ED 266 152	TM 860 044
AUTHOR	Hills, John R.; And Others
TITLE	An Investigation of the Feasibility of Using the Three-Parameter Model for Florida's Statewide Assessment Tests.
INSTITUTION	Florida State Univ., Tallahassee. Coll. of Education.
SPONS AGENCY	Florida Univ., Gainesville. Inst. for Student Assessment and Evaluation.
PUB DATE	Dec 85
NOTE	63p.; For a related document, see TM 860 045.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; Equated Scores; Feasibility Studies; Goodness of Fit; Guessing (Tests); Item Analysis; Item Banks; *Latent Trait Theory; *Mathematical Models; *Minimum Competency Testing; Sample Size; *State Programs; *Testing Programs
IDENTIFIERS	Florida; *Florida Statewide Student Assessment Tests; LOGIST Computer Program; Rasch Model; *Three Parameter Model

ABSTRACT

This study investigated the feasibility of using the three-parameter model in Florida's minimum competency testing program. LOGIST 4 was used to analyze 1984 Statewide Student Assessment Tests (SSAT)-II data, exploring possibilities that easiness of the test would cause problems in the estimation of the a and c parameters. These problems and results of the trial analysis are discussed in chapter 2 of this 7-chapter report. An elaborate investigation of the assumption of unidimensionality of SSAT scores is reported on in chapter 3. In the next two chapters, goodness to fit of the one-, two-, and three-parameter models to the SSAT-II data is discussed, and the results of a parameter estimation investigation are described, along with a consideration of the appropriateness of different models. Chapter 6 reports on studies of the optimum number of examinees to be used in analyzing the data available from Florida's tests, including an attempt to improve parameter estimates by oversampling the lower end of the achievement distribution. A brief chapter of conclusions completes the report. (LMO)

An Investigation of the Feasibility of Using the Three -Parameter Model for Florida's Statewide Assessment Tests

Prepared by

John R. Hills - Jacob G. Beard

Sawek Yotinprasert, Natalie R. Roca, and Raja G. Subhiyah

Educational Research and Evaluation Services
College of Education
Florida State University

December, 1985

Under Contract to the:

The Institute for Student Assessment and Evaluation
University of Florida
Gainesville, Florida

ABSTRACT

AN INVESTIGATION OF THE FEASIBILITY OF USING THE THREE-PARAMETER MODEL FOR FLORIDA'S STATEWIDE ASSESSMENT TESTS

The three paramet[illegible]em response theory model, a subject of research for th[illegible]ast three decades, has recently been successfully implemented in several large-scale testing programs. Other programs, including at least two statewide assessment programs, are currently considering its adoption. The purpose of this study was to investigate the feasibility of its use in Florida's minimum competency testing program. If it were to prove feasible for use, several aspects of Florida's SSAT program could be improved. The inclusion of parameters for guessing and for item discrimination make possible the development of highly efficient tests and more precise measurement.

An investigation of feasibility should proceed in several different phases. We have outlined a possible structuring of a feasibility investigation in a previous paper (Hills and Beard, 1984). This paper presents the results of the phase one studies proposed in that paper. The studies address the following questions.

1. Will existing IRT computer programs work satisfactorily using the SSAT-II data?
2. Is the assumption of unidimensionality valid for the SSAT-II data?
3. Do the two- and three-parameter models fit the SSAT-II data better than the Rasch or one-parameter model?
4. Are the guessing (c) parameters estimable for the SSAT-II data (using the LOGIST 4 computer program)?
5. How many examinees are needed to estimate the parameters?
6. Can the parameter estimates be improved by oversampling the lower end of the ability distribution?

TABLE OF CONTENTS

Chapter 1 Introduction.. 1

Chapter 2 Convergence of Logist 4 on Minimum Competency Test Data........................... 4

Chapter 3 Dimensionality of SSAT-II Mathematics and Communications............................ 8

Chapter 4 Comparison of the Fits of the One-, Two-, and Three-Parameter Models to Florida's SSAT-II Data.................................. 15

Chapter 5 Estimating C Parameters and Consideration of the Appropriate Model...................... 33

Chapter 6 Determining Appropriate Sample Size and Distribution Shape............................ 43

Chapter 7 Conclusions.................................... 54

References 56

CHAPTER 1

INTRODUCTION

The three parameter item response theory model, a subject of research for the last three decades, has recently been successfully implemented in several large-scale testing programs. Other programs, including at least two statewide assessment programs, are currently considering its adoption. The purpose of this study was to investigate the feasibility of its use in Florida's minimum competency testing program. If it were to prove feasible for use, several aspects of Florida's SSAT program could be improved. The inclusion of parameters for guessing and for item discrimination make possible the development of highly efficient tests and more precise measurement.

An investigation of feasibility should proceed in several different phases. We have outlined a possible structuring of a feasibility investigation in a previous paper (Hills and Beard, 1984). This paper presents the results of the phase one studies proposed in that paper. The studies address the following questions.

1. Will existing IRT computer programs work satisfactorily using the SSAT-II data?
2. Is the assumption of unidimensionality valid for the SSAT-II data?
3. Do the two- and three-parameter models fit the SSAT-II data better than the Rasch or one-parameter model?
4. Are the guessing (c) parameters estimable for the SSAT-II data (using the LOGIST 4 computer program)?
5. How many examinees are needed to estimate the parameters?
6. Can the parameter estimates be improved by oversampling the lower end of the ability distribution?

In order to answer the first question we analyzed actual SSAT-II data using LOGIST 4 (Wood, Wingersky, & Lord, 1976) to see what would happen. There was a possibility that the easiness of the SSAT-II tests would cause problems in the estimation of the a and c parameters. These problems and the results of the trial analyses are given in chapter two.

The second question is addressed in chapter three. All currently used operational measurement models assume unidimensionality; however, there is always the possibility that more than one dimension is being measured in an achievement test. The SSAT-II tests have been developed by specifying a number of more or less separate skills, each with several items. Such an approach could result in serious multidimensionality. On the other hand, if there are only a few items for each skill, if achievement of one skill is correlated with achievement of others, and if many skills are included in a test (as is the case here), then a single "superdimension" may account for most of the variance in the scores, and the test may function as a unidimensional measure of that superdimension. Previous analyses by King and Hills have already indicated that the scores of the SSAT II meet a widely-used criterion for unidimensionality. A more elaborate investigation of that assumption is described in chapter three.

The next inquiry explored the goodness of fit of the one-, two-, and three-parameter models to the SSAT-II data. The person and item parameters were estimated using LOGIST 4 and then used to estimate the responses to items of individuals, and the proportion correct for groups of individuals in several ability intervals. The differences between the estimated and actual results were summarized for each model in several ways. This study compares in a relatively direct way the fits of the three models to the SSAT-II data.

The fourth question is related to the first. If difficulties were to arise in applying the three-parameter model to the SSAT-II data it was believed that it would relate to the estimation of c parameters. Wingersky (personal communication, Summer, 1984) has suggested that since the items are so easy, there may be little guessing, and there may be no real need for precise estimates of c values. It might be sound to consider a two-parameter model, involving only difficulty and discrimination, or a modified three- parameter model using something like the reciprocal of the number of decoys in an item as the c value instead of estimating it from the data. The results of a parameter estimation investigation are described in chapter five along with a consideration of the appropriateness of different models.

Questions five and six are dealt with in chapter six. Studies of the optimum number of examinees to be used in analyzing the data available from Florida's tests were completed and are described in this chapter. The question has a great deal of importance because the more persons in an analysis, the more expensive it gets. For every additional person another parameter (ability) must be estimated. Current estimates in the literature range from a low of 700 persons to satisfactory results only with 3 or 4 times that many. Recommendations of sample size for SSAT-II mathematics and communications are made on the basis of the results of empirical studies described in this chapter.

An attempt was made to improve parameter estimates by oversampling the lower end of the achievement distribution. Heavier sampling was done among those of extremely low ability levels because that is where the estimation difficulties arise and where most data are needed. This was investigated simultaneously with the sample size and parameter estimation problems.

The above questions are fundamental and it appears that the problems associated with them can be resolved satisfactorily. They constitute the first phase of an empirical investigation into feasibility. Several further studies would seem appropriate for a second phase of the feasibility irvestigation.

CHAPTER 2

CONVERGENCE OF LOGIST 4 ON MINIMUM COMPETENCE TEST DATA

In evaluating the feasibility of using the three parameter model for the Statewide Assessment Tests, a basic question is whether the LOGIST 4 program would successfully analyze the Florida data. Eignor reported (D. W. Eignor, personal communication, Summer 1984) that in 1979 ETS tried to use LOGIST 4 to estimate the item parameters for the New Jersey minimum competency tests. He indicated that they encountered severe convergence problems, i.e., after many iterations the program would not reach the end. The problem seemed to involve estimating the c parameters when the items were so easy (as minimum competency tests tend to be.) Eventually they tried using a two-parameter model, ignoring guessing or fixing it at a specific value, but they still had problems with convergence. A basic problem then, was to whether LOGIST 4 would converge using Florida Statewide Assessment Data (SSAT II).

The data for the March, 1984, administration of the SSAT II were obtained. The file contained data for 127,033 cases, but from these we removed any who were taking the test for the second time, any who were not in the tenth grade, the normal time for taking the SSAT II, any who were classified as deaf, hard of hearing, physically impaired, emotionally handicapped, educably mentally handicapped, had specific learning disabilities, or were educably mentally retarded. The remaining group contained 94,261 cases, too many for our purposes. A systematic sample of 9000 was taken from this group.

The data from the tape reported each student's response to each item. From the data we determined whether the student had answered the item correctly, had answered incorrectly, had omitted the item, or had not reached the item. Omitted items are those which are not answered correctly or incorrectly before tne last item which is answered. Not reached items are items for which no answer has been made after the last item for which an answer is made. The responses were then recoded in the format required by LOGIST 4, with a correct answer recorded as 01, an incorrect answer as 00, an omitted response as 10, and a not reached item as 11.

The LOGIST 4 program available to us for use on the CDC Cyber computer is limited to 3000 cases. So we sampled once again, creating 5 different random samples of approximately 3000 cases, with replacement between sampling. That is, an individual might appear in more than one sample, but he could not appear in the same sample more than once.

The communications and mathematics item responses were analyzed for each of the five samples of 3000 cases by LOGIST 4. In each of the ten cases (2 tests times 5 samples), convergence was reached, usually in about 20 stages. Thus, the problem experienced by Eignor does not appear to occur with these data.

However, it is noticeable in these data that even with 75 items and 3000 cases, many of the c parameters were not estimated, i. e., the program when unable to estimate a c parameter gives the item the average c parameter of the items for which c parameters have been estimated. The program does not estimate the c parameter routinely when the quantity b - 2/a is less than -2. When this quantity is equal to or less than -2, the c parameters are poorly determined (Wood et al., 1976). Many fewer c parameters were estimated for the communications test than for the mathematics test, since the average difficulty levels of the communications items were much easier than the mathematics items. The means and standard deviations of the a parameters and b parameters for each data set, with standardization on theta, appear in Table 2.1. The numbers of c parameters estimated by LOGIST for the mathematics and communications tests for each data set appear in Table 2.2.

It is clear from these results that LOGIST 4 can be used to estimate a and b parameters for the SSAT II data in the Florida Statewide Assessment Program. However, it is also clear that the c parameters, which reflect the probability of a person with very low ability answering an item correctly (probably at least partly due to guessing), will require special attention. Routine use of LOGIST 4 does not result in estimates of many of the c parameters. The tests are so easy that few of the examinees are far enough down the ability scale to allow us to ascertain what the performance of a very low ability person would be. In a later chapter we will concentrate on developing procedures to estimate c parameters. If that is not successful, we will consider the use of a modified three-parameter model in which all c parameters have the same nonzero value, or a two-parameter model in which all c parameters are fixed at zero as though there were no guessing and a person of low ability would surely not be able to answer an item correctly.

Table 2.1

Means and Standard Deviations of a and b Parameters for Mathematics and Communications Tests for Each of the Five Samples

			a Parameter		
			Mathematics		
Mean	.93	.91	.93	.93	.91
S.D.	.33	.32	.32	.33	.32
			Communications		
Mean	.83	.80	.83	.83	.80
S.D.	.31	.28	.28	.31	.28
			b Parameter		
			Mathematics		
Mean	-1.32	-1.39	-1.32	-1.33	-1.39
S.D.	1.19	1.31	1.14	1.22	1.30
			Communications		
Mean	-2.09	-2.11	-2.06	-2.09	-2.17
S.D.	.84	.94	.82	.83	.80

Table 2.2

Numbers of Estimated c Parameters
for Mathematics and Communications Tests
for Each of the Five Samples

Sample	Mathematics	Communications
1	2	0
2	2	0
3	3	0
4	2	0
5	2	0

CHAPTER 3

DIMENSIONALITY OF SSAT II MATHEMATICS AND COMMUNICATIONS

Background

Dimensionality is an important aspect of item response theory. The unidimensional theory, which we are using in this project, assumes that only one dimension or characteristic of the examinee is determining whether he answers an item correctly, and that the same dimension is involved for all examinees. If that is correct, then the desirable attributes of item response theory can be expected to hold. If the test involves more than one dimension, the estimate of the level of an individual on the characteristic being measured may not depend on which items he is asked to answer, and the parameters of the items may not depend on which individuals are administered the test, (It might be useful to point out that classical test theory must make the same assumption of unidimensionality for scores to be meaningfully interpretable. This was largely overlooked for many years by all but those who were familiar with the concepts of factor analysis.)

Unidimensionality is assumed by the Rasch (one-parameter) model as well as by two- or three-parameter models, so evaluation of dimensionality is not critical to deciding whether to move from a one-parameter model to a two- or three-parameter model for Statewide Student Assessment purposes. What is found about dimensionality is relevant for interpretation of total scores on the SSAT tests, regardless of how they are analyzed. There is a difference, however, between what happens in a one-parameter model and in a more complex model in the presence of more than one dimension. The one-parameter model weights the dimensions present approximately according to the number of items on each. A two- or three-parameter model gives dominant weight to the first or strongest dimension (Reckase, 1981). In any case, it would seem pointless to move to a more complex model from a simple model if a major assumption of both were seriously violated. So checking of this assumption seems particularly appropriate at this time.

It is highly probable that given enough items and examinees, any set of data would be in violation of an assumption of unidimensionality (Hambleton & Swaminathan, 1985). As Lord (1968) indicated when discussing the College Board's Scholastic Aptitude Test, unidimensionality for the SAT cannot be strictly correct, but it is probably a tolerably good approximation. This was evidenced by a factor analysis by Coffman which found 11 factors, but most of the variance (16%) was on the first unrotated factor, with only about 2% of the variance on any other factor. Similarly with any test, the issue is not whether unidimensionality is violated, but whether it is violated to such an extent that the results are no longer useful.

Procedure would be simple if there were an easy and well-accepted procedure to test the assumption of unidimensionality. Unfortunately, that is not the case. In 1968, Lord, on the basis of a factor analysis by Coffman, inferred that the Verbal test on the SAT was sufficiently unidimensional to be worth study by item-response theory. Similarly, Lord and Novick (1968) said that a unidimensional space is most likely to be an "adequate approximation" for tests "that appear as though they ought to be homogeneous; for example, certain tests of vocabulary, reading, spelling, and some kinds of spatial ability. On the other hand, we should expect a mathematics test made up half of arithmetic reasoning items and half of plane geometry items to show at least k=2 dimensions" (p. 381). They go on to suggest that if the first latent root of a matrix of tetrachoric correlations with communalities in the diagonal is large, and the second latent root is nearly as small as the rest, "there is good reason to treat their data as arising from a one-dimensional latent space" (p. 382).

In 1980 Lord pointed out the "great need for a statistical significance test for the unidimensionality of a set of test items" (p. 21) and gave a rough procedure based on the latent roots of the tetrachoric item intercorrelation matrix with estimated communalities in the diagonals. Again, the procedure required comparing the size of the first latent root with the second, and the second latent root with the others.

Other authors have suggested using factor analysis also. Hulin, Drasgow, and Parsons (1983) suggest using the results of factor analyzing the matrix of interitem phi coefficients with principal axes extraction and squared multiple correlations as estimates of communalities. However, they point out that this procedure will result in spurious factors due to differences in item difficulties. They also indicate that the number of factors obtained by factor analyzing interitem tetrachoric correlations provides a valid estimate of dimensionality if the sample is sufficiently large. With tetrachoric correlations there are also problems of missing values and non-Gramian matrices, and if there is guessing on the items, the assumptions of the tetrachoric correlations are violated. Carroll is cited by Hulin, et al. (1983) as providing a correction for the guessing problem, but they found his correction to be inadequate to eliminate spurious factors. Hulin et al. (1983) finally recommend a different factor analysis procedure called "modified parallel analysis" based on work by Drasgow and Lissak as "being a way for researchers to determine whether in fact there is a latent variable that is strong enough to allow application of IRT in an item pool" (p. 261).

Some authors have not been willing to accept factor analysis of interitem correlations as an appropriate procedure for evaluating dimensionality. McKinley and Mills (1984), for example, after noting that "...there is no generally accepted procedure for testing the assumption of unidimensionality" (p. 1), go on to develop an approach based on analysis of the residuals derived from estimating the probability of a correct response based on an IRT model and the observed response to the item. They argue that if the data are one dimensional, the residuals will be random error. So, if the correlation matrix based on the residuals is factor analyzed, there should be no common factor. Analysis of residuals eliminates the problem of difficulty factors, since the residuals are on a continuous scale. Product-moment correlations can be used, so there is no problem with non-Gramian matrices. Divgi (1980) is reported to have used such a procedure on the Rasch model, but McKinley and Mills applied it to two- and three-parameter models. They found that on their synthetic data, while principal components analysis of tetrachoric interitem correlations would not have discriminated between one-dimensional sets of data and sets with several dimensions, the analysis of residuals was effective in making that discrimination. For the residuals, a ratio of greater than 1.2 to 1.0 between the first eigenvalue and the second was evidence of more than one dimension.

Hattie (1984) identified 87 different indices of unidimensionality. He then evaluated each of them in a Monte Carlo study of tests 15 items in length, with 1, 2, or 5 dimensions, various degrees of item discrimination, homogeneity, difficulty, and guessing. He simulated 500 cases, with 24 replications. The results were very discouraging for factor analysis approaches--as well as most of the rest of the 87 indices. For example, he states, "Indices based on component or factor analyses do not aid in determining unidimensionality" (p. 71), and "...indices based on tetrachorics cannot be recommended" (p. 72). Also, "Using the number of eigenvalues greater than one to estimate the number of factors appears to lack justification" (p. 72). "...the ratio of first and second eigenvalues [was]not [an] effective [index]" (p. 72) And, "Overall, linear factor analysis is not appropriate for determining unidimensionality" (p. 73). He does find a study of residuals to be useful for the two-parameter model, but in his study does not go as far as to recommend a cutoff value for deciding that data are unidimensional.

Reckase (1979) studied the problems of applying unifactor methods to multifactor tests and concluded that if the first principal component is large relative to the other factors "good ability estimates can be obtained from the (one parameter and three parameter) models even when the first factor accounts for less than 10 percent of the test variance, although item calibration results will be unstable. For acceptable calibration, the first factor should account for at least 20 percent of the test variance." (page 228). He suggests that sufficient unidimensionality for stable item parameter estimates should be present for 50-item tests when the first eigenvalue is 10 or greater.

Hambleton and Swaminathan (1985) note that Bejar (1980) suggested that one evaluate unidimensionality by splitting a test into subtests based on content, i.e., make a subtest of items that appear to test a different aspect of content than the rest of the items. For the items in that subtest, obtain item parameter estimates twice, once by themselves and once including them among the other items in the larger test. Compare the two sets of item parameter estimates.

Much earlier, in 1961, Lumsden (cited in Hambleton & Swaminathan, 1985) had suggested a similar idea for constructing tests that would be unidimensional. He advocated factor analyzing a set of test items and removing items that did not measure the dominant factor. The remaining items would be factored again, and again items not measuring the dominant factor would be removed. This would be repeated until a satisfactory solution was reached. Lumsden proposed that the ratio of first-factor variance to second-factor variance be used as the index of unidimensionality.

Martois, Rickard, and Stiles (1985) reported that in their analyses of data from the California Adult Student Assessment System, the criteria proposed by Bejar and Reckase were in serious disagreement. Out of twelve tests, Bejar's criterion found only one to be unidimensional, but all twelve of them met Reckase's criterion.

Procedure

Obviously the problem of determining whether a set of data is sufficiently unidimensional for scores from the data to be used as measures of a single characteristic of examinees has not been satisfactorily resolved. Many possible approaches have been advocated, and most of them have been criticized or found inadequate upon further study. Our approach was to pick one of the methods for evaluating dimensionality that has been widely used and generally accepted, even though criticized, and use it to evaluate unidimensionality. Then to go further, we attempted to purify the data along the lines suggested by Bejar and by Lumsden and to determine whether "purification" of what already appeared to be unidimensional would result in any important change in item parameters. If no appreciable change in item parameters resulted from purification, then we would conclude that the data were sufficiently unidimensional that item parameters resulting from its analysis would not be importantly affected by whatever multidimensionality remained.

The method for evaluating unidimensionality that we chose was the one recommended by Reckase (1979) of computing a principal components analysis on the phi correlations among items and examining the percent of variance contributed by the first unrotated factor and the relationships among the eigenvalues of the factors. If the first factor variance is 20% or greater, and if the first eigenvalue is large compared to the rest, with the remaining eigenvalues being similar in size to the second, the data would appear to be sufficiently unidimensional for the item parameters from the three-parameter logistic model to be satisfactory. While Martois et al. (1985) found that Reckase's and Bejar's methods did not agree, and Hambleton and Swaminathan pointed out the problems of using factor analysis in determining unidimensionality, when Hambleton and Swaminathan reported an actual study in the same text in which they had been critical, they used Reckase's criterion (p. 275). Combining the Reckase and Bejar approaches should establish a conclusion on a firmer foundation than either taken alone.

Data and Results

The data from the SSAT II administration of March, 1984, were factor analyzed using the principal components method, unities in the diagonals. Random samples of approximately 750 cases were used in each analysis. Analyses were done separately for the communications and mathematics sections. The eigenvalues and percentages of variance accounted for by each of the first 6 factors appear in Table 3.1, below.

It can be seen in Table 3.1 that in each case the percent of variance accounted for was near 20%, that the first eigenvalue was greater than 10 and very large compared to the second and subsequent eigenvalues, and that the first factor accounted for a considerably larger proportion of the variance than the second and subsequent factors. Since items were intercorrelated using phi coefficients, the second factor, if it exists at all, is most likely a factor related to the difficulty values of the items. Judging from these results, the SSAT II mathematics and communications scores appear to be reasonably unidimensional.

To explore dimensionality further, for the mathematics test the skill scores were factor analyzed using the principle components method. The seven skills which were least highly related to the first factor were deleted, creating a more homogeneous test. LOGIST 4 was used to estimate the item parameters and the abilities,

Table 3.1

Eigenvalues and Percentages of Variance Accounted For

Mathematics		Communications	
Eigenvalue	Percent	Eigenvalue	Percent
12.48	16.6	13.87	18.5
2.40	3.2	2.46	3.3
2.22	3.0	2.14	2.8
1.73	2.3	1.87	2.5
1.59	2.1	1.64	2.2
1.53	2.0	1.57	2.1

standardizing on abilities, for both the heterogeneous test using the odd-numbered items (to equalize length approximately) and the homogeneous test using only the approximately half of the items (40 items on 8 skills) which loaded most highly on the first factor. (Twenty items on the homogeneous test were among the odd-numbered items on the heterogeneous test.) The differences between the parameter values for these two analyses were calculated, and the means and standard deviations of the twenty differences obtained. Those means and standard deviations appear in Table 3.2, below. (The statistics for the c parameters are, of course, not very meaningful, since most c parameters are set at the average value of the estimated c parameters in these data.) Similarly, the means and standard deviations of the differences between abilities estimated from the two sets of items were calculated for a systematic sample of 158 cases. (Of the systematic sample of 158 cases, 13 persons who obtained perfect scores on either test were deleted leaving a reduced sample of 145 cases.)

It can be seen in Table 3.2 that the mean differences in item parameter values for the heterogeneous groups of items and the homogeneous groups of items, as defined above, are quite small, and their variability is not great. In other words, if content multidimensionality is present in these data, removing the half of the items that least reflects the major dimension makes only slight differences in the item parameters. Thus, whatever multidimensionality is present is of little practical importance in estimating item parameters. The mean difference for thetas was .08, but the standard deviation was .39. What this implies about dimensionality is not clear at this point. In general, we interpret these results to indicate that the Statewide

Table 3.2

Means and Standard Deviations of Differences In Item Parameters

	a	b	c
Mean	.Ø53	.Ø43	.Ø31
Standard Dev.	.Ø66	.111	.Ø31

Assessment Tests meet the criterion of unidimensionality sufficiently well that results from parameter estimation procedures which assume unidimensionality, such as the Rasch model or the three-parameter model, will be quite useful.

Attempts to compare item parameters and ability estimates from heterogeneous tests and homogeneous tests of communication skills were unsuccessful. LOGIST 4 would not converge using only 4Ø items and 3ØØØ cases on a test composed of items as easy as the communications items in SSAT II. The mean item p value for the mathematics items was .81, with 16 of the 75 items having p values less than Ø.75,and the smallest p being .33. The mean item p value for the communications items was .9Ø, with only 2 having p values below Ø.75, and those 2 were .74. The smallest p was Ø.74. This suggests that the mathematics items are on the borderline of being unanalyzable by LOGIST 4, and the communications items are so easy that one must have a large number, such as 75 as in the total test, in order for the program to function satisfactorily.

Conclusion

Dimensionality is not an easy assumption to evaluate. New approaches to it appear frequently. It may be that no set of data with more than a few items is completely unidimensional. The fundamental issue for item response theory analyses is whether the data are so multidimensional that misleading interpretations are produced. Our impression from the above analyses is that the SSAT data are sufficiently unidimensional that the advantages of item response theory should be pursued. Nothing here suggests that grossly misleading results will be produced due to multidimensionality.

CHAPTER 4

COMPARISON OF THE FITS OF THE ONE- TWO- AND THREE-PARAMETER MODELS TO FLORIDA'S SSAT-II DATA

The appropriateness of the one-parameter, or Rasch model, for mulciple choice achievement test data has been questioned since its introduction. The Rasch model is based on the assumption that the response to a test item is based only on the ability of the person and the difficulty of the item. Proponents of the three-parameter model argue that one item parameter is insufficient, and that test data are modeled more closely by including additional item parameters for guessing and for item discrimination. Advocates of the Rasch model claim that problems in estimating these two additional parameters negate potential advantages in their use (Wright, 1977).

Hambleton and Murray reviewed the literature on goodness of fit and proposed three ways for addressing the fit of test data to an item response model. (1983, p. 72)

a. Determine if the test data satisfy the assumptions of the test model of interest.
b. Determine if the expected advantages derived from the use of the item response model (for example, invariant item and ability estimates) are obtained.
c. Determine the closeness of fit between predictions and observable outcomes (for example, test score distributions) utilizing model parameter estimates and the test data.

In a separate paper (1985) Murray reviewed the literature on fit analysis and presented a rationale for the use of exploratory analytic techniques without tests of statistical significance. The statistical significance tests which have been previously used have been shown to be sensitive to sample size, with larger numbers of misfitting items being found when larger sample sizes were used. The exploratory techniques involve an examination of size and direction of residuals, descriptive statistics, and the identification of particular kinds of misfit.

This study will focus on the third method above, the closeness of fit between predictions and observable outcomes utilizing model parameter estimates and the SSAT-II data.

The analyses completed closely parallel those done by Hambleton, Murray, and Williams in an investigation of the fit of the one-, two-, and three-parameter models to item scores from the 1982 Maryland Functional Reading Test (1983). They found larger differences between the one- and two-parameter models than between the two- and three-parameter models. They concluded that the two-parameter model was adequately fit by the data. They speculated that the small improvement in fit gained by using the three-parameter model was related to the easiness of the tests which resulted in small amounts of guessing by the students who took the tests.

The fit of the SSAT-II data to the Rasch model has also been studied in previous investigations of item fit (Beard, Julian, and Roca, 1984) and of person fit (Beard, Julian, Richards, and Roca, 1984).

Method

The analysis of residuals has been an integral part of tests of fit of data to the Rasch model since the first BICAL programs were prepared by Wright and Panchapakesan (1969). More recently, the technique has been used to test the fit of the three-parameter model (Hambleton and Murray, 1983). The basic idea is to estimate parameters for a psychometric model and use the parameters to estimate item responses. The differences between the estimated and actual responses constitute the residuals. These residuals are then analyzed in various ways to show the amount and nature of misfit of the data to the model. Models which more closely fit the real data would be preferred over those which fit it less well, all other things being equal.

Sample

A random sample of 3,000 cases was selected from the population of 100,571 students who took the SSAT-II in the March 85 administration. The item responses were scored as right or wrong with omitted or "not reached" items also scored as incorrect. This sample provided the data for separate analyses of the mathematics and communications parts of the test. The numbers of cases in the analyses were reduced to 2929 for mathematics and to 2768 for communications when zero and perfect scores were removed by the computer program.

Analyses

The LOGIST 4 computer program (Wood et al., 1976) was used to generate the person abilities and item parameters for the one-, two-, and a modified three-parameter model from the 3,000 SSAT-II item response records. A modified three-parameter model was used because the LOGIST program could estimate c parameters for only a small number of the items. The modification consisted of fixing the "c" parameters at .20 for all items. The term "three-parameter model" will be used throughout this part of the paper to indicate the "modified three-parameter model".

These abilities and item parameters were then used to estimate responses and to compute average residuals for each item and across all 75 items by selected ability groups using a computer program entitled RESID developed by Hambleton and Murray (1983). The display of average residuals by ability levels shows the average departure of the data from the item characteristic curves at different intervals on the ability continuum. The mathematics ability continuum was divided into twelve intervals and the communications into ten groups in order to represent best the distribution of log ability estimates for the two sets of scores.

For each score interval a residual was found by computing the difference between the proportion of the examinees answering each item correctly (P) and the proportion expected to answer correctly based on the item characteristic curve ($\hat{P}$). Standardized residuals were computed by dividing each residual by the standard error of the proportion correct.

$$SR_{ij} = \frac{P_{ij} - \hat{P}_{ij}}{\sqrt{\frac{P_{ij}(1-P_{ij})}{N_j}}}$$

There was concern initially that the standardized residuals might be overly inflated by small standard errors of proportions resulting from easy test items. However, an examination of the results showed that easier items did not tend to yield the highest standardized residuals. This finding is consistent with that of Murray (1985) who compared the use of raw and standardized residuals extensively and concluded that standardized residuals yielded results which were similar to raw residuals and provided the additional benefit of taking into account sampling errors associated with the proportion correct (P-value).

The mean absolute values of raw and standardized residuals by ability intervals are shown in Table 4.1. The mean raw residuals are inversely related to the ability levels of the students; i. e., the residuals are greater for low ability and smaller for high ability students. This is true for the one-, two-, and three-parameter models. The raw residuals are greatest for the one-parameter model and

Table 4.1

Average Absolute Values of Raw and Standardized Residuals

Ability	Number of Examinees			Av. Absolute Valued Raw Residuals			Av. Absolute Valued Standardized Residuals		
Level	1-P	2-P	3-P	1-P	2-P	3-P	1-P	2-P	3-P
Mathematics									
-2.75	9	6	16	.183	.152	.116	2.359	1.198	1.048
-2.25	45	27	50	.169	.124	.062	3.595	1.638	.973
-1.75	137	131	111	.112	.043	.040	3.547	1.139	.935
-1.25	323	325	289	.067	.024	.022	3.036	.998	.860
-.75	454	482	467	.030	.019	.016	1.619	1.034	.861
-.25	541	556	571	.025	.016	.014	1.714	1.073	.986
.25	491	494	533	.030	.011	.009	2.596	.995	.878
.75	477	432	462	.032	.009	.008	3.249	.910	.822
1.25	291	254	231	.025	.011	.011	2.420	.934	.957
1.75	124	133	134	.019	.010	.010	1.434	.763	.818
2.25	74	68	66	.015	.018	.015	1.054	.996	.898
2.75	1	21	19	.023	.018	.022	.233	.657	.694
Overall Average	—	—	—	.061	.038	.029	2.238	1.028	.894
N	2967	2929	2949	—	—	—	—	—	—
Communications									
-3.00	18	20	17	.203	.120	.103	2.903	1.361	.960
-2.50	51	32	43	.173	.074	.062	3.401	.918	.871
-2.00	104	97	93	.121	.034	.031	3.249	.724	.677
-1.50	147	166	150	.068	.024	.023	2.406	.707	.685
-1.00	296	283	280	.034	.018	.017	2.049	.836	.776
-.50	464	492	484	.029	.010	.010	2.657	.768	.772
0.00	401	557	556	.031	.006	.006	3.258	.661	.650
.50	602	487	505	.026	.007	.007	4.344	.941	.863
1.00	349	341	327	.018	.007	.007	2.760	.967	.943
1.50	336	193	186	.011	.015	.016	1.862	1.418	1.464
Overall Average	—	—	—	.071	.032	.029	2.889	.930	.866
N	2768	2668	2641	—	—	—	—	—	—

least for the three-parameter model at almost all ability levels; however, the differences in mean residuals among the three models become trivial at the highest ability levels. These data also show the tendency of small ra`` residuals to appear large when standardized by dividing by standard errors from large P-values. A mean raw residual (one-parameter) of .Ø25 at an ability level of 1.25 yields a mean standardized residual of 2.42Ø; whereas a raw residual (two-parameter) of .Ø24 at an ability level of -1.25 yields a standardized residual of only .998.

Several descriptive analyses were done in order to evaluate (a) the residuals for each of the one- two- and three-parameter models for each item, (b) the distributions of standardized residuals, (c) the average residual for selected ability levels, (d) the association between residuals and content categories, (e) the association between residuals and difficulties of items, and (f) the association between the items' residuals and discrimination indices.

Results

The mean standardized residuals for the one-, two-, and three-parameter models are shown in Table 4.2 for Mathematics and Table 4.3 for Communications along with the classical item analysis indices of difficulty and discrimination. The proportions correct reflect the easiness of this minimum competency test. The residuals are smallest for the three-parameter model and largest for the one-parameter. In fact, none of the 15Ø standardized residuals for the two- or three-parameter models are as large as 2.ØØ which could be considered statistically significant. Sixty-nine of the 15Ø residuals for the one-parameter model exceed this arbitrary significance value. A casual inspection of these data would certainly indicate that use of more item parameters improves fit to the item responses.

The data in Tables 4.2 and 4.3 also confirm that in the one-parameter model difficult items show greater misfit than easier ones. For example, note item 66 in the mathematics test. It is the most difficult item (p = .34) and also has the poorest fit to the one-parameter model (1Ø.155). Its fit to the two- and three-parameter model is not significantly large (1.579 and 1.423 respectively).

Tables 4.2 and 4.3 also show that the items within most skill areas vary considerably in difficulty, discrimination, and fit to the IRT models.

The distributions of fit indices for the three models are shown in Table 4.4. The total set of indices consists of one index for each of 75 items for each of the score intervals, and can be arrived at by multiplying the number of items by the number of intervals, 12 for mathematics and 1Ø for communications. The residuals should be normally distributed if the data fit the model. Table 4.4 shows that

Table 4.2

Mathematics Item Statistics

Test Item	Skill	Prop. Correct	Pt.Biserial Correlation	Absolute-Valued Standardized Residuals 1-p	2-p	3-p
1	1	.986	.117	.838	.837	.593
2	1	.978	.120	1.257	.447	.480
3	1	.919	.227	2.384	1.067	1.070
4	1	.875	.389	1.455	.860	.999
5	1	.948	.370	1.127	1.099	.904
6	2	.927	.262	1.628	.627	.736
7	2	.594	.427	2.913	.845	.576
8	2	.878	.383	1.311	.869	.808
9	2	.896	.312	1.748	.778	.693
10	2	.487	.468	2.140	1.230	.849
11	3	.717	.470	1.422	.971	1.189
12	3	.971	.231	.779	.959	.784
13	3	.928	.288	1.310	1.214	.706
14	3	.902	.241	2.945	1.053	.660
15	3	.958	.298	1.018	1.020	1.099
16	4	.836	.485	.834	.913	.944
17	5	.829	.453	1.126	.883	.906
18	5	.897	.402	.921	.889	.745
19	5	.851	.309	3.177	1.223	1.013
20	5	.925	.440	1.169	.810	.820
21	5	.946	.332	1.299	1.117	.986
22	6	.964	.284	.910	1.126	.673
23	6	.896	.366	1.471	1.082	.813
24	6	.950	.407	1.216	1.125	.814
25	6	.880	.515	1.094	.832	.708
26	6	.964	.304	1.202	1.198	1.174
27	4	.847	.572	1.474	1.080	1.147
28	4	.867	.429	1.241	1.343	1.351
29	4	.611	.243	7.424	1.041	1.112
30	4	.921	.252	2.036	.846	.648
31	7	.680	.489	1.653	.893	.820
32	7	.802	.377	2.266	.947	.805
33	7	.854	.429	1.019	1.124	1.105
34	7	.881	.412	1.065	1.129	1.197
35	7	.820	.340	2.498	.997	1.142
36	8	.980	.307	1.430	.982	1.003
37	8	.877	.402	1.086	.802	.628
38	8	.831	.236	4.839	1.184	.905
39	8	.912	.261	2.125	.969	.866
40	8	.826	.461	1.057	.910	.832

continued

(Table 2 continued)

41	9	.749	.241	5.446	.940	.972
42	9	.748	.261	5.785	.655	.856
43	9	.735	.503	1.123	1.021	.598
44	9	.806	.442	1.438	1.432	1.467
45	9	.814	.551	1.451	1.404	1.109
46	10	.803	.436	1.597	.895	.896
47	10	.834	.235	4.414	.767	.885
48	10	.828	.320	2.951	.810	.785
49	10	.904	.287	2.050	.755	.713
50	10	.864	.260	3.270	.976	.895
51	11	.888	.377	1.143	.884	.651
52	11	.880	.437	.582	.777	.956
53	11	.881	.269	2.700	.736	.812
54	11	.601	.411	3.216	1.201	1.074
55	11	.681	.534	.941	1.059	1.008
56	12	.774	.432	1.561	.788	.832
57	12	.756	.343	3.333	.952	.841
58	12	.372	.190	8.145	1.468	1.023
59	12	.878	.337	1.740	1.051	.954
60	12	.669	.482	1.595	1.089	1.031
61	13	.485	.269	7.351	.960	.918
62	13	.787	.497	.989	.911	.776
63	13	.623	.477	1.845	.877	.709
64	13	.738	.395	2.732	1.041	.747
65	13	.776	.387	2.626	.796	.982
66	14	.340	.132	10.155	1.579	1.423
67	14	.806	.470	1.948	1.842	1.114
68	14	.707	.431	2.430	.488	.514
69	14	.641	.446	2.975	1.321	.876
70	14	.781	.454	1.845	1.336	1.095
71	15	.722	.540	.844	.882	.340
72	15	.692	.439	2.484	1.901	.690
73	15	.662	.455	2.320	1.485	.770
74	15	.763	.471	1.709	1.497	1.124
75	15	.822	.471	1.206	1.201	1.313

Table 4.3

Communication Item Statistics

Test Item	Skill	Prop. Correct	Pt.Biserial Correlation	Absolute-Valued Standardized Residuals 1-p	2-p	3-p
1	1	.878	.266	5.432	1.005	1.143
2	2	.930	.279	3.656	.681	1.007
3	2	.979	.323	1.651	.878	1.010
4	2	.978	.339	1.350	.688	.991
5	3	.970	.373	1.553	.710	1.008
6	5	.961	.440	.772	.860	.787
7	1	.848	.321	4.724	.505	.308
8	1	.851	.457	2.007	.775	.655
9	2	.932	.435	1.168	.936	.675
10	2	.956	.402	1.580	.562	.509
11	8	.981	.312	1.389	1.124	.746
12	8	.991	.267	1.145	1.174	.956
13	8	.963	.228	4.729	.676	.738
14	8	.910	.282	3.549	.907	.875
15	8	.946	.251	6.553	1.258	.707
16	5	.873	.411	2.759	.568	.649
17	1	.864	.505	1.279	.739	.778
18	7	.859	.383	3.122	.898	.496
19	7	.896	.530	.894	1.202	.947
20	5	.839	.394	3.057	.789	.691
21	5	.911	.472	1.708	.876	.986
22	5	.925	.544	.963	1.089	1.104
23	6	.865	.456	2.094	1.248	.988
24	7	.848	.473	1.325	1.226	.811
25	7	.811	.412	2.738	1.720	.964
26	6	.830	.359	4.160	1.057	1.149
27	10	.834	.399	3.841	.903	.717
28	9	.823	.398	3.479	1.378	1.432
29	9	.901	.337	3.679	.882	.857
30	9	.956	.330	6.124	1.139	.767
31	9	.832	.224	7.519	.899	.967
32	9	.973	.356	1.818	1.063	.699
33	6	.898	.507	1.104	.767	.744
34	6	.917	.454	.884	1.148	.927
35	4	.911	.372	3.008	.978	.730
36	11	.952	.507	1.027	.897	.931
37	3	.926	.482	1.086	1.134	.919
38	11	.936	.321	2.652	.458	.451
39	6	.881	.445	1.520	1.362	1.058
40	3	.907	.524	.667	.840	.614

continued

(Table 3 continued)

41	3	.896	.388	2.476	.777	.630
42	11	.955	.403	1.663	.672	.756
43	3	.901	.489	1.245	.988	.972
44	11	.737	.305	7.246	1.214	1.244
45	12	.791	.304	6.061	1.254	1.436
46	12	.837	.342	4.315	1.274	1.358
47	4	.921	.485	1.780	1.147	1.296
48	4	.942	.437	1.094	.748	.629
49	4	.896	.399	2.884	.905	.919
50	10	.928	.403	1.850	.637	.647
51	7	.924	.581	1.442	1.536	1.056
52	4	.939	.488	1.084	.855	.793
53	11	.957	.473	.725	.911	.818
54	1	.909	.555	.867	1.117	1.027
55	10	.888	.265	5.369	1.062	.785
56	10	.931	.408	1.363	.984	1.019
57	10	.939	.300	3.013	.947	1.016
58	13	.891	.341	3.832	.886	.739
59	13	.941	.361	1.917	.545	.792
60	13	.940	.352	2.576	.738	1.009
61	13	.960	.498	.767	.736	.786
62	13	.964	.334	1.516	.952	.919
63	14	.921	.294	3.532	.794	.755
64	14	.946	.315	6.689	.909	.741
65	14	.966	.441	1.358	.799	.619
66	14	.794	.318	5.907	1.048	1.249
67	14	.877	.320	4.016	.611	.799
68	15	.874	.293	5.057	1.006	.621
69	15	.736	.281	7.585	1.049	1.154
70	15	.808	.436	2.615	.552	.719
71	15	.910	.470	1.451	.782	.804
72	15	.852	.278	5.890	1.079	1.152
73	12	.920	.311	3.363	.842	.754
74	12	.910	.260	4.744	.701	.784
75	12	.789	.319	5.606	.698	.671

Table 4.4

Levels of Absolute Values of Standardized Residuals for the Three Logistic Test Models

Model	Number/Percentage of Absolute Values of St. Residuals 0 to 1	1 to 2	2 to 3	over 3	Total
Mathematics					
1-Parameter	369/41%	213/24%	101/11%	217/24%	900/100%
2-Parameter	528/59%	277/31%	67/ 7%	28/ 3%	900/100%
3-Parameter	578/64%	258/29%	55/ 6%	9/ 1%	900/100%
Communications					
1-Parameter	219/29%	191/26%	82/11%	258/34%	750/100%
2-Parameter	477/64%	213/28%	46/ 6%	14/ 2%	750/100%
3-Parameter	493/66%	205/27%	41/ 6%	11/ 1%	750/100%
Normal Dist.	68%	27%	4%	.25%	

the residuals for the two- and three-parameter models approximate normal distributions with slightly fewer small residuals and slightly more large residuals than normal. The residuals for the one-parameter model are obviously not normally distributed and tend toward rectangular distributions having many more large and fewer small residuals than would be expected for a normal distribution.

The average standardized residuals and the average absolute values of standardized residuals are shown in Table 4.5. These values show the relationship between amount of misfit for the three models and ability level of the examinees. The numbers of ability intervals are different for math and communications because of the different distribution shapes for the two sets of scores. When the communications data were analyzed with 12 intervals as for math, an empty ability interval resulted because of the particular raw to log ability score conversion. The number of students in each of the intervals is also shown in Table 4.5.

The misfit of the one-parameter model is larger than that of the two- and three-parameter models throughout the ability scale. The only exception occurs for a cell in which the average residual for the one parameter model was based on only one student.

Table 4.5

Mean Standardized Residuals at Twelve Ability Levels for the Three Logistic Models

Ability	Number of Examinees			Average Standardized Residuals			Av. Absolute Valued Standardized Residuals		
Level	1-P	2-P	3-P	1-P	2-P	3-P	1-P	2-P	3-P
Mathematics									
-2.75	9	6	16	1.879	.394	.117	2.359	1.198	1.048
-2.25	45	27	50	2.948	.433	.170	3.595	1.638	.973
-1.75	137	131	111	3.200	.430	.168	3.547	1.139	.935
-1.25	323	325	289	2.859	.608	.420	3.036	.998	.860
-.75	454	482	467	1.064	.145	.280	1.619	1.034	.861
-.25	541	556	571	-.933	-.306	-.114	1.714	1.073	.986
.25	491	494	533	-2.088	-.202	-.256	2.596	.995	.878
.75	477	432	462	-2.662	-.060	-.069	3.249	.910	.822
1.25	291	254	231	-1.659	.275	.192	2.420	.934	.957
1.75	124	133	134	-.518	.419	.402	1.434	.763	.818
2.25	74	68	66	.071	.385	.450	1.054	.996	.898
2.75	1	21	19	-.032	.248	.176	.233	.657	.694
Overall									
Average	—	—	—	.344	.231	.161	2.238	1.028	.894
N	2967	2929	2949	—	—	—	—	—	—
Communication									
-3.00	18	20	17	2.166	.339	.103	2.903	1.361	.960
-2.50	51	32	43	2.297	.405	.117	3.401	.918	.871
-2.00	104	97	93	1.754	.291	.144	3.249	.724	.677
-1.50	147	166	150	.415	.091	.036	2.406	.707	.685
-1.00	296	283	280	-.534	.086	.077	2.049	.836	.776
-.50	464	492	484	-1.791	-.003	.009	2.657	.768	.772
0.00	401	557	556	-2.942	-.050	.076	3.258	.661	.650
.50	602	487	505	-3.905	.126	.083	4.344	.941	.863
1.00	349	341	327	-1.960	.434	.420	2.760	.967	.943
1.50	336	193	186	-.421	.629	.587	1.867	1.418	1.464
Overall									
Average	—	—	—	.492	.235	.150	2.889	.930	.866
N	2768	2668	2641	—	—	—	—	—	—

BEST COPY AVAILABLE

The average standardized residuals, as contrasted to the average absolute values, may be viewed as indicators of bias in response estimation. If residuals, or errors, in response estimation are randomly distributed around zero we would expect the mean of the standardized residuals to be near zero for each ability level. The overall mean residuals for the one-parameter model are small, .344 and .492 for math and communications respectively. However, these low means are caused by a consistent pattern of positive residuals for low scorers and negative residuals for high scorers. This pattern is probably caused by items being guessed correctly by students at lower achievement levels. For the two- and three-parameter models there was a tendency toward negative residuals near the zero log ability level and positive ones near the low and high extremes. The same pattern was present in the Maryland Functional Reading Test data (Hambleton, Murray, and Williams, 1983), but was not discussed in that paper. The pattern would seem to result from overestimates of ability in the middle of the range. Further research is needed to explain this pattern of residuals.

The average absolute values of standardized residuals show a consistent pattern of larger values for the one-parameter models with slightly smaller values for the three- than for the two-parameter model. This pattern is generally consistent throughout the ability range and for both math and communications tests.

All three of the item response theory models used in this part of the study assume unidimensionality. This assumption is suspect when the model is applied to tests such as the SSAT-II. These are minimum competency tests, each based on 15 specific skills. If one or more of these skills measures a trait which is substantially different from the dominant trait of the total test, the difference might be reflected in misfit for those skills. In order to identify such misfit, the residuals for the three models were aggregated by skills. These mean residuals by skills are shown in Table 4.6 along with the corresponding mean p-values and point-biserial discrimination indices.

The data in Table 4.6 show a general decrease in mean residuals as one goes from the one- to the two- to the three-parameter models. None of the mean residuals for the two- or three-parameter models approaches the value 2.00 which might be considered significant. On the other hand, the mean one-parameter residuals of several of the skills in both mathematics and communications exceed this value. The mean residuals for the one-parameter are also considerably more variable than those of the two- and three-parameter models.

Table 4.6

Mean Standardized Residuals by Skill

Skill	Number of Items	Mean P-Value	Mean Pt.Biserial	Mean Standardized Residual One Parameter	Mean Standardized Residual Two Parameter	Mean Standardized Residual Three Parameter
Mathematics						
M1	5	.941	.245	1.412	.862	.809
M2	5	.756	.370	1.948	.870	.732
M3	5	.895	.306	1.495	1.043	.888
M4	5	.816	.396	2.602	1.045	1.040
M5	5	.890	.387	1.538	.984	.894
M6	5	.931	.375	1.179	1.073	.836
M7	5	.807	.409	1.700	1.018	1.014
M8	5	.885	.333	2.107	.969	.847
M9	5	.770	.400	3.049	1.090	1.000
M10	5	.845	.308	2.856	.841	.835
M11	5	.786	.406	1.716	.931	.900
M12	5	.690	.357	3.275	1.070	.936
M13	5	.682	.405	3.109	.917	.826
M14	5	.655	.387	3.871	1.313	1.004
M15	5	.732	.475	1.713	1.393	.847
Communications						
C1	5	.870	.421	2.862	.828	.782
C2	5	.955	.356	1.881	.749	.838
C3	5	.920	.451	1.405	.890	.829
C4	5	.922	.436	1.970	.927	.873
C5	5	.902	.452	1.852	.836	.843
C6	5	.878	.444	1.952	1.116	.973
C7	5	.868	.476	1.904	1.316	.855
C8	5	.958	.268	3.473	1.028	.804
C9	5	.897	.329	4.524	1.072	.944
C10	5	.904	.355	3.087	.907	.837
C11	5	.907	.402	2.663	.830	.840
C12	5	.849	.307	4.818	.954	1.001
C13	5	.939	.377	2.122	.771	.849
C14	5	.901	.338	4.300	.832	.833
C15	5	.836	.352	4.520	.894	.890

It appears that the misfit of the one-parameter model is adequately taken care of by the addition of the discrimination parameter. Efforts to relate this improvement in fit to the mean p-value and discrimination indices were not productive except for the tendency of one-parameter misfit to be associated with the more difficult skills.

The subject matter of the skills and the texts of their constituent items were examined for association with improvements in fit from the two- to 'he three-parameter model. No obvious anomalies in item form, subject matter content, or placement of the items within the test were observed.

The associations between difficulties of the items and the means of the absolute values of the residuals are shown in Table 4.7. The items for each of the tests were arbitrarily divided into two groups; easier and harder. The mean absolute values of the standardized residuals were then computed for each difficulty group for each IRT model. The results show again the larger mean residuals for the one-parameter model. Furthermore, the tendency of difficult items to misfit the one-parameter model is obvious in the larger mean residuals for the harder items than for the easier ones. This difference is less pronounced for the two-parameter model and is reversed for the three-parameter math results. It appears that the use of the two-parameter model would almost eliminate the association of difficulty and misfit found in the one-parameter model. That association would be virtually eliminated with the three-parameter model.

Table 4.7

Relationship between Item Difficulty and the Absolute-Valued Standardized Residuals for the SSAT-II

Difficulty Level	# of Items	1-Parameter		2-Parameter		3-Parameter	
		Mean	SD	Mean	SD	Mean	SD
Mathematics							
Hard(p<.8)	27	3.22	2.48	1.08	.31	.88	.23
Easy(p>.8)	48	1.68	.91	1.00	.23	.90	.21
Communications							
Hard(p<.9)	31	3.87	1.93	.98	.28	.90	.28
Easy(p>.9)	44	2.20	1.58	.89	.21	.84	.17

Table 4.8 shows the relationship between item point-biserial correlations and the mean absolute values of standardized residuals for items by IRT model. The chi-squared

Table 4.8

Relationship Between Item Point-biserial Correlations and Absolute-Valued Standardized Residuals for the SSAT-II

		Item Point-biserial Correlations		
Model	Residuals	0 to .20 (N=4)	.21 to .40 (N=35)	above .40 (N=36)
Mathematics				
One-	0 to 1.0	25%	5.7%	16.7%
Parameter	1.1 to 2.0	25%	37.1%	63.9%
Model	over 2.0	50%	57.1%	19.4%
Chi-squared=12.11	df=4	p=.016	eta=.340	
Two-	0 to 1.0	50%	60%	47.2%
Parameter	1.1 to 2.0	50%	40%	52.8%
Model	over 2.0	00%	00%	00%
Chi-squared=1.18	df=2	p=.550	eta=.093	
Three-	0 to 1.0	50%	80%	61.1%
Parameter	1.1 to 2.0	50%	20%	38.9%
Model	over 2.0	00%	00%	00%
Chi-squared=3.72	df=2	p=.156	eta=.107	
Communications				
One-	0 to 1.0	0.0%	0.0%	25.0%
Parameter	1.1 to 2.0	7.7%	23.3%	59.4%
Model	over 2.0	92.3%	76.7%	15.6%
Chi-squared=35.13	df=4	p=.000	eta=.668	
Two-	0 to 1.0	46.2%	70.0%	65.6%
Parameter	1.1 to 2.0	53.8%	30.0%	34.4%
Model	over 2.0	0.0%	0.0%	0.0%
Chi-squared=2.30	df=2	p=.320	eta=.175	
Three-	0 to 1.0	69.2%	66.7%	81.3%
Parameter	1.1 to 2.0	30.8%	33.3%	18.8%
Model	over 2.0	0.0%	0.0%	0.0%
Chi-squared=1.82	df=2	p=.403	eta=.156	

BEST COPY AVAILABLE

t .s show a statistically significant relationship for the one-parameter model and insignificant relationships for the two- and three-parameter models. The eta correlation coefficients show that the relationship is substantial for the one-parameter model and trivial for the two- and three-parameter models. The relationship between item discrimination and fit is further shown in Figure 1. Two things are obvious in this figure. First, the negative relationship between discrimination and fit for the one-parameter model is clearly shown. Second, the absence of such a relationship for the two- and three-parameter model is apparent. The inclusion of the a or "discrimination" parameter in the model effectively eliminates the misfit caused by differing item discriminations or slopes of item characteristic curves. The amount of misfit, but not the relationship, is reduced slightly further by the addition of the "c" or "guessing" parameter.

Conclusions

The item response data from the mathematics and communications parts of the SSAT-II fit the two- and three-parameter better than the one-parameter IRT model. The misfit of the one-parameter model appears to be related to the difficulty of the items and to varying discrimination indices. Most of the misfit can be eliminated by the addition of a discrimination parameter. The addition of a guessing parameter reduces misfit slightly, but the reduction may be so small that the consequences would be practically insignificant

The three-parameter model used in this study was in fact a modified three-parameter model in which fixed c or "guessing" parameters were supplied to the "LOGIST" parameter estimation program. The estimation of c parameters proved to be possible for only small numbers of these items. Therefore, the choice of models, from the standpoint of fit, would be between the two-parameter and the modified three-parameter models.

The failure of the three-parameter model to improve fit significantly over the two-parameter was probably a consequence of the easiness of the items. Guessing would be minimal with such easy items. Therefore, the two-parameter model can be recommended for the current SSAT-II test. If the test were made substantially more difficult, then the use of the c parameter might improve fit significantly and should be investigated further.

The results of this part of the study are clear. The use of the two- or three-parameter model would significantly improve fit over the one-parameter model. However, the practical consequences of the misfit are not precisely

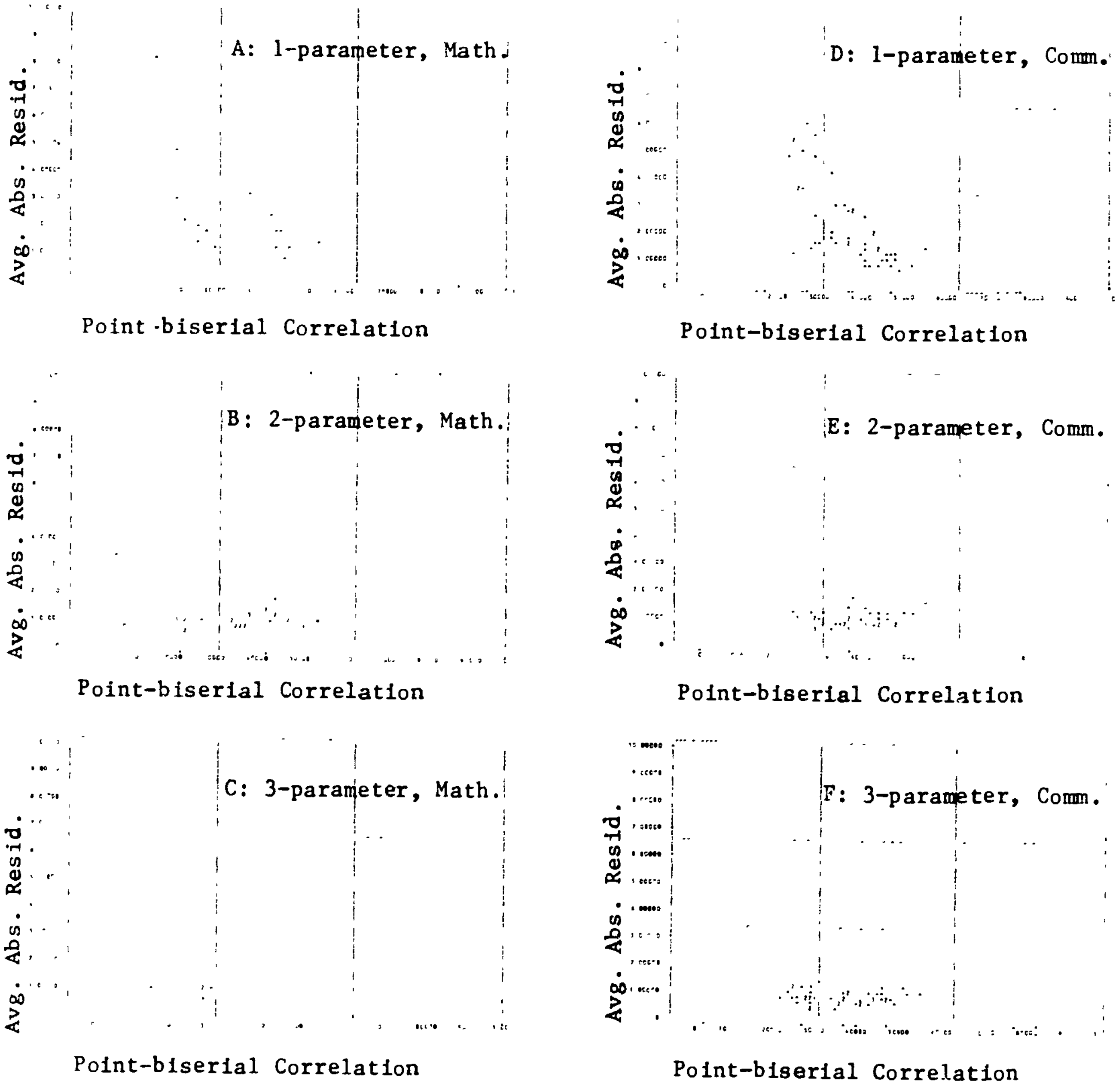

Figure 1. Relationship Between Item Discrimination and Fit

known. The standardization of residuals takes into account the sampling error of the estimated proportion correct; i. e., when the number of subjects is small the residual is divided by a larger number to arrive at the standardized residual. However, division by the standard errors also results in larger numbers, making small raw residuals appear large, especially when the items are very easy and the estimated proportions correct are near 1.00. Therefore, although the results clearly show better fit for the two- and three-parameter models, those results should be kept in perspective with other findings showing the effects of misfit on students' scores and the relative utility and complexity of the models.

CHAPTER 5

ESTIMATING c PARAMETERS AND CONSIDERATION OF THE APPROPRIATE MODEL

Problems in Estimating the c Parameter

The estimation of c parameters in the three-parameter item response theory model is an important issue. It has long been know that the accurate estimation of these parameters is a difficult problem, and this is especially true for easy items. For example, Lord (1968) wrote, "On easy items on which even the poorest examinees do better than chance, the value of c is poorly determined by the data. On such items, the estimated value of c was arbitraily (sic) fixed at the chance level, 0.20....Values of c below 0.20 may arise because examinees prefer a wrong distracter in an item to a right answer" (p. 1011). Later in the same article, Lord wrote, "It was found, however, that many or most values of c_i are rather poorly determined by the data. . . . In the case of the easier items, on which even examinees with low verbal scores performed above the chance level of .20, c_i was arbitrarily fixed at .20. The errors in this procedure are surely much smaller for such items than the sampling errors in the maximum likelihood estimates, which might turn out to have some unreasonable value such as $c_i = .60$ or $c_i = -.10$." (p. 1014).

The above quotations are from the first published report of use of the three-parameter model on a published test. Eight years later, Wood et al. (1976, p. 18) wrote that Lord had shown in 1975 that "the c_i's are poorly determined when $(b_i - 2/a_i)$ is ≤ -2." After further study and use of the model, estimation of the c parameters was still a problem. To quote Lord again, he more recently (1980) stated, ". . .b_i is unstable when P_i or Q_i is small. The situation is much worse when both a_i and b_i must be estimated simultaneously. If only a few examinees answer item i incorrectly, it is obviously impossible to estimate a_i with any accuracy" (p. 185) . He went on to say, "The problem is even more obvious for c_i, which represents the performance of low-ability examinees. If we have no such examinees in our sample, any reasonable value of c_i will be able to fit the data about as well as any other. If we arbitrarily assign some plausible value of c_i and then estimate a_i and b_i accordingly, we shall obtain a good description of our data" (p. 186).

Several years later, Hulin et al. (1983) summarize a study by Ree done in 1979 by stating, "Clearly, even long tests and large samples do not necessarily allow accurate estimation of c" (p. 100).

Hutten (1980) explored the question of whether c parameters would even be estimated by LOGIST 4, the most widely used computer program for estimating the parameters using the three-parameter model. As we shall see, that program does not make estimates of c parameters for individual items when it seems that those estimates might be subject to large errors or lack of stability. In those cases, the program gives the item a sort of average c value, and many unestimated items may be given that same value for the c parameter. Using twenty-five data sets of 1000 examinees on each of a number of widely used standardized tests, such as the Stanford Achievement Series, the Comprehensive Tests of Basic Skills, the Iowa Tests of Basic Skills, the California Achievement Tests, and the Scholastic Aptitude Test of the College Board, Hutten concluded that "Although there are many indicators that ICC's are characterized by non-zero (sic) lower asymptotes, only one-third of guessing parameters were estimable by the LOGIST program for samples of 1000 examinees" (p. 30). She suggested that ". . .it may be more practical to set the lower asymptote to some reasonable value without attempting more refined estimation" (p. 30).

Thissen and Wainer (1981) studied the standard errors that would be obtained for estimates of item parameters in the three-parameter model if unrestricted maximum-likelihood estimation procedures were used. They concluded that ". . . the use of an unrestricted maximum likelihood estimation for the three-parameter model either yields results too inexact to be of any practical use, or requires samples of such enormous size so as to make them prohibitively expensive. This problem arises for items that are easier than average. This effect is the result of the huge covariance. . . between location and lower asymptote. When an item is relatively easy there are few observations available to estimate the lower asymptote thus making its standard error very large. The large covariance between lower asymptote and location then causes this uncertainty to move partially to the estimate of location. . . . The two-parameter model has problems as well, but they are far less severe" (p. 7).

They go on to say, "The estimates [of location] are hopeless for easy items. This is fortunate, for it is on items like this where the guessing parameter is not required. This provides hope that a hybrid model that computes a lower asymptote for difficult items and doesn't for easy items may be useful when the assumption of no guessing does not seem to be plausible" (p. 9). The situation is not entirely hopeless, they point out, since even with these large standard errors, using the entire ICC to reproduce the data can be useful. However, the problem is most severe when the location parameters are to be used separately, such as their use in test equating/linking or in computerized adaptive testing. To display the magnitude of the problem, they point out that if the lower asymptote is not homogeneous, i.e., the c values for all the items are not the same, accurate easiness values for easy items are obtainable only with Ns of 100,000 or so.

As Thissen and Wainer point out, one solution to the problems of the large standard errors of unrestrained maximum-likelihood estimation is to abandon the unrestrained approach. A Bayesian scheme would be appropriate, with prior estimates of the parameter values constraining the posterior estimates. This is essentially what the computer program LOGIST 4 does. De Gruijter (1984) points out that this, in itself, may result in biased estimates, but that estimates are biased if the lower asymptote is fixed, also. However, De Gruijter concludes that ". . . it is clear that reasonable parameter estimates for the three-parameter model are possible, using prior and collateral information on the lower asymptote" (p. 272).

Wood et al. (1976) describe the constraints in LOGIST 4 as follows:

> To avoid wild fluctuations in the paramater estimates, the amount that each parameter can change is restricted as follows: . . . [the absolute value of the change in c_i is equal to or less than] CLAMBDA, where CLAMBDA is read in or set to the default value of .06. . . . Since many of the c's are poorly determined, any movement is severely restricted. The restrictions are based on an approximation to the standard error of c_i . . ., on an approximate median value for all the $\hat{c}$ denoted by $\bar{c}$, on the proportion correct adjusted for omits . . ., on the value of $(b_i - 2/a_i)$, on a minimum $c_{min} > 0$, and on a maximum $c_{max} < .5$. The restrictions on the range define bounds on c_i that c_i may exceed only if its standard error is small. The approximation to the standard error is the square root of the diagonal element for c_i in the inverse of T_i.
>
> An absolute minimum of 0 and maximum of .5 is placed on c_i. In Lord (1975) it is shown that the c_i's are poorly determined when $(b_i - 2/a_i)$ is ≤ -2. The c_i's are better determined when $-2 < (b_i - 2/a_i) < -1$, and the c_i's should be well determined when $(b_i - 2/a_i) > -1$. Therefore if $c < p_i$ then c_i is automatically held fixed at c if $b_i - 2/a_i$ becomes less than FEPS (-2, or a user supplied value) (pp. 16-18).

The $\bar{c}$ in that quotation is described several pages later in the same document as follows:

> A good value for c_{max} is 1/A (number of choices per item), for $\bar{c}$ is c_{max} - .05, and c_{min} should be the same distance below $\bar{c}$ as c_{max} is above $\bar{c}$. In each stage where all of the items are estimated, the program computes the median of the values (for the first iteration of each item) to which $\hat{c}$'s currently allowed to vary would move if they were completely unrestricted. If the new median differs from $\bar{c}$ by more than ± .005, $\bar{c}$ is changed to the new median and c_{min} is adjusted to keep c_{min} the same distance below the new $\bar{c}$ as c_{max} is above the new $\bar{c}$. (p. 20)

The criterion (b-2/a) is called the stability criterion, and it is described in the user's guide for a later version of LOGIST (Wingersky, Barton, and Lord, 1982) as follows:

> The stability criterion (b-2/a) is the ability level at which the proportion of correct responses is only about .03 above the lower asymptote ("c"); if there are few examinees with ability estimates below this level, then a stable estimate of "c" cannot be obtained. In other words, we cannot estimate well the "c" parameter for very easy items, or for moderately easy items with low "a" parameters. . . .A value of -3.5 for CRITFIXC is suggested for samples of 2000 to 3000 examinees; a value of -2.5 is reasonable for smaller samples. (pp. 20-21)

Nature of the SSAT II Tests

The tests of SSAT II are very easy tests. The distribution of scores for both the mathematics test and the communications test are highly negatively skewed, as can be seen in Figure 5.1. The skew index for communications was -2.4, and the index for mathematics was -1.1. This suggests that we will have difficulty estimating c values for these tests by means of LOGIST 4. Supporting that premonition is Eignor's experience (personal communication, Summer, 1984) with the New Jersey minimum competency tests. When an attempt was made to use LOGIST 4 to estimate item parameters for the New Jersey tests, severe convergence problems were encountered. Even fixing c parameters, or setting them to zero, did not alleviate the problem. If the program does estimate c values, it may simply give most or all of the

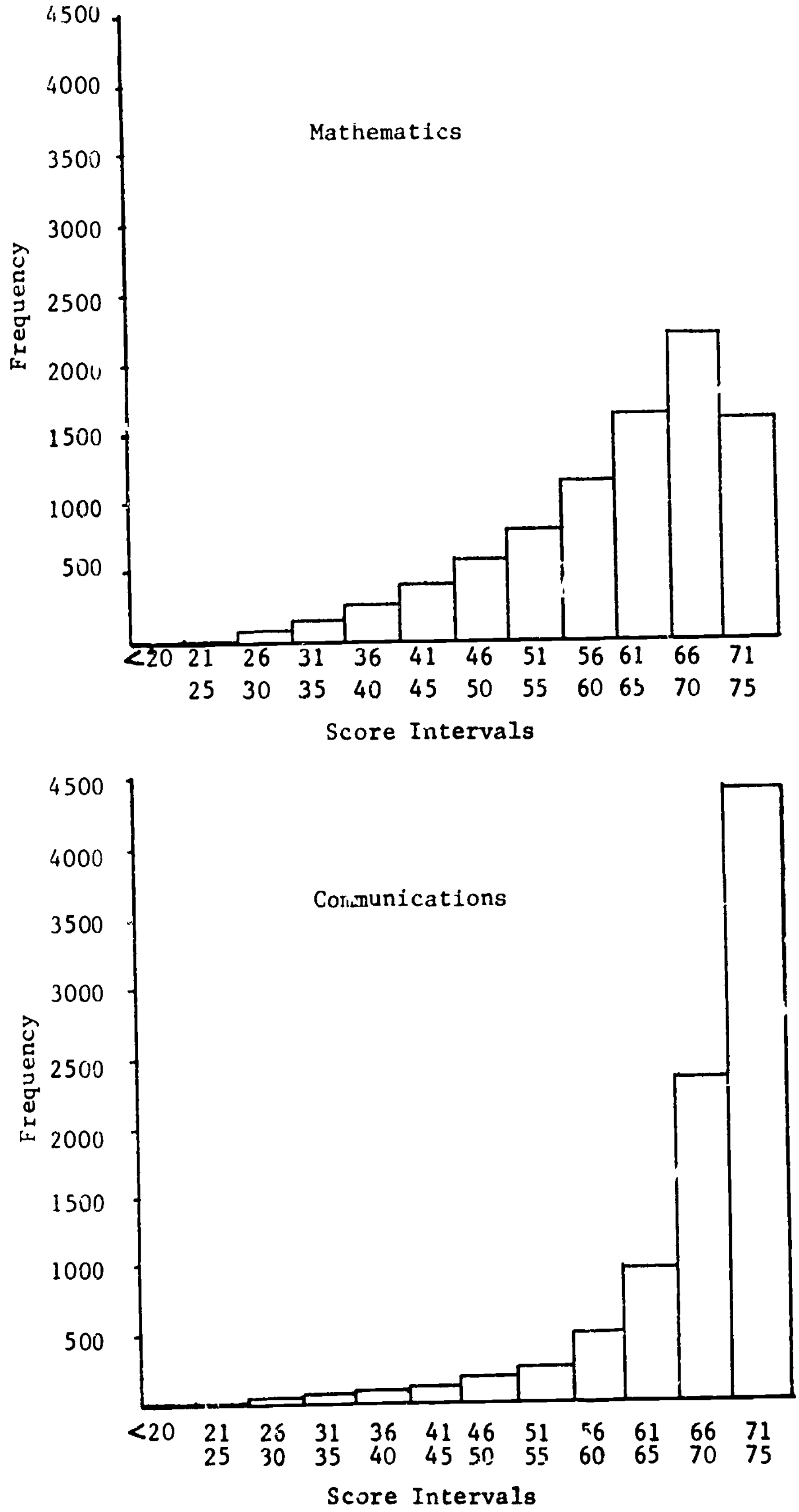
Mathematics
Frequency
4500
4000
3500
3000
2500
2000
1500
1000
500
<20
21 25
26 30
31 35
36 40
41 45
46 50
51 55
56 60
61 65
66 70
71 75
Score Intervals
Communications
Frequency
4500
4000
3500
3000
2500
2000
1500
1000
500
<20
21 25
26 30
31 35
36 40
41 45
46 50
51 55
56 60
61 65
66 70
71 75
Score Intervals

Figure 5.1

items the $\bar{c}$ value described ab^ve.

Several authors have suggested that one might oversample at the low end of the distribution in order to have large numbers of examinees with low scores and thereby sufficient data to estimate with some degree of accuracy the c parameter, the probability of someone of very low ability answering correctly, Lord (1968) suggested this, as did Thissen and Wainer (1981).

Another possibility that might be explored is to reduce the critical value of c below which $\bar{c}$ is used by the program instead of an estimate of c for the individual item. In fact, for LOGIST 5, Wingersky et al. (1982) suggest setting CRITFIXC at -3.5 for samples of 2,000 to 3,000 instead of using the value of -2 recommended for LOGIST 4.

One more possibility exists for obtaining about the same effect as lowering the critical value of c. That is to standardize on b parameters instead of on thetas. The usual procedure is to standardize on thetas, since the scale is arbitrary. However, one has the option of standardizing on b parameters. Since these items are very easy for the examinees, standardizing on thetas results in a negative average b parameter. If the average b parameter is already -1, not many of the easier items will fail to meet the critical value of (b-2/a), and thus few individual cs will be estimated. If we standardize on bs, the average b value is automatically zero, and more items might have estimated cs above the critical value.

Of course, the problem with these approaches is that if we simply arrange to have more c_is estimated but the estimates are unstable, as Lord (1975) has indicated, we may not have profitted much from the exercise. Still, as far as we know, explorations of these issues have not been reported in the literature, so their study here has not only practical importance for SSAT II analyses of the future, but also has a more general significance for others who might attempt to employ the three-parameter model using LOGIST 4 as the estimating program for the item parameters.

Therefore, this study was aimed at determining how effective the various strategies of:

1) oversampling at the low end of the distribution,
2) using a lower critical value of c, i.e.,(b-2/a < -2), and
3) standardizing on b rather than theta,

singly and in combination, would be at causing more c_is to be estimated and fewer $\bar{c}$s to be reported.

In addition to attempting to obtain more $\hat{c}_i$s, we wanted to evaluate how well the a, b, and c values agreed from one analysis to another. One might be distressed if he found that using one or more of these techniques resulted in a much higher proportion of estimates of c values for individual items, but the corresponding a values did not agree with a values from other procedures, or corresponding b values did not agree with b values using other procedures. Furthermore, it is also possible simply to fix the c values at zero, or at some value that more reasonably reflects the effects of guessing and other extraneous variables, such as 1/A - .05. The former would be a two-parameter model, with a parameters reflecting item quality and b parameters reflecting item difficulty, and an assumption that no guessing existed--that a person of very low ability had zero probability of getting an item correct. The latter might be called a modified three-parameter model, indicating that three parameters are used, but the third is not estimated for individual items but is set at a common fixed value for all items based on the number of alternatives.

Data

Data from the March, 1984, administration of SSAT II were used in these analyses. A random sample of 3000 cases was taken from a systematic sample of 9000 cases from the total population of 94,261 usable cases. Another sample was drawn for each test from the 9000 case sample by eliminating the lowest-scoring 50 cases (assuming that their scores are so low as to represent failures in test administration rather than sound estimates of ability), and then taking the lowest-scoring thousand cases and a random sample of 2000 from the rest of the distribution. This procedure was believed to provide samples heavily overrepresenting the examinees of low ability on each measure, mathematics and communication.

Analyses

For the representative sample, LOGIST 4 was used to estimate item parameters for each test, communications and mathematics. First, the standard options on the program were used. These options standardize on thetas and set the criterion for (b-2/a) at -2. Then the same analysis was done standardizing on b parameters. Next, the same analyses were done setting the criterion for (b-2/a) at -3 and, again, at -4. For these analyses, it was observed that the limit on the maximum a parameter of 2 was inhibiting. A large number of a parameter values were fixed at the maximum allowed (2.0). Therefore, the maximum on a was raised to 3. This same set of analyses was repeated using the data which overrepresented the low scorers.

The results of these analyses were evaluated to determine which of the above procedures, or combinations of them, seemed most effective at yielding increased numbers of estimates of c_i for individual items. The most effective procedure was used, then, to provide input for new LOGIST 4 analyses of the same data for comparisons with the results obtained from fixing c values at 0.20, (1.0/A - .05) representing chance, and from fixing c values at zero.

The results from the various procedures used to increase the number of c s indicated that for the communications test, no arrangement resulted in estimating more than nine of the c parameters. Apparently that test is so easy for the examinees that the population contains insufficient members of low enough ability to estimate the probability that a person of very low ability will answer these items correctly.

For the mathematics test, a number of c parameters were estimated using these devices. The number is still fewer than half (only about one-third), so the results are not entirely satisfying. Changing the standard for (b - 2/a) to -4, by itself, was about as effective as -3, but better than -2. Using a criterion of -2 resulted in estimates of 12 individual c parameters, or 16%, but using -3 resulted in estimates of 16 individual c parameters, or 21%. Standardizing on bs instead of thetas was about as effective as changing the standard to -3 or -4. Using the standard of -4 and weighting the low ability end of the sample by taking 1000 lowest scoring students and a random sample of 2000 from those scoring above these 1000 cases resulted in the largest number of c parameters being estimated. So the c parameters from using a standard of -4 and the nonrepresentative sample were input as fixed values to LOGIST 4 as one of the procedures to be compared. Since we were unable to find a way to estimate a useful number of c parameters for the communications test, it was not further studied.

To compare ICC parameters resulting from fixing cs at different values, we contrasted results from the following:

A. Use c parameters based on -4 as the criterion for (b - 2/a) and a nonrepresentative sample as fixed values for a LOGIST run--a three-parameter model.

B. Fix c parameters at (1.0/NCH - .05), i. e., at essentially a chance value--a modified three-parameter model.

C. Fix c parameters at zero, i. e., as they would be fixed in a two-parameter model.

Results

As can be seen in Table 1, the mean $\hat{a}$ values and mean $\hat{b}$ values and their standard deviations were very similar when the c parameters were fixed at values obtained from LOGIST and when c parameters were set at chance level (.2). The average $\hat{a}$ and $\hat{b}$ values and their standard deviations setting c equal to zero were noticeably different.

Table 5.1

Means and Standard Deviations of a and b Parameters When Fixed c is Input from LOGIST, Set at .2, and Set at Zero

	Logist Estimated $\hat{c}$		c set at .2		c set at 0	
	a	b	a	b	a	b
Mean	.915	-1.385	.912	-1.370	.831	-1.578
S.D.	.325	1.201	.324	1.190	.302	1.056

The correlations between estimates of a parameters using c parameters fixed in three different ways can be seen in Table 5.2 where "Input c" is obtained by fixing c at the level estimated by a previous run of LOGIST 4 setting the b-2/a criterion at -4 and oversampling low scoring examinees.

Table 5.2

Correlations between Parameter Estimates Using Different Procedures for Fixing c Parameters

	1	2	3
1. Input c		.97	.82
2. c = .2	.97		.87
3. c = 0.0	.82	.87	

The correlation between a parameters when c is set at 0.20 and when c is input from LOGIST 4 estimates is very high, .97, but neither of these procedures for estimating c parameters results in a parameters which are highly correlated with those obtained from setting c equal to zero.

Frequency distributions of the differences between $\hat{a}$ values based on the three procedures for determining c parameters, and frequency distributions of the differences between b parameters based on the three procedures appear in Tables 5.3 and 5.4. These data clearly support the conclusion that the results from the three-parameter model and the modified three parameter model are quite similar, and they are both quite different from results from the two-parameter model.

Table 5.3

Frequency Distributions of Differences Between a Parameters Based on Three Procedures for Fixing c Parameters

Interval Midpoint	$a_{\hat{c}}-a_{c.2}$	$a_{\hat{c}}-a_{c0}$	$a_{c.2}-a_{c0}$
<-.10	2	6	6
-.10	5	4	4
-.05	9	12	7
0	48	14	15
.05	0	6	12
.10	7	6	3
.15	1	8	9
.20	0	6	5
.25	0	2	2
>.25	3	10	12

Table 5.4

Frequency Distributions of Differences Between b Parameters Base[illegible] Three Procedures for Fixing c Parameters

Interval Midpoint	$b_{\hat{c}}-b_{c.2}$	$b_{\hat{c}}-b_{c0}$	$b_{c.2}-b_{c0}$
<-.10	5	8	9
-.10	9	3	0
-.05	40	1	2
0	10	4	0
.05	3	7	3
.10	0	8	4
.15	0	4	10
.20	3	3	5
.25	3	11	10
.30	1	6	8
.35	0	4	11
.40	0	4	7
>.40	1	12	6

Conclusion

From these data and analyses, it seems apparent that the c parameters obtained from setting the standard of b-2/a at -4 while using a nonrepresentative sample weighted at the low end and from fixing the c parameters at chance level are very similar, but the results from fixing c parameters at zero differ appreciably from the other two procedures. Since fixing the c parameters at zero flies in the face of the fact that the items are multiple-choice items in which it is unreasonable to expect that a person of very low ability would have probability of zero of answering correctly, the most sensible recommendation appears to be the simple one of setting the c parameters at a value such as 1/A-.05 for all the items and relieving LOGIST 4 from having to estimate the c parameters at all. This recommendation is sound for the SSAT II mathematics test based on the above analyses; it is sound for SSAT II communications because no procedure was found that enabled LOGIST 4 to estimate c parameters for a significant number of items.

It should not be surprising that using 1/A-.05 as estimates for c values corresponds fairly well with our best estimates of c parameters by means of the LOGIST 4 program. LOGIST 4 at best estimated c parameters for only about 1/3 of the mathematics items, and it set the c parameters to a common value for the rest. The common value was not far from .20. So the c values for most of the items were not far from the c values one would set based solely on knowledge of the number of alternative responses.

The program LOGIST 5 includes among its improvements over LOGIST 4 modifications in the procedure for estimating c parameters. The common c ($\bar{c}$) is estimated by maximum likelihood procedures; the only restrictions on c are bounds of 0.0 and 0.5, and the fixing of c at a common value if b-2/a is less than a critical value which has been set; and, the c parameters which have been set at a common value due to violation of b-2/a are re-estimated in the last step of the program. Whether these modifications in LOGIST would decrease the problems of estimating c values for minimum competency tests to an appreciable degree deserves exploration. The program BILOG uses a Bayesian procedure to estimate item parameters. Bayesian procedures set constraints on item parameters in terms of prior probability distributions instead of the less formal constraints set by LOGIST 4. It is possible that a Bayesian approach would result in better estimates of c parameters in minimum competency data. This possibility also deserves exploration.

CHAPTER 6

DETERMINING APPROPRIATE SAMPLE SIZE AND DISTRIBUTION SHAPE

The Significance of Sample Size and Distribution Shape

The number of examinees which must be used to obtain satisfactory estimates of item parameters is a significant consideration in determining whether it would be wise for the State of Florida to use one model or another of the family of item response theory models. The fewer parameters that are estimated, the smaller the number of examinees needed for stable estimates of parameters. Also, some parameters are easier to estimate stably than others. Item difficulty seems to be rather easy to estimate satisfactorily, but the guessing parameter seems to be quite recalcitrant. Since the current procedures are to use pretest samples of around 5ØØ cases, and that does not seem to be onerous, a relevant question is whether the use of a more complex model than the one-parameter model would be feasible with samples of about that size. If not, an issue in adopting a more complex model would be the feasibility of obtaining larger pretest samples on which to obtain item parameter estimates.

The usual minimum sample sizes stated for the three-parameter model are 1ØØØ examinees and 4Ø items. Some computer programs require more, but these numbers are those usually suggested for LOGIST 4. However, if one were to consider a modified three-parameter model in which the c parameters were fixed at a specified value, such as chance, the program would have one of its troublesome estimation problems removed as well as having a smaller number of parameters to estimate. As a result fewer examinees might be needed. If one were to consider a two-parameter model, with all c parameters set at zero, the program might also require fewer cases. Experience with these data seems to suggest that estimation of c parameters is impossible for the communications test data and largely unproductive for the mathematics data (only about 1/3 of c parameters estimated at best). Therefore, it is important to see what savings in numbers of examinees are feasible by eliminating estimation of the c parameter, and whether it would be feasible to use pretest samples of about the size now being used (5ØØ) when estimating parameters using LOGIST 4 and a modified three-parameter model or a two-parameter model.

Another way to attempt to obtain stable estimates of parameters is to oversample the extremes of the population distribution so that more data are available for evaluation of the asymptotes of the logistic curve. In a sense, the problem of estimating c parameters is that there aren't enough low-scoring examinees to fix accurately the shape of the logistic function at the lower end. Lord (1968), for example, says that an item characteristic curve can be considered as the regression of item score on ability, and the problem of estimating a_i and b_i can be considered as a problem of estimating a regression. He says, "It is well known that data having a normal or bell-shaped distribution of the independent variable are not nearly as efficient for estimating a regression curve as data containing many extreme values of the independent variable. Thus, item characteristic curves can probably best be estimated by using a group of examinees with rectangular or bimodal distribution of ability" (p. 1017). In his analysis of the SAT Verbal score he used 2,682 examinees chosen from a larger group of 5000, using more low-scoring than high-scoring examinees "since the sampling error of low scores is high, due to random guessing" (p. 1018). Even then, he suggested that a still more nearly rectangular, or even U-shaped distribution, might have been preferable.

The distributions of SSAT II scores are highly skewed, as has been noted earlier. The communications scores are exceedingly skewed, much more skewed than the mathematics scores. A result we have seen from this is that LOGIST 4 is unable to estimate c parameters for the communications items. The question remains, however, if we sampled from the available population of examinees so that the low end was greatly overemphasized, would we be able to obtain reasonably sound estimates of the item parameters using smaller numbers of examinees? If so, and if the State Department of Education can identify schools or locations in which large numbers of examinees at the low and high ends of the ability scale can be pretested conveniently, oversampling might be a feasible procedure for obtaining sound item parameters efficiently.

Procedure

Before any of the studies described in previous chapters had been done, the following procedure was planned for this analysis. A matrix of samples would be assembled, one dimension being sample size, the other being sample mix. Sizes such as 3000, 2000, 1500, and 1000 would be used. Three different mixes of abilities would be crossed with these sample sizes. One mix would be the abilities that naturally occur, a second would overemphasize the extremes of ability by a factor of 3, and the third would overemphasize the extremes of ability by a larger factor. It was planned that these factors might be modified depending on the kinds of distributions that naturally occur. The data would be analyzed in each cell using LOGIST 4 in an effort to locate the optimum sample size and mix for future analyses of Statewide Assessment Test data.

Experience with other phases of this project guided us in revising our procedure. First, our previous analyses suggested that the estimation of c parameters might not be practical for these data. Therefore, we decided not to pursue the estimation of c parameters throughout all levels of analyses for these steps unless the initial ones were encouraging. Second, it appeared that it would be necessary to locate fairly precisely the minimum number of cases required to make reasonably precise and stable estimates of a and b parameters. The logistics involved in the current trial administration of items argues for small samples. It is relevant to obtain information on whether the current sampling procedures for trial administrations can be followed unaltered while using a more complex model. Third, the extreme skew made oversampling of different degrees very difficult. In Figure 5.1, for instance, it can be seen that the lowest-scoring 1000 out of 9000 cases would include all the students who had answered correctly from zero to approximately 60 out of 75 items on the communications test while the highest-scoring 2000 would all have scores of 74 or 75. To use the lowest-scoring 1500 would include some rather high scores in the "low-scoring" group. The other alternative would be to sample from a larger number than 9000 which becomes rather cumbersome.

As a result of these considerations we did LOGIST analyses for the following sets of data:

> For mathematics, using the existing distribution we analyzed representative samples of 3000, 1000, 750, 500, and 250. We used both LOGIST-estimated c parameters and fixed c parameters (c parameters fixed at .20) for the samples of 3000 and 1000, but fixed c parameters only for the samples of 750, 500, and 250. We oversampled the low end, and did LOGIST analyses of 3000 cases with both estimated and fixed c parameters. For a sample of 1000 we did an analysis with estimated c parameters to verify our anticipation that few if any c parameters would be estimated even with oversampling at the low end.
>
> For communications, we did not believe that anything we could do would result in useful numbers of c estimates. Therefore, we did not bother with oversampling at all or with estimating c parameters at all. We merely tried to see what would happen to the estimates of a and b parameters as the sample size decreased, with c fixed, using samples of 3000, 1000, 500, and 250 cases.

Results

We compared the means, the standard deviations, and the intercorrelations of a and b parameters using different methods of estimating in order to see what happened as sample size became smaller and as we oversampled the lower end. We assumed that the analyses using 3000 examinees and estimating c parameters were the most sound, and other procedures would be evaluated in terms of how closely their results approximated those of N=3000. As a result of work on other aspects of this project, we had available one analysis on N=3000. We drew another random sample of 3000 from the 9000 sample so that two samples of N=3000 could be compared as a base line. These two samples are labeled N3000A c-est and N3000B c-est in Table 6.1. We also ran LOGIST 4 on these samples with c parameters fixed at .20. Those two analyses are labeled N3000A c-fix and N3000B c-fix. A final sample of 3000 cases was analyzed, N3000-S, c-est, in which the lower end of the distribution was dramatically oversampled following Lord's suggestion.

It can be seen in Table 6.1 that there is quite good agreement between random samples using the same method with Ns of 3000, and that estimating c parameters or fixing c parameters at .20 makes little difference in the means and standard deviations of a and b parameters. However, as one might expect, oversampling at the low end increases the mean b parameter. (Since the scale is fixed so that the mean ability of the group is zero, and since the mean ability is lower with greater sampling at the low end, the difficulty of the items scales out at a higher value, i. e., less negative.) The mean a parameter is also larger when low abilities are oversampled. Notice that the mean c parameter value is only .18 with low abilities oversampled, but, when c parameters are estimated, the mean c parameter is 0.24 when a representative sample is used. A lower c parameter allows the item characteristic curve to follow a steeper shape before becoming asymptotic, and the a parameter of the curve is thus larger. (Consider that if the c parameter were some large value such as .90, no a parameter could be very large. All slopes would be flat--and the item could do little discriminating, indeed, with all that guessing going on.) Thus, the baseline data on samples of 3000 behaves appropriately.

Now when we move to representative samples of 1000, the means and standard deviations of a and b parameters are much like those for representative samples of 3000, whether c parameters are estimated or fixed at .20. With oversampling at the low end, however, the mean a parameter is increased noticeably (from .9 to 1.2), and the a parameters are more variable (standard deviations increasing from .3 to .4.) More dramatically, the mean b parameters are increased noticeably more than they were for the oversampled group of 3000. Instead of the mean b parameter of -.8 we have a mean b parameter of about -.3 with reduced variability. So oversampling results not only in disagreement with representative sampling of N=3000, but also in disagreement with the oversampled data for N=3000. Oversampling the low end with N of 1000, thus, seems to produce results unlike

Table 6.1

Means and Standard Deviations of the Item Parameter Estimates

Sample	a	b	c
N3000A c-est	.93	-1.32	.24
	(.33)	(1.20)	(.02)
N3000B c-est	.91	-1.39	.24
	(.33)	(1.31)	(.02)
N3000A c_fix	.91	-1.37	.20
	(.32)	(1.19)	(.00)
N3000B c-fix	.89	-1.43	.20
	(.32)	(1.25)	(.00)
N3000-S c-est	1.02	-.85	.18
	(.35)	(1.06)	(.03)
N1000 c-est	.88	-1.48	.19
	(.31)	(1.29)	(.01)
N1000 c-fix	.87	-1.47	.20
	(.31)	(1.30)	(.00)
N1000-S c-est	1.15	-.30	.18
	(.42)	(.85)	(.05)
N1000-S c-fix	1.15	-.27	.20
	(.40)	(.85)	(.00)
N750 c-fix	.94	-1.38	.20
	(.35)	(1.35)	(.00)
N500 c-fix	.89	-1.52	.20
	(.34)	(1.60)	(.00)
N250 c-fix	.90	-1.89*	.20
	(.43)	(3.08)	(.00)

*Based on 74 items. Item 36 deleted from calculations due to the extreme value estimated for the b parameter.

Note: "c-est" means that LOGIST estimated the c parameter, and "c-fix" means that the researcher fixed the value at a constant of .20.

Note: "S" after the sample size indicates skewness different from parent population (oversampling of lower end of ability).

Note: "A" and "B" indicate two equal-sized random samples.

those obtained with larger samples and seems unattractive as a procedure for obtaining sound estimates of a and b parameters.

When sample size is reduced to 750, c parameters being fixed since they are unlikely to be estimated stably with such small samples, the mean a parameters are quite similar to those for N=3000. So are the a standard deviations, and the mean b parameters. The standard deviation of the b parameters is perhaps a little larger than for N=3000 (1.35 vs. 1.19 or 1.25). For N of 500 and N of 250 the mean a parameters are quite similar to the sample of 3000, and the standard deviation differs only for N of 250, and only a small amount there (.43 vs. about .33). However, the mean of the b parameters drops to -1.5 instead of about -1.4, and more noticeably, the standard deviation increases from about 1.2 to 1.6 for N=500 and 3.1 for N=250.

Table 6.2 shows the correlations among most of these estimates of a parameters. The two samples of N=3000, one with c estimated and the other with c fixed correlate .97, encouraging one to believe that one set of results could be substituted for the other with impunity. When representative sampling is used, the correlations with similar procedures between N=1000 and N=3000 are both .91. The same value results when oversampling is used but other procedures are similar between the sample sizes. When c is fixed, the correlation between the results from N=3000 and N=750 is 0.88. For N=500, the correlation is .84, and for N=250 it falls to .75. One might wonder if the relationship is satisfactory when N is below 750, or, at most 500.

Table 6.3 shows similar statistics for the b parameters. Here the correlations remain satisfactorily high until N becomes as small as 250.

Table 6.4 shows the correlations among the c parameter estimates for those samples in which c parameters were estimated. When representative sampling was used, the correlations among c parameters are very high (.99) for large samples (both N=3000). But between N of 3000 and N of 1000, the correlations drop to .61 and .55. The correlations between analyses using oversampling at the low end for one analysis and representative sampling for the other are disappointingly low, all below .40. These results add to the results of other steps in suggesting that for these minimum competency tests, use of an estimated c parameter may not be very helpful.

From these analyses we might conclude, then, that for the mathematics test, oversampling at the low end seems to result in parameter estimates that disagree markedly with those from taking a large representative sample. Using LOGIST 4 estimates of the c parameter seems unpromising for

Table 6.2

Pearson Correlation Coefficients for Mathematics a Parameters

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	3000	3000	3000	1000	1000	1000	1000	750	500	250
	A	A	S			S	S			
	est	fix	est	est	fix	est	fix	fix	fix	fix
(1)	1.00									
(2)	.97	1.00								
(3)	.92	.90	1.00							
(4)	.91	.86	.91	1.00						
(5)	.90	.91	.93	.96	1.00					
(6)	.79	.79	.85	.78	.82	1.00				
(7)	.74	.76	.87	.79	.82	.90	1.00			
(8)	.85	.88	.85	.85	.89	.75	.75	1.00		
(9)	.85	.84	.85	.92	.94	.77	.77	.83	1.00	
(10)	.76	.75	.69	.75	.79	.68	.59	.84	.81	1.00

Note: All coefficients are significant (p=.001)
Note: Only samples "A" of N=3000 are shown.

these minimum competency test scores. For the b parameters, one might comfortably use representative samples of as few as 250, but for the a parameters he might more safely stay above 750, or at least 500, examinees.

For the communications scores, remember that we were unsuccessful at estimating c parameters using LOGIST 4 for these very easy items. So we did not oversample at the low end for these analyses, nor did we estimate c parameters. We merely fixed c at .20, and then did analyses on samples of 3000, 1000, 500, and 250 cases. The results for a and b parameters appear in Table 6.5. For these data the mean a and b parameters were near the same value for Ns of 3000, 1000, and 500, but changed appreciably for N of 250. However, Table 6.6 indicates that the correlation between a parameter estimates becomes rather low when N decreases below 1000. A similar result appears in Table 6.7 for the b parameters.

Table 6.3

Pearson Correlation Coefficients for Mathematics b Parameters

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	3ØØØ	3ØØØ	3ØØØ	1ØØØ	1ØØØ	1ØØØ	1ØØØ	75Ø	5ØØ	25Ø
	A	A	S			S	S			
	est	fix	est	est	fix	est	fix	fix	fix	fix
(1)	1.ØØ									
(2)	.99	1.								
(3)	.98	[illegible]	[illegible]							
(4)	.95	.96	.99	1.ØØ						
(5)	.95	.96	.99	.99	1.ØØ					
(6)	.95	.96	.97	.96	.96	1.ØØ				
(7)	.95	.96	.97	.97	.97	.99	1.ØØ			
(8)	.86	.87	.89	.91	.91	.91	.91	1.ØØ		
(9)	.95	.94	.92	.91	.91	.87	.87	.77	1.ØØ	
(1Ø)	.83	.81	.84	.85	.85	.73	.74	.68	.9Ø	1.ØØ

Note: All coefficients are significant (p=.ØØ1).
Note: Only samples "A" of N=3ØØØ are shown here.

From these analyses, it seems that with these very easy minimum competency tests (skew indices below -2.2, and minimum item p value of .74), not only are LOGIST 4 estimates of c not feasible, but estimates of a and b parameters are shaky unless based on Ns of at least 1ØØØ.

Conclusions

We set out to determine whether by fixing c parameters or by oversampling at the lower end of the distribution it would be possible to estimate item parameters more accurately or with smaller numbers of cases than are needed with a full three-parameter model. We found slightly different results for mathematics and communications. For mathematics, oversampling at the low end has a noticeable effect on both a and b parameters. Both a and b parameters increase, on the average, and a parameters become more variable. The correlations between c parameters obtained with and without overrepresentation of low ability are quite

BEST COPY AVAILABLE

Table 6.4

Pearson Correlation Coefficients for Mathematics c Parameters

	N3000A	N3000B	N3000S	N1000	N1000S
N3000	1.00				
N3000B	.99 (p=.001)	1.00			
N3000S	.39 (p=.001)	.38 (p=.001)	1.00		
N1000	.61 (p=.001)	.55 (p=.001)	.27 (p=.001)	1.00	
N1000S	.18 (p=.122)	.18 (p=.117)	.70 (p=.001)	.09 (p=.437)	1.00

Table 6.5

Means and Standard Deviations (in parentheses) of Communications ICC Parameter Estimates*

Sample	a	b
N3000	.831 (.309)	-2.090 (.837)
N1000	.841 (.293)	-2.102 (.776)
N500	.855 (.313)	-2.076 (.763)
N250	.911 (.374)	-1.825 (.825)

*c parameter statistics not included because c parameters were fixed at 0.20 for all Communications LOGIST runs. Therefore the mean is .20 with a SD of zero in all cases.

Table 6.6

Pearson Correlation Coefficients
for Communications a Parameters

	N3000	N1000	N500	N250
N3000	1.00			
N1000	.90	1.00		
N500	.76	.81	1.00	
N250	.73	.71	.62	1.00

All coefficients statistically significant (p=.001)

Table 6.7

Pearson Correlation Coefficients
for Communications b Parameters

	N3000	N1000	N500	N250
N3000	1.00			
N1000	.96	1.00		
N500	.85	.86	1.00	
N250	.88	.84	.77	1.00

All coefficients are statistically significant (p=.001)

low, below .40. Thus, if one accepts the results from representative sampling of N=3000 as being the most sound, oversampling results in unsatisfactory results. It is possible and plausible that the results from oversampling are the most sound, however, and that the results from representative sampling are inferior. Which is most sound seems best answered by studies using simulated data with known item parameters like those found for these items. Such a procedure was not included in this research.

For mathematics, decreasing sample size without oversampling seems quite satisfactory down to N=1000, and even to N=750. The correlation among a parameter estimates (.88) is reasonably high, and the estimation of b parameters is quite satisfactory with N=750. However, the standard deviation of the b parameters starts to increase, and the correlation between a parameters to decrease (.84) at N=500, causing one to consider samples of that size too small for estimating parameters upon which to do such things as equate item difficulties for a pool of items. Since that is one of the main anticipated uses of item parameters, these results suggest that for mathematics such parameters be estimated on no fewer than 750 cases representing the target population.

For communications, not only were c parameters rarely individually estimatable, but oversampling did not help, and when sample sizes became smaller than N=1000, correlations of both a and b parameters with their counterparts at N=3000 were too low to be encouraging. Thus, for communications one would recommend larger experimental groups than for mathematics for obtaining item parameters, with Ns of 1000 or more being appropriate. Presumably this increased sample size requirement is a function of the fact that the communications items are so very easy.

For both communications and mathematics, these data lead us to the conclusion that estimating c parameters is probably not worth the trouble with these minimum competence data. Most of the c parameters will be set a common value which will be near the value at which one would fix them at in advance. Thus, little is to be gained by estimation.

CHAPTER 7

CONCLUSIONS

This set of studies was directed at answering some basic questions about the feasibility of using a three-parameter IRT model for the Florida Statewide Assessment Tests in place of the one-parameter model currently in use. Several basic questions were apparent. First, would the available computer program, LOGIST 4, converge when applied to these data which are so highly skewed due to the very easy items appropriate for minimum competency tests? The answer to that question is yes.

Second, we wondered whether the data were unidimensional as is assumed clearly by item response theory and implicitly by conventional theory. Again, the answer is in the affirmative. We see no indication that the data are sufficiently multidimensional to pose a problem.

Third, the data should theoretically fit a model more closely if the model contained more parameters. Indeed, these data fit a model which includes the discrimination parameter appreciably better than one which ignores that parameter. Including a guessing parameter, however, provided little gain in fit, perhaps because the items are so easy that there is little guessing and because estimation of guessing parameters is hazardous on such easy items.

Fourth, we tried several approaches to improve the estimation of the guessing parameters of the items--at least to result in individual guessing parameters being estimated for larger numbers of items. These approaches were generally unsuccessful. For mathematics, at best guessing parameters were estimated for only one third of the items. For communications, rarely were guessing parameters estimated for individual items. Thus, estimation and use of different guessing parameters for individual items is not really feasible for these tests using LOGIST 4. Using a common guessing parameter for all items, based on the number of alternatives contained in the item, is the sound way to approach this aspect of use of the three-parameter model. It appears wiser to do this than to resort to a two-parameter model in which it is assumed that an individual has no chance of answering a multiple-choice item like these by guessing. It may be that LOGIST 5 or BILOG would be more successful in estimating guessing parameters, but to our knowledge no one has tried those procedures on data similar to these.

A fifth question is whether modest sized samples, 500 or so, as are currently used in obtaining estimates of item parameters for the one-parameter model in the Statewide Assessment Tests, can be used satisfactorily for estimating item parameters in a more complex model. The answer seems to be negative. One might use as few as 750 for the mathematics test, but for the communications test we could not recommend using fewer than 1000.

Finally, it has been suggested that better estimates of the parameters might be obtained by oversampling low abilities. Our analyses were not encouraging. A serious problem is the extreme skew of the data. Unless one uses a huge number of examinees, there simply are not enough cases at the low end of the scale to be very helpful. Particularly for the communications test, oversampling at the low end did not result in more individual c parameters being estimated. A more satisfying approach to this problem might be followed, that of using synthetic data patterned after the data of the Statewide Assessment Tests but with known parameters. What procedure best estimates known parameters remains to be determined.

In sum, a more complex model is feasible. It will fit the data better. A modified three-parameter model would be recommended at this point, using the difficulty and discrimination parameters but setting the guessing parameters to a common value based on the number of alternatives. The samples used to estimate the parameters should be of approximately 1000 cases.

REFERENCES

Bejar, I. I. (1990). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17, 283-296.

Divgi, D. (1980, April). Dimensionality of binary items: Use of mixed models. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

De Gruijter, D. N. M. (1984). A comment on standard errors in IRT. Psychometrika, 49, 269-272.

Hambleton, R.K. & Swaminathan, H. (1985) Item response theory: Principles and applications. Boston: Kluwer-Nijhoff.

Hambleton, R. K., & Muray, L. N. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 71-94). Vancouver: Educational Research Institute of British Columbia.

Hambleton, R. K., Murray, L. N., & Williams, P. (1983). Fitting item response models to the Maryland Functional Reading Test results. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavior Research, 19, 49-78.

Hills, J. R. & Beard, J. G. (1984). An investigation of the feasibility of using the three-parameter IRT model in Florida's Student Assessment Program. Tallahassee, FL: Florida State University, College of Education.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

Hutten, L. R. (1980, April). Some empirical evidence for latent trait model selection. Paper presented at the meeting of the American Educational Research Association, Boston, MA.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28, 989-1020.

Lord, F. M. (1975). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. (Research Bulletin 75-33.) Princeton, N. J.: Educational Testing Service.

Lord, F. M. (198Ø) Applications of item response theory to practical testing problems. Hillsdale, N. J.: Lawrence Erlbaum Associates.

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

Lumsden, J. (1961). The construction of unidimensional tests. Psychological Bulletin, 58, 122-131.

Martois, J. S. Rickard, P. L. & Stiles, R. L. (1985, April). Item-response-theory-based life skills reading tests: their unidimensionality and the temporal stability of item difficulty estimates. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

McKinley, R. L. & Mills, C. N. (1984, June). The use of analysis of residuals to assess the unidimensionality of data. Paper presented at annual meeting of Psychometric Society, Santa Barbara.

Murray, L. N. (1985). Using residual analyses to assess item response model-test data fit. Unpublished doctoral dissertation. University of Massachusetts, Amherst, MA.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 2Ø7-23Ø.

Reckase, M. D. (1981) A comparison of procedures for constructing large item pools. (Report No. 81-3). Columbia, MO: Tailored Testing Research Laboratory, University of Missouri.

Thissen, D. & Wainer, H. (1981). Some standard errors in item response theory (Report No. 81-12). Princeton, NJ: Educational Testing Service.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). Logist user's guide: Logist 5, Version 1.Ø. Princeton, NJ: Educational Testing Service.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976) LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (Report RM-76-6). Princeton, NJ: Educational Testing Service.

Wright, B. D. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.

Wright, B. D. & Stone, M. H. (1977). Best test design. Chicago, IL: Mesa.