DOCUMENT RESUME

ED 266 153 TM 860 045

AUTHOR Hills, John R.; Beard, Jacob B.
TITLE An Investigation of the Feasibility of Using the Three Parameter IRT Model in Florida's Student Assessment Program.
INSTITUTION Florida State Univ., Tallahassee. Coll. of Education.
SPONS AGENCY Florida State Dept. of Education, Tallahassee. Bureau of Program Support Services.
PUB DATE Sep 84
NOTE 46p.; For a related document, see TM 860 044.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Equated Scores; Feasibility Studies; Guessing (Tests); Item Analysis; Item Banks; *Latent Trait Theory; *Mathematical Models; Minimum Competency Testing; *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 the use of the three-parameter item response theory (IRT) model in Florida's minimum competency testing program. The paper includes the following sections: (1) a description of the procedures currently being used by the assessment program, with an emphasis on procedures currently involving the Rasch model; (2) a technical discussion of the differences between the two approaches; (3) a description of commercial and governmental testing programs now using the three parameter approach; (4) a description of the availability and adequacy of computer procedures for implementing the three parameter approach, (5) an analysis of the effect on the procedures currently used by the program if a move to the three-parameter program were made; and (6) a plan for implementing a try-out of the three parameters approach on Florida's assessment program data. (Author/LMO)

An Investigation of the Feasibility of Using the Three Parameter IRT Model in Florida's Student Assessment Program

Prepared by

John R. Hills - Jacob G. Beard

College of Education
Florida State University

Under Contract to the:

Assessment Section
Bureau of Program Support Services
Florida Department of Education

September, 1984

ABSTRACT

AN INVESTIGATION OF THE FEASIBILITY OF USING THE THREE PARAMETER IRT MODEL IN FLORIDA'S STUDENT ASSESSMENT PROGRAM

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. However, these improvements would entail some additional costs and complexities. The nature of these costs and complexities, and the potential advantages and disadvantages in using the model are explored in this paper.

The paper includes the following sections: a) a description of the procedures currently being used by the assessment program, with an emphasis on procedures currently involving the Rasch model, (b) a technical discussion of the differences between the two approaches, (c) a description of commercial and governmental testing programs now using the three parameter approach, (d) a description of the availability and adequacy of computer procedures for implementing the three parameter approach, (e) an analysis of the effect on the procedures currently used by the program if a move to the three parameter program were made, and (f) a plan for implementing a try-out of the three parameter approach on Florida's assessment program data.

AN INVESTIGATION OF THE FEASIBILITY OF USING THE THREE PARAMETER IRT MODEL IN FLORIDA'S STUDENT ASSESSMENT PROGRAM

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. However, these improvements would entail some additional costs and complexities. The nature of these costs and complexities, and the potential advantages and disadvantages in using the model are explored in this paper.

The paper includes the following sections: (a) a description of the procedures currently being used by the assessment program, with an emphasis on procedures currently involving the Rasch model, (b) a technical discussion of the differences between the two approaches, (c) a description of commercial and governmental testing programs now using the three parameter approach, (d) a description of the availability and adequacy of computer procedures for implementing the three parameter approach, (e) an analysis of the effect on the procedures currently used by the program if a move to the three parameter program were made, and (f) a plan for implementing a try-out of the three parameter approach on Florida's assessment program data.

Description of Current Procedures

The Florida Statewide Student Assessment Test (SSAT) Program currently uses the Rasch model of Item Response Theory (IRT) as the mathematical framework for its testing procedures. (The term "Rasch model" and "one parameter model" will be interchanged at various places in the paper. The use of one term or the other may be used to show origin of research discussed, a particular perspective, or may be a random choice of terms.) The Rasch model was chosen for use in the SSAT program for practical reasons rather than psychological. For example, from a psychological perspective there is little interest in assuming that the SSATs measure a unidimensional latent trait, an assumption of the Rasch and most other item response theory models. The tests have been interpreted as criterion-referenced, minimum

competency tests. Total scores are generated for each major subtest such as mathematics or communication skills; however, individual skill objective scores based on short tests of approximately five items are also reported to students.

From a practical perspective an IRT model such as that of Rasch has considerable appeal, and the Rasch model has been introduced into the program as an instrument for achieving certain goals. Prominent among these goals has been the generation of equated scaled scores for the SSAT Part II (SSAT-II). The Rasch model has been gradually introduced into the program alongside the classical measurement model. Whenever new uses have been proposed for the Rasch model, applied research has been done, using local data, to determine its applicability for the purpose. Its use has been rejected for some purposes, such as pre-equating for score reporting. However, it is used for:

1. Calibrating item difficulties.
2. Statistical descriptions of item performance in item banks.
3. Selecting items for inclusion in tests.
4. Equating test scores.

These tasks are done on a regular and routine basis. In addition, regular (non-Rasch) analyses of item data are made in order to detect potentially biased items. The procedures used in accomplishing each of the tasks are given in the following sections.

Calibrating Item Difficulties

Items proposed for use in the SSATs are first administered as experimental forms to samples of students in conjunction with the operational forms to determine their statistical characteristics. Rasch log difficulty and fit indices are computed along with classical indices of difficulty and discrimination. Frequency of responses for each alternative are also computed. The experimental forms contain from five to fifteen items each and are administered to approximately three to five hundred students in one to three schools. An adaptation of the BICAL computer program (Wright, Mead, and Bell; 1980) is used to do the analysis. The link to the base year is usually an indirect one composed of several intermediate links. For example, the scores of 1981 might be linked to those of 1980 which might have been equated to the 1978 reference year. The linking pattern is established at the time the test is built. The procedure of using common items for linking serves both the item calibration and test score equating function.

While items are initially calibrated in experimental forms, all items of each operational administration are

recalibrated to the base year scale. The recalibrations are done using the same set of common items and linking patterns as the experimental forms. The values from the operational administrations of forms are considered to be more stable than those from the experimental forms because the motivation of some students to perform on the experimental forms may be lower than on the operational form. The recalibration values are recorded in the item bank along with all previously obtained values. These initial and recalibration values are used in the selection of items for new forms of the SSATs.

The item calibration task has been done through contracts with the Florida State University (FSU). Tapes of item scores or responses are provided to FSU by the Department of Education (DOE) and FSU provides the DOE with the item calibration information and a report describing the procedures used.

Statistical Descriptions in Item Banks

The Rasch log difficulty and fit values are recorded in the item bank along with the classical difficulty (p), discrimination values, and frequency choosing each alternative. When items are re-administered in operational forms, their new statistical values are added to the bank information file.

Selecting Items

When a new form of the test is built, items are selected from the bank according to skill objective, item content, Rasch difficulty, item P value, discrimination, pattern of responses to alternatives, and fit to the Rasch model. The selection process has been informal, but the criteria are probably considered in the order given above. The item must necessarily match the identified skill objective, and in addition items are sought which a. are not redundant in content with those already included, b. provide variation in item difficulties for the skill, c. correlate highly highly with total scores, and d. fit the Rasch model. The staff have been uncertain about the criterion of model fit: however, the DOE has commissioned a study, currently underway, which should yield relatively specific guidelines for fit consideration.

The selection of items for new forms of the test considers both the classical and the Rasch IRT mathematical framework as well as the criterion referenced interpretational framework. These nearly orthogonal dimensions for classifying testing procedures are discussed by Bejar (1983). The mathematical framework is the set of assumptions about the statistical properties of the data. The interpretational framework refers to procedures used to interpret and validate scores obtained from the test. An

important implication of this dual basis for test construction is that item selection must consider item statistical quality as well as content and skill representation.

Equating Test Scores

The Statewide Student Assessment Test Part II(SSAT-II) scores are reported on an equated score scale. The scores are equated using Rasch methods and transformed to a scale in which the passing score is 700. The equating table is generated by the following procedure. As soon as a substantial majority of answer sheets has been scored by the scoring contractor, a tape containing the item responses is sent to the University of Florida (UF). The UF generates an equating constant based on a random sample of students. The constant is used to adjust the item difficulties of the test to the item bank scale. These adjusted item difficulties are then used to produce a raw score to log ability transformation table. The log abilities are transformed to the SSAT-II scale by the following expression.

$$Y=25(B_i-B_c) + 700$$

where Y= equated scaled score

B_i = log ability of student

B_c = log ability at minimum

passing score

The score transformation table is sent by UF to the scoring contractor who proceeds to prepare score reports.

The scores of each grade level are also equated to a base year by linear methods in order to assess longitudinal changes in statewide mean achievement. This equating also uses a common item design and the same common item links used in calibrating the items. The mean scores for each test and grade level are compared with those of previous years back to 1976 or 1977, depending on the grade level.

Item Bias Review. The items contained in Florida's Statewide Student Assessment Tests and Teacher Certification Tests are routinely reviewed for bias against racial and ethnic minority groups. The reviews consist of two major steps, in addition to those steps taken during the item writing process. The first of these occurs before the test is printed. The items are inspected for possible bias by a carefully constructed committee composed of representatives of the major and minor racial and ethnic groups living in the state of Florida. The second step consists of statistical analyses done after the tests have been administered. These analyses of item bias follow procedures

used by Hills and King (1981) in a study of item bias in Florida's Statewide Student Assessment Program. The procedures used are described in the following paragraph.

The method used is a form of the "transformed item difficulty" (TID) method. Ability is controlled by dividing the minority group's score distribution into ten or so intervals, and choosing at random from the same intervals among the majority group a number of students equal to the number of minority students in each interval. The items' difficulties are separately computed for each of the groups being compared. The difficulties for the two groups are cross-tabulated, and scatterplots of the data are constructed. A line is drawn through the center of the ellipse. The line is usually close to the 45 degree line since the groups have been roughly equalized on ability. Potentially biased items are those most distant from the hand-drawn line. Items identified as potentially biased are inspected by judges who attempt to explain the discrepancies.

Scatterplots are made of both item P values and delta values for the Student Assessment and Teacher Certification program data. However, there are insufficient numbers of the various minority groups taking the Teacher Certification test to permit experimentally controlling for ability level.

A foil analysis is also completed for both testing programs. The percentages of majority and minority students choosing each foil of each item are compared. Items having discrepant patterns of percentages are identified as potentially biased and referred to the panel of judges for inspection and ultimately for acceptance, rejection, or revision.

While the item bias review procedures are not currently done using Rasch procedures, there are implications for this area of activity in considering the three parameter model. These implications will be discussed in a subsequent part of the paper.

Additional Uses of the Rasch Model.

The Rasch model was used to pre-equate new tests for the 1984-85 standards, and for estimating frequency distributions of scores before the test was administered. These estimated frequency distributions of scores were used in estimating the proportions of students who would pass each of the tests.

The passing scores for the SSAT-II are currently stated in terms of Rasch log ability scores. The passing log scores are equivalent to a given percentage correct for a reference year administration.

Differences Between the One Parameter and Three Parameter Models

The purpose of this section is to describe the differences between the one and three parameter models. The description will focus on the purposes, assumptions, mechanics of use, and the results of use of the two models. Some of the differences discussed will extend beyond the technical formulations and will deal with the perspectives on each arising from their separate origins and uses.

Purposes

Rasch Model. The original purpose of Rasch's work with the one parameter model was to derive ability scores for persons which were independent of the particular test items used to generate the score, and to derive difficulty values for items which were independent of the particular persons taking the test. Rasch referred to this concept as specific objectivity and considered it analogous to Maxwell's analysis of the concepts of mass and force (Rasch, 1960). The concept has been interpreted by Wright (1968) as follows. "The model says that the outcome of such an encounter [between a person and an item] is governed by the product of the ability of the person and the easiness of the item and nothing more!" (p. 1) Choppin defined specific objectivity with the following example. "In the context of mental testing, it means that the comparison of two individuals who have been tested should be independent of which items were included in the tests." (Choppin, 1982, p. 1). In order to achieve specific objectivity a test developer would eliminate items from consideration that did not contribute to this purpose; for example, items reflecting excessive guessing and extreme discrimination. This search for a measurement model that would provide specific objectivity, an objective measurement system, was the driving force behind Rasch's work.

Three Parameter Model. The three parameter model was based on a theory of latent traits. This theory supposes that an individual's behavior can be explained through knowledge of the individual's status with regard to a number of traits. The purpose of testing is to make inferences about the examinee's trait or traits on the basis of responses to the test items. The relationship between trait status and response to a test item is described by an item characteristic curve(ICC); the relationship between trait status and responses to a multiple-item test is described by a test characteristic curve. Much of the research on the three parameter model is focused on the fitting of ICCs to test data.

Generally, one would expect to describe an ICC more accurately with three parameters, or two, than with one parameter as is done with the Rasch model. In fact, the

Rasch or one-parameter model can be considered to be a special case of the three-parameter model, the case in which there is no guessing and all items are equal in discrimination. Regardless of their separate origins and initial purposes, the Rasch and three parameter models can be used to perform similar functions. While the three parameter model does not achieve specific objectivity in the way Wright interprets it, i. e., guessing and differences in item discrimination, along with person ability and item easiness, determine a person's score, the benefits of item-free person measurement and person-free item measurement are obtained with the three parameter model. Furthermore, both serve as theoretical models for the construction of tests, and can be used to equate test scores, calibrate test items, examine test bias, and perform most analytic functions in testing programs.

Assumptions

The basic assumption of the Rasch model is that a person's response to a test item depends only on the ability of the person and the difficulty of the item. An implication of this assumption is that items and persons can each be uniquely ordered in terms respectively of their difficulties and their abilities. Several assumptions are implied by the basic one. The most frequently discussed are unidimensionality, local independence, equality of discriminations, and no random guessing. (Rasch did not attend to the latter two, probably because he did not consider use of the model with multiple-choice test items. The significance of these assumptions became clear when the relationship between the one- and three-parameter models became clear.)

The three parameter model assumes unidimensionality and local independence. The assumptions of equality of discriminations and no random guessing, required by the Rasch model, are unnecessary because the three parameter model incorporates these elements as parameters. Therefore, the three parameter model requires fewer assumptions than the one parameter or Rasch model.

Unidimensionality. The assumption of unidimensionality has similar implications for both models. The addition of the second and third parameters is not intended to account for multidimensionality in data except insofar as someone might think of guessing or variations in discrimination among items as "dimensions" or variables other than ability and item easiness which influence success on items.

While the assumption of unidimensionality appears plausible for stable general traits of individuals, some researchers have questioned its validity in the area of achievement testing (Choppin, 1982). Several recent studies have focused on the dimensionality problem in achievement

test data (for example, Phillips and Mehrens, 1984; Cook, Eignor, and Taft, 1984). It appears that the notion that achievement of basic school objectives may produce substantial multidimensionality in data is not supported. Both models appear to be robust with respect to the level of multidimensionality found in basic skills tests. Of course, it is possible that exceptions might be found, and it would be wise to investigate the dimensionality of data sets to which the models are to be applied.

The dimensionality assumption is primary in both the Rasch and three parameter models. In fact, most of the other assumptions can be considered to be implications of the unidimensionality assumption. Within the Rasch model, the concept of specific objectivity implies unidimensionality. Within the three parameter model, unidimensional latent traits are also assumed. This assumption is the basis for the fact that in the three-parameter model the item parameters do not depend on the group tested, and the ability estimates do not depend on which items are administered (Rasch's concept of specific objectivity). Furthermore, the classical measurement models also assume unidimensionality. Therefore, the assumption of unidimensionality is a basic and pervasive assumption in measurement and is not likely to be ignored in any measurement model. Elementary logical reasoning does not permit the placement of an object in two dimensions with one number.

When several dimensions are present and models which assume a single dimension are used, the one-parameter and three-parameter models do not necessarily behave similarly. Reckase (1981) has pointed out that the one and three parameter models yield scores that may differ in the constructs measured. If only one dimension is present, both models measure that dimension. If several dimensions are present, neither or both models may fit. If fit statistics are judged to be satisfactory, according to Reckase (1981), "...the one-parameter model yields ability estimates that have meaning equivalent to that of the number correct raw scores, but that are on a transformed scale. This results in an ability estimate that is based on the sum of the various components of the test, where the components are weighted by the number of items measuring the component. The three-parameter model yields an ability estimate with a different interpretation. Since the estimates are based on a weighted sum of item responses, the weights being the item discrimination parameters, and since the discrimination parameters are related to the first factor loading of a test (Lord & Novick, 1968), the three-parameter based ability estimates are only related to the largest component of the test" (p. 7). This difference may have important implications for Florida's Statewide Assessment program. Where a single score is reported, logically one dimension is being measured. One must ask whether there is but one large

factor among the items, and if it is satisfactory for the test score to measure the dominant factor? Or would it be more satisfactory for the single score to measure the amalgam of whatever factors are present among the items?

Local Independence. Local independence is assumed by both models and is an implication of the more basic assumption of unidimensionality. Local independence is violated when the response to one item affects the response to a second. A common violation of this assumption occurs when a mathematical test item requires the result of a previous item for its solution. An incorrect response to the previous item insures failure on the dependent item, thus the second item is dependent on the first and the assumption of local independence is violated. Another common violation occurs in reading comprehension tests when several different questions are asked about a single paragraph which has been read. A speeded test violates the assumption of unidimensionality in that failure on the last items is due to slow work as well as to ability to respond correctly given sufficient time. In this sense, the assumption of local independence is equivalent to that of unidimensionality in that a second factor or dimension other than the ability intended is being measured.

No Guessing. The assumption of no guessing, required by the Rasch model, may also be considered an implication of the assumption of unidimensionality. When one student consistently achieves higher scores than lower students of the same ability level by guessing items correctly, then another dimension, "propensity to guess", is affecting the score on the test. The three parameter model controls for the average amount of guessing on an item across individuals, but does not control for differences among individuals in their "propensity to guess." Errors in ability estimation will still be induced by the tendency of one student to guess more than another. For a discussion of this viewpoint see Traub (1983).

Item Discrimination. A major difference between the one and three parameter models is the inclusion of an item discrimination parameter in the three parameter model. The presence of items which discriminate differently can influence several aspects of measurement. First, the one parameter model may reject items as not fitting the model if their ability to discriminate between low and high performers on the test is much worse are much better than the average for the items in the test. The test developer who adhered to the one parameter model might eliminate from his test any very highly discriminating items, a practice which would be regarded as ineffective and detrimental by a follower of classical practice or a user of the three parameter model. This may not be as much of a problem as it would seem. In theory, all items must discriminate equally to fit the Rasch model, but in practice one must decide just

how deviant in discrimination items can be and still be included. Typically this has been done on the basis of goodness of fit tests. For example, Curry, Bashaw, and Rentz (1978) found that variation in item discriminations affected fit to the Rasch model little except when average discrimination for tests differed. It turns out (George, 1979) that the widely-used goodness of fit tests for the Rasch model are insensitive to items which deviate from the model by discriminating particularly well. (Some authors question whether the readily available tests of goodness of fit associated with the BICAL program are adequate. Hambleton and Murray (1983) suggest a large number of approaches to evaluating goodness of fit by checking model assumptions, expected model features, and model predictions of test results. It is not clear whether use of these tests of goodness of fit would also find little effect from violation of the assumption of equal discrimination.)

In addition, several authors have concluded that when guessing is not present, the one parameter model performs satisfactorily. It "maintains high relative efficiency until the range of discrimination became large" (Reckase, 1977, based on a study done by Hambleton and Traub, 1971). Panchapakesan (1969) found that the Rasch model item discrimination parameters could vary by .20 from the mean for the test and the model could be adequately used. (The BICAL program computes item discrimination parameters even though the model does not include an item discrimination parameter.) The discrimination parameters can be evaluated to see whether differences in discrimination may be causing the data to misfit the model. George (1979) has shown that highly discriminating items have a greater tendency than lower discriminating ones to fit, rather than misfit, the Rasch model.

However, some things can be done better in test development using three parameters than one. For example, one can plan a test characteristic curve that he wishes to obtain using the three parameter model, and considering the item characteristic curves, can choose items which will result in the desired test characteristic curve. This might be done in order to obtain a desired distribution of scores, given a known distribution of ability, as for example, a distribution of scores with a dip or hollow near a cutoff score, when the distribution of ability is normally distributed. The one parameter model cannot take into consideration the effectiveness of discrimination of items in trying to achieve the same objective, since it does not take this characteristic into consideration.

The three parameter model also permits measurement of ability using more of the information from items. The one parameter model simply says that the raw score, or number right, provides all the information that is needed about a person's performance. Everyone with the same number correct

is judged to be of the same ability level. The three parameter model says it matters which items a person gets correct. If a low ability person gets very difficult items correct, he should not get credit for them because he has probably done this be guessing. However, if a high ability person gets difficult items correct, the weight given those items should correspond to the ability of those items to discriminate. So a person's ability is estimated by the three parameter model according to the pattern of responses, not merely the number of correct responses.

One may argue that not much information is added by considering the pattern of responses. Dinero and Haertel (1976) concluded that number correct ability estimates were satisfactory for the one and two parameter models when the item discrimination values were not highly heterogeneous. They found correlations between the number correct scoring and scoring in which each response was weighted by the item's discrimination to vary from .81 to .98 depending on the shape of the distribution of item discrimination values. Lord (1968) similarly noted that the use of pattern scoring might add only slightly to validity, an effect of an improvement on the order of .01. However, he argues that such a slight improvement is not to be ignored. If one were to keep the scoring improvement that yields an increase in validity of .01 and shorten the test to bring the validity back to its former level, with a reliability of .90 and a validity of .51, he could throw away a random 30 percent of the test items. If another .01 can be spared, the test length could be cut in half using the improved scoring.

Mechanics of Use

Computer Programs. An issue in using either the one or three parameter model is the availability of computer programs to compute the required analyses, or the ease with which programs could be written. Computer packages for implementing each model are readily available and their characteristics well known. These packages include BICAL (Wright, Mead, and Bell, 1980), the basic parameter estimation program, as well as several adaptations of the program used for item calibration, equating, and item bias. Several programs are available within the state of Florida for performing support functions in the analysis environment; for example, pre-equating.

Computer programs for implementing the three parameter model are less readily available. Most analyses are currently being done with the LOGIST 4 program authored at Educational Testing Service (ETS) (Wood, Wingersky, and Lord, 1976). This program is complex and relatively expensive to run (Hutten, 1980). (However, in relationship to the total scope of expenditures of a state-wide testing program, the additional costs of running LOGIST versus BICAL may be relatively insignificant.) A revised version of this

program (LOGIST V) has been prepared. However, it contains sections written in machine language and has been implemented at few sites other than ETS to this date. It is assumed that it could be implemented given sufficient resources. The LOGIST program contains many options for performing such tasks as item calibration and equating. However, it has not been adapted for specialized local uses, for example, calibration and equating, to the extent that BICAL has. LOGIST 5 will perform many of these functions directly and LOGIST 6 is being designed for several common computer systems and will incorporate features to give it even greater flexibility of use. (Computer programs are described in a later section of this paper.) Use of the currently available LOGIST program will generally require considerable orientation and training and the development of supporting computer software.

Costs. The three parameter model costs more to implement than the Rasch model. The greater costs are associated with computer and personnel time. Hutten (1980) found the costs for estimating parameters with the three parameter model to be about three times that required for the one parameter model. One analysis might cost $5 using BICAL and $15 using LOGIST 4, for example. LOGIST 5 is said to reduce the costs of a 3 parameter model analysis by about 1/3. Also, computer costs have been decreasing and their direct costs may not be a crucial issue in the near future. Related increases in costs would stem from the weighted scoring of student responses, and from the employment of more highly skilled technical personnel.

The cost factors given by Hutten are for parameter estimation only and do not represent the full costs associated with the adoption of the three parameter model. A more comprehensive description of tasks involved when a three parameter model replaced traditional procedures in a large scale testing program was given by Yen (1983). Yen described the implementation of the three parameter model for form U of the Comprehensive Tests of Basic Skills (CTBS/U) published by CTB/McGraw Hill. She discussed item parameter estimation, bias analysis, model fit, anchor tests, estimating P values, item bank information, choosing items, test construction, equating, and test scoring. She included evaluative statements about the major tasks and about the overall undertaking. In summary, Yen noted that the change to the three parameter model from traditional procedures was a tremendous amount of work but the pay-off in clarification of psychometric issues made it worthwhile. Of course, less dramatic changes would be involved in changing from a one-parameter model to a three-parameter model since many of the major tasks Yen discussed are already included in use of the one-parameter model.

Item Selection. Both models offer significant advantages by making it possible to design tests of maximum

efficiency for particular purposes. Wright and Stone (1979) suggest a procedure for selecting items that is similar in concept to the one described below which was suggested earlier by specialists in the three-parameter approach.

Lord (1977) gives a procedure for selecting items for a new test using the three parameter model. He assumes that the items have been calibrated and that item information curves are available for each item.

1. Decide on the shape desired for the test information function.
2. Choose items having information curves that will fill areas under the target test information curve.
3. Cumulatively add up the item information curves.
4. Continue selecting items until the area under the target test information curve is satisfactorily approximated.

Reckase (1981) points out that the classical procedure of selecting items with high discrimination indices and appropriate difficulty values should also work well. However, one problem with the classical procedures for mastery tests is that the classical p values are not on the ability scale, so it is much more difficult to determine how to select the items that will measure most efficiently at the cutoff score. Further, the classical p values depend on the group tested. Item-response-theory difficulty estimates do not depend on the group as long as the test measures only one dimension.

Items are also selected for both models on the basis of item fit. However, the effects that selection of items on the basis of item fit will have on the characteristics of a test are not clear. Reckase (1981) reviewed several studies which investigated the relationship between item fit and test psychometric characteristics. He summarized the studies and the results of his own work as follows, "To summarize these results, there seems to be no good procedure for selecting items with the one-parameter logistic model. Not only do the fit statistics not work well, but no reason can be thought of for selecting items with discrimination parameters equal to the mean discrimination on the pool. Typically, use of the best items in a pool would seem desirable, ..." (p. 4). Our own investigation of the use of item fit in item selection confirms Reckase's conclusion (Beard, et al, 1984). Items may be selected on the basis of content and traditional psychometric indices with little detriment to the resulting scores.

Item Calibration. The calibration of test items is one of the most frequent tasks done when using the one or three parameter model, and it is in item calibration that several major differences between the two models occur. The most

basic and obvious difference is that the one parameter model requires only the estimation of item difficulties whereas the three parameter model requires the estimation of item difficulty, discrimination, and guessing parameters. The difficulties inherent in the estimation of the discrimination and guessing parameters constitute major issues in item calibration. In fact, Wright and Stone (1979) state, regarding the estimation of these two parameters, that, "...there are good reasons to expect that all such attempts must, in principle, fail." (p. 11). However, item parameters for discrimination and guessing are routinely estimated with the major problems appearing to be the cost and the large number of subjects and items required. Lord (1980) recommended sample sizes larger than 2,000, but others are willing to use smaller numbers. Hambleton (1982) suggested that sample sizes in excess of 600 or 700 appear to be needed, with disproportionately large numbers in the lower end of the ability scale to facilitate the estimation of the guessing (c) parameter. Reckase (1977) found that discrimination parameters estimated with N=382 were not significantly different from those estimated with N=2997. He found that N=763 gave excellent results for estimates of the difficulty parameters, though he suggested that in his data even 2997 cases might be none too many. Reckase found that there was considerable loss in calibration precision between N=382 and N=150 for the one-parameter model. Tinsley and Dawis (1975) cite Anderson et. al. (1968) as pointing out that generally samples of 500 or more are needed to obtain stable easiness and ability estimates. However, the general consensus seems to be that sample sizes of approximately 300 are minimal for calibrating items with the one parameter model and approximately 1,000 for the three parameter model.

Parameter estimation methods which use Bayesian techniques are able to function with smaller numbers of examinees under the right conditions. The Bayesian approach is to start from a prior distribution other than a "diffuse prior," the term applied by Bayesians to the assumption made by traditional statisticians that we have no knowledge about distributions before the data for the current investigation are collected. Bayesians might, for example, assume that rather than a diffuse or rectangular distribution, examinees' abilities are distributed normally. The Bayesian then collects data which he adds to the assumption and uses to modify the assumption he makes prior to his next analysis. Using the assumption is equivalent to having more data, and thus fewer examinees are needed if the prior assumption about distribution shape is correct. If it is not correct, the Bayesian may be worse off than the traditional approach which assumed that any outcome was equally likely, and the prior distribution was, thus, diffuse (or rectangular). In the case of the Statewide Assessment Data, the assumption of normally distributed abilities might be more sound than assuming that they are rectangularly

distributed. Probably some form of Beta distribution would be even more appropriate, and this could be determined on the basis of previous years' data. With a well-established prior such as that, the amount of data required for effective estimates might be appreciably reduced. Swaminathan is reported to be in the process of developing programs for Bayesian three-parameter models, and the approach taken by Owen to estimating abilities uses Bayesian Modal Estimators rather than Maximum Likelihood Estimators. Thus there are available already some techniques for reducing LOGIST's requirement for large samples, and further developments appear to be in the offing.

An additional problem is that it appears to be difficult to estimate guessing parameters for easy items unless special efforts are made. One cannot expect to have good estimates of the guessing parameter unless a large number of items are used (40 or more is recommended), and a reasonable number of persons, with special emphasis on including persons of low ability so adequate data are available to indicate the level of performance to be expected of persons of very low ability. (That is what the so-called guessing parameter really indicates.) Hutten (1980), using LOGIST 4, was able to estimate c parameters for only one-third of the items of 25 verbal and quantitative achievement tests using sample sizes of 1,000 persons. Yen, using LOGIST 5, appeared to have little problem in estimating the c parameters for the CTBS/U using sample sizes of 5,000 persons (Yen, 1983). A consideration for Florida's assessment program is the relative easiness of the tests' items, which serves to decrease the amount of guessing done on the test items. On the other hand, since the tests are so easy, few "low ability" pupils are ordinarily available to enable sound estimates of the probability of a person of very low ability answering the item correctly.

The calibration of items from two tests to a common scale is relatively easy to do with the LOGIST program, provided there are items common to the tests to be linked. The item responses of the two tests are analyzed in one combined data file with the items not taken by each student coded as not reached. The resulting item calibrations will then be on the same scale.

Using LOGIST 5, if a test contains a subset of pre-calibrated items, the remainder of the items in the test can be calibrated to the item bank scale or equated to the bank reference scale by holding the parameters of the pre-calibrated items fixed for the analysis. These options in the LOGIST program make the calibration of items and the equating of test scores quite manageable.

Another method of calibrating two sets of items to the same scale when a subset of the items have been administered

to both groups is as follows. The linear relationships between the difficulty and discrimination parameters from the two groups can be represented by scatterplots and a prediction equation for predicting the second set of parameters from the first can be established. This prediction equation can be used to predict item parameters for those items not administered to the first group. This procedure will provide estimates of item parameters on a common scale (Hambleton, 1982).

To equate the scores of two or more tests using LOGIST 5, the items would be calibrated to the same scale and the calibration values for the items of the test of interest would be held fixed while the abilities of the students were estimated. The resulting scores would be equated to the first test.

These procedures compare with those of the Rasch model which consist of finding the mean difference in difficulty of the common items from two separate analyses, and applying this difference as a constant to put the difficulties of the two item sets on the same scale, or the abilities of the two groups on the same scale. Alternatively, a common-person procedure could be used in which the items of two tests are administered to a group of persons. The entire set of items would be calibrated in one computer run and the resulting item calibrations would be on the same scale.

Item Bias. Both the one and three parameter models can be used to detect item bias. One procedure using the Rasch model is to calibrate a set of items using separate groups for which bias is being investigated. The two calibration values for each item are then compared using a standardized difference or by examining scatterplots of the two values for the entire set of items. Another procedure which provides more information regarding the particular nature of item bias is to estimate the item difficulties and person abilities for the two groups of interest in one combined analysis. These abilities and difficulties are then used to calculate residuals separately for the two groups. The residuals can be plotted against the difference between ability and difficulty (b-d). If an item is biased against a particular group, the residuals will take negative values (Mead, 1976). Several variations of the residual analysis may be found in the literature.

The general approach for detecting item bias with the three parameter model is to compare item characteristic curves generated by the two groups of interest. This procedure essentially facilitates the comparison of the expected proportion of successes in each group after ability level has been controlled. It also enables one to notice items whose characteristic curves cross, indicating that they favor one group at the low end but the other group at the high end of the ability scale. The one parameter model

does not permit these observations since it assumes that all item discrimination values are equal and therefore such crossing of item characteristic curves cannot occur.

The procedures used by the one and three parameter models are based on the same theory, that of comparing actual and expected responses for persons at the same ability level from two different groups. However, Hambleton (1982) points out that, "Since the three-parameter model often provides a somewhat better fit to test data at the lower end of the ability continuum (Hambleton et al., 1982) than less general logistic models, the three-parameter model may be more useful than otner logistic models for studying bias" (p. 5.24). Empirical studies indicate that the two models do not agree very well on which items are biased (Douglas, 1981, for example). In some cases the disagreement is because measures of goodness of fit have been used to detect item bias for the one parameter model while for the three parameter model measures of differences in item characteristic curves have been used to detect bias. To some extent the differences may occur because the two models treat items which are not reached differently.

Results Yielded By The Two Models

Scores. The scores yielded by the one and three parameter models may not measure the same construct. The one parameter or Rasch model yields scores which are transformations of the number of items answered correctly. This procedure is typically referred to as number correct (NC) scoring. When using NC scoring, a person's ability "is assumed to be expressed in his responses to the set of items he takes as summarized in the unweighted count of the number of items he gets correct." (Wright and Stone, 1979, p. 45). The BICAL UCON ability estimation procedure "responds in detail to the distribution of item difficulties $\{d_i\}$ and so estimates a measure b_r which is completely freed of whatever distribution of item difficulties characterizes the test." (p. 143) However, NC scoring does not mean that test ability scores increase in equal increments when additional items are answered correctly. Rather, it means that all persons who answer a given number of items correctly will receive the same ability measure. Items on the difficult end of the item difficulty continuum essentially contribute more to ability scores than easy items, provided that all easier items have been answered correctly.

The three parameter model uses information from items more completely than does the one parameter model. The three parameter model gives much more weight to a difficult item answered correctly by a person of high ability than a person of low ability who answers the same item correctly (perhaps by means of sheer guessing). The most discriminating items receive the most weight for high ability examinees. Only the easy items are accorded much

weight for persons of low ability (Lord, 1968).

In order to take advantage of all information about the items when using the three parameter model, it is necessary to use item pattern (IP) scori g procedures. Lord (1980, p. 58-59) gives procedures for deriving maximum likelihood estimates of trait values using item response vectors. These procedures are complex and expensive compared to NC scoring procedures, although Yen (1983) has developed computer routines which speed up the trait estimation process. She also developed a procedure and computer software for obtaining maximum likelihood trait estimates based on NC scoring. She found the NC scores to be unbiased estimates of IP scores and highly correlated with the maximum likelihc d estimates although the NC scores had larger standard e rors of measurement.

A frequently mentioned problem with IP scoring is that it is theoretically possible for an examinee with a lower raw score than another to receive a higher IP score. Personnel at CTB report that this has not been a real problem in their experience with the CTBS/U (Yen, 1984, personal communication). Another major difference in scoring between the two models is that LOGIST has the capability of treating incorrect, omitted, and not reached items differently. (One can enter the data with "not reached" items recorded as "wrongs" to avoid this option.) BICAL treats omitted and not reached items as incorrect responses. In LOGIST the ability of a person may be based only on those items which were reached, and partial credit is given for omitted items. When "not reached" items are called wrong, cultural groups with different attitudes toward completing tests will cause items near the end of a test to appear to be "culturally biased."

Yen reported that omitted items were scored as wrong in the standardization phase of the CTBS/U development, and not reached items were treated as not reached (Yen, 1983). Both omitted and not reached items were scored as wrong answers in scoring the tests for norming and for reporting scores to users. She provided the following rationale for this treatment of omitted and not reached items. "The decision to treat omitted and not reached items as wrong answers was based on two primary considerations: (1) If a student is not penalized for not reached items and this fact is known to teachers and students, it becomes possible to manipulate scores through test directions and changes in test-taking strategies. (2) Unless omitted and not reached items are treated as wrong answers, the hand scoring of tests becomes complicated." (p. 135). (The second of her reasons appears to be of no consequence for statewide assessment testing.) The implementation of the three parameter model would require several decisions about the handling of omitted and not reached items, and, depending on the choices made, adaptations of LOGIST or other computer

programs might have to be made.

Item Fit. More items will fit the model, perhaps substantially more, when calibrations are done with the three rather than the one parameter model. Whether this is a major advantage or merely a problem with the assessment of item fit would require research on the particular application under consideration. Many investigators have raised serious questions about the goodness of fit tests ordinarily used with the Rasch model. (See, for example, Divgi, 1984, especially George, 1980). It is reasonable that more items can fit a model in which it is not assumed that items are equal in discrimination and that there is no guessing--that the probability of a person of low ability answering correctly is zero.

Test Efficiency. An advantage of using either the one or three parameter models is the possibility of systematically increasing the efficiency of tests, which theoretically makes possible the use of shorter tests to achieve the same quality of measurement. Procedures have been developed for both the Rasch and the three parameter models which facilitate the design and construction of optimally efficient tests. These were discussed earlier under the heading "Item Selection."

It would appear that the procedures of the three parameter model would enable the more precise construction of tests to specifications. However, the implementation of either procedure should result in efficient tests and be considerably better than using neither procedure. It should also be remembered that the Florida Assessment Program is based on a mastery model of score interpretation which places content constraints on the selection of items. Some items must be placed in the test to conform to pre-determined test specifications, even if they are not optimal for overall test efficiency as defined by Lord or Wright. Careful test design might, however, permit emphasis of measurement precision near the cutting score while adequately conforming to the pre-determined specifications.

Testing Programs Now Using the Three-Parameter Approach

California Test Bureau

One major test publisher has developed a battery of standardized tests for classroom use based on the three-parameter model. California Test Bureau used the three-parameter model in developing the California Tests of Basic Skills, Form U (CTBS/U). The procedure is described in a chapter by Yen (1983). After reading that description and several recent papers by Yen, it became possible for one of the investigators (Hills) to visit Yen and her colleagues, D. R. Green and G. Burket, at the California Test Bureau

Offices in Monterey, California. The description below reflects Yen's paper and the discussion during that visit. Some detail of their procedures is given to convey an idea of the tasks involved in converting to a three parameter mathematical framework from the conventional approach.

CTBS/U is an achievement battery containing tests in reading, language, mathematics, reference skills, science, and social studies. It is available in 9 levels spanning grades K-12. In developing the tests, item-response-theory procedures were used in various phases. First, the items were written and tried out on a heterogeneous sample of students in public and private schools across the United States. Items appropriate for one grade were tried out at adjacent grades. At least 400 "standard" students, and 200 black students (from schools with predominantly black enrollments) were used in the tryout phase. To obtain item-parameter estimates, LOGIST 5 was run separately on each test in each tryout book. The values of the c parameter were not estimated on these small samples. They were set at 1/A - 0.05, where A is the number of alternatives for the multiple-choice questions.

Using the b parameters from this analysis, bias was evaluated by comparing among ethnic groups using a chi-square statistic presented by Lord (1980). LOGIST was rerun without items that appeared to be biased in order to get "final tryout item parameter estimates." Items were then evaluated in terms of fit to the model using a procedure developed and described by Yen.

Items were selected for final forms based on the item parameters from the tryout analyses. Data were available from the California Achievement Tests which had been given to the same students, and these data, analyzed by LOGIST, were used to assist in choosing items of appropriate difficulty for the CTBS/U. Also, national p values were estimated, distractor information was examined, and those plus an overall quality index were used in choosing the items for the final forms. A computer program to choose the items for a statistically "ideal" test provided a starting place for final item selection and revision. The program selected items in the desired subranges of difficulty, maximized the overall "quality" of the items, and maximized discrimination among items with the same overall quality. The editorial requirement for a minimum number of items for each category objective was also met.

Standardization data were collected on from 7000 to 45000 students for each test level. These data were then used with LOGIST to estimate final item parameters that are used in banking the items, in scoring, and equating. In these analyses, all three parameters of the items were estimated. Interlevel equating was done by analyzing together successive pairs of shelf and linkage books using

LOGIST and pooling data from common items while using LOGIST's "not reached" option for the items of the second test of each pair not common with the first. (A "shelf" test is one to be published for the market. "Linkage tests" are used only in the process of equating.)

Two scoring procedures are offered. One is based on the number correct: the other on item-patterns, i.e., the optimal scoring procedures that are part of LOGIST. Separate norms are provided for these two kinds of scores. Objective mastery scores use the scores on an objective along with scores on the rest of the test in a Bayesian approach which produces smaller standard errors for mastery scores than are available from the raw mastery scores. A separate program was written for scoring so that the rest of LOGIST would not be involved simply for that purpose.

Yen (1983) evaluated the use of IRT in this program, and drew the following conclusions. The IRT parameters seem to be more meaningful, and the bias evaluations seem satisfactory although larger Ns might be a good idea. Equating produced the interesting result that standard deviations decreased with level instead of increasing, as occurs with Thurstonian absolute scaling procedures. In scoring, IRT allows standard errors of measurement to vary with ability. Clearly, the change to IRT procedures involved a lot of work.

In our meeting with the CTBS group, Yen, Green, and Burket were exceedingly helpful. It was clear that they were very happy with the switch to the three-parameter IRT approach. When asked why they chose the three-parameter instead of the one-parameter approach, the response was that three advantages were clear: (a) Measurement is much more accurate at the low ability end when guessing is included in the model. They estimated roughly 25% more accurate. (b) Many more items fit the model when it is not assumed that items are the same in discrimination and that there is no guessing. (c) Items with high discrimination values are not eliminated for lack of fit, as they would be in a one-parameter model which assumes that all items are equal in discrimination.

On the other hand, one always asks whether the change was worth it. Could it be justified in dollars and cents? They don't have an answer, but if they were allowed to choose which way to go, they would not return to the classical procedures or try to make do with a one-parameter model. They noted that one problem if one wants continuity over the years is dealing with items which have data only in terms of classical parameters. If the data are available, those items can be reanalyzed as time permits.

CTB uses LOGIST 5 on its own IBM computer. It pays ETS for use--an amount of about $5000 per year for operational

use. Lord and Wingersky have been very helpful in getting the program operational at CTB and working with them on its use. CTB paid something like $2000 per year during the period when they were using LOGIST 5 for research. They have made some modifications to the program for their own purposes, have written some auxiliary programs for such things as scoring, preparing item characteristic curves, test characteristic curves, item and test information curves, displays for their item writers, etc. They are not in a position to offer those programs to anyone else since they developed them for in-house use without detailed documentation. To produce even flow charts now would require considerable effort. So to obtain similar programs would require new programming to fit the local needs.

CTB is heavily involved in equating. They feel that for horizontal equating it really does not matter much how you equate using the well-known methods. However, for vertical linking they feel that the best way is to use the 3 parameter model, and even that is not really adequate.

The CTB staff recognize that applying IRT to achievement tests is a problem. The b parameter values may change markedly from day to day even with effective instruction. However, this problem is not unique to IRT; IRT simply calls it to your attention. Their solution is simply to lump all students in a particular year together, from both Fall and Spring data collection and run the IRT analysis on the conglomerate. Multidimensionality is a problem with achievement tests also, and there is no good solution to it at present. Again, the same problem exists regardless of the method of analysis; IRT makes you think about it.

To summarize, California Test Bureau uses the three-parameter IRT model to develop, analyze, and score their major test offering, the California Tests of Basic Skills. They think it is a big step forward in testing, they recognize that the change requires a lot of work, they feel that they are still exploring and that there is a lot of exploring still to be done, but they would not change back if they were offered the opportunity. No insurmountable problems have arisen.

Educational Testing Service

Educational Testing Service (ETS) has a number of testing programs making use of the three-parameter model in one form or another. Dr. Al Beaton (personal communication, Summer, 1984) told us of the use of the three parameter model with National Assessment of Educational Progress (NAEP). ETS has had the NAEP contract for just over a year now. They inherited a large file of test items with no data on the items. Apparently the only data used by NAEP previously was difficulty level in terms of proportion

passing. By the nature of the test, there is no total score, so with traditional methods there was no way to obtain, or sense, to such things as item discrimination indices. ETS is using the three parameter model and obtaining data on (a) the level of ability required to have a probability of half way between chance and 1.00, (b) the slope of the item characteristic curve (which is possible because students who are not administered the item are treated as though they had not reached it), and (c) the level of performance expected by someone with minimum ability (the chance level). ETS has now (June, 1984) tested 90,000 examinees with NAEP instruments, and it will use the item data for maintaining an item bank and for using items as anchors in future tests. Dr. Beaton recognizes that ETS is more or less the focus of the three parameter model, so it would be unreasonable to expect it to use other models. However, he indicates that he is very happy with the analyses he is obtaining and would not be satisfied with anything else. He could identify no particular difficult or insurmountable problems.

Dr. Marilyn Hicks and Dr. Nancy Peterson are working with the Test of English as a Foreign Language (TOEFL). TOEFL seems to have been the first test at ETS that began using IRT in general, instead of just for research purposes. TOEFL has been using the three parameter model of IRT since 1978, and Dr. Hicks (personal communication, Summer 1984) regards their use of it as very successful. IRT is used for item banking and equating, but the TOEFL scores which are reported to candidates are based on the number correct. Item banking is done using both IRT and traditional item statistics. Test design and assembly are still done on the basis of the traditional item deltas and point-biserial correlations that have been used for many years at ETS.

The item analyses that are done for this program using IRT use the LOGIST 4 program. LOGIST 5 is not yet available to them. They think that most results will be very similar from the two programs, with occasional substantial differences. LOGIST 5 estimates the c parameters differently, and the size of the c parameters influences the size of the a parameters. They anticipate that LOGIST 5 will be more efficient, i. e., take less time for an analysis, and it will be easier to use because the mnemonics and input procedures are much more "user friendly."

Dr. Hicks says that they have had no serious unforeseen problems. Their main problem is getting suitable populations for pretest data. When asked what the main benefits of the three-parameter model over the one-parameter model would be, Dr. Hicks said first many items would fit the three-parameter model that would not fit the one-parameter model, and fit would generally be better. She also felt that the scores provided by the three-parameter model would be more accurate. She felt that equatings of

tests horizontally would not be very different from the two models.

Dr. Neil Dorans and Dr. Frank McHale (personal communication, Summer 1984) use the three parameter model in work with the Scholastic Aptitude Test (SAT) and the Preliminary Scholastic Aptitude Test-National Merit Scholarship Qualification Test (PSAT-NMSQT). They use the three-parameter IRT model for equating and item calibration. Scores are based on raw scores corrected for guessing rather than using the weighted scores derived from IRT. Item banking uses IRT statistics for large scale sets of data, but uses deltas and point-biserial rs for small data sets. They are gradually recalibrating items that only have deltas and point biserials so that IRT parameters will be available for all or nearly all items in the future. Test design and test assembly are currently being done based on deltas and point biserials, but the plan is to move to IRT statistics in 1986.

Dr. McHale reported that he has used only LOGIST 5. The differences between programs that he is aware of are that LOGIST 4 gives a common c parameter to more items than LOGIST 5, and LOGIST 5 is much quicker. For example, LOGIST 4 took 3 hours to process 290 items on 10,000 examinees. LOGIST 5 does the same job in 2 hours, about 1/3 less time.

Dr. McHale finds that the most difficult problem is getting all the item parameter estimates on the same scale. A number of procedures are available, but problems have arisen with each of them. Now a procedure being developed by Stocking and Lord is being investigated. It was reported on at the New Orleans AERA meeting.

Drs. Martha Stocking and William Ward are working with computerized adaptive testing (CAT) based on the three-parameter model at ETS. The tests they are developing are to measure basic skills for college level academic placement. Dr. Stocking explained that the tests will be taken on microcomputers (the IBM PC). Each test will start with an item of middle difficulty level. The next item will be the hardest item in the pool. After one item is correct and one incorrect, an ability can be estimated for the examinee. Then the next item administered will depend on which item in the pool will provide the maximum information at that ability level. Certain controls will assure that a sound mix of item type and content is maintained. All tests will be of the same length. An attempt to terminate testing when a fixed standard error of measurement was reached did not work very well. Scores will be in terms of estimates of the number correct that would have been obtained on the entire pool (120 items). The idea of maximum likelihood estimates of ability for scores was rejected for two reasons: (a) such scores are not familiar to users, and (b) sometimes maximum likelihood estimates give outliers that require special attention.

Obtaining parameter estimates for new items is a problem. They are not sure yet how they will accomplish that, but a group including Lord, Bock, Levine, and Samejima are working on it in connection with the Armed Services Vocational Aptitude Battery (ASVAB), and the solution achieved there will be evaluated for use in connection with these college placement tests.

Stocking is enthusiastic about this use of the three-parameter IRT approach. She says the problems are less severe than was expected. The hardware problems have been tractable. Since the items are to be multiple choice, she feels that the one-parameter model which assumes no guessing would be entirely inappropriate. She has no reason to believe that the items would be equal in discrimination, so for her that would be another inadequacy of the one-parameter model.

Dr. Dan Eignor (personal communication, Summer 1984) of ETS is not involved in a testing program there that is currently using IRT, but in 1979 ETS had the contract for the New Jersey minimum competency tests and tried to use LOGIST 4 to estimate item parameters for it. He indicated that they had severe convergence problems, i. e., after many iterations the program would not reach the end. The problem seemed to have to do with estimating the c parameters when the items were so easy (as minimum competency items tend to be). Eventually they tried using a two-parameter model, ignoring guessing or fixing it. They still had problems with convergence. When the proportion getting an item correct is .96 or .97, the item discrimination cannot be very high, and it may be that even students with very limited ability compared to those in this group would have a relatively high probability of answering the item correctly. In such cases, IRT parameters are difficult to estimate. Possible solutions are to fix the c parameters at a common value or to try starting values for cs other than those usually used in LOGIST. Eignor suggested that Bock and Mislevy have an approach to estimating item parameters, BILOG, that is Bayesian in nature. It allows one to set prior distributions, rather than assuming that distributions are normal. Eignor wonders if that might be of assistance, though he has never seen or used the BILOG program. Mislevy is scheduled to join the ETS staff in September, and he may have an opportunity then to help solve some of these problems. ETS rarely has minimum competency data to deal with, so they have little experience with tests consisting almost entirely of very easy items.

Air Force

Dr. Bryan Waters (personal communication, Summer 1984) of the Human Resources Research Office in Washington, D. C., described briefly a long-term (five to eight year)

development in the Air Force of computer administered tests of performance in Air Force jobs. The jobs are to be analyzed in terms of the cognitive skills and the cognitive tasks involved, and tests will be administered by computer. These will not necessarily be computerized adaptive tests--they may not be adaptive in the sense that will be described in the next section on the developments at the Naval Personnel Research and Development Center. However, instead of transforming a single battery of tests to the adaptive mode, the Air Force is considering administering all of its tests by means of computer terminals, and is developing plans for a thorough psychological analysis of the competencies which will be tested. Implementation, if all goes well, is anticipated around 1990.

Naval Personnel Research and Development Center

The Naval Personnel Research and Development Center (NPRDC) at San Diego has been using the three-parameter logistic item response model in connection with development of the Computerized Adaptive Testing System (CAT). The Navy will administer the Armed Services Vocational Aptitude Battery (ASVAB) adaptively on computers starting in 1986. Extensive studies are being done now to determine whether the CAT maintains or improves the reliability, the validity, the predictive efficiency, and the factorial structure of the ASVAB while reducing by about half the number of items administered to each examinee, eliminating security problems which arose when test booklets and answer sheets were stolen, providing suitable questions for the wide range of ability faced in testing the general population, and eliminating clerical errors in scoring, score conversion, and score recording.

Studies completed so far indicate that CAT tests have higher reliabilities than conventional tests for numbers of items up to 30, the difference being of the order of a 9-item CAT test having the same alternate forms reliability (.80) as a 19-item conventional test. The validity of the CAT also exceeds that of the conventional test up to 30 items in length, with about 10 CAT administered items reaching the validity of about 30 conventionally-administered items. Studies of the factorial nature of CAT administered tests indicate that they measure the same factors as their conventionally-administered counterparts, and the correlations between CAT administered tests and conventionally-administered tests are higher than the correlations between test and retest with the conventionally-administered forms. Studies of the validity of CAT administered ASVAB for predicting success in specialty schools in the Navy (specialties such as electronics technician, radioman, sonarman, mess manager, hospital corpsman, and hull maintenance technician) indicated no noticeable difference between the CAT tests and conventional tests. Dr. J. R. McBride, Head of the Personnel

Systems Research Department of NPRDC, indicated (personal communication, July 30, 1984) that studies are still needed on such issues as differences between CAT and conventional scores for females and for minorities, effect of computer experience on scores, whether items should be calibrated in the CAT mode instead of in the conventional administration, as well as on development of the hardware and software to make CAT effective on a large scale. On the latter point, nothing in the information received from NPRDC indicates the strategy that the Armed Forces will use to administer the CAT-ASVAB on a large scale. One wonders whether an entire typewriter keyboard will be used as a response device, since only a few keys may be needed. Will a micro-computer be used for each examinee, or will one micro serve a center, or will all centers be tied into a national large-scale computer by dedicated telephone lines? People more sophisticated in such problems can probably ask better questions than these, but no answers to these questions or considerations of other more appropriate ones are available in the material that NPRDC is providing to the public.

McBride seems completely confident that the CAT administration will be installed, and that the three-parameter model will be used successfully with it. No discussion appears comparing the possibilities of using a one-parameter model or a two-parameter model instead. No difficulties are noted with use of the three-parameter model in their reports. It may be that with the large and general population available to them, and no attempt to develop mastery or minimum competency tests, problems such as estimating the c parameter do not arise. Their research basis seems to depend on the efforts of Urry and Owens, using OGIVIA and Owens' Bayesian approach to ability estimation, rather than the chain of efforts based on Lord's work at ETS on LOGIST. (In three articles, only one reference to any work of Lord appears.)

Computerized adaptive testing is a step beyond merely changing from use of a one-parameter model to a three-parameter model. However, CAT has the potential of solving, or opening up different avenues for solution, of many problems in statewide testing. For example, the printing of test booklets and their distribution would be eliminated. Test security problems would be reduced, or at least dramatically altered. Scoring could be done and reported immediately. Testing could be done whenever a student was deemed ready, rather than on a single day each year. In fact, it might be useful to consider all the changes that are foreseeable from CAT of statewide tests, and to analyze how it could be done with the current state of the art of hardware. At some point, it seems highly likely that the problems will be offset by the advantages, and the State of Florida would probably want to move to CAT at that time.

Availability and Adequacy of Computer Procedures for the Three Parameter Approach

The available programs for the three parameter approach now seem to be LOGIST 4, LOGIST 5, ANCILLES, and BILOG. In reverse order, BILOG is a program being developed by Darrell Bock at the University of Illinois. It is reported to be superior to LOGIST 4 and 5 in some respects. Since it does not estimate abilities until item parameters have been estimated, Lord (1984) thinks it may function with fewer items than LOGIST. (With only 10-15 items, LOGIST obtains biased estimates of ability parameters, especially at low ability levels. This causes item parameters to be misestimated, even with large numbers of examinees.) Wendy Yen is working on a comparative study between LOGIST and BILOG. She reports, however, that BILOG is still in the process of development, and she has not received a final version to evaluate (personal communication, Summer, 1984). For the present, that makes BILOG unattractive for a major testing program. Bock is being used as a consultant to the State of California in its assessment program which uses BILOG along with multiple-matrix sampling to determine the level of performance in California's schools. An inquiry to the head of that program, Dr. Dale Carlsen, for information on their activities was fruitless.

ANCILLES and OGIVIA

Similarly, an inquiry to Dr. Urry for additional information on ANCILLES, a Bayesian Modal Estimation procedure, was to no avail. The computerized adaptive testing efforts of the Navy, reported above, may be based partly on use of ANCILLES or its predecessor, OGIVIA, along with Bayesian estimates of abilities based on a procedure developed by Owen, but the reports are not clear on what estimation procedure is being used. Swaminathan is reported to be developing a Bayesian Modal Estimation procedure program. Bayesian estimates are useful if the proper prior frequency distribution can be chosen, and in Florida's case that should be easy to do from previous years' data. Bayesian estimates of ability are more precise (smaller mean squared error) because they are based on more information than maximum likelihood estimates. The Bayesian estimates take into account the characteristics of the group to which one belongs, while the maximum likelihood methods ignore that information. However, since the Bayesian estimates do take into account the group of which one is a member, they are, by definition, not independent of the group, and that independence is one of the attractive characteristics of ability estimates and item parameter estimates from item response theory.

Comparisons of LOGIST and ANCILLES

McKinley and Reckase (1980) compared LOGIST and ANCILLES for analyzing data from 4000 examinees who took the Iowa Test of Educational Development (ITED). They found that using ANCILLES more items had unsatisfactory fit to the model. The average b values were not significantly different; the average a values were significantly higher for LOGIST, and the average c values were significantly lower for LOGIST. The mean abilities were not significantly different, and the correlation between abilities from the two programs was .987. Since the ANCILLES b values correlated highly (.68) with the chi square goodness of fit values, it appeared to the authors that ANCILLES had difficulty fitting items with extreme b values. (The correlation between b value and chi square for LOGIST was zero.) Poor fit in the case of LOGIST seems to be related to low a values. The authors conclude that LOGIST is the method of choice in spite of its additional cost.

Swaminathan and Gifford (1983) compared ANCILLES and LOGIST on artificial data, using tests of different length, different sizes of examinee population, and different distributions of ability. Both procedures resulted in poor estimates of the a parameter with short tests, but the LOGIST estimates were superior to those of ANCILLES. In general, the a parameter was poorly estimated by ANCILLES, with a tendency to overestimate its size. LOGIST also tended to overestimate a, but to a lesser degree. The b parameter was estimated well by both procedures, with more accurate estimates coming from LOGIST which did surprisingly well with small numbers of items and examinees. ANCILLES produced very poor estimates of the c parameter, while LOGIST estimates were close to the true values. Although no differences appeared between the ANCILLES and LOGIST estimates of ability for longer tests, when there were fewer than 15 items, LOGIST fared better, especially when distributions of ability were skewed. LOGIST tended to underestimate cs while ANCILLES estimates of the c parameter were extremely high when the ability distribution was skewed. The authors conclude that in general, the LOGIST procedure was superior to the ANCILLES procedure, especially in estimating the a and c parameters. However, the difference between them becomes negligible when the number of items and the number of examinees increases sufficiently.

LOGIST 4 and LOGIST 5

The most widely used, most completely developed, and certainly best documented computer program to use at present is LOGIST. LOGIST is currently available in two forms, LOGIST 4 and LOGIST 5. Wingersky (personal communication, September, 1984) has informed us that a new LOGIST 6 is being planned for development over the course of the next

year. Each new form is an improvement over the previous form. The plan for LOGIST 6 is to develop a form of LOGIST that will be readily usable for operations, rather than research, will be user-friendly, and will be designed to be readily translatable for computers other than those of IBM. LOGIST 4 is now available and usable on computers such as the Control Data Corporation (CDC) computer at Florida State University. Dr. Hambleton assures us that to the best of his knowledge the translation of LOGIST 4 for use with CDC computers, done under his direction, was complete, and all the usual options are available. So it appears that LOGIST 4 is immediately available locally. That has advantages since no royalty is involved in its use for research or operational purposes. LOGIST 5, however, is reported to be more efficient and more effective at estimating the c parameter. Estimation of that parameter for items is anticipated to be a problem since the Florida assessment tests are so easy, i.e., few people get low scores reflecting low ability so it is difficult to estimate the probability of a very low scorer (a person of very low ability) answering an item correctly. On the other hand, since the tests are easy for the population, little guessing may occur, and fixing all c parameters at a common value may yield satisfactory results. Dr. Eignor alerted us to his frustrating experience in using LOGIST 4 with minimum competency tests.

LOGIST 5 is reported to have two other advantages, also. First, it is reported to be easier to use, i.e., input has been simplified, and the instructions are easier to follow. Second, it takes less computer time, and should, therefore, be less expensive to use, by about one third. A typical LOGIST 4 run for 1000 examinees and 40 items might cost about $10 plus costs for printing the output which vary depending on how much output one wants. Saving $3.33 seems like a trivial amount, but if many LOGIST 4 runs are made trying different options or on different sets of data, the cost differential could easily amount to hundreds of dollars. It is hard to imagine that it would amount to thousands of dollars a year.

So, LOGIST 5 would seem to have some important advantages. Its disadvantages are that it has not yet been translated for the CDC computers, nor were we successful in running it on the IBM computers at the Northwest Data Center or the University of Florida. Furthermore, royalties must be paid for its use in an operational program. The standard fee for commercial use of LOGIST 5 is $5000 per year. We do not know whether that fee would apply to a State agency which was not deriving income from the use of the program. As far as we know, no State agency is now using LOGIST 5 in its testing programs. Wingersky, the primary source for information about LOGIST at ETS, has estimated that doing the necessary programming to adapt LOGIST 5 for the CDC would cost about $5000 in consulting and travel expenses,

plus computer costs. It would take about two weeks. She would be willing to contract for that undertaking, but would not be available until Spring, 1985. By that time, LOGIST 6 might be available, and it might be easier to adapt to varied computers, as well as having other useful features. Other programs now under development, such as ASCAL and BILOG, should, perhaps, be evaluated at that time also.

ASCAL

C. David Vale, President of Assessment Systems Corporation, St. Paul, Minnesota, has informed us (personal communication, Fall 1984) that Assessment Systems Corporation has developed a three-parameter logistic item calibration program called ASCAL. Vale reports that the program is similar to LOGIST although it restricts its parameter estimates in a different way, using Bayesian priors on the estimates. The program is being written to function on the PDP-11 and IBM PC computers. The PC version is scheduled to be available in October. Its price is anticipated to be $900. The Statewide Assessment Tests may have difficulty using ASCAL since it uses as the Bayesian prior for ability a normal distribution, and the scores from the Florida SSATs are highly skewed, being minimum competency tests. Further, nothing in the information received from Dr. Vale indicated that person abilities are estimated by ASCAL. The necessary options for equating tests, for putting item parameters from different testings on the same scale, etc. are not mentioned either. So it is not clear at present whether ASCAL will be complete enough for use by the Statewide Assessment Testing program. The fact that ASCAL can run on an IBM PC is attractive, but one must note that on the PDP-11 1.5 hours running time was required to process 30 items on 1000 people.

Effects of Moving to the Three Parameter Model on the Data Analysis Procedures

This section presents an analysis of the effects on the data analysis procedures currently being used by the assessment program if a move to the three parameter model is made. The impact of the move on the following procedures will be examined: a. the calibration of items to the anchor year scale, b. equating and the computation of students' scaled scores, c. the maintenance of item banks, d. test design and construction, e. item bias review procedures, and f. special tasks. Also, some possible implications for the future will be considered.

Calibration of test items to the reference year scale. Cook and Eignor (1983) discussed several procedures for placing item parameters from different administrations of a test on the same scale. The calibration of test items could proceed from the general anchor test design that is

currently used if analyses were done using LOGIST 5. The inclusion of common items in each new form of the test would be done as it is now. The common items would have been previously calibrated to the reference scale using the three parameter model. The new items and the common items would be analyzed together in one computer run with the calibration values of the common items held fixed. The resulting calibration values for the noncommon items would be on the reference scale.

Parameter estimates from two separate calibration runs may also be placed on the same scale using a linear transformation. Slope and intercept parameters for putting common item difficulties of a second test on the scale of a first test are determined. The transformation equation can then be used for placing the set of item difficulty parameters for the entire second test onto the scale of the first test. The same parameters of the transformation equation can be used to place the item discrimination and ability parameters onto the scale of the first test. The guessing parameter need not be transformed because it is expressed in terms of probabilities, which are already on a common scale.

We know of no procedure for directly transforming Rasch difficulty values to substitute for those of the three parameter model. However, the correlation between Rasch and three parameter item difficulty values would probably be above .90. A linear transformation could be used to put the existing Rasch values on the LOGIST item difficulty scale. Such a procedure might be satisfactory, depending on how "not reached" items are to be handled.

Greater numbers of persons are required for calibrating items for the three parameter than for the one parameter model. However, that should not be a problem for the assessment program since large numbers of students are involved in the program. Some adjustment in the current procedures would be required. Items are currently calibrated to the reference scale using responses to "experimental forms" as well as intact tests. The numbers of students taking each of these experimental forms has been as low as 300, which is acceptable for the Rasch model but would be unacceptably low for the three parameter model. However, an alternative procedure would be to consider the experimental forms as pilot tests and to use Rasch or classical methods of analysis for these data. The items could be calibrated to the reference scale upon their administration as a regularly scheduled test.

Another alternative would be to insure that each experimental form is administered to a minimum number of students. This minimum remains to be determined, but would probably be on the order of 1000. This minimum might be lowered in the event that Bayesian parameter estimation

procedures were adopted.

It has already been mentioned that Eignor (personal communication) found that when Logist 4 was applied to the New Jersey minimum competency test data, LOGIST didn't converge, and the "common c" was used for most items. However, if the examinees are so able, relative to the item difficulties, that few guess, Wingersky (personal communication) feels that good estimates of c are unimportant. Hambleton, Murray, and Williams (1983) compared the fits of the one, two, and three parameter models to the Maryland Functional Reading Tests. They did not mention difficulties with parameter estimation. However, they concluded that the guessing parameter improved fit very little over the two parameter model because the tests were easy, 77.8% correct. On the other hand, they found the discrimination parameter to improve fit substantially. In any case, the estimation of these two parameters for the Florida data should be carefully investigated before a decision to change models is made. Among these investigations should be an investigation of the possible improved capability of Logist 5 over that of LOGIST 4 in estimating guessing parameters.

Computation of Student Scaled Scores. The computation of scaled scores for the SSAT-II involves two major steps. First, the scores must be equated to the reference scale, and second, they must be transformed to the established SSAT-II scale where a passing score is set equal to 700. The topic of equating is dealt with by Lord (1977, 1980) and by Cook and Eignor (1983). Several procedures are discussed. The method of choice would depend on the overall design of the equating study. For example, one could choose to use a common person design or a common item design. The common item design in current use by the Florida Assessment Program would probably be continued. In this case the overall design for equating would remain basically unchanged. The item parameter estimates would be placed on the reference scale by using the procedures described under "item calibration" in an earlier section of this paper. Once the item calibrations have been placed on the reference scale, and fixed for an analysis, the resulting ability scores would be on the reference scale.

If LOGIST 5 were used, the equating of the scores could be accomplished by including a set of common items for which parameters are known. These pre-existing parameters would be fixed for the analysis, and the resulting scores would be equated to the reference scale.

The ability scores yielded by LOGIST are scaled to have a mean of zero and a standard deviation of one. It would be a relatively straightforward matter to re-scale these scores to have the characteristics of the current SSAT-II scale or of any other scale desired.

Item Banking. The most obvious effect of going to the three parameter model would be that wherever one item parameter has been recorded and used there would now be three. Users would have to attend to 3 statistics of items-especially discrimination, in addition to difficulty.

At a more substantive level, the procedures for initial calibration of items would have to be modified as already discussed. In addition, and perhaps the consequence involving the most work, would be the re-calibration of items already in the item bank to the three parameter model. We know of no procedure for estimating these parameters short of re-analyzing samples from the existing item response data files using LOGIST. The procedure of transforming Rasch item difficulties to the three parameter scale, as previously discussed, would not apply to the guessing or discrimination parameters, since there are no existing estimates for these parameters.

The problem of re-estimation of parameters would be essentially solv if the program adopted the procedure of calibrating ite s using only data from operational administrations of the tests. In that case the existing item bank data could be used as initial pilot test data in selecting items for the test forms with parameter estimation occurring after the item's administration in an operational form.

Test Design and Construction. The design and construction of tests could be done as it is now. However, several things should be done to capitalize on the advantages of the three parameter model. Items should be selected which would maximize the information in the test scores and minimize the standard error of measurement, while retaining the editorial requirements for a given number of items per skill and objective.

Yen (1983) described the procedures used by CTB in constructing the CTBS. They developed a computer program to assist in the test design process. The program selected an initial set of items to fit the criteria given above, and printed a listing of the items selected along with a statistical description of the proposed test. The statistical description included a curve showing the standard error of measurement at each ability level and an estimate of the mean proportion correct for the national population.

In addition, to construct the examinations the editorial staff employed an interactive computer program together with item cards. The item cards included statistics or ratings of bias, fit, discrimination, adjusted item difficulty, estimated national P values, and answer choice statistics. As items were replaced in the test the

staff received immediate feedback on the statistical properties of the revised test from the interactive program..

While such computer programs facilitate test construction, they would not be absolutely required. However, if the advantages of the three parameter model are to to be gained, then the use of such techniques would be strongly recommended. They make it possible to construct tests having optimal psychometric characteristics. For example, tests could be constructed which have smaller standard errors of measurement in the area of the passing score while preserving reasonable accuracy at all other points on the score continuum.

Item Bias Review Procedures. Evaluation of item bias must include review by sensitive judges along with comparisons of performance by examinee groups. Ordinarily, the judges review items before they are administered. After administration, statistical procedures are used to identify items that need further scrutiny. The item bias review procedures currently in use may be quite satisfactory for the statewide assessment tests. However, the three parameter model would make possible more precise statistical investigations of item bias. Lord (1980) presented a strategy for studying item bias which was essentially followed by Yen and others in the development of the CTBS.

1. Estimate approximately the item parameters for all groups combined, standardizing on the b and not on θ.
2. Fixing the c at the values obtained in step 1, re-estimate a and b, separately for each group, standardizing on the b.
3. For each item, compare across groups the item response functions or parameters obtained in step 2 (Lord, 1980, p. 217).

The comparison of item response functions for two groups would permit the evaluation of bias at different points along the ability scale and would show any differences in the item's discrimination for the two groups compared. Ability would be controlled in such an analysis. Tests of significance for these comparisons have been developed and are described in Lord (1980, pp. 217-223).

While no changes from current procedures would be absolutely necessary, the three parameter model would make possible state-of-the-art statistical analyses which provide considerably more information than those currently possible.

Other Effects. The Rasch model has contributed to several special tasks which would have been difficult to achieve with the classical measurement model. For example, pre-equating has been used to estimate frequency

distributions of scores before the tests were administered, and log ability scores have been used in the setting of passing scores for the SSAT-II. Tasks such as these could be done using the three parameter model, assuming the necessary background work to convert the item bank to the three parameter model had been done and suitable computer programs had been written. In some cases the use of three parameters might make tasks more difficult, but more precise information might be gained.

Adoption of the three parameter model might, in due time, also contribute to new directions for the testing program. For example, the Army, Air Force, and the Navy are either using, or investigating the use of, computerized testing, and computerized adaptive testing. Each of these efforts is based on use of the three parameter model, and on the decreasing costs of computers. As computers become more widely available in Florida's schools, questions regarding their use in testing are bound to arise. It already appears that computer assisted testing will soon be a reality in Broward County. Therefore, an effect of adoption of the three parameter model might be that it positions the Florida Student Assessment Program to take advantage of other testing technologies.

Implementing a Tryout of the Three Parameter Approach

Phase One

A number of things should be evaluated before a full-scale tryout of the three parameter approach is conducted. Perhaps the first is to attempt to analyze actual Florida Statewide Assessment Program data using LOGIST 4 to see what happens. We can anticipate difficulty, judging from Eignor's experience, but we don't know how much since we do not know how the New Jersey testing program he worked with in the 1970's compares with the current program in Florida. If problems are severe, the next logical move might be to analyze Florida data with LOGIST 5. If neither program provides satisfactory results, additional study is in order to determine whether the problems could be corrected easily.

Several aspects deserve consideration. One might inquire about the goodness of fit of the three parameter model to Florida's data. As noted previously, Hambleton and Murray (1983) have suggested a number of avenues for exploring goodness of fit which have not been tried on the Florida test data. Very reassuring would be findings that using all the techniques suggested by Hambleton and Murray, the fit of the three parameter model was satisfactory. Fit of the three parameter model might be compared with fit of the one parameter model using the Hambleton and Murray techniques. If the fit is found to be lacking for the one parameter model but satisfactory for the three parameter

model, Florida would be in a much more defensible position if it used the model which best fit the data when carefully and thoroughly analyzed.

Another possible source of difficulty could be multidimensionality. All currently used measurement models assume unidimensionality, but there is considerable danger that in achievement tests more than one dimension is being measured. The tests have been developed by specifying a number of more or less separate skills, each with several items. That might result in serious multidimensionality. On the other hand, if there are only a few items for each skill, and if achievement of one skill is correlated with achievement of others, and if many skills are included in one 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. In fact, that concept is implied by having a total score for such a test instead of having many subscores, each based on a much larger number of items than is now the case. A suitable exploration, then, would be of the dimensionality of the Statewide Tests. Previous analyses by King and Hills have already indicated that the scores of the SSAT II meet a widely-used criterion for unidimensionality. More elaborate investigation of that test might be appropriate.

Another possible problem is that the easiness of the Florida tests may make estimation of the c parameters difficult. 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. Eignor (personal communication, Summer, 1984) did not find that using the two parameter model solved the convergence problem for his New Jersey data. He suggested trying to fix the cs at a common value or trying other starting values for cs than the standard starting values used by LOGIST 4.

Studies of the optimum number of examinees to be used in analyzing the kind of data available from Florida's tests would be helpful. Current estimates in the literature range from a low of 700 persons to satisfactory results only with 3 or 4 times that many. The more persons in an analysis, the more expensive it gets because for every additional person another parameter (ability) must be estimated.

Study of just what kind of sample of Florida's examinees should be used in analyzing items by LOGIST would be appropriate. A random sample is not the most desirable. Heavier sampling should be done among those of extreme

ability level because that is where the estimation difficulties arise and where most data are needed.

Phase Two

The above questions are fundamental. If the problems associated with them cannot be resolved satisfactorily, the three parameter model does not seem feasible for Florida's Statewide Student Assessment Testing Program. They might constitute the first phase of an investigation into feasibility. If the results of these studies are satisfactory, then several further studies would seem appropriate for a second phase.

While equating procedures with LOGIST seem straightforward, several different methods have appeared in the literature. Comparisons of the results of different methods on SSAT data, and of those results with equating by means of a one parameter model or linear and curvilinear (equipercentile) methods should be of value.

It would be important to make sure that item calibration procedures to put the future test items on the same scale of the former items are tried out and found to be satisfactory. The task of converting the data from previous forms which were analyzed using the one parameter model to the parameters of the three parameter model must be accomplished.

California Test Bureau has developed procedures for estimating trait levels from number correct scores, and has reported considerable success with it. It might be reasonable to analyze a sample of Florida's test data with LOGIST to establish item parameters. However, as many as 100,000 students might take a test during an administration. Most of those person's data would not be needed to estimate item parameters, but each of their abilities must be estimated. Exploration of the CTB procedure for estimating the abilities of the examinees not used in the analysis of items might be appropriate. Development of computer programs to estimate the abilities of the examinees not used in the item analysis would be necessary regardless of whether the CTB procedure could be used.

The three parameter model opens new avenues for detecting items which might be biased. One has been described above. There are others which might be useful especially with small minorities which would not provide 1000 or so examinees at a single testing for use with the standard LOGIST analysis. These procedures should be explored on real SSAT data.

Each of the studies in the proposed second phase involves writing of new computer programs. The studies of the first phase do not involve that aspect.

Phases Three and Four

If the second phase yields encouraging results, two more phases logically follow. The first is a phase of planning and describing the writing of the specific computer programs that will enable the Statewide Assessment Staff to make the maximum use of the information from the three parameter model, including such activities as test development, item banking, equating, and bias analyses. The second phase consists of the planning and describing of the training necessary for effective use of the three parameter model by the Statewide Testing Office staff.

Time estimates for the first two phases are about seven months for each. The first phase should precede the second, and the second should be revised in light of the results of the first. Phases three and four would only be conducted if the results of phases one and two were satisfactory. Some of the activities in each of the first two phases could be concurrent, and some of the analyses for one part could be used in another part.

References

Anderson, J., Kearney, G. E., & Everett, A. V. (1968). An evaluation of Rasch's structural model for test items. The British Journal of Mathematical and Statistical Psychology. 21, 231-238.

Beard, J. G., Julian, E. R., Richards, L., & Roca, N. (1984). Effects of deleting misfitting persons on the calibration of items for Florida's Student Assessment Program (Contract No. 084096). Tallahassee, Florida: Florida Department of Education.

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 1-23). Vancouver, British Columbia: Educational Research Institute.

Choppin, B. (1982). The Rasch model for item analysis. (Grant No. NIE-G-80-0112, p3). Los Angeles, California: Center For The Study of Evaluation, University of California.

Cook, L. L. & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 175-195). Vancouver, British Columbia: Educational Research Institute.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1984). A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Curry, A. R., Bashaw, W. L., & Rentz, R. R. (1978). Invariance of Rasch model ability parameter estimates over different collections of items. Paper presented at the meeting of the American Educational Research Association, Toronto, Canada.

Dinero, T. E. & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. Applied Psychological Measurement. 1(4), 581-592.

Divgi, D. R. (1984). Reading tests and the Rasch model: A study of model-data misfit. Paper presented at the meeting of the American Educational Research Association. New Orleans.

Douglas, J. B. (1981). Item bias, test speededness, and Rasch tests of fit. Paper presented at the meeting of the American Educational Research Association, Los Angeles.

George, A. A. (1979). Theoretical and practical consequences of the use of standardized residuals as Rasch model fit statistics. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Hambleton, R. A. (1982). The three-parameter model. (Report No. 126). Amherst, MA: Laboratory of Psychometric and Evaluative Research.

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

Hambleton, R. A. & Murray, L. (1983). Some goodness of fit investigations for item response models. In R. A. Hambleton (Ed.), Applications of Item Response Theory. British Columbia: Educational Research Institute.

Hambleton, R. A. & Traub, R. E. (1971). Information curves and efficiency of three logistic test models. British Journal of Mathematical and Statistical Psychology, 24, 273-281.

Hills, J. R. & King, F. J. (1981). Evaluation of a method for detecting potentially biased items in the Florida Statewide Assessment Tests. (Contract No. 081188) Tallahassee, Florida: Florida Department of Education.

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

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28(4), 989-1020.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14(2) 117-138.

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

Lord, F. M. (1984). Technical problems arising in parameter estimation. Paper presented at the meeting of the American Educational Research Association, New Orleans.

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

McKinley, R. L. & Reckase, M. D. (1980). A comparison of the ANCILLES and LOGIST parameter estimation procedures for the three-parameter logistic model using goodness of fit as a criterion. (Report No. 80-2). Columbia, MO: Tailored Testing Laboratory, University of Missouri.

Mead, R. (1976). Assessing the fit of data to the Rasch model. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Panchapakesan, N. (1969). The simple logistic model and mental measurement. Unpublished doctoral dissertation, University of Chicago.

Phillips, S. E. & Mehrens, W. A. (1984). Detecting impacts of curricular differences and curricular multidimensionality in achievement tests. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut. (reprinted by the University of Chicago Press, 1980)

Reckase, M. D. (1977). Ability estimation and item calibration using the one and three parameter logistic models: A comparative study. (Report No. 77-1). Columbia, MO: Tailored Testing Research Laboratory, University of Missouri.

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.

Swaminathan, H. & Gifford, J. A. (1983). Estimating item parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), New horizons in testing (pp. 14-30). New York: Academic Press.

Tinsley, H. E. A. & Dawis, R. V. (1975). An investigation of the Rasch simple logistic model: Sample-free item and test calibration. Educational and Psychological Measurement. 35, 325-339.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), Applications of item response theory. (pp. 57-70). British Columbia: Educational Research Institute.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum No. RM-76-6). Princeton: Educational Testing Service.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In Proceedings of the 1967 invitational conference on testing problems. Princeton: Educational Testing Service.

Wright, B. D., Mead, R. J.,& Bell, S. R. (1980). BICAL: Calibrating items with the Rasch model. (Research Memorandum No. 23c). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D. & Stone, M. H. (1979). Best test Design. Chicago: MESA Press.

Yen, W. (1983). Use of the three-parameter model in the development of a standardized achievement test. In R. K. Hambleton (Ed.) Applications of item response theory (pp. 123-141). British Columbia: Educational Research Institute.