CALIBRATION OF AN ITEM POOL FOR THE ADAPTIVE MEASUREMENT OF ACHIEVEMENT

Isaac I. Bejar
David J. Weiss
G. Gage Kingsbury

RESEARCH REPORT 77-5
SEPTEMBER 1977

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

Prepared under contract No. N00014-76-C-0627, NR150-389
with the Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Research Report 77-5	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Calibration of an Item Pool for the Adaptive Measurement of Achievement		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Isaac I. Bejar, David J. Weiss, and G. Gage Kingsbury		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 September 1977 13. NUMBER OF PAGES 31
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 the 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	sequential testing	programmed testing
achievement testing	branched testing	response-contingent testing
computerized testing	individualized testing	automated testing
adaptive testing	tailored testing	item characteristic curve theory

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The applicability of item characteristic curve (ICC) theory to a multiple-choice test item pool used to measure achievement is described. The rationale for attempting to use ICC theory in an achievement framework is summarized, and the adequacy for adaptive testing of a classroom achievement test item pool in a college biology class is studied. Using criteria usually applied to ability measurement item pools, the item difficulties and discriminations in this achievement test pool were found to be similar to those used in adaptive testing pools for ability testing. Studies of the dimen-

sionality of the pool indicate that it is primarily unidimensional. Analysis of the item parameters of items administered to two different samples reveals the possibility of a deviation from invariance in the discrimination parameter but a high degree of invariance for the difficulty parameter. The pool as a whole, as well as two subpools, is judged to be adequate for use in adaptive testing. It is also concluded that the ICC model is not inappropriate for application to typical college classroom achievement tests similar to the one studied.

CONTENTS

Introduction 1

Alternative Psychometric Bases for Adaptive Testing 2
Classical Test Theory 2
Order Theory 2
Item Characteristic Curve Theory 2

Objective 3

The Achievement Measurement Context 3

The Course and Examination Procedures 3

The Item Pool 4

Applicability of the ICC Model 5

The ICC Model 6

Estimation of Item Parameters 6
Procedure 6
Evaluation of the Estimation Procedure 7
Criteria for Excluding Items 8

Results 9
Excluded Items 9
Item Pool Characteristics 9
Midquarter Subpools 11

Conclusions 12

Dimensionality of the Item Pool 12

Factor Analysis 13
Method 13
Results 14

Equality of ICC's Based on Content Areas and Total Test 16
Rationale 16
Method 16
Results 17

Conclusions 20

Sampling Invariance of Item Parameter Estimates 20

Method 20

Results 22

Conclusions 24

Conclusions 24

References 26

Appendix: Supplementary Tables 28

Calibration of an Item Pool for the Adaptive Measurement of Achievement

The majority of research in adaptive testing to date has been concerned with ability testing (Weiss, 1973, 1976). Very little adaptive testing research has addressed itself to the unique problems of achievement measurement (Weiss, 1973, pp. 40-41). Although frequently treated as if they are highly similar in approach (e.g., English, Reckase, & Patience, 1977), the adaptive measurement of ability and achievement can present quite different problems. These differences arise, in part, from the different kinds of item pools which are available for the measurement of ability vs. achievement.

In the measurement of ability, the test constructor defines the nature of the item pool. Once the ability domain is specified, large numbers of test items can be generated; and the item pool can be defined to have whatever characteristics are deemed by the test constructor to be psychometrically desirable. Thus, ability tests can be designed to be unidimensional by eliminating from the item pool those items which measure extraneous dimensions. Similarly, if an item pool is being developed for adaptive testing, the ability test constructor can construct a unidimensional pool which consists of items with a wide range of difficulties and high discriminations (e.g., McBride & Weiss, 1974). Based on the availability of such a pool, there is little question of the applicability of such unidimensional models as those from latent trait theory (e.g., Lord & Novick, 1968) or the strategies of adaptive testing which have been designed to measure individual differences within a unidimensional framework (Weiss, 1974).

In most practical achievement testing settings, however, test constructors do not have the freedom to contruct the kinds of ideal item pools that are possible in ability measurement. In the achievement testing environment, where the purpose is to measure what students have learned as a result of some instructional exposure, the nature and extent of an item pool is largely dictated by the content covered in the course. Thus, a course might convey information on a variety of topics which are part of the larger content area defining the course but are not so highly correlated with each other that they can be considered to be one dimension. Similarly, because these separable content areas may be limited in scope, it may not be possible for the test constructor to generate large numbers of test items in each content area or to generate a pool of items large enough to meet the requirements of some adaptive testing strategies.

Since adaptive testing in the ability domain has been shown to have considerable promise (Lord, 1977; Urry, 1977; Weiss, 1976), it is appropriate to determine whether it will be similarly useful in applications to the unique problems of achievement measurement. However, because of the differences in the characteristics of the item pools, it is necessary first to examine typical pools of achievement test items; in this way it can be determined whether they can meet the criteria necessary for the implementation of

currently available adaptive testing models or whether new models will be required to implement the adaptive measurement of achievement. This report is addressed to that question.

Alternative Psychometric Bases for Adaptive Testing

There are three general psychometric models on which the adaptive measurement of achievement can be based: classical test theory (Gulliksen, 1950), order theory (Cliff, 1975, 1976), and item characteristic curve (ICC) theory (Lord, 1974).

Classical test theory. In general, classical test theory cannot provide an adequate psychometric framework for an adaptive achievement testing system. The objective of an adaptive testing system is to individualize the test for each testee by selecting test items on the basis of the testee's responses to previously administered items. As a result, different testees respond to different items. Since classical test theory uses as its scoring system the total number of correct answers to test items, testees of different levels of achievement will be indistinguishable from one another if their adaptive tests are scored in this way.

The only method that classical test theory has at its command for dealing with an incomplete response matrix is multiple-matrix sampling (Lord & Novick, 1968). However, although this technique is designed to estimate the mean achievement level of persons in a group, it cannot efficiently estimate an individual's achievement score (Lord, 1977). Furthermore, matrix sampling assumes that each individual in the sample takes a goup of items selected at random from the pool. This assumption runs counter to the philosophy of adaptive testing in which the objective is to select items for each testee in a deliberately non-random manner.

Order theory. One method to circumvent the problems caused by different persons completing different test items is called order theory (Cliff, 1975, 1976). This theory is based on the formation of a triangular matrix which orders individuals using their responses to some subset of items from an item pool. One assumption of order theory is that all items are Guttman items, i.e., items which are perfectly discriminating. However, although this assumption will yield greatly reduced test lengths, it is doubtful that Guttman items will appear in typical achievement testing situations. By basing its procedures on Guttman items, order theory also makes very strong assumptions about unidimensionality--considerably stronger than those made by either classical test theory or ICC theory. Order theory as a general system for the measurement of individual differences is quite new, and many of its basic problems and procedures have yet to be adequately articulated. Perhaps at a later date it will become a useful system for the adaptive measurement of achievement.

Item characteristic curve theory. Item characteristic curve (ICC) theory or item response theory, which has been used to provide a psychometric basis for the adaptive measurement of ability (e.g., Lord, 1976; McBride & Weiss, 1974; Urry, 1976; Vale & Weiss, 1975a,b), may also provide an appropriate model for the adaptive measurement of achievement.

Two properties of ICC theory are especially relevant in this context. First, ICC theory provides a means for obtaining scores on the same metric for persons who have completed different test items. As indicated earlier, this is an essential requirement for adaptive tests. Second, under the assumptions of ICC theory, the resulting score metric is invariant with respect to population. Thus, if a set of data from a given group of testees can be shown to meet the assumptions of ICC theory, it is possible to score all individuals on the same equal interval scale regardless of the subgroup of the population to which they belong.

With these two advantageous properties, ICC theory provides the promise of measurement which is not dependent upon either the set of test items a person has answered or his/her population subgroup membership. There is, in addition, a third advantage of ICC theory: it provides a flexible psychometric framework for the development of criterion-referenced achievement tests. As Hambleton & Cook (1977) note, there is likely to be a great degree of homogeneity among items covering a single criterion-referenced instructional objective. As a result of this homogeneity, the basic assumption of unidimensionality required by ICC models is very likely to be satisfied.

Because of the degree of articulation of ICC theory and the development of means for its implementation, it appears to be a viable approach to the adaptive measurement of achievement. Furthermore, it is possible to test the fit of a set of data to the theory prior to its use for the development of an adaptive testing system.

Objective

Within the context of a practical achievement testing problem, this report is concerned with the applicability of ICC theory to the measurement of achievement. Specifically, its purpose is to 1) evaluate the fit of the item characteristic curve model to items on a multiple-choice achievement test; 2) investigate the dimensionality of an achievement test item pool with respect to the unidimensionality assumption of latent trait theory; and 3) determine whether the item parameters of ICC theory, within the context of an achievement test, are invariant across different subgroups from a population.

The Achievement Measurement Context

The Course and Examination Procedures

This study used data from Biology 1-011, an introductory biology course open to all students at the University of Minnesota. Both majors and non-majors in the natural sciences enroll in this course. Biology 1-011 is offered every quarter. Quarterly enrollment ranges from 1000 to 1500 students, with the fall quarter tending to have the highest number of students. Students are generally freshmen, but a substantial number of sophomores and a few juniors and seniors enroll in the course. The sexes are about equally represented. According to the course staff, there seem to be no important changes in the demographic composition of the student body from quarter to quarter. Instruction in the course is by means of videotaped lectures which are shown on closed circuit television. The lectures do not change from

quarter to quarter but are revised every two years. In addition to the lectures, there is a compulsory laboratory.

Students are given two midquarter examinations and a final examination each quarter. All examinations use multiple-choice items. The first midquarter examination includes 55 questions and each student is required to answer only 50 of them. It covers the areas of 1) chemistry, 2) the cell, and 3) energy. The second midquarter examination also includes 55 questions, of which 50 must be answered. It covers two additional content areas: 4) genetics and 5) reproduction and embryology. The final examination includes 110 items, of which only 100 must be answered. It covers the five previous content areas plus two additional ones: 6) ecology and 7) evolution.

Table 1
Content Areas and Item Number Ranges

Content Area Number	Content	Item Numbers
1	Chemistry	3000-3200
2	The Cell	3201-3400
3	Energy	3401-3600
4	Heredity/Genetics	3601-3800
5	Reproduction and Embryology	3801-4000
6	Ecology	4001-4200
7	Evolution	4201-4400

The Item Pool

The basic item pool for this study consisted of item responses on the two midquarter examinations and the final examination for winter and spring quarters of 1976. Items were classified by content areas; items in each content area were assigned numbers within the range shown in Table 1.

Table 2
Number of Items in the Item Pool by Test and Content Area

	Content Area							
Test	1	2	3	4	5	6	7	Total
W1	21	22	12					55
S1	19	25	11					55
W2				36	19			55
S2	2			35	18			55
WF	9	14	7	18	9	28	25	110
SF	9	12	6	17	11	30	25	110
Total	60	73	36	106	57	58	50	440
Unique	53	60	33	101	48	52	47	394

Table 2 shows the number of items in the item pool by source and content area. In the first column of Table 2, the letters S and W refer to spring and winter quarters, while 1, 2, and F refer to the test from which the items were taken: the first midquarter, the second midquarter, and the final examination, respectively. Since some of the items were repeated between the two quarters, Table 2 also shows the number of unique items in each content area. The repeated items were used to test the invariance assumption of ICC theory across population subsamples.

Table 3 shows the number of unique items obtained from each of the exams and the average number of testees who answered each of these items in the tests used for calibration of the item pool.

Table 3
Number of Unique Items and Average Number of Testees for Each Test

Test	Number of Unique Items	Average Number of Testees
W1	48	998
S1	46	838
W2	52	934
S2	48	760
WF	99	888
SF	101	638

The initial goal of these analyses was to form two item pools for later adaptive testing research. Each of these pools was to be designed for use with one of the midquarter examinations. The dimensionality analyses reported below are thus confined to these midquarter item pools. The applicability analyses and the invariance analyses, however, utilized items from the final examinations.

Applicability of the ICC Model

An initial question to be answered in the use of ICC theory in a multi-content achievement test is whether application of the procedures of the unidimensional ICC model to such test items would yield estimates of item parameters which would be useful for adaptive testing. Since adaptive tests function best when items span a wide range of difficulties and have relatively high discriminating power (Urry, 1976; Vale & Weiss, 1975b), it is possible that typical achievement test items might not meet even these minimal requirements. For example, it is possible that because of the varying content in the item pool, item discriminations would be so low as to indicate a great deal of heterogeneity in the test items. Therefore, the first set of analyses of the item pool involved the determination of item parameter estimates for each item in the pool and the examination of the resulting estimates with regard to their utility for the construction of adaptive tests.

The ICC Model

Because the items were multiple-choice, a three-parameter ICC model for dichotomous item responses was appropriate. This model has been described in detail by Hambleton & Cook, 1977; Lord & Novick, 1968, Ch. 17; and McBride & Weiss, 1974. The model assumes that the item characteristic curve for an item can be completely described by three parameters: a, the discriminating power of the item, which is proportional to the maximum slope of the ICC at its point of inflection; b, the item difficulty, which specifies the location on the underlying trait continuum at the point of inflection of the ICC; and c, the "guessing" parameter, which is the probability of a correct response to the item for a testee of infinitely low trait level and is sometimes described as the probability of a correct response by random guessing.

Estimation of Item Parameters

Procedure. The process of estimating item parameters in ICC test theory is essentially a curve-fitting procedure. An item characteristic curve is fit for each item based on the item responses of a group of testees. Because "best fit" may be defined in several ways, there are different estimation procedures (see Hambleton & Cook, 1977, p. 89). The procedure used here was based on a logistic ICC model using a minimum χ^2 definition of fit, as operationalized in Urry's ESTEM program (see Urry, 1976, p. 99).

As defined by Urry, the best-fitting curve is the one that minimizes the criterion

$$K_g = \sum_{j=0}^{m-1} [r_j - n_j P'_g(j)]^2 [n_j P'_g(j) Q'_g(j)]^{-1} \qquad [1]$$

where r_j = the number of testees at score j, who correctly answer item g,

n_j = the number of testees who obtain a score of j,

$P'_g(j)$ is the expected proportion of correct responses to item g, among those with a score of j,

$Q'(j) = [1-P'(j)]$,

m is the number of items in the test.

Urry's computing algorithm consists of two stages. During the first stage, for a given item the procedure increments the value of c (the guessing parameter) from .02 to .30. At each increment, values of a and b consistent with c are found. That is, several trial ICC's are generated. Then, for each of these trial ICC's, Equation 1 is computed. The parameters corresponding to the equation that yield a minimum value of χ^2 are taken as initial estimates. These estimates are refined by a method known as ancillary estimation, which was developed by Fisher (1950). They are refined further at the second stage, which is identical to the first, except that a Bayes

modal estimate of trait level (Samejima, 1969) is used as the metric, rather than the standarized raw scores used in the first stage.

Evaluation of the estimation procedure. The accuracy and efficiency of the ESTEM program has been tested in computer simulations with synthetic data (Gugel, Schmidt & Urry, 1976; Urry, 1976), using sample sizes ranging from 500 to 3000 and test lengths ranging from 50 to 100 items. In these studies two criteria have been used in evaluating the estimates yielded by the program. The first evaluative criterion was the root mean square (RMSE) which was defined as

$$\text{RMSE} = \sum_{g=1}^{n} \frac{(\hat{\alpha}_g - \alpha_g)^{\frac{1}{2}}}{n} \qquad [2]$$

where $\hat{\alpha}_g$ is an estimated parameter value for the g^{th} item,

α_g is the known parameter value from which the synthetic data were generated,

n is the number of items.

Their second evaluative criterion was simply the Pearson product-moment correlation between the estimated parameter value and the known parameter value.

Root mean square error is a measure of the discrepancy between the value of the parameter estimate and the numerical value of the generating parameter; it includes both sampling fluctuations and bias. Its usefulness is limited to comparing estimates of the same parameter across different situations since it is scale dependent. The correlation coefficient, on the other hand, is scale free and can be used in intra- as well as inter-parameter comparisons.

The simulation studies by Gugel, Schmidt, & Urry (1976) provide some data with which to evaluate the applicability of ESTEM's item parameter estimation procedures for the data base available in the present study (i.e., testee groups of between 600 and 1,000 persons and test lengths of 50 or 100 items). Table 4 shows results from the simulation studies of a 50-item test for 500 and 1,000 simulated testees.

Table 4
RMSE and Correlation of Estimate and Parameter Values for the *a*, *b* and *c* Parameters for 50 Items and Two Sample Sizes
[From Gugel, Schmidt and Urry (1976)]

	RMSE			Correlation		
N	*a*	*b*	*c*	*a*	*b*	*c*
500	.472	.259	.077	.780	.989	.454
1000	.326	.209	.078	.908	.990	.492

As Table 4 shows, for a 50-item test (similar to the midquarter examinations used in this study) more accurate estimates of the parameters were generally obtained with the larger group of simulated testees. For example, the RMSE values for the final estimates of the a parameter were .472 for N=500 and .326 for N=1,000. The corresponding correlations were .780 and .908. The improved accuracy of estimation as N increased occurred for the b and c parameters as well. It should be noted, however, that for 50-item tests for the two sample sizes the b parameter is very accurately estimated regardless of sample size, the a parameter is fairly well estimated, and the c parameter is poorly estimated (r=.454 and .492).

Table 5 shows the results of the Gugel et al. simulation study corresponding to the maximum sample size used in the present study (N=1,000). The test lengths in Table 5 vary from 50 to 100 to reflect the lengths of the midquarter and final examinations used here. As Table 5 shows, for a fixed number of persons, increases in the number of items do not generally result in more accurate parameter estimates. For the b parameter, which is very accurately estimated with 1,000 cases, the accuracy improves from r=.990 to .996. The c parameter, which is poorly estimated at N=1,000, shows increases from r=.492 to .627. For the a parameter there is no clear trend in the correlations, with the highest accuracy at 50 items (r=.908) and the lowest at 60 items (r=.842). The results for the three parameters, using the RMSE criterion, show no clear trends either.

Table 5
RMSE and Correlation of Estimate and Parameter Values for Parameters a, b and c for a Sample Size of 1000 at Three Test Lengths
[From Gugel, Schmidt and Urry (1976)]

Number of Items	RMSE			Correlation		
	a	b	c	a	b	c
50	.326	.209	.078	.908	.990	.492
60	.322	.144	.062	.842	.995	.558
80	.261	.166	.073	.879	.993	.550
100	.240	.162	.062	.863	.996	.627

The results from Table 4, together with those from Table 5, show that with the numbers of testees and numbers of items used in this study, the b parameter (item difficulty) is very accurately estimated, while the a (discrimination) and c (guessing) parameters are less well estimated by this procedure.

Criteria for excluding items. Urry's item calibration program does not report ICC item parameters for an item if the calculated parameters meet any of the following criteria:

1. a less than .80
2. b less than -4.00 or greater than 4.00
3. c greater than .30.

These rejection criteria are applied to the items only in the first phase of the calibration procedure. The final parameters of the items that are not excluded in the first phase are allowed to vary unrestrained in the second

phase of calibration. Those items that were rejected in the first phase of the program were excluded from further analyses.

Results

Excluded items. Table 6 shows the number and percentage of items in each content area which did not meet the criteria specified by Urry's calibration program. Of the 394 unique (i.e., non-repeated) items in the pool, 85 (or 22%) met one or more of Urry's exclusionary criteria. The percentage of items lost by content area varied from 9% for content area 3 (energy) to 33% for content area 6 (ecology). Almost without exception, the items which were excluded by the calibration program had very low point-biserial correlations with total score. This indicates that most of the rejected items were excluded because of low estimates of the a parameter for these items.

Table 6
Number of Items Lost in the Calibration Process by Test and Content Area

	Content Area							
Test	1	2	3	4	5	6	7	Total
W1	8	5	2					15
S1	4	4	1					9
W2				5	6			11
S2	1			4	3			8
WF	1	2		2	1	4	4	14
SF	2	2		2	3	13	6	28
Total	16	13	3	13	13	17	10	85
Percent of Unique Items	30	22	9	13	27	33	21	22

Item pool characteristics. ICC item parameter estimates for all the items in the pool which survived the calibration procedure are shown in Appendix Table A, along with the sources from which they were taken. Table 7 shows the mean, standard deviation (S.D.), and range of values for each ICC parameter estimated for the items in each content area. The final line in Table 7 contains the same statistics, computed for the 309 items in the final pool.

As Table 7 shows, the mean discrimination (a) within content areas varied from 1.09 to 1.32. The lowest a values were .63 and the highest was 4.68. The difficulties within content areas were generally centered around zero, with the exception of content area 3, which had items of relatively high average difficulty ($\bar{b}$=.92). The item difficulties within content areas ranged from about -1.75 to about 2.50, with some differences among content areas. The c parameters for these four-choice items averaged between .24 and .34 and ranged from .00 to .65.

Table 7
Mean, Standard Deviation, and Range of Item Parameter Estimates by Content Area for Total Item Pool

Parameter and Statistic	Content Area 1	2	3	4	5	6	7	Total Item Pool
Number of Items	38	47	29	87	36	35	37	309
a (discrimination)								
Mean	1.20	1.23	1.32	1.17	1.26	1.09	1.16	1.20
S.D.	.35	.60	.80	.41	.60	.39	.36	.50
Low	2.40	3.54	4.68	3.66	3.88	2.03	2.22	4.68
High	.75	.67	.65	.63	.73	.63	.63	.63
b (difficulty)								
Mean	-.24	.06	.92	.17	.15	-.46	.13	.10
S.D.	1.03	1.26	1.06	1.15	1.18	1.29	1.28	1.22
Low	2.48	2.49	3.02	3.21	2.62	2.55	2.70	3.21
High	-1.76	-1.77	-1.56	-1.80	-1.74	-1.88	-1.69	-1.88
c (guessing)								
Mean	.28	.25	.34	.32	.32	.24	.29	.29
S.D.	.09	.09	.13	.12	.14	.11	.12	.12
Low	.51	.44	.60	.65	.64	.47	.58	.65
High	.14	.00	.00	.12	.06	.11	.11	.00

Urry (1977) has suggested the following guidelines, developed through a series of simulation studies (Urry, 1971, 1977), to assure that an adaptive testing item pool will improve the quality of ability measurement:

1. The *a* parameters of the items in the pool should exceed .80.
2. The *b* parameters of the items should be widely and evenly distributed from -2.00 to +2.00.
3. The *c* parameters of the items should be less than .30.
4. There should be at least 100 items in the pool.

As the data in Table A show, less than 12% of the items fell below .80 for the *a* parameter. Table 7 shows that the average estimate of the *a* parameter was above 1.00 for all content areas and 1.20 across all items in the pool. Thus, the vast majority of the items in this achievement test pool meet Urry's minimum criterion of a=.80.

The *b* parameter estimates in this pool show the wide range suggested in the guidelines, except for a slight deficiency of easy items. With the exception of content area 3 and, to some extent, content area 6, the mean values of *b* were near zero; and the standard deviations were over 1.0. For the total pool mean *b* was .10, and the range of *b*'s was -1.88 to 3.21.

The *c* parameter estimates averaged .29, narrowly meeting Urry's guidelines; the *c* parameters of 140 items failed to meet the .30 cutoff. This failure was probably caused in part by the inherent instability of the *c*

parameter estimates, in part by the use of four alternative multiple-choice items (in which a correct response could be achieved by random guessing with p=.25), and in part by the requirement that a student omit five items from each test. The total parameterized item pool consisted of 309 items drawn from an initial pool of 394 unique items.

Midquarter subpools. The total item pool described above was used for the creation of two smaller pools. One pool (MQ1) included all of the items from the first three content areas covered in the course; the other pool (MQ2) included all items from the fourth and fifth content areas covered. These two subpools were also evaluated using Urry's criteria for adaptive testing item pools.

Table 8
Distribution of a and c Parameters for Selected Ranges of the b Parameter for Items in Each of Two Midquarter Sub-Pools

	Range of b		No. of	a				c			
						Range				Range	
Pool	Low	High	Items	Mean	S.D.	Low	High	Mean	S.D.	Low	High
MQ1											
	-1.77	-1.50	8	1.20	.61	.79	2.67	.31	.13	.17	.56
	-1.49	-1.00	15	1.15	.41	.77	2.40	.27	.11	.14	.51
	-.99	-.50	15	1.23	.29	.80	1.81	.24	.08	.16	.41
	-.49	.00	15	1.32	.56	.65	2.31	.25	.08	.12	.39
	.01	.50	20	1.09	.29	.66	1.66	.27	.09	.13	.54
	.51	1.00	14	1.14	.30	.71	1.72	.33	.09	.12	.45
	1.01	1.50	9	1.76	1.18	.89	4.68	.35	.17	.00	.60
	1.51	2.00	6	1.32	1.10	.68	3.84	.25	.14	.00	.38
	2.01	3.02	12	1.28	.70	.67	2.77	.35	.09	.17	.52
Total	-1.77	3.02	114	1.24	.59	.65	4.68	.28	.11	.06	.60
MQ2											
	-1.80	-1.50	8	1.21	.31	.81	1.58	.33	.15	.21	.65
	-1.49	-1.00	13	1.17	.26	.79	1.53	.26	.16	.14	.64
	-.99	-.50	22	1.21	.27	.82	1.79	.27	.13	.13	.60
	-.49	.00	20	.95	.27	.63	1.53	.31	.12	.12	.53
	.01	.50	13	1.15	.23	.78	1.57	.33	.11	.12	.56
	.51	1.00	19	1.18	.33	.65	1.90	.31	.08	.19	.47
	1.01	1.50	13	1.04	.31	.68	1.69	.37	.08	.24	.48
	1.51	2.00	6	1.72	1.21	.89	3.88	.31	.16	.06	.53
	2.01	2.50	5	1.71	1.16	.81	3.36	.37	.11	.24	.52
	2.51	3.21	4	1.66	.54	.95	2.11	.52	.13	.39	.65
Total	-1.80	3.21	123	1.19	.47	.63	3.88	.32	.13	.06	.65

Table 8 shows the distributions of the three ICC parameters for the two testing pools. As the "Total" lines in Table 8 show, discrimination parameters (a) for the two pools varied from .65 to 4.68 for MQ1 (114 items) and from .63 to 3.88 for MQ2 (123 items) with means of a=1.24 and 1.19, respectively. In the MQ1 pool 13% of the items had a values less than .80; in the MQ2 pool only 11% were below this value. The b parameters were centered around 0.0 for each pool ($\bar{b}$=.18 and .16) and ranged from -1.77 to 3.02 for MQ1 and -1.80 to 3.21 for MQ2. Mean c parameters were .28 and .32, respectively.

Table 8 shows that, in accordance with Urry's recommendations, these pools had difficulties which were generally rectangularly distributed, at least in the range of b=-1.50 to +1.50. There was a lack of easy items in both pools (b<1.50), and the MQ2 pool had relatively fewer difficult items (b>1.50) than did the MQ1 pool. Table 8 also reveals a tendency for the higher difficulty items to also have higher discriminations. A positive correlation between item difficulties and discriminations was also reported in the context of ability measurement by McBride & Weiss (1974) and Lord (1975). There was no general tendency in these data for the c parameters to covary with difficulty level, with the exception that highest average values of c tended to occur for the most difficult items.

Similar to the total item pool, however, these subpools generally met Urry's recommendations for adaptive testing item pools. Each pool included more than 100 items, most items had discrimination values greater than .80, item difficulties were reasonably rectangularly distributed and wide-ranging, and typical c values were not unreasonably high.

Conclusions

It is apparent from these data that a three-parameter ICC model is applicable to college classroom achievement test items. Almost 80% of the items in the initial pool obtained parameter estimates in usable ranges. The resulting calibrated pool of items, as well as two subpools, met general recommendations for the construction of adaptive testing item pools in the ability testing domain. The subpools deviated somewhat from these criteria in terms of a lack of very easy and very difficult items, as well as in c parameters which were slightly higher than desirable. Whether these high c parameters are a result of unstable estimates, unique characteristics of the achievement testing pool, or the testing instructions is unknown. Further research in other achievement testing contexts will be necessary to answer this question.

Dimensionality of the Item Pool

Traditionally, the hypothesis that a single factor accounts for performance on a set of test items has been investigated by examining the dimensionality of the matrix of inter-item tetrachoric correlations by factor analytic methods (e.g., Indow & Samejima, 1966; McBride & Weiss, 1974; Prestwood & Weiss, 1977). However, factor analyses of such matrices will, on occasion, result in more than one factor when only one dimension is present in the data.

Bock and Lieberman (1970), for example, fitted a two-parameter normal ogive model to a unidimensional set of five test items. The fit of the model (and, therefore, unidimensionality) was tested by comparing the observed and predicted response frequency of every possible response vector. By this test the unidimensional model was found to fit very well. However, factor analysis of the inter-item tetrachoric correlation matrix rejected the hypothesis of a single factor.

Apparently, in the Bock and Lieberman data unidimensionality was not evident in the factor analysis because of problems introduced by computation of the tetrachoric correlation coefficient. Thus, in computing such a matrix, irregularities may be introduced which prevent unidimensionality from emerging, even if it is present in the data. In the present study, therefore,the factor analysis was supplemented by additional analyses to further examine the unidimensionality of the data.

Factor Analysis

Method. The factor analytic approach was used with two of the tests available: the first midquarter administered in winter (W1) and the second midquarter administered in spring (S2). The first step of the analysis was to compute a 55x55 matrix of inter-item correlations. The tetrachoric routine in the Statistical Package for the Social Sciences (SPSS; Nie, Hull, Jenkins, Steinbrenner, & Bent, 1970) was used. Since students were instructed to answer only 50 of the 55 questions, there was considerable non-systematic missing data. The program was instructed to compute a correlation between any two test items, excluding cases for which the responses to one or both items were missing (i.e., "pairwise deletion"). Since items were probably omitted on a non-random basis, an unknown amount of bias may have been introduced as a result of this procedure.

The resulting correlation matrices were factor analyzed by the principal axis method. The initial communality estimate for each item was chosen to be the largest off-diagonal correlation. These estimates were then iterated (with a limit of 25 iterations) until the difference between communality estimates on two successive iterations was negligible. The correlation matrices for the two tests with iterated communalities are shown in Appendix Table B.

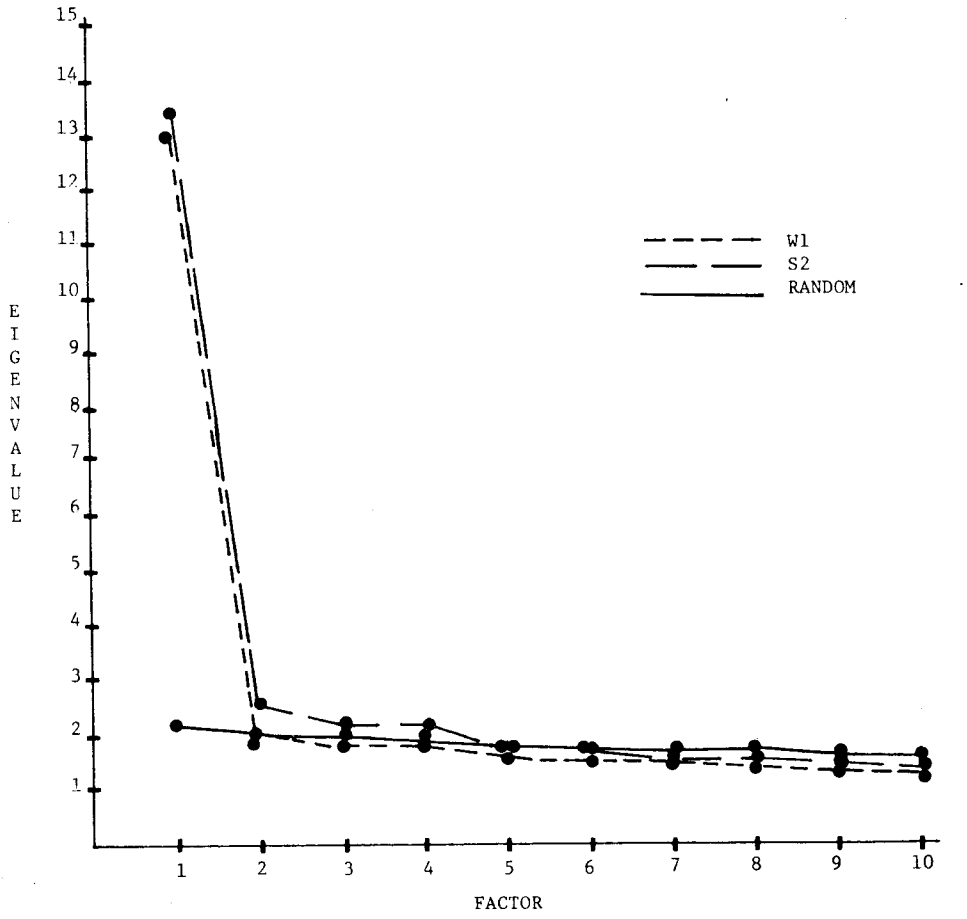
Following the procedures suggested by Horn (1965) and used by McBride and Weiss (1974) and Prestwood and Weiss (1977) to determine the number of factors in the real data matrix, a matrix of random data for 55 variables and 1,000 hypothetical testees was generated. These random data were intercorrelated and factor analyzed employing the same procedures as for the two real data matrices. The eigenvalues from the random data were used to compare with those of the real data in order to determine the number of factors in the real data.

Predictions about the factor structure to be obtained if the data are unidimensional can be made in a manner parallel to that used by McBride and Weiss (1974). In this instance, the predictions to be made are as follows:

1. The first factor extracted from each of the real data sets should be a general unipolar factor; the random data set should not exhibit this factor.
2. All factors, other than the first factor, from each of the real data sets should be of approximately equal magnitude and should be bipolar (that is, they should have as many negative loadings as positive loadings).
3. All factors extracted from the real data, except for the first factors, should be indistinguishable from the factors extracted from the random data.

Results. Figure 1 shows the factor contribution (eigenvalue) plots for the two sets of real data and the random data. From this figure it can be seen that both real data sets included a relatively strong first factor and that all of the remaining factors had low factor contributions restricted to a narrow range. It is also clear that the random data set lacked the strong first factor evident in the real data. Finally, all of the factors extracted from the real data, with the exception of the first factor, had factor contributions that were very similar in magnitude to the factor contributions of the factors extracted from the random data. The factor contribution data show that in the W1 data there was clearly one factor; in the W2 data there was a very strong first factor and a suggestion of two or three very weak secondary factors.

Figure 1
Eigenvalues for W1 data, S2 Data and Comparable Random Data



The first factor extracted from the W1 data accounted for 23.3% of the total variance in the 55 items with a factor contribution of 12.8; the first factor from the S2 data accounted for 24.4% of the total variance with a factor contribution of 13.4. No other factor extracted from either the real data or the random data accounted for more than 4.5% of the total variance of the test items.

Table 9 reports the factor loadings from each of the three data sets for the first four factors extracted from each matrix. The first factor obtained from each of the two real data sets had a large number of loadings which were higher than those in the random data; all these high loadings were unipolar. The first factor obtained from the random data was weak and bipolar. The second, third, and fourth factors obtained from all data sets were weak bipolar factors. Although the second factor from W1 had a factor contribution (1.96) indistinguishable from the corresponding factor (1.98) of the random data, it had two loadings which were higher in absolute value than those of the random data. Factor 2 from S2, which had a factor contribution (2.49) slightly higher than that of the random data (1.98), had three loadings greater than the highest in the random data. For factors 3 and 4 the factor contributions for the W1 data (1.81 and 1.75, respectively) were lower than for those of the random data (1.90 and 1.83); for the S2 data the corresponding factor contributions were higher (2.24 and 2.22). None of the loadings of the W1 factors 3 and 4 exceeded the highest loading in the random data, while two of the S2 loadings on factor 3 and one loading on factor 4 exceeded the corresponding random data loadings in absolute value.

These results suggest that factors 2, 3, and 4 from S2 and W1 are similar to factors of random data and , in all probability, represent trivial factors. In general, then, these results tend to support the existence of a single major factor in these achievement test data.

Equality of ICC's Based on Content Areas and Total Test

Rationale. In addition to implying that there is one factor in the item responses, the assumption of unidimensionality implies that ICC's will be linearly related across samples of items from the same domain of content. One way to examine this assumption is to compare the ICC's based on the total set of 55 items within a given midquarter with the ICC's computed within the content areas comprising that midquarter. If the total test measures a single dimension, parameterization of items within content areas should result in ICC parameters which are highly correlated with those obtained across all content areas. If this result is not found, it can be concluded that the content area is measuring a dimension which is not predominant in the total set of items and that the test items are not unidimensional.

A more stringent criterion for unidimensionality is that the item parameter estimates for items parameterized within a content area should be numerically the same as the parameter estimates obtained for those same items when all the content areas are calibrated together. This is equivalent to saying that the metric defined by items in a given content area is interchangeable with the metric defined by all the items. This criterion of unidimensionality implies that 1) the regression of the two sets of parameter estimates should be linear; 2) the slope of the regression line should be 1.0 within sampling error; and 3) the intercept of the regression line should be 0.0.

Method. Using Urry's ESTEM item calibration program, ICC item parameter estimates were computed within each content area for each of the four midquarter examinations. Item parameter estimates within content areas (shown

Table 9
Unrotated Factor Loadings for the First Four Factors of
W1 Data, S2 Data and Comparable Random (Ran) Data

	Factor 1			Factor 2			Factor 3			Factor 4		
Item	W1	S2	Ran	W1	S2	Ran	W1	S2	Ran	W1	S2	Ran
1	.27	.46	-.06	.13	.10	.07	-.09	.09	.05	-.09	-.04	-.06
2	.43	.43	.39	.12	.05	.02	.13	.07	.10	.02	.06	.00
3	.48	.37	-.28	-.40	-.08	-.09	.05	.04	.28	-.03	-.06	-.10
4	.50	.48	-.02	-.01	.05	.16	-.17	-.07	.17	-.03	-.10	.00
5	.43	.53	.14	-.36	.12	.20	.09	.08	-.18	-.17	-.23	-.14
6	.26	.59	-.11	-.15	.08	.08	.16	-.12	.13	.04	-.11	-.12
7	.58	.06	.00	-.02	-.09	.11	.11	-.12	-.01	.00	-.08	-.20
8	.58	.53	-.09	.08	.13	-.05	-.12	-.06	.01	-.03	.14	-.14
9	.51	.55	.06	-.07	.09	.07	-.18	-.12	.26	.12	-.42	-.03
10	.63	.61	.04	.02	.08	.04	-.23	.03	.19	-.11	-.70	-.01
11	.55	-.04	.08	.02	-.13	.03	.00	-.37	.03	.07	-.25	-.05
12	.55	.50	.00	.05	.23	.00	.05	-.04	.08	.16	-.14	.00
13	.54	.53	.12	-.02	.27	.20	-.17	.09	.16	-.23	.07	.00
14	.48	.17	.12	-.48	.18	.06	-.31	-.19	.12	.03	.10	.10
15	.22	.45	.13	.14	.17	-.12	-.02	-.04	.06	.04	-.08	-.02
16	.28	.47	-.16	-.01	.25	.05	-.08	.09	.17	.03	.11	.07
17	.47	.55	.24	.09	.32	-.01	-.03	-.04	-.09	.09	.04	.06
18	.66	.66	.06	.10	.27	-.18	.07	.11	-.03	.05	-.02	-.06
19	.58	.59	-.02	.08	.25	-.27	-.09	-.12	-.09	-.11	.03	.08
20	.28	.50	-.03	.10	.21	.00	.19	.04	.09	.16	.10	.17
21	.33	.51	-.15	-.03	.35	.09	-.13	-.21	-.02	.17	.07	.02
22	.41	.46	.04	.17	.27	.14	-.19	-.03	-.02	.10	.12	-.10
23	.41	.50	.06	.22	-.02	.25	-.01	-.01	-.01	-.16	-.18	-.18
24	.37	.49	-.06	.12	-.14	.01	.06	.03	.05	-.08	.07	.06
25	.38	.40	-.03	-.13	.00	.17	.07	.00	.11	-.13	.24	-.10
26	.54	.49	-.13	-.26	-.04	.08	-.17	-.08	.03	.29	.27	.02
27	.59	.15	-.30	-.14	.08	.14	.20	.11	-.08	-.20	.07	.26
28	.59	.46	.00	.04	.14	-.13	.19	-.04	.24	.21	-.11	-.11
29	.34	.35	.27	.15	.07	-.15	.22	.13	-.01	.08	.22	.34
30	.49	.62	-.04	.02	.03	-.02	.10	-.08	-.06	-.23	.03	.02
31	.50	.64	.02	-.08	-.07	.09	-.28	-.20	.09	-.15	.16	.14
32	.65	.32	.21	.05	-.02	-.10	.03	-.07	-.02	.17	.13	.06
33	.38	.34	-.18	.13	.10	-.19	-.12	.14	.08	-.24	.25	.06
34	.64	.64	.04	-.05	-.12	.16	.10	.14	-.03	-.15	.14	.01
35	.44	.63	.15	.22	-.09	-.13	-.19	.13	.15	.10	.03	.21
36	.34	.46	.15	.18	-.07	.11	-.08	.10	-.04	-.28	.18	.17
37	.66	.47	-.07	-.07	-.30	-.20	.08	.12	.06	.02	.24	-.02
38	.46	.47	.07	-.09	.08	-.10	.11	.14	-.03	.03	-.09	.07
39	.28	.19	-.09	.07	-.04	-.38	-.08	.01	.02	.02	-.09	.04
40	.49	.65	.12	-.06	-.13	.20	.44	-.04	-.01	-.12	.06	.04
41	.47	.55	.00	-.16	-.10	-.04	.02	.02	.19	-.05	-.10	.12
42	.30	.49	.04	.07	-.08	.11	.12	-.22	-.06	.07	-.16	.02
43	.49	.56	-.03	-.27	-.08	.08	.16	.08	-.04	-.17	.14	-.30
44	.63	.56	-.06	.16	-.54	-.12	-.03	-.65	.42	.13	-.08	-.27
45	.57	.32	-.04	.07	.13	.12	.00	.19	-.26	.10	.22	.04
46	.68	.37	.42	.13	-.05	-.08	.00	-.28	-.03	.04	.16	-.06
47	.32	.36	-.07	-.03	-.08	-.06	.06	.07	.03	.08	.03	.28
48	.27	.38	.21	-.17	-.02	-.01	-.23	.03	-.10	.25	-.14	-.18
49	.27	.32	.13	.02	-.06	-.34	.10	-.31	-.08	.22	-.18	-.29
50	.50	.53	.35	.11	.06	-.04	-.11	-.14	.15	-.20	-.16	.00
51	.08	.55	.02	.12	-.46	.28	.04	-.48	.23	-.02	-.16	.21
52	.40	.60	-.21	.20	-.36	.02	-.09	-.38	-.07	-.11	.00	-.05
53	.42	.59	-.17	.27	-.52	-.14	.06	-.37	-.14	.06	.07	.11
54	.52	.48	-.08	-.07	-.18	.03	.18	-.02	-.36	.10	.11	-.06
55	.37	.47	.07	-.03	-.12	.13	.04	-.06	.03	.26	.17	.08
Factor Contribution	12.84	13.44	2.11	1.96	2.49	1.98	1.81	2.24	1.90	1.75	2.22	1.83

in Appendix Table C) were then correlated with those determined earlier using all the items in each examination. Item parameter estimates for content area ICC's and total test ICC's were correlated for the a and b parameters separately and within each examination. The significance of linear and polynomial trends was also tested in these data using program BMD02V from the Biomedical Computer Program Package (Dixon, 1975). In addition, the slope and intercept of the regression lines were determined and tested for statistical significance. Because the c parameter was poorly estimated by Urry's program with the numbers of testees and items available in this study, these analyses were confined to the a and b parameters.

Results. Fifty-one items were rejected, using the criteria in Urry's calibration program. Approximately half were excluded by the program in both the total test calibration and the content area calibration. Only one item was excluded in the content area calibration that was not excluded in the total test calibration.

Table 10 shows the Pearson product-moment correlations of the a parameter estimates for the content areas and the total test. It also shows the significance levels of the first through fourth degree polynomials in the prediction of the a parameter estimates for items in each content area by the total test a parameters. Correlations varied from .18 to .95. These linear trends were statistically significant ($p \leq .05$) in 7 of 10 instances. As Table 10 and Appendix Table D show, non-linear quadratic trends were significant in only two instances; none of the cubic and quartic trends were statistically significant. In test S1 there was no significant relationship between the two sets of parameters for content area 3; it was the only content area which did not exhibit a significant trend in one of the two quarters.

Table 10

Product-Moment Correlations and Level of Significance for Polynomial Trends in the Prediction of Content Area a Parameter Estimates From Total Test a Parameter Estimates for Four Tests

Test	Content Area	No. of Items	r	Significance of Polynomial Trends			
				Linear	Quadratic	Cubic	Quartic
W1	1	13	.69	$p \leq .005$	NS*	NS	NS
	2	18	.77	.001	NS	NS	NS
	3	10	.24	NS	.05	NS	NS
S1	1	12	.43	NS	.05	NS	NS
	2	14	.72	.005	NS	NS	NS
	3	9	.18	NS	NS	NS	NS
W2	4	31	.93	.001	NS	NS	NS
	5	11	.86	.001	NS	NS	NS
S2	4	30	.95	.001	NS	NS	NS
	5	12	.74	.01	NS	NS	NS

* NS indicates that the polynomial was not statistically significant at the .05 level. Significance was determined by the use of an F-statistic. The sums of squares used for calculating the F-value are shown in Appendix Table D.

Table 11 shows the correlations and tests of polynomial trends for the b parameter. These correlations ranged from .86 to .99; all but two were .94 or above. Table 11 and Appendix Table E show that the linear trends for all 10 instances were significant at the $p \leq .001$ level. None of the non-linear trends were statistically significant.

Table 11
Product-Moment Correlations and Level of Significance for Polynomial Trends in the Prediction of Content Area b Parameter Estimates From Total Test b Parameter Estimates for Four Tests

Test	Content Area	No. of Items	r	Significance of Polynomial Trends			
				Linear	Quadratic	Cubic	Quartic
W1	1	13	.99	.001	NS*	NS	NS
	2	17	.94	.001	NS	NS	NS
	3	10	.95	.001	NS	NS	NS
S1	1	12	.98	.001	NS	NS	NS
	2	14	.99	.001	NS	NS	NS
	3	9	.91	.001	NS	NS	NS
W2	4	31	.97	.001	NS	NS	NS
	5	11	.98	.001	NS	NS	NS
S2	4	30	.99	.001	NS	NS	NS
	5	12	.86	.001	NS	NS	NS

*NS indicates that the polynomial was not statistically significant at the .05 level. Significance was determined by the use of an F-statistic. The sums of squares used for calculating the F-value are shown in Appendix Table E.

The data in Tables 10 and 11 show that the relationship between the ICC item parameters computed within content areas and those computed when the items were embedded within the total test were linear for the b parameter and primarily linear for the a parameter. The data from the spring quarter tests tended not to fit the predictions as well as that from the winter quarter tests, since there was no significant relationship in the a parameter data for content area S1. This is the same content area which also had one of the lowest correlations in the b parameter data.

Strong inferences concerning the unidimensionality assumption can be drawn from an examination of the slope and intercept of the regressions of the content area and total test ICC parameters. These data are shown in Table 12. The results for the slope of the a (discrimination) parameter were in accordance with the prediction of slope of 1.0 in only one instance. The intercept of the a parameter exceeded twice its standard error in only three of the ten instances.

For the b parameter, Table 12 shows that the slope of the regression line deviated significantly from its predicted value in content area 3 for W1 and S1 and content area 1 for W1; the remainder of the slopes did not

Table 12
Slopes and Intercepts and Their Standard Errors (S.E.) for the Bivariate Regression of Content Area Item Parameters and Total Test Item Parameters

Test and Content Area	No. of Items	Slope			Intercept		
		Slope	S.E.	Pred.[1]	Int.	S.E.	Pred.[2]
a (discrimination) Parameter							
W1							
1	13	.54	.17	N	.43	.30	Y
2	15	.56	.11	N	.14	.19	Y
3	8	.20	.22	N	.67	.40	Y
S1							
1	12	.13	.09	N	.83	.15	N
2	14	.77	.21	Y	-.16	.36	Y
3	7	.15	.23	N	.76	.47	Y
W2							
4	29	.82	.07	N	.12	.09	Y
5	19	.51	.10	N	.31	.17	Y
S2							
4	30	.37	.15	N	.63	.19	N
5	12	.22	.06	N	.66	.10	N
b (difficulty) Parameter							
W1							
1	13	.94	.03	N	.00	.03	Y
2	15	1.08	.06	Y	-.41	.09	N
3	8	.73	.08	N	.46	.13	N
S1							
1	12	1.03	.07	Y	-.16	.08	Y
2	14	.93	.04	Y	-.31	.06	N
3	7	.72	.12	N	.11	.20	Y
W2							
4	29	.97	.05	Y	-.07	.06	Y
5	19	.97	.06	Y	.01	.07	Y
S2							
4	30	1.05	.07	Y	.06	.07	Y
5	12	.77	.14	Y	-.21	.13	Y

[1] Y indicates that the value of the slope was as predicted, i.e., did not differ from the predicted value of 1.0 by more than twice its standard error; N otherwise.

[2] Y indicates that the value of the intercept was as predicted, i.e., did not differ from the predicted value of 0.0 by more than twice its standard error; N otherwise.

differ from 1.0 by more than twice their standard errors. The intercepts for the *b* parameter deviated significantly from zero for content areas 2 and 3 in W1 and content area 2 in W2. There were no deviations from the predicted values for either slope or intercept of the *b* parameters for the second examination (W2 or S2).

Conclusions

The factor analysis strongly supported the belief that only one real factor was present in each of the two tests analyzed. Every other factor fell at or near the level of the factors extracted by the same methods from random data and had loadings which were largely similar to those in the random data.

The analysis of the ICC parameters estimated in the context of the total test and individual content areas also lent credence to the hypothesis of unidimensionality. Although there were some deviations from predicted relationships, content area estimates were primarily linearly related to total test parameter estimates. The regression slopes and intercepts tended to follow the predicted patterns, particularly for the *b* parameter. For the *a* parameter the slope of the regression did not generally follow the predicted pattern, but the results were generally in accord with the predictions for the intercept of the regressions.

Thus, even though there were some deviations from strict unidimensionality, the two types of evidence indicate that the assumption of essential unidimensionality is valid.

Sampling Invariance of Item Parameter Estimates

According to Lord and Novick (1968, p. 380), ICC item parameter estimates determined in two subgroups are invariant if :

1. the regression of the *b* parameter estimates for two population subgroups is linear with a slope equal to $\sigma_1(\theta)/\sigma_2(\theta)$, where $\sigma_1(\theta)$ and $\sigma_2(\theta)$ are the standard deviations of θ in the two population subgroups, and the intercept is equal to the difference in the mean ability level between the two groups
2. the regression for the *a* parameter estimates is also linear and has a zero intercept, and the slope is equal to $\sigma_1(\theta)/\sigma_2(\theta)$.

Similar predictions could be made for the *c* parameter. However, similar to the previous analyses, these analyses of sampling invariance were confined to the *a* and *b* parameters and were not applied to the *c* parameter.

Method

In the two quarters used for item calibration, 46 items were administered to two different groups of students. Since these items were administered to different groups in the context of different tests, a comparison of the parameters obtained from the two calibrations of these items will serve as a strong

test of the invariance of the item parameters. If invariance is observed, it can be interpreted as additional evidence for the applicability of ICC theory in an achievement measurement setting.

Of the 46 items which had been administered to two groups of students, 25 items were used by the sampling invariance analysis. Items were included in the analysis if they had been administered at the same point in the course during both quarters (e.g., items administered at W1 and S1 or WF and SF were used, whereas an item administered at W1 and SF was not used).

For each item administered, item parameter estimates were obtained in each of the samples within the context of the calibration of the total set of items. Parameter estimates obtained from the second administrations were regressed on those obtained from the first administration; these regressions were tested for polynomial trends. In addition, the slopes and intercepts of the regression equations were compared with predicted values.

Table 13
Parameter Estimates for Items Used in Study of Sampling Invariance

Item Number	First Administration			Second Administration		
		Parameter			Parameter	
	Test	*a*	*b*	Test	*a*	*b*
3002	WF	.82	.13	SF	.87	.12
3034	W1	1.01	.37	S1	.85	-.29
3038	W1	1.58	-.56	S1	1.20	-1.06
3201	W1	1.07	-1.34	S1	.85	-1.74
3206	W1	.74	1.51	S1	.75	1.57
3216	W1	1.27	-.62	S1	1.17	-.60
3218	W1	.82	.58	S1	.80	.34
3229	W1			S1		
3237	WF	1.54	-.37	SF	1.58	-.11
3241	W1	1.12	2.48	S1	.91	2.09
3243	W1			S1		
3414	W1	.88	2.29	S1	1.40	1.96
3612	WF			SF	1.12	.75
3651	W2	.81	2.27	S2	.95	2.31
3812	W2	.74	-.66	S2	.82	-.63
3909	W2	1.34	.77	S2	.90	1.12
4005	WF			SF	1.28	2.76
4006	WF	.84	-.59	SF	1.05	-.19
4025	WF			SF		
4026	WF			SF		
4036	WF	1.24	-.61	SF	.95	-1.30
4044	WF	.80	-.12	SF	.80	-.60
4203	WF			SF		
4229	WF	1.36	-.45	SF	1.64	-.92
4238	WF	.83	1.54	SF	.83	1.47

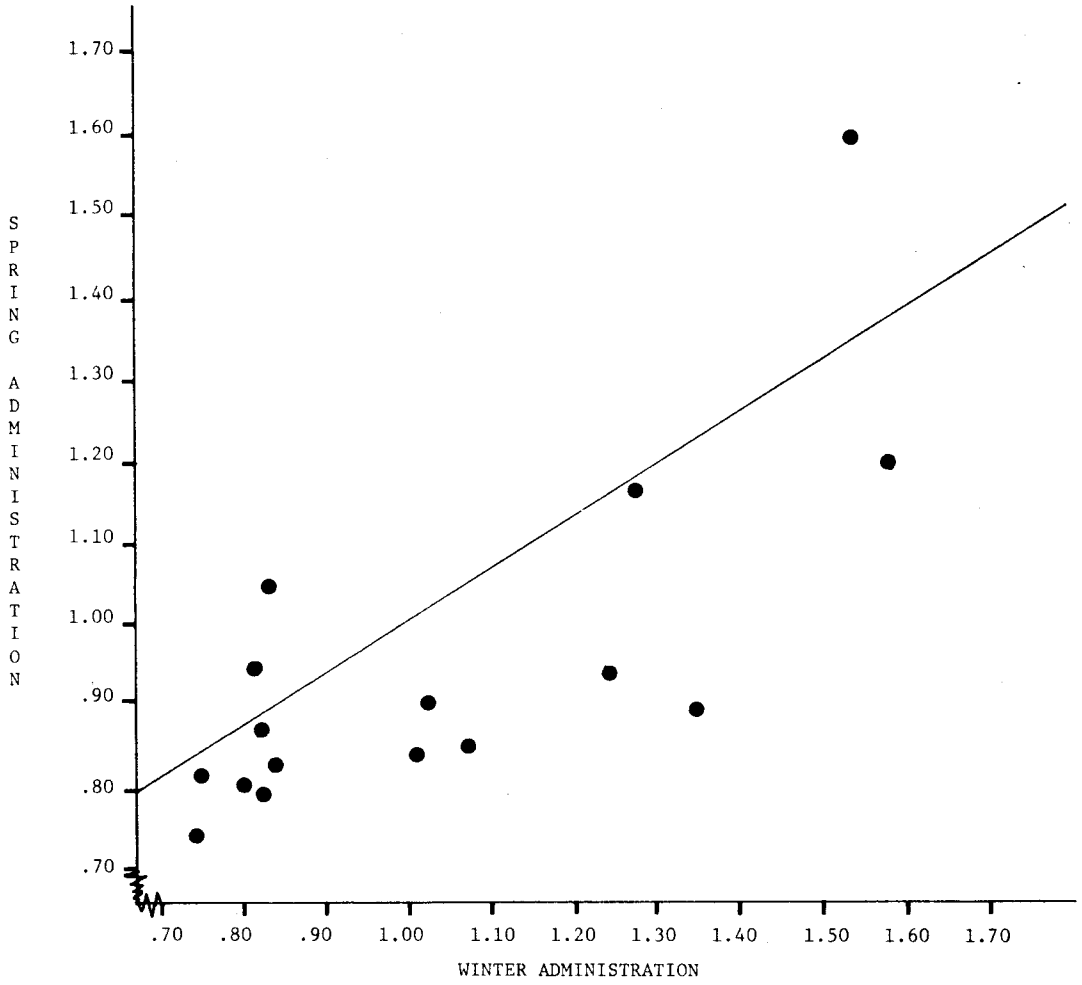
Note. Blank item parameters indicate that the item was rejected by the parameterization program.

Results

The items used in this phase of the analysis and their parameter estimates are shown in Table 13; these items had a fairly representative range of a and b values and included items from each content area. Of the 25 items available, seven were rejected by Urry's exclusionary criteria in one of the two groups. Five of these items were rejected at both calibrations.

Figure 2 shows a plot of the a parameter estimates obtained for the 18 items for which parameter estimates were available both quarters; results of the linearity test are in Table 14. As Figure 2 shows, the slope of the linear regression line was .61 with a standard error of .19. The predicted value of the slope of the linear regression was .97, based on the ratio of the standard deviations of the total test θ estimates obtained in the winter and spring quarter data. Thus, the slope did not deviate from its predicted value by more than twice its standard error. The intercept of the regression line was .38 with a standard error of .21; it, too, did not deviate from its predicted value (0.0) by more than twice its standard error.

Figure 2
Plot of a Parameters of Items Calibrated Twice



The data shown in Table 14 indicate that the regression of the two sets of parameter estimates was linear. The Pearson product-moment r of .63 was statistically significant at $p \leq .005$; none of the curvilinear trends was statistically significant.

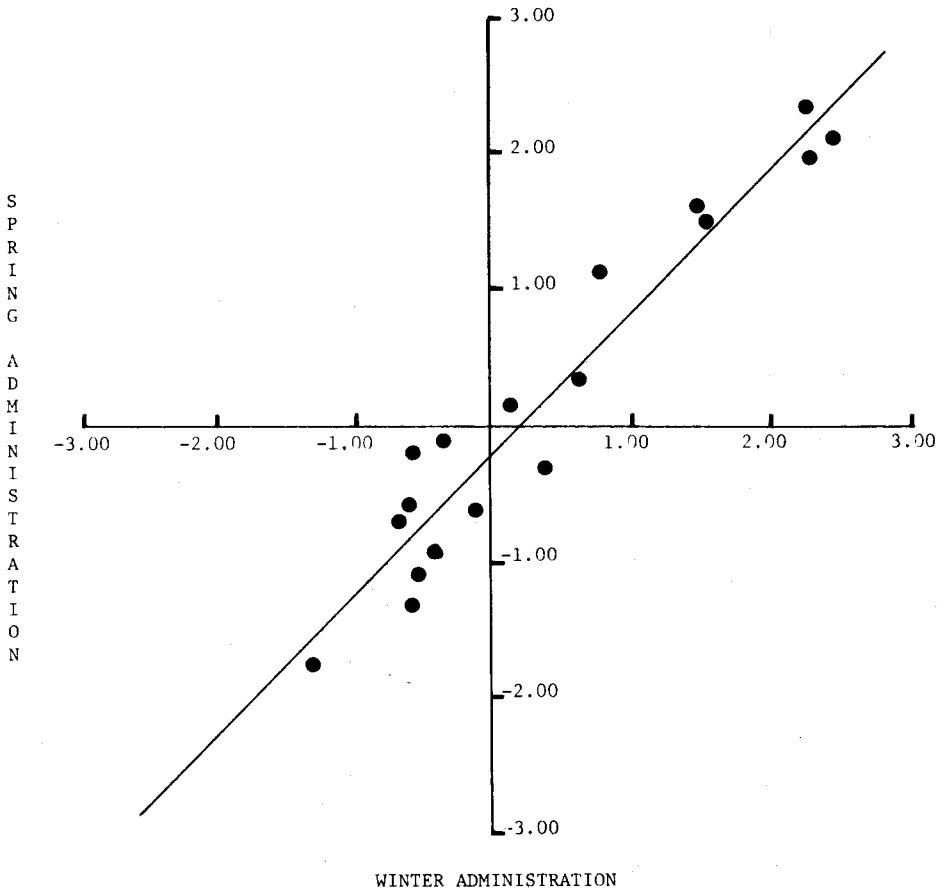
Table 14
Product-Moment Correlations and Level of Significance of the Contribution of Each Term of a Fourth Degree Polynomial Expression to the Prediction of the a and b Parameter Estimates Obtained During Spring Quarter Testing from Those Obtained During Winter Quarter Testing

Parameter	r	Significance of Polynomial			
		Linear	Quadratic	Cubic	Quartic
a	.63	.005	NS*	NS	NS
b	.96	.001	NS	NS	NS

*NS signifies that significance level of p=.05 was not attained.

Figure 3 shows the bivariate plot of the b parameter estimates for the data from the two quarters. The linear regression line fitted to these points had a slope of 1.02 with a standard error of .07. Thus, it did not differ from its predicted value of .97 by more than twice its standard error. The

Figure 3
Plot of b Parameters of Items Calibrated Twice



mean differences in θ estimates obtained from the winter and spring groups was -.09. The intercept of the regression in Figure 3 was -.18 with a standard error of .08. Thus, the observed slope for the *b* parameters did not differ from the predicted slope by more than twice its standard error.

As shown in Table 14, the linear correlation between the two sets of parameter estimates was .96, which was highly significant; none of the non-linear trends was statistically significant.

Conclusions

These results strongly support the invariance characteristics of the *a* and *b* ICC parameters across subgroups from the same population. Results for both parameters showed linear relationships between the parameter estimates derived in two samples of persons, when the items were in the context of different subsets of items in each sample. In addition, the results from the linear regression met the strong criteria of sampling invariance predicted by the ICC model. These results strongly support the application of the ICC *a* and *b* parameter estimates in an achievement testing context.

Conclusions

Answers can now be given to the questions which guided this research:

1. *Do achievement test item pools permit calibration by ICC models and result in an item pool suitable for adaptive testing?*

   Of the 394 unique items, 309 survived ICC calibration procedures to form a total pool of wide-ranging difficulty with moderate to high discriminations. Except for the high values of the *c* parameter, this pool met and exceeded reasonable standards set for an item pool for use in adaptive testing. The two midquarter examination subpools also were suitable for adaptive testing. The two pools contained 114 and 123 items with mean *a*-values of 1.24 and 1.19, respectively. Difficulty (*b*) parameter values were relatively rectangularly distributed in the range of -1.75 to about +1.75; items were also available with *b* values as high as 3.21. However, there was a lack of items in the very low difficulty range.

2. *Are responses to achievement test items reasonably unidimensional?*

   Both the factor analytic study and the study of item parameter estimates for content areas and the total test support the unidimensionality assumption. There was some indication that deviations from unidimensionality existed in the data, but they appeared to be minor compared to the major factor in the data.

3. *Do item parameter estimates remain invariant across samples?*

   Both the *a* and *b* parameters were consistently estimated across two samples. Both met strong criteria of invariance in terms of linearity

of the estimates and predicted values of the regression slopes and intercepts. These results are particularly meaningful, considering that the items studied appeared in the two tests in the context of other items which were not generally the same in both groups of students.

The primary results of these studies indicate that ICC theory can be applied to a classroom achievement test item pool. This is an extension of the application of ICC theory, which has been primarily limited to ability testing until now. If these results replicate in other areas of the achievement testing domain, it will be possible to link ICC theory with computerized adaptive test administration. This combination will yield a more thorough and efficient system for measuring achievement and for evaluating the effectiveness of training programs.

References

Bock, R. D., & Lieberman, M. Fitting a response model for N dichotomously scored items. Psychometrika, 1970, 35, 179-197.

Cliff, N. Complete orders from incomplete data: Interactive ordering and tailored testing. Psychological Bulletin, 1975, 82, 289-302.

Cliff, N. Incomplete orders and tailored testing. In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing. Washington DC: US Civil Service Commission, 1976, pp. 18-23.

Dixon, W. J. Biomedical computer programs. Los Angeles: University of California Press, 1975.

English, R. A., Reckase, M. D., & Patience, W. M. Application of tailored testing to achievement measurement. Behavior Research Methods and Instrumentation, 1977, 9, 158-161.

Fisher, R. A. Contributions to mathematical statistics. New York: Wiley, 1950.

Gugel, J. F., Schmidt, F. L., & Urry, V. W. Effectiveness of the ancillary correction procedure. In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing. Washington DC: US Civil Service Commission, 1976, pp. 103-106.

Gulliksen, H. Theory of mental tests. New York, NY: Wiley, 1950.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Horn, J. L. A rationale and test for the number of factors in factor analysis. Psychometrika, 1965, 30, 179-185.

Indow, T., & Samejima, F. On the results obtained by the absolute scaling model and the Lord model in the field of intelligence (Third report). Tokyo: Keia University, The Psychological Laboratory on the Hiyoshi Campus, 1966. (In English)

Lord, F. M. Individualized testing and item characteristic curve theory. In Krantz, Atkinson, Luce, & Suppes (Eds.), Contemporary developments in mathematical psychology. San Francisco, CA: W. H. Freeman, 1974.

Lord, F. M. The ability scale in item characteristic curve theory. Psychometrika, 1975, 40, 205-217.

Lord, F. M. A broad-range tailored test of verbal ability. In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing. Washington DC: US Civil Service Commission, 1976, pp. 75-78.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

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. (AD 781894)

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. Statistical Package for the Social Sciences. New York, NY: McGraw-Hill, 1970.

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. (AD A041084)

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No. 17, 1969.

Urry, V. W. A Monte Carlo investigation of logistic mental test models. (Doctoral dissertation, Purdue University, 1970). Dissertation Abstracts International, 1971, 31, 6319B (University Microfilms No. 71-9475).

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: US Civil Service Commission, 1976, pp. 97-102.

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Vale, C. D., & Weiss, D. J. A study of computer-administered stradaptive ability testing. (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975(a). (AD A018758)

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, 1975(b). (AD A020961)

Weiss, D. J. Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973. (AD 757788)

Weiss, D. J. Strategies of adaptive ability measurement. (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1974. (AD A004270)

Weiss, D. J. Final report: Computerized ability testing, 1972-1975. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976. (AD A024516)

APPENDIX: SUPPLEMENTARY TABLES

Table A
ICC *a*, *b* and *c* Item Parameter Estimates for Items in the Final Pool

Item Number	*a*	*b*	*c*	Test	Item Number	*a*	*b*	*c*	Test	Item Number	*a*	*b*	*c*	Test	Item Number	*a*	*b*	*c*	Test
3000	124	52	36	WF	3254	228	-17	27	SF	3646	119	82	33	W2	3913	131	-131	19	S2
3002	82	13	14	WF	3255	114	-72	26	SF	3647	79	1114	37	W2	3914	98	-39	16	S2
3003	96	-176	34	S1	3256	231	-33	26	SF	3648	159	-96	33	S2	3915	108	-61	16	S2
3005	143	11	39	S1	3257	98	-102	25	WF	3649	132	11	22	SF	3916	139	114	47	SF
3006	77	-37	33	SF	3258	124	81	36	WF	3651	95	231	52	S2	4001	147	-114	13	WF
3008	96	-175	18	S1	3259	69	-41	20	S1	3653	83	-51	33	SF	4002	78	-153	12	WF
3011	132	-86	20	W1	3260	71	84	28	S1	3654	151	84	21	W2	4003	70	-129	11	WF
3012	75	80	38	SF	3402	83	244	36	W1	3655	137	-90	60	WF	4004	139	-56	26	WF
3013	100	-97	39	S1	3403	99	18	19	W1	3656	63	-31	34	W2	4006	84	-59	16	WF
3014	86	-124	14	S1	3404	65	-29	35	WF	3657	81	-174	34	W2	4007	81	-150	42	WF
3017	99	-58	16	WF	3405	140	55	32	WF	3658	125	32	38	S2	4009	84	-54	31	WF
3018	89	125	45	S1	3406	131	248	52	SF	3659	137	67	29	S2	4010	88	-182	23	WF
3019	131	29	29	WF	3407	102	241	29	S1	3660	78	-39	14	S2	4011	90	-46	14	SF
3020	123	-128	17	S1	3408	251	105	31	SF	3661	190	68	32	WF	4012	125	-157	14	WF
3021	196	-49	21	WF	3409	468	128	00	S1	3662	154	93	27	WF	4013	176	-188	16	WF
3022	101	-48	30	SF	3410	130	134	31	W1	3663	69	-17	33	W2	4015	203	-162	12	WF
3023	240	-115	36	SF	3411	136	123	99	WF	3664	111	160	35	WF	4016	70	44	30	WF
3027	167	-138	40	SF	3412	112	19	54	SF	3665	119	54	22	W2	4019	105	-20	31	SF
3028	112	-126	51	SF	3413	140	76	37	S1	3666	68	141	30	S2	4020	91	-113	14	WF
3029	113	-150	28	WF	3414	88	229	32	W1	3668	97	-87	14	W2	4022	81	-174	13	WF
3031	147	-33	39	W1	3415	85	-96	41	W1	3669	81	227	42	W2	4027	136	-65	28	WF
3032	77	-106	27	W1	3417	267	302	56	SF	3670	80	111	35	W2	4028	63	-52	34	WF
3033	154	244	36	W1	3419	123	148	25	W1	3671	151	-14	26	W2	4029	191	-128	12	WF
3034	101	37	28	W1	3420	68	162	38	W1	3672	157	-80	15	W2	4030	115	-43	14	WF
3035	90	68	28	S2	3421	117	115	52	S2	3673	151	111	31	S2	4031	89	-110	15	SF
3036	92	-118	16	S1	3422	147	150	60	S2	3674	172	63	26	S2	4032	160	255	47	WF
3038	171	-93	21	WF	3423	66	16	27	WF	3675	121	40	28	W2	4033	90	223	38	SF
3039	112	12	34	WF	3425	136	17	23	S2	3676	89	151	25	SF	4036	95	-130	17	SF
3041	151	23	37	W1	3426	68	07	22	S2	3679	121	-94	17	S2	4037	145	137	42	SF
3042	115	37	27	W1	3427	92	151	26	W2	3680	133	-101	16	W2	4039	91	-112	12	WF
3044	87	-142	15	S1	3428	90	-156	40	W2	3681	103	154	36	SF	4042	66	-14	33	SF
3045	102	248	27	S1	3429	125	124	28	WF	3682	133	-72	34	WF	4043	187	245	39	WF
3046	118	24	22	W1	3430	115	-30	29	S2	3683	85	-131	15	W2	4044	80	-12	38	WF
3047	116	44	29	W1	3431	70	28	20	S2	3684	86	-85	14	S2	4046	127	-28	16	SF
3048	135	66	33	W1	3432	172	67	45	W2	3685	119	-101	16	W2	4047	82	-171	31	SF
3049	115	-71	18	W1	3433	135	86	30	S2	3686	126	-88	29	SF	4048	84	163	31	SF
3050	112	35	18	S1	3601	104	127	38	S2	3690	336	236	24	S2	4049	135	-158	23	SF
3051	129	21	28	S1	3602	109	-137	49	WF	3692	153	-128	36	SF	4050	86	197	36	SF
3201	107	-134	23	W1	3603	121	56	33	S2	3693	113	-24	24	SF	4051	84	-110	15	SF
3202	181	-99	21	W1	3605	122	57	34	W2	3695	109	-173	21	W2	4201	152	260	58	WF
3204	114	166	36	SF	3606	71	-22	14	WF	3696	68	-35	21	W2	4202	128	153	37	WF
3205	125	-153	19	S1	3607	138	09	37	S2	3697	156	321	65	W2	4204	104	75	41	WF
3206	74	151	21	W1	3608	104	-78	16	SF	3698	211	282	62	W2	4205	70	82	33	WF
3207	70	46	28	WF	3609	78	18	41	SF	3700	84	85	30	S2	4207	103	05	39	SF
3208	76	-16	12	WF	3610	80	-133	14	SF	3701	82	-15	42	S2	4208	63	-75	32	WF
3209	277	229	29	S1	3611	122	39	32	SF	3801	80	-17	45	S2	4209	100	71	41	WF
3210	104	-122	40	S1	3612	112	75	47	SF	3804	95	142	45	WF	4210	96	-64	14	SF
3211	88	01	13	W1	3613	86	-174	33	S2	3805	250	238	38	SF	4211	169	263	35	WF
3213	93	52	40	WF	3614	79	46	39	S2	3806	157	48	36	W2	4214	154	-101	20	SF
3214	112	03	23	S1	3615	169	117	29	W2	3807	152	-110	17	W2	4216	97	11	25	WF
3215	159	-82	23	WF	3616	86	62	25	W2	3808	99	-100	30	WF	4217	138	52	38	SF
3216	127	-62	18	W1	3617	79	-111	14	W2	3809	127	-61	53	SF	4218	102	67	24	SF
3217	106	-48	14	S1	3618	64	-05	35	WF	3810	92	220	27	W2	4219	118	269	36	WF
3218	82	58	12	W1	3620	204	297	65	W2	3811	115	22	56	SF	4220	105	-133	18	SF
3219	123	62	21	W1	3621	92	-09	33	W2	3812	82	-63	13	S2	4221	134	270	54	SF
3220	179	-03	26	WF	3622	95	253	42	SF	3813	120	-97	17	S2	4222	190	05	23	SF
3221	125	-52	17	WF	3623	133	-100	18	S2	3814	126	-32	38	WF	4223	101	-08	14	SF
3224	80	-50	27	S1	3624	80	-19	12	WF	3815	95	58	38	W2	4224	133	-66	27	SF
3226	109	-98	20	WF	3625	98	166	39	W2	3819	76	53	42	SF	4225	131	-59	26	SF
3228	67	249	31	W1	3626	65	52	25	WF	3820	92	38	12	S2	4226	79	-107	11	SF
3230	90	87	41	WF	3627	103	107	48	SF	3831	90	-92	43	S2	4227	119	59	41	WF
3234	354	173	00	W1	3628	98	51	27	W1	3823	100	-07	53	WF	4228	222	105	38	WF
3235	115	-140	28	S1	3629	111	-03	37	W1	3825	109	-138	34	SF	4229	164	-92	17	SF
3236	126	-120	33	SF	3630	78	-24	43	S1	3827	87	135	46	W2	4230	99	-152	13	WF
3237	154	-37	18	WF	3631	153	-18	38	S1	3821	388	196	06	WF	4231	87	-169	20	SF
3238	82	-106	21	S1	3632	123	27	37	S1	3832	99	-174	32	S2	4234	137	-23	39	SF
3239	104	-113	21	WF	3633	94	-08	40	S1	3901	155	262	39	WF	4235	86	95	20	WF
3240	98	-28	15	W1	3634	179	-58	30	WF	3902	73	149	29	W2	4237	65	04	36	WF
3241	91	209	17	S1	3635	117	66	44	S1	3903	121	-43	31	W2	4238	83	147	43	SF
3242	94	240	41	SF	3636	124	-63	27	SF	3904	345	158	28	SF	4239	82	-142	11	WF
3244	135	-44	23	S1	3637	129	-73	28	S1	3905	98	35	20	W2	4240	154	-01	35	WF
3245	134	-96	21	W1	3638	135	-154	21	S2	3906	87	-66	14	S2	4242	100	-65	13	WF
3246	110	-72	28	SF	3639	147	-180	40	W2	3907	143	-108	64	SF	4243	91	-153	18	SF
3247	82	242	43	S1	3640	143	-69	39	S2	3908	115	07	31	W2	4244	73	-77	17	SF
3249	91	-169	17	S1	3641	120	-65	22	S2	3909	134	77	38	W2	4245	130	-158	22	SF
3250	91	194	29	W1	3642	111	111	24	WF	3910	158	-159	21	W2	4246	140	143	45	SF
3251	260	239	44	SF	3643	140	-50	25	W2	3912	95	70	19	S2					
3252	79	-177	35	S1	3644	88	125	40	SF										

Note. Two decimal places assumed throughout.

Table B

Tetrachoric Correlation Matrices for Items in W1 and S2 Tests

Item	Item 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	h^2 W1	h^2 S2
1		26	-00	08	23	06	21	16	14	26	10	16	13	13	17	04	03	25	19	09	-04	17	05	15	11	10	14	11	09	19	25	15	12	13	20	18	11	04	09	05	07	02	05	20	14	27	07	02	04	14	-02	15	11	14	15	27	40
2	23		19	20	15	16	22	18	20	25	33	19	20	04	23	12	20	38	29	22	05	16	28	18	21	20	26	38	11	12	25	22	11	28	19	16	29	26	17	22	14	31	24	22	28	32	16	09	12	15	05	12	14	14	17	38	39
3	26	24		21	31	13	36	19	16	34	38	19	33	47	-02	09	14	33	30	20	29	16	14	31	19	46	30	25	16	15	29	30	07	40	11	09	23	27	01	38	32	18	31	21	21	23	01	10	05	23	-12	10	10	25	16	47	36
4	14	15	20		18	07	33	38	32	40	25	26	31	34	16	09	21	45	33	13	25	22	15	13	15	28	35	30	09	19	30	27	19	21	34	22	38	13	07	20	26	13	33	31	30	19	10	12	15	27	12	22	22	18	13	45	42
5	23	20	12	27		18	33	22	17	27	13	34	18	46	08	08	12	19	20	04	08	19	19	13	19	28	36	38	07	22	21	24	11	32	-02	16	41	16	07	27	22	19	39	21	14	22	16	04	04	25	-02	16	-01	20	13	46	48
6	40	24	26	42	41		23	19	15	13	13	03	09	19	06	07	16	24	15	02	12	08	04	-01	10	12	18	12	06	14	10	19	12	08	04	02	15	07	10	29	23	18	19	07	11	15	15	07	11	04	-01	03	15	27	16	29	43
7	-07	11	-03	-03	08	16		37	34	33	32	26	33	06	09	15	22	40	29	19	31	29	18	21	24	34	34	38	24	34	26	31	22	38	26	09	35	20	14	33	25	23	36	33	34	35	19	09	12	30	-10	28	25	30	28	40	22
8	22	20	13	19	15	40	-03		39	50	31	33	37	29	04	15	26	43	34	12	24	26	22	26	25	24	32	31	21	17	26	32	22	32	34	21	35	15	12	34	24	14	20	31	35	41	20	17	04	32	07	25	31	32	16	50	46
9	24	19	36	32	36	26	08	27		53	24	38	28	30	08	19	27	33	27	12	24	26	01	12	13	26	23	25	08	25	23	32	22	25	19	15	35	24	15	18	21	14	34	37	28	26	25	29	09	22	-01	19	20	31	21	53	74
10	26	32	28	40	49	35	09	24	74		36	31	39	28	07	23	31	40	37	08	20	32	27	21	24	36	31	31	17	31	32	34	30	43	37	32	34	19	12	21	27	22	35	38	24	37	21	13	18	35	05	28	20	26	23	53	74
11	04	00	00	02	07	-02	-12	-05	06	06		36	27	41	19	20	19	33	29	20	18	32	22	19	22	17	30	32	26	18	30	35	17	39	30	14	29	18	19	33	20	18	26	32	27	33	21	21	15	30	05	25	33	36	17	40	43
12	33	12	04	27	37	37	08	46	38	43	-07		21	25	13	10	36	34	26	23	17	25	25	22	19	28	34	47	19	24	20	46	20	36	21	16	34	29	22	25	24	15	17	38	31	33	10	14	19	17	06	21	24	28	26	47	46
13	30	17	13	26	37	37	01	31	28	14	-04	42		27	12	17	30	27	38	03	08	18	28	23	22	28	32	29	10	34	42	30	28	37	22	27	34	30	18	23	34	08	17	27	28	32	30	14	11	35	-03	26	21	23	18	42	44
14	-05	-04	11	06	12	04	01	13	10	03	-15	21	07		22	24	27	19	30	05	29	03	10	-02	25	33	35	14	05	08	42	26	20	28	23	-02	37	21	22	17	30	-02	17	30	24	32	09	23	14	25	06	13	02	28	15	47	24
15	10	17	18	16	29	27	06	35	29	33	07	28	39	04		06	01	14	20	09	05	18	06	08	-11	04	06	22	05	14	13	15	06	10	13	11	27	04	04	14	12	18	02	20	07	15	10	04	04	12	10	15	18	08	03	27	41
16	24	27	10	11	37	32	-08	38	17	19	-09	29	37	08	20		02	24	17	04	05	20	08	11	09	19	07	08	16	11	14	16	20	07	07	07	24	14	13	14	09	05	21	28	21	18	07	03	14	26	02	09	05	14	21	28	51
17	31	13	18	31	36	39	07	30	31	33	-05	36	40	20	26	36		39	29	23	21	22	18	18	19	14	25	28	14	20	20	27	22	36	23	13	34	24	18	16	20	14	16	37	33	29	13	15	22	24	06	21	19	23	18	39	43
18	40	37	14	30	34	26	07	37	43	51	-11	41	44	18	33	51	39		45	21	23	32	28	18	23	40	41	38	24	31	30	47	23	39	18	25	41	23	19	39	30	20	28	51	38	43	12	08	25	25	13	31	20	31	23	51	51
19	25	25	21	27	22	37	00	36	31	33	-06	29	31	16	36	40	42	46		15	24	24	32	13	24	35	28	25	12	27	23	42	26	41	23	26	42	25	18	25	25	17	27	41	25	23	10	04	15	39	02	34	27	23	16	45	46
20	33	30	13	26	32	26	-08	42	20	19	-01	42	43	17	22	25	24	41	33		01	09	05	14	01	17	11	21	13	12	07	24	07	16	06	06	19	12	06	12	07	10	20	29	15	20	12	-06	11	11	14	12	14	20	16	29	43
21	25	22	20	24	28	33	00	29	37	31	-19	37	37	24	41	27	34	30	39	32		17	15	16	08	03	14	22	11	05	14	26	11	14	19	-01	19	13	11	05	21	16	06	29	25	15	14	15	09	10	-09	18	18	10	09	31	42
22	25	29	15	20	30	34	-04	21	25	23	-06	09	25	18	24	20	37	39	32	23	42		37	08	15	33	13	18	13	10	23	28	22	21	20	12	24	15	12	02	14	23	10	30	27	33	14	14	05	22	01	18	15	15	18	37	42
23	19	22	22	27	21	34	07	41	36	35	07	30	18	12	24	22	32	26	37	25	22	19		22	06	04	27	24	14	23	17	25	23	23	17	25	30	26	25	19	21	11	19	26	21	29	04	12	18	25	06	27	20	08	09	37	39
24	16	31	17	26	27	27	-10	26	31	24	-07	10	18	04	19	20	18	33	32	14	20	17	26		23	18	21	21	22	12	23	26	25	25	18	14	26	21	13	20	15	10	11	20	20	32	11	11	03	27	05	12	23	24	04	32	38
25	19	20	16	17	17	14	07	24	17	04	-09	14	29	14	21	12	27	15	25	23	28	23	22	30		25	33	14	16	19	20	26	22	30	10	06	26	24	07	23	18	04	26	20	24	25	14	02	-04	15	09	15	14	26	17	33	33
26	19	16	17	27	19	26	-01	30	17	03	-06	24	33	11	23	27	21	30	26	25	20	24	25	29	28		16	39	09	19	36	42	10	35	19	10	39	36	18	08	17	10	33	27	31	49	14	45	20	14	11	19	23	23	30	49	40
27	20	19	16	08	-01	-03	-05	15	01	12	03	09	13	-03	-02	02	14	16	07	20	06	17	10	-03	04	21		35	22	34	29	35	19	45	18	24	40	25	13	38	35	19	43	38	39	38	16	11	24	26	02	23	19	31	13	45	27
28	15	33	16	21	18	41	-01	30	38	43	05	32	31	00	31	26	33	35	38	29	21	16	29	15	12	10	-27		22	26	25	49	13	33	23	19	46	36	15	25	28	19	24	40	35	40	27	11	17	17	15	19	26	31	49	43
29	19	20	11	20	20	07	05	21	08	13	-11	18	15	17	16	15	30	34	15	20	32	13	17	19	20	17	11	28		19	07	21	22	19	17	-07	27	15	05	32	16	16	04	31	18	29	17	01	12	19	11	19	16	20	20	32	38
30	23	27	14	33	33	41	-06	40	30	35	-01	32	41	04	29	16	43	36	25	29	32	35	34	34	21	29	10	37	12		31	40	26	22	14	23	26	07	22	18	20	26	02	40	39	19	30	13	14	03	31	04	15	22	27	11	40	60
31	21	21	23	32	33	29	05	41	26	22	-03	29	34	14	32	28	37	36	36	31	40	27	39	29	27	40	06	22	25	60		29	29	27	27	26	35	16	18	10	25	14	23	29	30	38	17	15	05	26	01	21	13	21	10	42	60
32	24	22	07	12	22	23	05	13	06	11	-09	14	18	-12	05	13	24	14	17	13	22	21	18	15	18	25	-01	10	20	32	31		19	45	37	18	52	30	21	24	26	10	23	44	35	47	20	17	23	34	13	27	20	38	31	52	32
33	19	28	21	14	15	11	-02	25	16	-02	-01	09	20	09	10	15	13	19	35	29	16	14	06	16	16	15	08	24	23	22	18	11		35	14	16	25	15	17	09	12	05	24	27	18	23	04	-01	-05	22	10	20	20	12	35	35
34	32	39	26	26	22	29	01	34	30	32	03	14	24	13	24	25	34	28	35	29	25	33	30	35	29	39	14	32	29	41	48	20	27		27	21	40	29	19	42	30	14	35	35	41	48	19	15	09	28	09	23	28	30	19	48	50
35	33	14	29	33	31	43	-02	23	26	37	06	25	28	10	27	32	35	40	29	29	27	30	39	35	19	33	04	26	20	47	40	18	32	50		20	19	11	13	28	17	09	12	24	29	31	09	28	11	23	-01	21	44	23	23	44	55
36	12	26	18	24	17	30	09	25	27	18	-09	20	24	07	13	29	27	39	22	24	19	09	12	33	25	29	06	22	19	23	24	21	26	34	32		13	20	12	17	17	12	20	10	22	24	07	09	04	33	12	20	18	15	09	33	39
37	11	20	11	24	10	28	08	30	14	07	03	23	12	02	26	16	27	21	25	21	08	10	08	35	21	35	02	14	22	24	29	22	14	35	29	37		26	11	36	29	22	36	37	29	44	27	15	23	42	02	19	27	36	29	52	58
38	25	26	18	21	31	30	02	22	17	33	02	19	32	07	28	26	28	37	40	27	23	16	26	23	16	09	13	08	13	23	26	07	22	31	33	18	20		22	18	30	13	29	24	29	35	16	16	11	15	07	11	20	37	18	40	40
39	01	06	02	16	11	04	-01	07	11	27	01	-11	-01	-05	12	05	00	09	20	02	17	14	20	13	07	14	06	02	-03	09	13	08	00	15	11	10	14	21		13	09	06	07	23	12	12	07	13	10	18	-03	14	16	15	10	25	27
40	35	27	29	40	24	33	-03	31	33	40	-01	27	31	07	18	22	30	35	37	26	35	26	31	29	33	38	14	15	22	44	46	24	22	45	37	26	28	29	13		20	18	31	30	26	34	17	02	19	22	-01	06	28	36	11	42	50
41	25	09	26	30	30	39	01	27	33	42	-07	12	31	05	23	20	19	33	39	19	25	22	35	29	13	29	19	06	15	33	37	16	13	29	39	23	34	36	22	39		17	30	27	26	31	20	24	11	20	01	19	15	26	18	35	44
42	27	10	16	21	38	31	03	21	26	41	-01	15	33	14	18	19	26	29	23	21	11	22	29	30	18	17	14	09	01	22	18	13	07	45	36	13	18	31	26	38	38		14	25	08	22	09	09	07	17	00	10	19	18	11	31	50
43	22	20	20	27	25	35	-02	27	16	21	-08	26	17	19	12	30	37	44	25	19	24	24	29	19	26	26	19	28	26	34	35	10	27	42	39	25	27	23	12	50	33	41		15	27	26	26	16	12	20	03	16	18	29	17	43	50
44	21	15	32	10	37	15	03	22	15	39	43	31	32	-22	15	13	04	29	03	24	01	08	33	30	25	13	03	23	38	31	31	15	17	47	55	33	60	31	09	38	38	50	44		34	54	18	20	19	41	06	26	21	28	27	54	58
45	17	07	10	13	17	06	08	17	08	13	-10	07	30	16	12	19	16	26	22	19	20	33	04	15	25	20	13	03	21	16	14	02	17	25	20	10	24	10	08	32	12	30	21	14		49	30	21	23	31	00	19	36	33	17	49	46
46	13	22	21	18	13	18	01	16	14	12	05	14	16	03	14	20	20	28	17	20	20	25	08	12	18	09	-00	19	18	23	14	05	22	20	26	31	23	15	04	20	16	10	22	48	46		09	18	20	38	12	35	29	38	18	54	48
47	09	17	15	12	19	15	-00	08	18	13	-04	10	19	-05	27	19	23	20	31	06	12	14	21	24	08	18	-03	12	00	24	15	20	15	29	20	28	21	25	01	18	30	19	17	20	17	34		16	09	10	02	19	13	16	23	30	37
48	23	04	06	20	26	21	04	19	29	44	05	11	14	11	21	14	14	26	27	13	11	13	18	17	18	22	-03	18	06	27	29	26	17	32	19	23	15	25	15	26	18	22	-00	25	13	14	37		08	09	-07	-07	06	15	13	45	44
49	27	29	12	13	29	21	08	08	25	27	24	27	17	01	11	30	14	32	12	12	01	-04	26	26	14	22	14	06	09	10	04	-10	07	19	04	15	21	17	-06	16	08	24	20	39	10	22	19	08		16	02	02	13	21	19	25	39
50	23	22	21	30	37	37	07	22	35	46	-09	19	20	07	28	17	34	35	37	28	27	26	25	21	15	31	08	18	08	39	34	13	11	24	32	11	20	35	08	33	44	20	34	17	23	10	17	21	14		02	21	16	23	15	42	46
51	26	11	23	24	29	38	22	24	32	37	-13	28	17	15	20	01	14	21	30	23	21	11	30	24	18	23	09	19	06	33	43	16	01	27	35	18	24	17	-00	43	27	27	22	28	02	07	21	20	25	35		09	-04	06	03	15	84
52	23	31	21	22	26	29	11	29	42	29	-06	25	17	19	26	20	25	29	30	25	26	19	29	28	25	38	-03	28	10	39	48	19	11	36	34	24	40	18	20	46	30	22	36	32	04	16	19	18	07	37	70		22	22	05	35	70
53	23	25	27	31	19	38	15	25	31	23	-13	19	20	10	17	25	16	23	27	20	18	13	23	38	23	29	00	26	16	33	46	18	09	35	28	26	36	17	08	49	34	18	35	41	09	26	29	23	07	26	83	69		30	10	44	84
54	21	23	19	18	18	27	06	32	23	29	-03	23	23	-03	14	21	12	31	19	23	14	23	25	31	20	28	07	23	23	25	32	15	21	38	39	27	33	19	-03	31	31	18	29	28	18	18	09	04	12	26	34	37	43		24	38	43
55	13	24	04	14	30	21	-06	32	17	24	-11	16	15	00	18	29	23	30	22	14	23	27	22	34	23	38	03	21	22	28	41	16	22	29	30	26	31	23	13	32	16	21	32	27	06	20	21	14	08	18	42	30	30	28		31	42

Note. Correlations above the diagonal are from W1. Correlations below the diagonal are from S2. Communalities (h^2) are final iterated values.

Table C

ICC *a*, *b* and *c* Item Parameter Estimates Within Content Areas for Items from W1, W2, S1 and S2

Test	Content Area	Item No.	Parameter Estimate *a*	*b*	*c*
W1	1				
		3033	2.38	2.66	.63
		3032	1.29	-1.04	.35
		3046	1.39	.20	.26
		3031	1.59	-.51	.40
		3041	2.04	.16	.38
		3042	1.47	.37	.31
		3047	2.11	.39	.31
		3000	1.76	.87	.37
		3048	1.32	.68	.36
		3049	1.39	-.67	.30
		3034	1.45	.62	.41
		3038	2.28	-.66	.32
		3011	1.88	-.78	.35
W1	2				
		3234	2.53	3.01	.59
		3250	.99	2.36	.41
		3206	1.07	1.81	.38
		3216	1.96	-.23	.39
		3218	1.24	1.19	.40
		3230	1.47	2.46	.62
		3238	1.39	-.76	.30
		3236	1.35	.03	.46
		3219	2.23	.96	.43
		3240	1.54	.39	.46
		3241	2.18	2.62	.48
		3202	2.15	-.66	.44
		3201	1.74	-.94	.36
		3228	1.27	2.78	.54
		3211	1.48	.63	.43
		3227	.97	-1.04	.37
		3245	2.07	-.80	.32
W1	3				
		3419	2.20	1.49	.42
		3403	2.77	.06	.29
		3425	4.03	-.49	0.00
		3414	1.64	1.98	.47
		3402	1.68	2.17	.55
		3420	.94	1.10	.43
		3423	1.26	.27	.38
		3415	4.13	-2.27	.12
		3410	2.00	1.36	.48
		3406	1.47	2.35	.53
W2	4				
		3657	.87	-1.67	.38
		3783	.94	-1.22	.18
		3641	1.32	-.74	.17
		3663	.72	-.10	.36
		3642	1.06	1.17	.26
		3640	1.42	-.67	.40
		3624	.86	-.14	.14
		3685	1.25	-.98	.18
		3606	.77	-.27	.13
		3671	1.49	-.23	.22
		3665	1.45	.62	.28
		3669	1.05	2.27	.50
		3675	1.30	.48	.33
		3680	1.59	-.85	.21
		3695	1.23	-1.31	.32
		3694	0.00	0.00	0.00
		3696	.83	-.51	.14
		3610	.98	-1.10	.17
		3654	1.98	.83	.26
		3620	1.92	2.83	.66
		3668	1.16	-.74	.17
		3672	1.88	-.72	.18
		3630	.68	-.52	.38
		3697	1.98	2.53	.63
		3698	2.27	2.45	.60
		3609	.89	.18	.43
		3646	1.28	.89	.37
		3656	.67	-.40	.32
		3613	1.31	-1.62	.27
		3602	1.15	-1.29	.54
		3670	.78	1.03	.33
W2	5				
		3827	.93	.98	.44
		3810	1.61	2.33	.52
		3806	2.28	.30	.34
		3815	1.47	.56	.44
		3807	3.01	-1.04	.18
		3910	2.47	-1.47	.43
		3902	.92	1.74	.40
		3903	1.31	-.62	.29
		3905	1.69	.71	.41
		3909	1.79	.76	.43
		3908	1.69	.14	.39
S1	1				
		2014	1.35	-.85	.24
		3051	2.21	.27	.35
		3044	1.13	-1.19	.23
		3005	1.65	-.13	.33
		3050	2.28	.69	.39
		3045	3.00	2.70	.65
		3003	1.47	-1.66	.32
		3020	1.61	-1.09	.27
		3013	1.39	-1.12	.29
		3036	1.31	-.51	.42
		3008	1.40	-1.34	.26
		3018	.80	1.02	.42
S1	2				
		3249	1.28	-1.31	.26
		3247	2.33	2.37	.74
		3241	1.51	2.74	.50
		3244	1.79	-.28	.32
		3205	1.72	-1.29	.28
		3214	1.94	.28	.37
		3209	2.63	2.93	.72
		3210	1.74	-1.31	.28
		3235	1.97	-1.24	.26
		3217	1.66	.01	.35
		3252	1.00	-1.64	.40
		3259	.95	.28	.42
		3260	1.24	1.33	.51
		3224	1.14	-.07	.42
S1	3				
		3431	1.34	.03	.39
		3428	3.35	-1.46	.58
		3433	1.57	.66	.41
		3413	2.22	.60	.52
		3409	2.46	2.91	.71
		3429	2.85	.92	.33
		3426	2.28	-.05	.49
		3427	1.97	1.70	.51
		3407	1.43	2.88	.56
S2	4				
		3618	.98	.13	.41
		3684	1.05	-.75	.22
		3608	1.30	-.66	.24
		3648	1.89	-1.08	.32
		3632	1.39	.16	.36
		3638	1.70	-1.42	.34
		3603	1.29	.41	.28
		3611	1.34	.26	.29
		3700	.99	.84	.33
		3701	.77	-.61	.34
		3673	1.34	1.01	.28
		3631	1.73	-.86	.28
		3623	1.54	.74	.32
		3628	1.17	.46	.27
		3615	1.74	1.12	.30
		3660	.99	-.34	.20
		3614	.66	.33	.38
		3617	.99	-.99	.23
		3601	1.08	1.30	.41
		3658	1.31	.36	.40
		3690	3.31	2.38	.32
		3651	1.23	2.09	.56
		3659	1.37	.62	.29
		3674	1.66	.66	.28
		3622	1.08	2.48	.46
		3625	.79	1.53	.33
		3605	1.23	.30	.30
		3679	1.42	-.89	.27
		3666	.70	1.21	.27
		3626	.66	.94	.36
S2	5				
		3812	.95	-.25	.28
		3820	1.30	.52	.26
		3813	1.47	-.87	.27
		3832	1.75	-1.51	.38
		3814	1.47	-.67	.33
		3801	.92	-1.01	.25
		3821	1.13	-1.52	.31
		3912	1.41	1.02	.41
		3913	2.41	-1.05	.25
		3914	1.79	-.07	.30
		3915	2.53	-.33	.24
		3906	1.08	-.53	.21

Table D
Sum of Squares and Degrees of Freedom Accounted for by Each of the First Four Terms in the Polynomial Expression Used to Predict Content-Based a Parameter Estimates from Total Test-Based a Parameter Estimates for each Content Area Included in Each of Four Tests

		Content Area									
		1		2		3		4		5	
Test	Source of Variation	df	SS	df	SS	df	SS	df	SS	df	SS
W1											
	Linear Term	1	.83	1	3.80	1	.61				
	Quadratic Term	1	.02	1	1.02	1	2.91				
	Cubic Term	1	.32	1	.08	1	.02				
	Quartic Term	1	.14	1	.23	1	1.29				
	Deviation from Linearity	8	.43	13	1.03	5	6.17				
	Total	12	1.75	17	6.18	9	11.03				
W2											
	Linear Term							1	5.87	1	3.01
	Quadratic Term							1	.02	1	.03
	Cubic Term							1	.00	1	.09
	Quartic Term							1	.00	1	.13
	Deviation from Linearity							25	.86	6	.79
	Total							29	6.67	10	4.05
S1											
	Linear Term	1	.72	1	1.64	1	.11				
	Quadratic Term	1	1.01	1	.18	1	.01				
	Cubic Term	1	.03	1	.12	1	.17				
	Quartic Term	1	.20	1	.00	1	.01				
	Deviation from Linearity	7	1.92	9	1.20	4	3.30				
	Total	11	3.89	13	3.14	8	3.59				
S2											
	Linear Term							1	6.76	1	1.68
	Quadratic Term							1	.01	1	.11
	Cubic Term							1	.01	1	.12
	Quartic Term							1	.02	1	.71
	Deviation from Linearity							25	.65	7	.41
	Total							29	7.44	11	3.03

Table E
Sum of Squares and Degrees of Freedom Accounted for by Each of the First Four Terms in the Polynomial Expression Used to Predict Content-Based b Parameter Estimates from Total Test-Based b Parameter Estimates for Each Content Area Included in Each of Four Tests

		Content Area									
		1		2		3		4		5	
Test	Source of Variation	df	SS	df	SS	df	SS	df	SS	df	SS
W1											
	Linear Term	1	11.33	1	30.43	1	16.70				
	Quadratic Term	1	.02	1	.16	1	.77				
	Cubic Term	1	.01	1	.84	1	.01				
	Quartic Term	1	.00	1	.16	1	.04				
	Deviation from Linearity	8	.10	13	2.94	5	.95				
	Total	12	11.47	17	34.51	9	18.47				
W2											
	Linear Term							1	41.96	1	12.36
	Quadratic Term							1	.18	1	.02
	Cubic Term							1	.58	1	.00
	Quartic Term							1	.01	1	.02
	Deviation from Linearity							25	2.34	6	.39
	Total							29	45.12	10	12.80
S1											
	Linear Term	1	16.69	1	31.74	1	13.25				
	Quadratic Term	1	.09	1	.08	1	.19				
	Cubic Term	1	.05	1	.07	1	.22				
	Quartic Term	1	.00	1	.03	1	.10				
	Deviation from Linearity	7	.60	9	.56	4	2.22				
	Total	11	17.43	13	32.48	8	15.98				
S2											
	Linear Term							1	29.69	1	4.74
	Quadratic Term							1	.01	1	.10
	Cubic Term							1	.00	1	.10
	Quartic Term							1	.01	1	.04
	Deviation from Linearity							25	.61	7	1.44
	Total							24	30.32	11	6.41