	.66	-.78
65	1.02	-2.71	124	1.09	-2.64	94	.56	-1.02	241	.57	-1.05
68	1.01	-2.47	181	1.02	-2.58	546	.56	-.80	108	.54	-1.16
105	.98	-2.63	126	.96	-2.27	141	.48	-1.21	371	.44	-.92
80	.86	-2.25	198	.80	-2.50	26	.36	-1.02	615	.44	-.86
121	.74	-2.82	16	.75	-2.95	Mean	.89	-.93	Mean	.90	-.98
184	.73	-2.19	89	.72	-2.49	S.D.	.36	.18	S.D.	.37	.19
131	.60	-2.58	17	.72	-2.89	Stratum 4 (18 items)			Stratum 4 (18 items)		
93	.52	-2.18	82	.54	-2.31	501	1.20	-.55	270	1.22	-.14
81	.44	-2.39	151	.44	-2.65	91	1.13	-.20	128	1.07	-.36
73	.43	-2.69	135	.43	-2.79	522	1.06	-.39	143	1.04	-.15
201	.31	-2.97	507	.39	-2.74	130	.95	-.44	332	.97	-.40
Mean	1.34	-2.50	Mean	1.29	-2.53	154	.87	-.12	37	.86	-.24
S.D.	.98	.23	S.D.	.91	.22	46	.84	-.36	156	.84	-.17
Stratum 2 (18 items)			Stratum 2 (17 items)			221	.82	-.28	123	.82	-.56
13	1.89	-1.55	196	2.13	-1.79	203	.82	-.38	33	.80	-.39
190	1.82	-1.44	138	1.73	-2.02	390	.80	-.26	211	.77	-.24
84	1.70	-1.64	27	1.43	-1.68	183	.73	-.45	535	.77	-.37
125	1.24	-1.88	129	1.27	-1.35	307	.70	-.33	110	.70	-.54
22	1.20	-1.97	101	1.17	-1.40	224	.68	-.26	293	.67	-.57
96	1.13	-1.72	44	1.15	-1.41	234	.65	-.13	222	.65	-.50
127	1.08	-1.35	158	1.08	-2.00	53	.64	-.48	117	.62	-.66
134	1.07	-1.94	186	1.07	-1.34	58	.59	-.38	287	.52	-.65
90	.94	-1.31	83	.88	-1.45	218	.41	-.13	588	.47	-.46
34	.83	-1.58	66	.87	-2.02	157	.38	-.25	155	.40	-.57
311	.75	-1.43	262	.77	-1.93	136	.32	-.56	142	.31	-.54
5	.75	-2.16	31	.72	-2.14	Mean	.75	-.33	Mean	.75	-.42
63	.69	-2.14	88	.71	-1.33	S.D.	.24	.14	S.D.	.24	.17
106	.67	-2.01	255	.64	-2.18	Stratum 5 (17 items)			Stratum 5 (18 items)		
202	.62	-2.17	76	.62	-1.75	630	3.00	.28	272	1.96	.22
95	.56	-1.71	559	.62	-1.68	568	1.63	.29	599	1.63	.16
214	.48	-1.49	643	.49	-2.03	161	1.38	.13	329	1.42	.18
276	.45	-1.53	Mean	1.02	1.23	56	1.11	.14	503	1.06	-.09
Mean	.99	-1.72	S.D.	.44	.31	144	.91	.29	104	.94	.05
S.D.	.44	.29				551	.90	.34	365	.88	-.11
Stratum 3 (18 items)			Stratum 3 (18 items)			670	.87	.20	209	.87	.07
194	1.79	-.96	191	1.75	-1.26	52	.84	.21	382	.86	-.01
51	1.43	-1.04	36	1.64	-.79	145	.79	.09	207	.79	-.04
40	1.24	-1.03	87	1.24	-.76	369	.77	.30	502	.73	.22
109	1.11	-.70	43	1.11	-.86	50	.69	.32	645	.67	.24
515	1.08	-.71	199	1.09	-1.09	444	.62	.06	597	.62	-0
103	1.06	-1.00	47	1.04	-.96	292	.61	.01	205	.60	-.02
239	.94	-.71	85	.93	-.67	355	.51	.10	318	.53	.31

(continued)

Table E (continued)

Item Discrimination (*a*) and Difficulty (*b*) Parameter Estimates for Vocabulary Items by Stratum and Midquarter Subpool (*c*=.20 for All Items)

Midquarter 1			Midquarter 2		
Item No.	*a*	*b*	Item No.	*a*	*b*
Stratum 5 (continued)			Stratum 5 (continued)		
116	.49	.33	137	.50	-.06
622	.44	.20	176	.42	-.11
54	.38	.20	354	.33	.15
Mean	.94	.20	547	.30	.22
S.D.	.62	.10	Mean	.84	.08
			S.D.	.45	.14
Stratum 6 (18 items)			Stratum 6 (18 items)		
283	3.00	.49	296	3.00	.67
347	3.00	.49	264	2.28	.55
266	2.12	.51	340	1.92	.65
342	1.59	.54	315	1.85	.52
265	1.57	.55	301	1.38	.47
60	1.23	.64	386	1.25	.54
538	1.18	.52	582	1.20	.35
59	1.09	.60	113	1.06	.68
146	.93	.47	271	.89	.80
633	.71	.47	506	.81	.58
139	.61	.79	267	.65	.77
377	.59	.39	133	.57	.56
590	.54	.62	593	.56	.55
629	.53	.42	519	.53	.44
324	.52	.77	289	.48	.69
252	.42	.47	549	.43	.35
372	.35	.56	165	.38	.56
Mean	1.18	.55	Mean	1.13	.57
S.D.	.84	.11	S.D.	.75	.13
Stratum 7 (17 items)			Stratum 7 (19 items)		
288	3.00	1.26	162	3.00	1.25
337	3.00	1.18	562	3.00	1.22
583	3.00	1.07	541	3.00	1.16
294	3.00	1.07	321	3.00	1.00
617	2.78	1.17	114	3.00	.96
299	1.77	1.16	586	1.54	1.31
306	1.32	1.20	601	1.32	1.10
523	1.21	.88	581	1.26	1.21
598	1.08	1.04	526	1.17	.92
592	1.01	1.20	666	1.00	.85
215	.91	1.07	304	.89	1.34
302	.85	.85	231	.87	1.19
375	.83	.93	111	.82	.94
164	.69	1.14	238	.76	1.13
341	.63	1.28	397	.65	1.34
576	.43	1.13	259	.37	1.29
333	.35	1.34	516	.35	1.12
Mean	1.52	.12	308	.34	1.31
S.D.	1.01	.14	Mean	1.46	1.15
			S.D.	1.03	.16
Stratum 8 (19 items)			Stratum 8 (19 items)		
652	3.00	1.66	360	3.00	1.71
665	3.00	1.62	595	3.00	1.58
Stratum 8 (continued)			Stratum 8 (continued)		
378	3.00	1.48	263	3.00	1.47
174	3.00	1.46	120	3.00	1.46
140	3.00	1.38	254	3.00	1.37
253	2.32	1.44	534	2.52	1.61
561	1.72	1.42	383	2.11	1.52
291	1.64	1.35	572	1.29	1.43
217	1.25	1.38	400	.93	1.68
587	.87	1.36	168	.91	1.55
147	.83	1.47	660	.83	1.37
610	.79	1.57	159	.77	1.56
367	.72	1.40	521	.75	1.70
107	.69	1.59	216	.67	1.40
616	.61	1.76	525	.57	1.51
119	.53	1.73	403	.54	1.76
505	.50	1.43	242	.52	1.57
368	.46	1.42	213	.43	1.43
172	.38	1.36	350	.32	1.53
Mean	1.49	.49	Mean	1.48	1.54
S.D.	1.05	.13	S.D.	1.08	.12
Stratum 9 (18 items)			Stratum 9 (18 items)		
528	3.00	2.82	545	3.00	2.82
604	3.00	2.82	585	3.00	2.82
180	3.00	2.63	274	3.00	2.72
245	3.00	2.63	353	3.00	2.51
381	3.00	2.36	118	3.00	2.41
564	3.00	2.29	273	3.00	2.14
319	3.00	2.14	609	3.00	2.14
445	3.00	2.07	193	3.00	2.07
573	3.00	1.86	115	3.00	2.02
591	3.00	1.80	228	2.94	2.41
316	1.16	2.68	260	.71	1.82
504	.64	1.81	247	.65	2.06
533	.63	2.15	577	.61	2.00
167	.42	2.16	374	.56	1.99
603	.38	1.80	569	.40	2.40
531	.35	1.92	362	.34	2.06
511	.32	2.28	243	.32	2.61
357	.31	2.68	580	.32	2.64
Mean	1.90	2.27	Mean	1.88	2.31
S.D.	1.28	.36	S.D.	1.29	.32
Total, Midquarter 1			Total, Midquarter 2		
Mean	1.22	.01	Mean	1.20	.02
S.D.	.88	1.49	S.D.	.86	1.51