DOCUMENT RESUME

ED 206 642 TM 810 530

AUTHOR Reckase, Mark D.
TITLE To Use or Not to Use--(The One- or Three-Parameter Logistic Model) That Is the Question.
INSTITUTION Missouri Univ., Columbia.
SPONS AGENCY Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE Apr 81
CONTRACT N00014-77-C-0097
NOTE 13p.: Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests: Comparative Analysis: *Latent Trait Theory: Scores: *Test Construction: Test Reliability
IDENTIFIERS *One Parameter Model: *Three Parameter Model: Vertical Equating

ABSTRACT

Definition of the issues to the use of latent trait models, specifically one- and three-parameter logistic models, in conjunction with multi-level achievement batteries, forms the basis of this paper. Research results related to these issues are also documented in an attempt to provide a rational basis for model selection. The application of the latent trait models is evaluated in terms of: (1) the assistance they can lend to test construction: (2) their use in the vertical equating of test scores; and (3) their use in comparing score scales. It is suggested that the target information function of the latent trait models can give control over test construction and precision, especially with the three-parameter model: but, there appears to be no acceptable way to select test items using the one-parameter model. The latter model is criticized for perpetuating the idea that all items are equal. Even when the test battery is sorted into unidimensional subtests, it is reported that neither the one- nor the three-parameter model is adequate for vertical equating purposes. It is concluded that both models have advantages and disadvantages. The recommendation that both models be used in conjunction with traditional procedures is suggested. (AEF)

To use or not to use--- [the one- or three-parameter logistic model] that is the question

Mark D. Reckase
University of Missouri-Columbia

In the two decades since 1960, a revolution has taken place in the areas of test theory and its applications. The one- and three-parameter logistic models that were obscure curiosities in 1960 have become widely researched, sometimes applied, and often debated components of measurement technology. Computer programs have been developed to determine estimates of the parameters of these models (Wright & Panchapakesan, 1968; Wood, Wingersky & Lord, 1976), place the estimates on the same scale (Reckase, 1979), and estimate the ability of individuals for sets of items (Owen, 1976). The one-parameter model (Rasch) has been applied to reading tests (Woodcock, 1974; Rentz & Bashaw, 1977) and to state testing programs (Forbes, 1976). The three-parameter model has had fewer but similar applications (Cowell, 1979; Lord, 1968).

Because of the growing acceptance of latent trait models as an improvement over traditional procedures, serious thought is now being given to applying the various latent trait models to major testing programs. The one-parameter model is being considered for use in the analysis and equating of the widely used Stanford Achievement Test, and the three-parameter logistic model is being considered as the basis for tailored testing administration of the Armed Services Vocational Aptitude Battery (ASVAB). Unfortunately, these procedures are often accepted for use without a thorough evaluation of competing procedures. When competing procedures are considered, the selection is often made on the basis of philosophical distinctions rather than empirical evaluations. The purpose of this paper is to review the data available related to the application of one- and three-parameter logistic models and to try and give some rational basis for model selection. The particular orientation used in the evaluation of models will be towards use in the development of multi-level achievement batteries.

Characteristics Required of the Measurement Models

Before trying to compare the statistical models that might be used in the development and application of multi-level achievement test batteries, some consideration must be given to what is required in the development and use of such batteries. Three general areas were defined which require the use of test analysis procedures. They include (a) assistance in test construction, (b) vertical equating of test forms, and (c) formation of multi-level score scale. The

---

Paper presented at the meeting of the American Educational Research Association, Los Angeles, April 1981. This research was supported by the Office of Naval Research under contract number N00014-77-C-0097 from the Personnel and Training Research Programs of the Office of Naval Research.

test construction component can be further subdivided into (a) formation of an item pool, (b) item selection, and (c) evaluation of test quality. In order to make recommendations concerning which model to use with multi-level achievement tests, the applications of the models to each of these areas will be reviewed and recommendations developed. These area recommendations will then be synthesized and a global recommendation given.

Evaluation of the Applications of Latent Trait Models

Assistance in test construction

The first step in the construction of any test is usually the production and tryout of a large number of test items. The traditional analyses performed on the tryout data are the computation of item analysis statistics such as difficulty and discrimination indexes. The sheer number of items required in a multi-level achievement battery make it impossible for all items to be administered to all subjects. This fact unfortunately means that the item statistics from different groups are not comparable. Latent trait theory approaches overcome this lack of comparability by making it relatively easy to translate item statistics determined using different groups onto the same scale. The process has been labeled "linking" in the current literature.

Very little has been done to evaluate linking procedures in the latent trait research literature. One paper by Reckase (1979), however, has discussed some of the basic issues in item pool linking. In that paper, various techniques for item pool linking were reviewed, and linked item calibration results from a series of 50 item tests were compared to the calibration of the full 357 items available, as calibrated on a sample of 4,000 examinees. Sample size and number of common items in the linked tests were variables of interest in the study. The results generally showed that the one-parameter model could be used to link calibrations fairly well if five or more items were in common between tryout tests and if samples of 300 or more were used. For the three parameter model, sample sizes of 1,000 or more were required for acceptable linking and the use of the major axis method with the ANCILLES calibration program (Urry, 1978) or maximum-likelihood linking with the LOGIST program (Wood, Wingersky & Lord, 1976) were found to yield the best results. The three-parameter procedure requires a larger sample size because three parameters need to be estimated and placed on the same scale, rather than just one parameter. If sample size were the only issue, the one-parameter model would be selected at this point because it gives adequate results using fewer cases. However, sample size is not the only issue.

Once an item pool is generated and tried out, the next step in the construction of the test battery is the selection of the actual test items to be used on the final forms. The traditional procedure used to select items is to pick the items with high discrimination indices and with item difficulty in the appropriate range.

Essentially the same process can be followed if the three-parameter logistic model is the basis for item calibration. Items with high discrimination parameter estimates and appropriate difficulty parameter estimates can be selected for the use on the tests. Alternatively, a target information function can be specified and then items can be selected so that their item information functions sum to the target function. Lord (1977) describes this process in more detail.

Item selection using the one-parameter model is typically done differently. In conjunction with the item calibration procedure, a test of the fit of the one-parameter model to the item data is performed, usually using some type of squared difference between observed and expected frequencies (Wright & Stone, 1979). These squared differences are assumed to have a chi-square distribution, and probabilities of fit are determined accordingly. The basis for item selection is to pick items for which the model fits, while deleting the rest. In theory, three violated assumptions contribute to lack of fit with this model. The violations are variation in discrimination, nonzero guessing, and multi-dimensionality. Thus selecting items on the basis of one-parameter logistic model fit should yield tests with low guessing, moderate and equal discrimination, and unidimensionality.

A number of studies have tried to determine whether selecting on fit will really yield tests with the above characteristics. Reckase (1979), in a study of 150 tests items from a series of classroom tests, found lack of fit to be strongly related to guessing level. Brooks (1964) found that selecting items that were fit by the model yielded lower test reliabilities, as measured by the KR-20 formula, than did traditional methods. Hambleton (1969) and Panchapakesan (1969) used simulation procedures in their dissertations to determine the effect of variation in discrimination and guessing on fit. They found that items that differed from the average discrimination parameter by more than .20 tended to cause lack of fit. Guessing was also found to have a strong influence on fit. Thus, variation in discrimination and nonzero guessing do seem to have something to do with lack of fit, but avoiding violations of the assumptions did not seem to improve reliability. In some cases (Anderson, Kearney & Everett, 1968), lack of fit did not seem to be related to any item characteristic.

One possible cause of the ambiguous results concerning the selection of items on the basis of model fit is the basic inadequacy of the chi-square fit statistic. Forster & Karr (1980) summarize these inadequacies as follows:

> "First, it is sensitive to differences in sample size---generally low for small samples, generally high for large samples. Second, it disproportionately weights differences near the top or bottom of the item curve. Finally, it does not provide an adequate indication of where the actual and ideal distributions do not fit."

As an alternative, they suggest looking at the empirical and theoretical item characteristic curves to check fit. Our own experience has shown the fit

statistics to be relatively meaningless. Sometimes more than half of the items were found not to be fit by the model (e.g., 29 out 50 items from the Iowa Tests of Educational Development with a sample of 2,000).

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 in the pool. Typically, use of the best items in a pool would seem desirable, as opposed to using the mediocre items as suggested by selection on the basis of one-parameter model fit.

After the items are selected from the item pool for use on the forms of a test, an overall measure of test quality is usually determined. Traditionally, a test reliability coefficient is computed based on the tryout sample. Unfortunately, this type of statistic is highly sample dependent and only yields an average measure of precision over the full range of ability measured by the test. The test information function (Birnbaum, 1968) used with the latent trait models has substantial advantages over the reliability coefficient because it gives an indication of test precision at all points of the ability scale, and because it is somewhat sample independent. Since the information function can be defined for either the one- or three-parameter models, its availability does not give an advantage to either model, although the availability of the information function does argue for the use of latent trait models. The three-parameter model does tend to give higher values for the information function because of the ability to select on the basis of the discrimination and guessing parameters (Koch & Reckase, 1978).

One other concept that should be kept in mind during the process of item selection is the content validity of the test formed. Most achievement tests are produced using a table of specifications to help insure content validity. This procedure does not usually insure that the test produced will have one dimension---a basic assumption of the latent trait models. The use of a number correct score ignores this issue by simply summing the varied content areas in whatever proportion they happen to appear in the test. The one-parameter model treats the possible multidimensionality in the same way, since the raw score is a sufficient statistic for the one-parameter ability estimate. The three-parameter model yields an ability estimate with a different interpretation because its ability estimates are based on a weighted sum of item scores, with the weights being the discrimination parameter estimates. This weighting procedure has the effect of emphasizing the items measuring one factor in the test while ignoring the others. Thus, for the most part, the three-parameter model ability estimates do not contain information from every component in the test. They only emphasize the largest component (see Reckase, 1979 for a more thorough discussion). This should be kept in mind in selecting the model to be used in obtaining test scores.

Vertical Equating

When a multi-level achievement battery (a battery with forms at several grade levels) is produced, it is usually desirable to equate the scores on the various levels so that scores on different levels can be compared. Traditionally, some type of equipercentile approach was used to equate the test levels. That is, tests at two different levels are administered to the same group and scores with the same percentile rank in the groups are said to be equivalent. Angoff (1971) does a good job of summarizing the traditional procedures for equating.

In recent years latent trait theory approaches have been suggested as an alternative to the time honored procedures because of several theoretical advantages. These advantages include the sample independent nature of item and ability parameter estimates, the capability of getting ability estimates on the same scale regardless of the set of items administered, and the possibility of equating tests using common items as opposed to administering two tests to the same sample.

Both one-parameter (Wright & Stone, 1979) and three-parameter (Marco, 1977) logistic model based procedures have been developed for vertical equating, but only recently have these procedures been evaluated. Both models typically use a procedure that has items in common between the tests to be equated. The items in the tests are then calibrated separately, and a linear equation is determined to translate the item difficulty estimates for the common items from one calibration to the other. This same transformation is used to translate the ability scale of one test to the scale of the other. Alternatively, two tests can be administered to the same sample of people and the linear transformation can be found to equate the ability estimates.

Several studies have been done to evaluate the quality of the vertical equating done using the one-parameter model (Slinde & Linn, 1978, 1979; Gustafsson, 1979; Loyd & Hoover, 1980). Slinde & Linn (1979) equated easy and difficult tests constructed from 60 reading and vocabulary items from the SRA Achievement series. Three different ability groups, high, middle, and low, formed on the basis of Comprehensive Tests of Basic Skills scores were used for the calibration. They found that "For extreme comparisons which involve widely separated groups and tests of substantial different difficulties, the Rasch model does not seem to result in adequate vertical equating of existing tests." They felt that guessing was a major cause of the poor results. Despite the poor results, they felt that the one-parameter model would work reasonably well with groups that are closer in ability and tests closer in difficulty.

Loyd and Hoover (1980) also obtained negative results when evaluating vertical equating using the one-parameter model. Their study was somewhat more realistic than the Slinde & Linn (1979) study because they used three existing test levels of the Iowa Test of Basic Skills instead of constructing easy and difficult tests specifically for the study. They also used sixth, seventh, and eighth grade groups rather than forming different ability groups on the basis of test scores. Their results showed that equating using the one-

parameter model was "inconsistent." The equated scores tended to be higher than the scores that would be obtained if the original test had been taken. In an example of translating scores from one level to another, they demonstrated that the differences could be quite large. On the basis of their results, they conclude that "While latent trait methods show a great deal of promise for improving horizontal equating [linking] of tests, results of the present study, and others, indicate that the use of the Rasch model in vertical equating should be approached with extreme caution." Loyd & Hoover (1980) feel that the major cause of the equating problems is the change of course content with the change in test level.

The results of these studies on vertical equating certainly do not seem positive for the one-parameter model. Unfortunately, similar evaluations of the vertical equating procedures based on the three-parameter model do not yield much better results and are much harder to find. Three relatively obscure studies could be found that evaluated three-parameter based vertical equating. Kolen (1981) compared equipercentile, three-parameter, and one-parameter based procedures on the basis of consistency in a cross-validation study and found that the equipercentile and three-parameter based procedures worked best, while the Rasch model procedure worked poorest.

Patience (1981), in an unpublished paper, reported slightly different results. He compared the equated score scale from an easy, middle, and hard test with the score scale obtained from the three tests administered together as one. The three tests were produced from the verbal subtests of the Iowa Tests of Educational Development in such a way that items were in common between the tests. Patience (1981) found that the equipercentile procedure worked best, followed by the one-parameter procedure and then the three-parameter procedure. He attributed the poor showing of the three-parameter model to unstable estimates of the item parameters despite a relatively large sample size (1,000).

Marco, Petersen & Stewart (1980), in a very complicated and elaborate study, evaluated the equating of easy, medium and hard tests formed from the verbal subtest of the Scholastic Aptitude Test using samples of 1,577 cases. They used a series of statistics based on the difference between actual and equated scores. Five different types of equating procedures were used including procedures based on the one- and three-parameter logistic model. Their results indicated that "For most equatings, the model with the smallest total error was the 3-parameter ICC model." The one-parameter model did relatively poorly in this study. Equipercentile equating was the next best procedure after the three-parameter model based procedures.

The overall trend of the results reported here seem to indicate that the use of the one-parameter model results in serious problems when it is applied to vertical equating, while the value of the three-parameter model is yet to be determined in this area because of the varied results available. The safest conclusion might be to recommend equipercentile based procedures, since they

worked well in both the Kolen (1981) and Patience (1981) studies.

Score Scales

The underlying purpose of vertically equating the set of tests in a multi-level battery is to form a single score scale to which all scores on all test levels can be transformed. Such a multi-level score scale will allow comparison to be made across all test levels. Tracing the improvement in performance of a student is one possible application of such a scale. In theory, the development of such a scale should be easily accomplished on the basis of item characteristic curve models, since any set of items can be used to get scores on the same scale once all of the item calibrations have been linked. Unfortunately, as described in the previous section, the use of latent trait models for vertical equating cannot currently be considered as an acceptable procedure.

With the extensive research being done with latent trait models, there is hope that the problems in vertical equating will be solved in the future. Therefore it is important to look at the implications of the use of latent trait models for the interpretation of the resulting multi-level score scale. The necessity of the analysis of the possible scales is further motivated by the difference in the meaning of ability estimates obtained through the use of the one-parameter and three-parameter logistic models. As mentioned earlier, the one-parameter model yields ability estimates that have meaning equivalent to that of the 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.

In many cases, this distinction in the interpretation of the score scale produced by the latent trait models will have little practical significance. Reckase (1979) has found that for many tests there is a dominant first factor. The one-parameter and three-parameter ability estimates then correlate in the high .90's. However, if the dimensions present in the tests contained in a multi-level battery change with the test level, as suggested by Loyd & Hoover (1980), the ability scales defined by the two procedures might be quite different. If, for example, the major factor in the tests at the various levels remained the same, but the actual proportion of various content areas changed, the three-parameter logistic based estimates would maintain the same meaning, while the one-parameter estimates would change in meaning with the level tested. Of course, the same change in meaning would also occur in the raw scores.

On the other hand, if the major component of the test changed across levels, both procedures would result in ability estimates that had different meanings at the different levels. Thus, changes in the dimensionality of the test battery at the different levels can be seen to be a critical issue in score scale development. The same problem plagues procedures based on raw scores and equi-percentile methods, but it has been mostly ignored until recently because the assumptions of the methods were not clearly stated.

Discussion

In the course of this paper, an attempt has been made to define the issues related to the use of latent trait models in conjunction with multi-level achievement batteries, and to summarize some of the research results related to those issues. The issues identified included (a) the use of the models in test construction, (b) the use of the models for vertical equating, and (c) the comparability of the score scales obtained using the models. The one- and three-parameter logistic models were concentrated on in the summary because the majority of work has been done with these two models.

In the area of test construction, the availability of the concepts of item and test information give the latent trait models a clear advantage over traditional methods. Target information functions can give much greater control over test construction, putting test precision where it is desired. Also, the information function indicates the precision at each ability level.

For some reason the concept of the information function has been embraced by users of the three-parameter model, while being largely ignored by the users of the one-parameter model. The result is that there seems to be no acceptable way to select items for a test using the one-parameter model. The available research seems to indicate that selecting on fit is not an adequate procedure, and since all of the items are assumed to have equal discrimination and no guessing, the only source of selection information is the estimate of item difficulty. The one-parameter model perpetuates the myth that all items are equal---an idea that has been accepted as long as raw scores have been used. This is a serious problem with the use of the one-parameter model.

The one-parameter model does have the advantages of simplicity of estimation and smaller sample size requirements than the three-parameter model. But this should not be a major issue for large group tests. Adequate numbers of cases for the requirements of the three-parameter model have been used for the analysis of the tests in the past.

The issue of how to vertically equate the tests in a multi-level battery is a serious one, and not only because of the use of latent trait models. The major consideration is whether the tests have the same content and emphasis for the content at all of the levels. If not, the meaning of the ability estimates obtained from the tests changes over levels, and none of the equating procedures available is appropriate.

Under the circumstances, the best procedure would be to sort the battery into unidimensional subtests that measure the same dimension at all difficulty levels. These subtests could then reasonably be equated over levels unless all levels measured the various different content areas in exactly the same proportions.

Even under the above ideal circumstances, the one-parameter model does not seem to be adequate for vertical equating purposes. The Slinde & Linn (1979) and Loyd & Hoover (1980) papers show that the techniques being used have serious problems. Unfortunately, the three-parameter logistic based procedures have not consistently demonstrated to be any better. At this point the equipercentile based procedures seem to be the most acceptable.

If unidimensional subsets can be formed and equated, the score scale issue is not a serious one. Either model can be used since in the unidimensional case the models will yield ability estimates that are correlated .95+. Since the one-parameter model is cheaper to use, it should probably be selected.

Recommendations

It is unfortunate that on the basis of the data presented, no clear choice can be made. The three-parameter model based procedures seem to be more desirable because of the better item selection procedures, the possibly better vertical equating procedures, and some possible advantages in the meaning of the ability estimates. But these same procedures require larger sample sizes, are more complex, and sometimes yield unstable parameter estimates. The one-parameter model has the advantages of simplicity and smaller sample size requirements.

Perhaps the solution is to avoid taking sides in the model selection controversy and suggest that both models be used in conjunction with traditional procedures. The tests in the battery could be designed using target information functions and items could be selected using linked calibration results from the three-parameter logistic model. By selecting the highly discriminating items, the resulting tests would have dominant first factors that were common to the tests at all levels, making vertical equating reasonable.

The vertical equating is implicit in the linking of the item pool of the test, since transforming the item difficulty parameters is the same as transforming the ability scale. However, due to the uncertainty of the value of three-parameter vertical equating procedure, it could be verified using traditional equipercentile equating. Finally, the multi-level score scale could be based on the one-parameter model, because by then the items selected would have reasonably similar discrimination parameter estimates and low guessing parameter estimates. The resulting score scale would than reflect all of the content areas in the tests.

References

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

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) Educational measurement (2nd Ed.). Washington, D. C. American Council on Education, 1971.

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

Brooks, R. D. An empirical investigation of the Rasch ratio-scale model for item-difficulty indexes. (Doctoral dissertation, University of Iowa). Ann Arbor, Michigan: University Microfilms, 1964. No. 65-434.

Cowell, W. R. ICC preequating in the TOEFL testing program. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1979.

Forbes, Dean W. The use of Rasch logistic scaling procedures in the development of short multi-level arithmetic achievement tests for public school measurement. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, April 1976.

Forster, F. & Karr, C. Using the Rasch model to increase the power of item analysis. Paper presented at the meeting of the American Educational Research Association, Boston, April 1980.

Gustafsson, J. The Rasch model in vertical equating of tests: a critique of Slinde and Linn, Journal of Educational Measurement, 1979, 16(3), 153-158.

Hambleton, R. K. An empirical investigation of the Rasch test theory model. Unpublished doctoral dissertation, University of Toronto, 1969.

Koch, W. R. & Reckase, M. D. A live testing comparison study of the one- and three-parameter logistic models (Research Report 78-1). Columbia, MO: University of Missouri, Educational Psychology Department, June 1978.

Kolen, M. J. Comparison of traditional and item response theory methods for equating tests, Journal of Educational Measurement, 1981, 18(1), 1-11.

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

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

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

Loyd, B. H. & Hoover, H. D. Vertical equating using the Rasch model. Journal of Educational Measurement, 1980, 17(3), 179-194.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14(2), 139-160.

Marco, G. L., Peterson, N. & Stewart, E. A test of the adequacy of curvilinear score equating models. Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis, Minnesota: University of Minnesota, Department of Psychology, September, 1980.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70(350), 351-356.

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

Patience, W. M. A comparison of latent trait theory and equipercentile methods of vertically equating tests. Unpublished manuscript, 1981.

Reckase, M. D. Unifactor latent trait models applied to multifactor tests: results and implications. Journal of Educational Statistics, 1979, 4(3), 207-230.

Reckase, M. D. Item pool construction for use with latent trait models. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1979.

Rentz, R. R. & Bashaw, W. L. The national reference scale for reading: an application of the Rasch model. Journal of Educational Measurement, 1977, 14(2), 161-180.

Slinde, J. A. & Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating, Journal of Educational Measurement, 1978, 15(1), 23-25.

Slinde, J. A. & Linn, R. L. A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measuremer , 1979, 159-165.

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

Woodcock, R. W. Woodcock Reading Mastery Test. Circle Pines, Minn.: American Guidance Service, 1974.

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

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