ITEM RESPONSE THEORY: THE THREE-PARAMETER LOGISTIC MODEL

Ronald K. Hambleton

CSE Report No. 220
1982

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

1. Definition and Background

In a few words, item response theory postulates that (a) examinee test performance can be predicted (or explained) by a set of factors called traits, latent traits, or abilities, and (b) the relationship between examinee item performance and the set of traits assumed to be influencing item performance can be described by a monotonically increasing function called an item characteristic function. This function specifies that examinees with higher scores on the traits have higher expected probabilities for answering the item correctly than examinees with lower scores on the traits. In practice, it is common for users of item response theory to assume that there is one dominant factor or ability which explains performance. In the one-trait or one-dimensional model, the item characteristic function is called an item characteristic curve (ICC) and it provides the probability of examinees answering an item correctly for examinees at different points on the ability scale. In addition, it is common to assume that item characteristic curves are described by one-, two-, or three-parameters. The interpretation of these parameters will be described in section 3. In any successful application of itemresponse theory, parameter estimates are obtained to describe the test items, ability estimates are obtained to describe the performance of the examinees, and there is evidence that the chosen item response model, at least to an adequate degree, fits the test data set (Hambleton,

Murray, & Simon, 1982).

Item response theory (or latent trait theory, or item characteristic curve theory as it is sometimes called) has become a very popular topic for research in the measurement field. There have been numerous published research studies, conference presentations, and diverse applications of the theory in the last several years (see for example, Hambleton et al., 1978; Lord, 1980; Weiss, 1980). Interest in item response models stems from two desirable features which are obtained when an item response model fits a test data set: Descriptors of test items (item statistics) are not dependent upon the choice of examinees from the population of examinees for whom the test items are intended, and the expected examinee ability scores do not depend upon the particular choice of items from the total pool of test items to which the item response model has been applied. Invariant item and examinee ability parameters, as they are called, are of immense value to measurement specialists.

Today, item response theory is being used by many of the large test publishers, state departments of education, and industrial and professional organizations, to construct both norm-referenced and criterion-referenced tests, to investigate item bias, to equate tests, and to report test score information. In fact, the various applications have been so successful that discussions of item response theory have shifted from a consideration of their advantages and disadvantages in relation to classical test models to consideration of such matters as model selection, parameter estimation, and the

determination of model-data fit. Nevertheless, it would be misleading to convey the impression that issues and technology associated with item response theory are fully developed and without controversy. Still, considerable progress has been made since the seminal papers by Frederic Lord (1952, 1953). It would seem that item response model technology is more than adequate at this time to serve a variety of uses (see, for example, Lord 1980) and there are several computer programs available to carry out item response model analyses (see Hambleton & Cook, 1977).

The purposes of this paper are to address (1) the measurement philosophy underlying item response theory, (2) the assumptions underlying one of the more popular of the item response models, the three-parameter logistic model, (3) the strengths and weaknesses of the three-parameter model, and present gaps in our knowledge of the model, (4) several promising three-parameter model applications, (5) extensions and new applications of the model, and (6) several controversies.

2. Measurement Philosophy

There are many well-documented shortcomings of standard testing and measurement technology.[1] For one, the values of such useful item statistics as item difficulty and item discrimination depend on the

[1] "Standard testing and measurement technology" refers to commonly used methods and techniques for test design and analysis.

particular examinee samples in which they are obtained. The average level of ability and the range of ability scores in an examinee group influences the values of the item statistics, often substantially. This means that the item statistics are only useful when constructing tests for examinee populations which are very similar to the sample of examinees in which the item statistics were obtained. Another shortcoming of standard testing technology is that comparisons of examinees on an ability measured by a set of test items comprising a test are limited to situations where examinees are administered the same (or parallel) test items. But, a problem is that many achievement and aptitude tests are (typically) suitable for middle-ability students and so the tests do not provide very precise estimates of ability for either high- or low-ability examinees. Increased test score validity without any increase in test length can be obtained if the test difficulty is matched to the approximate ability level of each examinee. But, when several forms of a test which vary substantially in difficulty are used, the task then of comparing examinees becomes more complex because test scores, only, cannot be used. For example, two examinees who perform at a 50% level on two tests which differ substantially in difficulty cannot be considered equivalent in ability, but how different are they in ability? And, how can the ability levels of two examinees be compared when they receive different scores on tests which vary in their difficulty?

Another shortcoming of standard testing technology is that it provides no basis for determining what a particular examinee might do

when confronted with a test item. Such information is necessary, for example, if a test designer desires to predict test score characteristics in one or more populations of examinees or to design tests with particular characteristics for certain populations of examinees. In addition to the three shortcomings of standard testing technology mentioned above, standard testing technology has failed to provide satisfactory solutions to many testing problems: For example, the design of tests, identification of biased items, and the equating of test scores. For these and other reasons, psychometricians have been investigating and developing more appropriate theories of mental measurements.

Item response theory purports to overcome the shortcomings of classical or standard measurement theory by providing an ability scale on which examinee abilities are independent of the particular choice of test items from the pool of test items over which the ability scale is defined. Ability estimates obtained from different item samples for an examinee will be the same except for measurement errors. This feature is obtained by incorporating information about the items (i.e., their statistics) into the ability estimation process. Also, item parameters are defined on the same ability scale. They are, in theory, independent of the particular choice of examinee samples drawn from the examinee pool for whom the item pool is intended although errors in item parameter estimation will be group dependent. More will be said about this point later. Again, item parameter invariance across samples of examinees differing in ability is achieved by incorporating information about examinee ability levels into the item parameter estimation process. Finally, by deriving

standard errors associated with the ability estimates, another of the criticisms of the classical test model can be overcome.

In summary, the goal of item response theory is to provide both invariant item statistics and ability estimates. These features will be obtained when there is a reasonable fit between the chosen model and the data set. Through the estimation process, items and persons are placed on an ability scale in such a way that there is as close a relationship as possible between the expected examinee probabilities for success on test items obtained from the estimated item and ability parameters and the actual probabilities of performance for examinees positioned at each ability level. Item parameter estimates and examinee ability estimates are revised continually until the maximum agreement possible is obtained between predictions based on the ability and item parameter estimates and the actual test data.

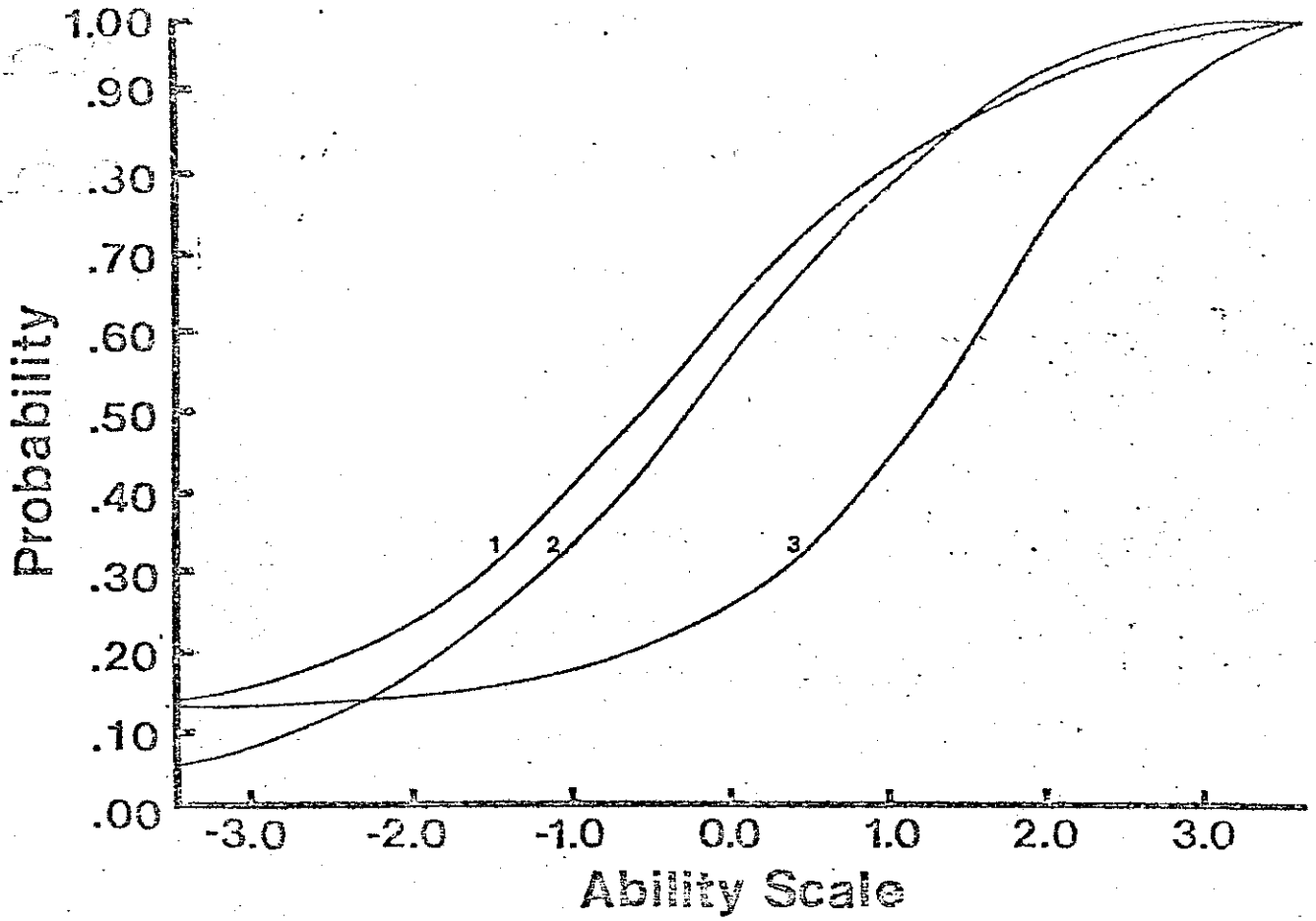
The feature of item parameter invariance can be observed in Figure 1. In the upper part of the figure are three item characteristic curves (ICCs); in the lower part are two distributions of ability. When the chosen model fits the data set the same ICCs are obtained regardless of the distribution of ability in the sample of examinees used to estimate the item parameters. Notice that an ICC provides the probability of examinees at a given ability level answering each item correctly but the probability value does <u>not</u> depend on the number of examinees located at the ability level. The number of examinees at each ability level is different in the two distributions. But, the probability value is the same for examinees in each ability distribution or even in the combined distribution. Of course suitable item parameter estimation will require a heterogeneous

distribution of examinees on the ability measured by the test. It is possible that to some researchers the property of item invariance may seem surprising and unlikely to be obtained in practice, but it is a property which is obtained whenever we study, for example, the linear relationship (as reflected in a regression line) between two variables, X and Y. The hypothesis is made that a straight line can be used to connect the average Y scores conditional on the X scores. When the hypothesis of a linear relationship is satisfied, the same linear regression line is expected regardless of the distribution of X scores in the sample drawn. Of course proper estimation of the line does require that a suitably heterogeneous group of examinees be chosen. The same situation arises in estimating the parameters for the item characteristic curves which are also regression lines (albeit, non-linear).

3. Assumptions

When fitting an item response model to a test data set, assumptions concerning three aspects of the data set are commonly made (Lord, 1980; Wright & Stone, 1979). These three assumptions will be introduced next.

Dimensionality. It is commonly assumed that only one ability is being measured by a set of items in a test. Of course, this assumption cannot be strictly met because there are always many cognitive, personality, and test-taking factors which impact on test performance, at least to some extent. These factors might include level of motivation, test anxiety, ability to work quickly, knowledge of the correct use of answer sheets, and other cognitive skills in addition to the dominant one measured by the set of test items. What



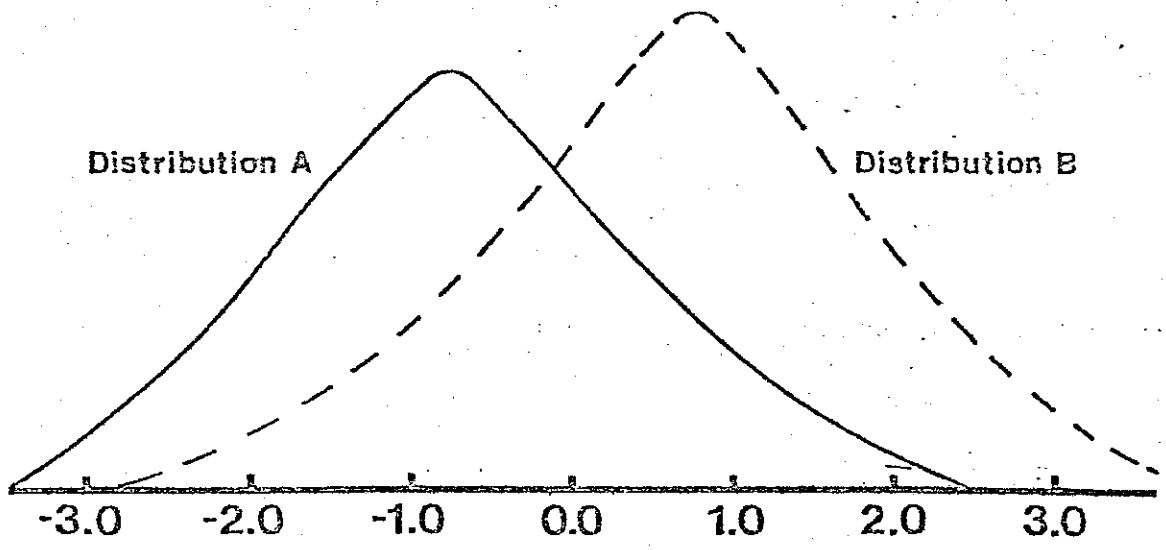


Figure 1. A diagram showing the independence of the shape of item characteristic curves from the underlying ability distribution.

is required for this assumption to be met adequately by a set of test data is a "dominant" component or factor which influences test performance. This dominant component or factor is referred to as the ability measured by the test. This is the ability on which examinees are being measured. All other contributing factors to test performance are defined as errors.

Item response models in which a single ability is presumed sufficient to explain or account for examinee performance are referred to as unidimensional models. Those models in which it is assumed that more than a single ability is necessary to account for examinee test performance are referred to as multi-dimensional models. These latter models are complex, and to date, not well-developed.

Principle of local independence. There is an equivalent assumption to the assumption of unidimensionality known as the assumption of the principle of local independence[1] (Lord & Novick, 1968; Lord, 1980). In words, the assumption requires that the probability of an examinee answering an item correctly (obtained from a one-dimensional model) is not influenced by his/her performance on other items in a test. When an examinee learns information from one test item which helps him or her on other test items the assumption is violated. What the assumption means then is that only the examinee's ability and the characteristics of the test item related to the dominant trait measured by the test influence performance.

Suppose we let u_j be the response of a randomly chosen examinee on items j (j=1, 2, ..., n), and u_j=1, if the examinee answers the

[1] Actually the equivalence only holds when the principle of local independence is defined in the one-dimensional case.

item correctly, and $u_j=0$, if the examinee answers the item incorrectly. Suppose also we let the symbols, P_j, and Q_j ($Q_j=1-P_j$) denote the probability of the examinee answering the item correctly and incorrectly, respectively. The assumption of the principle of local independence in mathematical terms can then be stated in the following way:

$$\text{Prob}\ (U_1 = u_1,\ U_2 = u_2,\ \ldots\ ,\ U_n = u_n)$$

$$= P_1^{u_1}\ Q_1^{1-u_1}\ P_2^{u_2}\ Q_2^{1-u_2}\ \ldots\ P_n^{u_n}\ Q_n^{1-u}$$

$$= \prod_{j=1}^{n} P_j^{u_j}\ Q_j^{1-u_j}$$

In words, the assumption of local independence in the one dimensional case requires that the probability of any response pattern occurring for an examinee is given by the product of probabilities associated with his/her successes and/or failures on the test items. The probabilities are obtained from a one-dimensional model.

Mathematical form of the item characteristic curves. An item characteristic curve is a mathematical function that relates the probability of success on an item to the ability measured by the set of items contained in the test. There is no concept comparable to the notion of an item characteristic curve in standard test technology. A primary distinction among different item response models is in the mathematical form of the corresponding item characteristic curves. It is up to the user to choose one of the many mathematical forms for the shape of the item characteristic curves. In doing so, an assumption

about the items is being made which can be verified later by how well the chosen model "explains" the observed test results.

Each item characteristic curve for a particular item response model is a member of a family of curves of the same general form. The number of parameters required to describe the item characteristic curves in the family will depend on the particular item response model. With the three-parameter logistic model, statistics which correspond approximately to the notions of item difficulty and discrimination (used in standard testing technology), and the probability of low-ability examinees answering an item correctly, are used. The mathematical expression for the three-parameter logistic curve is:

$$(1) \qquad P_g(\theta) = c_g + (1-c_g)\frac{e^{Da_g(\theta - b_g)}}{1+e^{Da_g(\theta - b_g)}}, \quad g=1, 2, \ldots, n,$$

where:

$P_g(\theta)$ = the probability that an examinee with ability level answers item g correctly,

b_g = the item g difficulty parameter,

a_g = the item g discrimination parameter,

c_g = the lower asymptote of an ICC representing the probability of success on item g for low-ability examinees,

D = 1.7 (a scaling factor),

and

n = the number of items in the test.

The parameter c_g is the lower asymptote of the item characteristic curve and represents the probability of examinees with low ability correctly answering an item. The parameter c_g is included in the model to account for test response data at the low end of the ability continuum, where among other things, guessing is a factor in test performance. It is now common to refer to the parameter c_g as the pseudo-chance level parameter in the model.

Typically, c_g assumes values that are smaller than the value that would result if examinees of low ability were to guess randomly to the item. As Lord (1974) has noted, this phenomenon can probably be attributed to the ingenuity of item writers in developing "attractive" but incorrect choices. For this reason, c_g is no longer called the "guessing parameter". To obtain the two-parameter logistic model from the three-parameter logistic model, it must be assumed that the pseudo-chance level parameters have zero-values. This assumption is most plausible with free response items but it can often be approximately met when a test is not too difficult for the examinees. For example, this assumption may be met when competency tests are administered to students following effective instruction. Perhaps the most popular of the present item response models is the one-parameter logistic model (or commonly named the "Rasch Model" after Georg Rasch, the discoverer of the model). It can be obtained from the three-parameter logistic model by assuming that all items have pseudo-chance level parameters equal to zero and by assuming all items in the test are equally discriminating. Also, the one-parameter model, or Rasch model as it is commonly referred to, can be produced from a different set of measurement principles and assumptions.

Readers are referred to Choppin (1983) for an alternate development of the Rasch model. The viability of these assumptions is discussed by Hambleton et al. (1978).

Item characteristic curves for the latent linear model[1] and the one-, two-, and three-parameter logistic models are shown in Figure 2. Readers are referred to Hambleton (1979), Lord (1980), and Wright and Stone (1979) for additional information about logistic test models.

4. Strengths, Weaknesses, and Gaps

The exploration of item response models and their application to eductional testing and measurement problems has been under study for about fifteen years now. Certainly there are many problems requiring resolution but enough is known about item response models to use them successfully in solving many testing problems (see Lord, 1980; Hambleton, 1983). Item response models, when they provide an accurate fit to a data set, and in theory, the three-parameter logistic model, will fit a data set more accurately than a logistic model with fewer item parameters, can produce invariant item and ability parameters described earlier. Some of these promising applications will be described in the next two sections (also see, Hambleton, 1983).

On the negative side, the three-parameter model is based upon several strong assumptions. (Of course, the one- and two-parameter logistic models are based on even stronger assumptions.) When these assumptions are not met, at least to an approximate degree, desirable

---

1. The item characteristic curves for the latent linear model are of the form:

$$P_g(\theta) = b_g + a_g \theta .$$

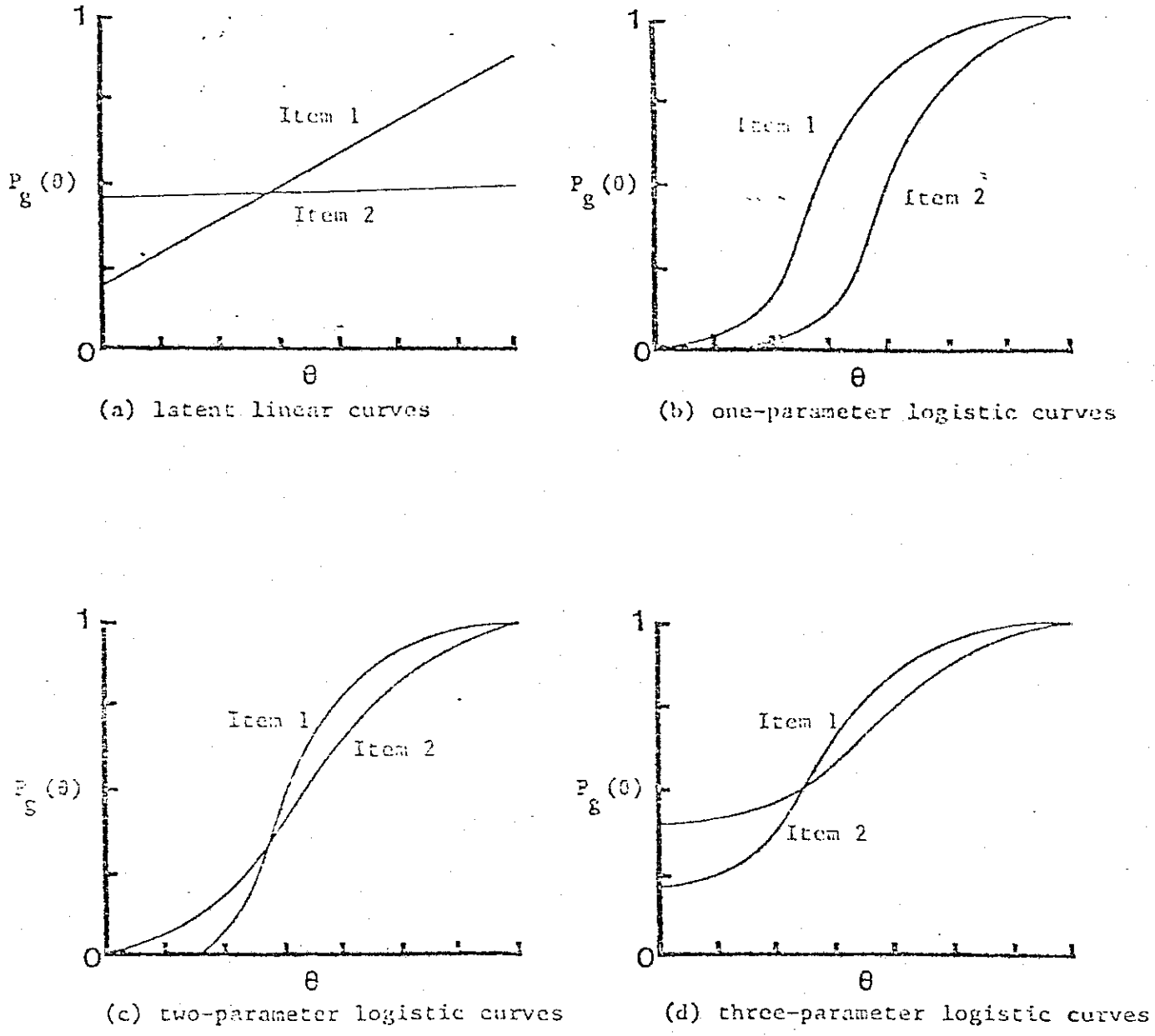


Figure 2. Examples of item characteristic curves.

features expected from applying the three-parameter model will not be obtained. Other weaknesses, presently, of the three-parameter model are (1) the need of rather large numbers of items and examinees for proper item parameter estimation, (2) the relatively high computer costs for obtaining item and ability parameter estimates, and (3) the difficulties inherent in interpreting a complex model for test practitioners.

On the first point, Lord (1980) suggested examinee sample sizes in excess of 2,000 are needed. Perhaps Lord is overly conservative in his figure but it does appear that sample sizes in excess of 600 or 700 are needed, and a disproportionate number of examinees near the lower end of the ability scale so that the c parameters can be estimated properly. Because of the required minimum sample sizes, small scale measurement problems (e.g., teacher-made tests) cannot properly be addressed with the three-parameter model. With respect to the second point, it is common to report high costs associated with using LOGIST although there is evidence that the LOGIST program will run substantially faster and cheaper on some computers. Hutten (1981) reported an average cost of $69 to run 25 data sets with 1,000 examinees and 40 test items on a CYBER 175 ($800/hour for CPU time). Finally, the untrained test developer will have difficulty working with three statistics per item but as CTB/McGraw-Hill has shown in building the latest version of the California Tests of Basic Skills, test editors can be trained to successfully use the additional information provided by the three-parameter model (Yen, 1983).

There is (at least) one practical shortcoming of the three-parameter model and its applications: There does seem to be a

shortage of available computer programs to carry out a three-parameter logistic model analysis. The most readily available program is LOGIST, described by Wingersky (1983) and Wingersky, Barton, and Lord (1982). The most readily available version of this program runs on IBM equipment although there is evidence that the program may run substantially faster on other computers. Additional investigation of this finding is needed along with on-going studies to try and speed up the convergence of estimates. In addition, there may be other ways to improve the estimation process. Swaminathan and Gifford (1981) have obtained very promising results with Bayesian item and ability parameter estimates. Their results compare favorably with results from LOGIST and they can be obtained considerably faster and more cheaply than the same estimates obtained with LOGIST.

There are (at least) three areas in which we lack full understanding of item response models. First, additional robustness studies with the one- and two-parameter logistic models are needed and with respect to a number of promising applications. What is the practical utility of the three-parameter model in comparison to the one- and two-parameter models? Second, appropriate methods for testing model assumptions and determining the goodness of fit between a model and a data set are needed. Hambleton and his colleagues (Hambleton, 1980; Hambleton, Murray, & Simon, 1982) have made a promising start by organizing many of the present methods and developing several new ones. Much of their work involves the use of graphs, replications, residual analyses and cross validation procedures. More work along the same general lines would seem

desirable. Third, there is a great need for persons to gain experiences with the three-parameter model and to share their new found knowledge and experiences with others.

5. Applications[1]

In this section, several promising applications of the three-parameter logistic model will be described briefly: Item banking, test development, criterion-referenced testing, item bias, and adaptive testing. Other applications of the three-parameter model are discussed by Hambleton et al. (1978), Lord (1980), and Hambleton (1983).

Item banking. The development of criterion-referenced testing technology has resulted in increased interest in item banking (Choppin, 1976). An item bank is a collection of test items, "stored" with known item characteristics. Depending on the intended purpose of the test, items with desired characteristics can be drawn from the bank and used to construct a test with known properties. Although classical item statistics (item difficulty and discrimination) have been employed for this purpose, they are of limited value for describing the items in a bank because these statistics are dependent on the particular group used in the item calibration process. Latent trait item parameters, however, do not have this limitation, and consequently are of much greater use in describing test items in an item bank (Choppin, 1976). The invariance property of the latent trait item parameters makes it possible to obtain item statistics that are comparable across dissimilar groups. Since the item parameters depend on the ability scale, it is not possible to directly compare

---

[1] Some of the material in this section is taken from and/or edited from a paper by Hambleton et al. (1978).

latent trait item parameters derived from different groups of examinees until the ability scales are equated in some way. Fortunately, the problem is not too hard to resolve since Lord and Novick (1968) have shown that the item parameters in the two groups are linearly related. Thus, if a subset of calibrated items is administered to both groups, the linear relationship between the estimates of the item parameters can be obtained by forming two separate bivariate plots, one establishing the relationship between the estimates of the item discrimination parameters for the two groups, and the second, the relationship between the estimates of the item difficulty parameters. Having established the linear relationship between item parameters common to the two groups, a prediction equation can then be used to predict item parameters for those items not administered to the first group. In this way, all item parameters can be equated to a common group of examinees and corresponding ability scale. One large test publishing company, the California Test Bureau/McGraw-Hill, presently customizes tests for school districts wih items calibrated using the three-parameter logistic model.

Test development. The three-parameter model is presently being used by a number of organizations in test development (e.g., CTB/McGraw-Hill, ETS). The three-parameter model provides the test developer with not only sample invariant item parameters but also with a powerful method of item selection (Birnbaum, 1968). This method involves the use of information curves, i.e., items are selected depending upon the amount of information they contribute to the total amount of information supplied by the test (Lord, 1980)[1]. One of the

[1] Readers are referred to Hambleton (1979) for an introduction to item and test information and efficiency curves.

useful features of item information curves is that the contribution of each item to the test information function can be determined without knowledge of the other items in the test. When standard testing technology is applied the situation is very different. The contribution of any item to such statistics as test reliability cannot be determined independently of the characteristics of all the other items in the test.

Lord (1977) outlined a procedure for use of item information curves to build a test to meet any desired set of specifications. The procedure employs a pool of calibrated items, with accompanying information curves, such as might be obtained from the item banking methods described earlier. The procedure outlined by Lord consists of the following steps:

1. Decide on the shape of the desired test information curve. Lord (1977) calls this the target information curve.
2. Select items with item information curves that will fill up the hard-to-fill areas under the target information curve.
3. After each item is added to the test, calculate the test information curve for the selected test items.
4. Continue selecting test items until the test information curve approximates the target information curve to a satisfactory degree.

An example of the application of this technique to the development of tests for differing ranges of ability (based on simulated data) is given by Hambleton (1979).

Criterion-referenced testing. A principal use of a criterion-referenced test is to estimate an examinee's level of mastery (or "ability") on an objective. Thus, a straightforward application of the three-parameter model would produce examinee ability scores. Among the advantages of this application would be that items could be sampled (for example, at random) from an item pool for each examinee, and all examinee ability estimates would be on a common scale. A potential problem with this application, however, concerns the estimation of ability with relatively short tests.

Since item parameters are invariant across groups of examinees, it would be possible to construct criterion-referenced tests to "discriminate" at different levels of the ability continuum. Then, a test developer might select an "easier" set of test items for a pretest than a posttest, and still be able to measure "examinee growth" by estimating examinee ability with the three-parameter model at each test occasion on the same ability scale. This cannot be done with classical approaches to test development and test score interpretation. If we had a good idea of the likely range of ability scores for the examinees, test items could be selected so as to maximize the test information in the region of ability for the examinees being tested. The optimum selection of test items would contribute substantially to the precision with which ability scores were estimated. In the case of criterion-referenced tests, it is common to observe substantially lower test performance on a pretest than on a posttest; therefore, the test constructor could select the easier test items from the domain of items measuring an objective for the pretest and more difficult items

could be selected for the posttest. This would enable the test constructor to maximize the precision of measurement of each test in the region of ability where the examinees would most likely be located. Of course, if the assumption about the location of ability scores was not accurate, gains in precision of measurement would not be obtained.

The results reported in Tables 1 and 2 (from Hambleton, 1979) show clearly the advantages of "tailoring" a test to the ability level of a group. Of course, the potential improvements depend on the validity of a test developer's assumption about the examinee ability distribution. If he or she uses an incorrect prior distribution as a basis for designing a test, the resulting test will certainly not have the desired characteristics.

Item bias. Identifying biased items in a test usually involves comparing the performance of the subgroups of interest (e.g., Blacks, Hispanics, and Whites) on the test items. The problem that arises is that differences among the subgroups due to bias is confounded with any true differences in abilities among the subgroups. Needed is an item bias detection method that can control for true ability differences. Via a three-parameter model analysis, it is possible to compare corresponding item characteristic curves. At each ability level, independent of the proportion of examinees in each subgroup who are located at the ability level, the expected proportion of successes in each subgroup,obtained from the ICCs, can be compared. The ICCs estimated in each group, in theory, do not depend upon the underlying ability distributions. Any differences in the curves, beyond the

Table 1

Test Information Curves and Efficiency for Three Criterion-Referenced Test Designs From a Domain of Items of Equal Discrimination and Pseudo-chance Levels Equal to .20

Ability Level	Test Information Curves			Efficiency (Relative to the "Wide Range Form")		Change in Effective Test Length	
	"Wide Range Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"
-3.0	.22	.36	.07	1.63	.31	63%	-69%
-2.0	.86	1.31	.36	1.53	.42	53%	-58%
-1.0	2.08	2.81	1.31	1.35	.63	35%	-37%
0.0	3.04	3.29	2.81	1.08	.92	8%	-8%
1.0	2.76	2.28	3.29	.82	1.19	-18%	19%
2.0	1.69	1.12	2.28	.66	1.35	-34%	35%
3.0	.79	.46	1.12	.59	1.42	-41%	42%

Table 2

Test Information Curves and Efficiency for Three Criterion-Referenced Test Designs From a Domain of Items with Varying Discrimination Indices and Pseudo-chance Levels Equal to .20

Ability Level	Test Information Curves			Efficiency (Relative to the "Wide Range Form")		Change in Effective Test Length	
	"Wide Range Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"
-3.0	.24	.37	.08	1.58	.35	58%	-65%
-2.0	.86	1.27	.37	1.48	.44	48%	-56%
-1.0	2.02	2.71	1.27	1.35	.63	35%	-37%
0.0	2.94	3.18	2.71	1.08	.92	8%	-8%
1.0	2.65	2.16	3.18	.81	1.20	-19%	20%
2.0	1.59	1.06	2.16	.67	1.36	-33%	36%
3.0	.75	.46	1.06	.61	1.41	-39%	41%

usual sampling errors, can be attributed to differential subgroup responses to the items, i.e., bias. It is becoming routine practice for several large test publishers to investigate bias in test items with the aid of the three-parameter logistic model. 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 other logistic models for studying bias.

Adaptive testing. Possibly the first and most well-developed application of the three-parameter logistic model to date is adaptive testing (Lord, 1980; Weiss, 1980). In adaptive testing each examinee is administered a set of test items "tailored" or "adapted" to his/her ability level. Clearly, total test scores cannot provide an adequate basis upon which to compare examinees. Some examinees will be administered sets of test items which are substantially more difficult (or easier) than the test items administered to other examinees. By calibrating test items using the three-parameter logistic model in advance of the actual testing, and using the three-parameter model to estimate examinee ability levels, examinees can be compared even though the test items administered to different examinees may differ substantially in difficulty. Because of the ready availability of the computer, scoring difficulties associated with the use of the three-parameter model can be overcome easily.

The U.S. military is firmly committed to the use of adaptive testing with the three-parameter model in many of its testing programs. Presently a feasibility study is being conducted along

with the preparation of plans for adaptive testing implementation and evaluation of the total adaptive testing system.

6. Possible Extensions/New Applications

Numerous researchers are presently addressing the development of new item response models. For example, Samejima (1979) is exploring the development of multidimensional models in which item options are ranked based on their relationship to ability, and characteristic curves are produced for each option. McDonald (1982) has provided a general formulation for generating a wide range of multidimensional linear and non-linear polychotomous item response models. Bock, Mislevy, and Woodson (1982) have described a two-parameter item response model which can handle continuous data and where the unit of analysis can be a group (e.g., the classroom or a school). This model will be especially useful in program evaluation investigations. A minor variation of the three-parameter model which appears to have some utility is a model in which a common value of the c parameter is used for all test items (Wingersky, 1983). This revised three-parameter model will receive some use in the coming years. A four-parameter logistic model has also been suggested (the fourth parameter is the upper asymptote) but it appears to have very limited practical usefulness. All of these new models can be viewed as modifications/extensions of the three-parameter logistic model and they will undoubtedly receive study from researchers in the coming years.

Because of the newness of the IRT area, all applications of the three-parameter model might legitimately be classified as new. For the purposes of this paper, "new applications" will be those which to

date have not been published. Two new applications, then, of the three-parameter model to the problems of item selection (Hambleton & de Gruijter, 1983) and score prediction (Hambleton & Martois, 1983) will be described briefly next.

Item selection. Item response models appear useful to the problem of item selection because they lead to item statistics which are referenced to the same scale on which examinee abilities are defined. In addition, it should be noted that IRT provides a procedure for placing a cut-off score which is normally set on a proportion-correct scale defined over a domain of items on the same scale as the test items and the examinees (Lord, 1980). Therefore, the usefulness of a test item for measurement at any point on the ability scale can be assessed.

Hambleton and de Gruijter (1983) described a nine step procedure for selecting test items using three-parameter model item statistics, and via a computer simulation study showed the advantages, at least in the absence of errors associated with item parameter estimates, of item selection with the aid of IRT over a standard item selection procedure.

Test score predictions. The concept of item banking has attracted considerable interest in recent years from school districts, state departments of education, and test publishing companies. When item banks consist of test items which are technically sound and validly measure the objectives or competencies to which they are referenced, the task of producing high quality tests is made considerably easier. Item banks are most often used to construct

criterion-referenced tests (CRTs) or mastery tests or competency tests, as they are sometimes called. What is not commonly available for use with these CRTs are derived scores such as percentiles. Derived scores are not always valued but on occasion they are required by school districts who receive federal funds (e.g., Title I) for they must evaluate their funded programs with national norms (e.g., percentile scores).

In theory, the problem faced by school districts who require information for (1) diagnosing and monitoring student performance in relation to competencies and (2) normative scores for the comparison of examinees is easy to solve. Teachers can use their item banks to build classroom tests on an "as-needed" basis, and when the need arises, they can administer any necessary commercially available standardized norm-referenced tests. But this solution has problems: (1) the amount of testing time for students is increased, and (2) the financial costs of school testing programs is increased. On the other hand, when testing time is held constant, and norm-referenced tests are administered, there is less time available for instructionally relevant testing (i.e., CRTs). A more satisfactory solution would allow teachers to administer test items measuring objectives of interest in their instructional programs, and at the same time, allow for normative scores to be estimated from the test items which are administered. An often used solution of selecting a norm-referenced test to provide normative scores and criterion-referenced information through the interpretation of examinee performance on an item by item basis is not very suitable criterion-referenced measurement and will not insure that all competencies of interest are measured in the test.

Hambleton (1980) suggested a possible item response model solution to the problem of providing both instructonal information and normative information from a single test. A latent ability scale to which a large pool of test items are referenced can be very useful in obtaining normative scores from tests constructed by drawing items from the pool. A norms table can be prepared from the administration of a sample of items in the pool. Then the norms table can be used successfully with any tests which are constructed by drawing items from the pool. Local norms can be prepared by districts who build their own item banks. A test publishing company probably would prepare national norms for selected tests constructed from their item banks.

Hambleton and Martois (1983) recently finished a study in which it was found that both the one- and the three-parameter logistic models resulted in excellent predictions of how examinees performed on a norm-referenced test. Predictions were made from tests with items that were easier, comparable to, or harder than items in the normed test. Similar results were obtained in three subject areas at two grade levels. Further research along the same general lines seems highly desirable because of the importance of the problem area.

7. Controversies

Perhaps like any emerging area, item response theory has generated considerable controversy and strong emotional feelings in support of one model versus another. Much of the debate has centered on the choice between the one- and three-parameter logistic models. There has also been some controversy surrounding the utility of

Bayesian estimators (Samejima versus Novick and Swaminathan) and the appropriateness of item response models for the analysis of aptitude versus achievement tests. On this latter point there is some feeling that items on achievement tests are instructionally sensitive and therefore item response model statistics will not be invariant in pre- and post-instructional groups.

With respect to the choice of the one- versus the three-parameter logistic model, a number of questions have arisen:

1. What is the effect of boundary constraints placed on item and ability parameter estimates obtained with LOGIST?

2. What is the practical utility of the three-parameter model? In most practical settings, won't the two models produce highly similar results?

3. What is the additional cost of running a three-parameter model analysis and is the practical utility of the gains that accrue worth the financial costs and the added complexity which results?

4. Since examinees can guess the answers to multiple-choice test items, the three-parameter model should be selected on the basis of this a priori consideration (Traub, 1983).

5. How well do the item response models fit any data sets? This point is in dispute because many of the present goodness of fit statistics have been found to be inappropriate (e.g., see papers by Wollenberg, 1980; Divgi, 1981).

These and other questions will undoubtedly be addressed in the coming years. Answers will contribute to our knowledge of the three-parameter logistic model and the situations in which the model should be used.

REFERENCES

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

Bock, R.D., Mislevy, R., & Woodson, C. The next stage in eductional assessment. Educational Researcher, 1982, 16, 4-11.

Choppin, B.H. Recent developments in item banking: A review. In D.N.M. deGruijter & L.J.Th. van der Kamp (Eds.), Advances in psychological and educational measurement. New York: Wiley, 1976.

Choppin, B.H. The Rasch model for item analysis. CSE Report No. 218. Los Angeles: UCLA Center for the Study of Evaluation, 1983.

Divgi, D.R. Does the Rasch model really work? Not if you look closely. Paper presented at the Annual Meeting of NCME, Los Angeles, 1981.

Hambleton, R.K. Latent trait models and their applications. In R. Traub (Ed.), Methodological developments: New directions for testing and measurement (No. 4). San Francisco: Jossey-Bass, 1979.

Hambleton, R.K. Latent ability scales, interpretations, and uses. In S. Mayo (Ed.), New directions for testing and measurement: Interpreting test performance. San Francisco: Jossey-Bass, 1980.

Hambleton, R.K. Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia, 1983.

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

Hambleton, R.K., & de Gruijter, D.N.M. Application of item response models to criterion-referenced test item selection. Journal of Educational Measurement, 1983, 20, in press.

Hambleton, R.K., & Martois, J. Evaluation of a test score prediction system based upon item response model principles and procedures. In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia, 1983.

Hambleton, R.K., Murray, L., & Simon, R. Applications of item response models to NAEP mathematics exercise results. Final Report. Submitted to the Educational Commission of the States, 1982.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., & Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Hutten, L. Fitting the one- and three-parameter models to a variety of tests. Laboratory of Psychometric and Evaluative Research Report No. 116. Amherst, MA: School of Education, University of Massachusetts, 1981.

Lord, F.M. A theory of test scores. Psychometric Monograph No. 7, 1952.

Lord, F.M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-75.

Lord, F.M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

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

Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

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

McDonald R.P. Linear versus non-linear models in item response theory. Applied Psychological Measurement, 1982, 6, in press.

Samejima, F. A new family of models for the multiple-choice item. Office of Naval Research, Research Report 79-4, 1979.

Swaminathan, H., & Gifford, J.A. Bayesian estimation in the three-parameter logistic model. Laboratory of Psychometric and Evaluative Research Report No. 119. Amherst, MA: School of Education, University of Massachusetts, Amherst, 1981.

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

Weiss, D. (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis, MN: University of Minnesota, 1980.

Wingersky, M.S. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia, 1983.

Wingersky, M.S., Barton, M.A., & Lord, F.M. LOGIST user's guide. Princeton, NJ: Educational Testing Service, 1982.

Wollenberg, A.L. van den. On the Wright-Panchapakesen goodness of fit test for the Rasch model. Internal Report 80-MA-02. Nijmegen, The Netherlands: Katholieke Universiteit, Vakgroep Mathematische Psychologie, Psychologisch Laboratorium, 1980.

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

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