DOCUMENT RESUME

ED 300 456 TM 012 506

AUTHOR Kingston, Neal M.; Stocking, Martha L.
TITLE Psychometric Issues in IRT-Based Test Construction.
PUB DATE Aug 86
NOTE 12p.; Paper presented at the Annual Meeting of the American Psychological Association (Washington, DC, August 22-26, 1986).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Goodness of Fit; *Latent Trait Theory; *Psychometrics; *Scoring; *Test Construction
IDENTIFIERS *Educational Testing Service; Rights and Formula Scoring; Unidimensional Scaling

ABSTRACT

Psychometric issues confronted when implementing a system of item response theory (IRT) tools for test development at the Educational Testing Service (ETS) are discussed. These issues include selecting and assessing the appropriateness of IRT models, choosing methods of IRT scaling for item pools, considering test scoring strategies, and applying IRT tools within the context of different test scoring strategies. Since all items currently used in ETS testing programs are scored either right or wrong, this paper deals only with models for binary-scored items and focuses on unidimensional models. Use of IRT tools with theta-hat, number-right or formula, and scaled scoring methods is addressed. The following test development steps are listed: (1) collect pretest data; (2) select an IRT model; (3) assess the IRT model's appropriateness; (4) place all item parameters on a single scale; (5) choose IRT tools appropriate for the test scoring method used; (6) develop one or more forms of the test; and (7) assess the test development effort's success. (TJH)

PSYCHOMETRIC ISSUES IN IRT-BASED TEST CONSTRUCTION[1]

Neal M. Kingston[2]
Martha L. Stocking

Educational Testing Service

[1] A paper presented as part of the symposium, "Item Response Theory Based Test Development," at the Annual meeting of the American Psychological Association, August 24, 1986.

[2] The authors are listed alphabetically.

INTRODUCTION

Item response theory (IRT) provides the test developer (for programs that pretest their items) with powerful tools that in theory allow the construction of better tests while simultaneously increasing the efficiency of the test development process. With IRT, multiple forms of a test can be constructed to be more parallel than could be done with conventional item statistics. Test reliability, conditional standard errors of measurement, and test equating functions could all be estimated before a test was ever administered as a whole.

The success of the application of such tools depends upon two important factors: the extent to which IRT is an appropriate approach to the analysis of the test and items in question, and the extent to which the theoretical usefulness of the available tools can be realized in the practical environment of the test developer. In exploring these factors, we have structured this paper based on issues confronted when implementing a system of IRT tools for test development at ETS. These issues include:

- selecting and assessing the appropriateness of IRT models,
- choosing methods of IRT scaling for item pools,
- a consideration of test scoring strategies, and
- a consideration of IRT tools and their applicability in the context of different test scoring strategies.

ASSESSMENT OF MODEL FIT

Before one uses any statistical model for any purpose, it is obviously necessary to insure the model chosen is appropriate for the data. Statistical models, such as item response theory, are based on assumptions. The assumptions of item response theory model are embodied in formulae that express the relationship between the probability of selecting a particular response to an item with the statistical characteristics of the item and the abilities of a person.

The most common type of test item is typically scored either right or wrong. IRT models appropriate for such binary-scored items differ in the number of parameters used to describe item characteristics; typically one or more of three parameters -- item difficulty (b), item discrimination (a), and lower asymptote (c) -- is used. Models may also differ in the number of ability dimensions; typically one dimension is used, but some researchers have been exploring the use of multidimensional IRT models (for example, Reckase, 1985). Other IRT models exist for more complex item-scoring strategies, such as the graded response model (Samejima, 1969) and continuous response model (Samejima, 1972) for responses that can be ordered on the basis of correctness, the nominal response model (Bock, 1972) that utilizes information contained in incorrect responses, and partial credit model (Masters, 1982). The first step in using item response theory to develop a

test is to choose one or more IRT models for consideration, based on your knowledge of the test items and the assumptions of each model. The next step is to test the assumptions of the chosen model for each testing program for which you are considering its use.

This paper deals only with models for binary-scored items since all items currently used in ETS testing programs are scored either right or wrong. However, some of the methods described here would be appropriate, with some modification, for other types of item scoring strategies and corresponding models. The models discussed here will be unidimensional ones; there are many unresolved issues regarding multidimensional models.

There are two assumptions of the most commonly used item response models: unidimensionality, and particular shape of the item response function. There are a number of ways that unidimensionality has been assessed in the item response theory literature. Hambleton and Swaminathan (1985, Chapter 8) and Hambleton and Rogers (1986) have written comprehensive reviews of the IRT model-fit literature, and we will not try to replicate their efforts here

There is one point with regard to dimensionality assessment that we would like to emphasize, however. The apparently straightforward approach of factor analyzing binary test data is far more complicated than it appears. Phi coefficients are not appropriate for the factor analysis of binary data. Tetrachoric correlations are difficult to estimate accurately and, when based on responses to multiple-choice items, should be corrected for guessing. Also, in item response theory, the relationship between item responses and the underlying trait is not linear, while most factor analytic methods require a linear relationship. What is needed is a factor analytic method based on item response theory. Such a method is now available.

Full-information factor analysis (Bock, Gibbons, & Muraki, 1985), as implemented in the program TESTFACT (Wilson, Wood, & Gibbons, 1984) uses the marginal maximum likelihood method (Bock & Aitken, 1981) to estimate (reparameterized) discrimination and difficulty parameters for multidimensional IRT models. The lower asymptote for each item is treated as a known value that is input by the program user. TESTFACT allows a stepwise factor analysis to be performed. First a one-factor solution is obtained, then a two-factor solution. The difference between likelihood ratio chi-squares for the two solutions is used to test the statistical significance of the added factor. A third, fourth, or even more factors can be added, but computer time and expense increase exponentially with the number of factors.

The only truly unidimensional set of mental test data is one that has been artificially generated using a unidimensional model. Given a sufficiently large sample, any real set of data can be shown to be multidimensional. For practical applications the important question is not "Are these data unidimensional?" but instead, "How strong is the first factor

compared to the other factors?" Hand-in-hand with the answer to the latter question is the question, "How robust are IRT methods to violations of unidimensionality?" Several researchers have looked at this question in the context of stability of parameter estimates and accuracy of equating (for example, Dorans & Kingston, 1985; Drasgow & Parsons, 1983; Kingston, Leary, & Wightman, 1985; Reckase, 1979). In general, these studies have shown that parameter estimates and equatings are very stable in the face of minor departures from unidimensionality. There has been no direct research, however, on the robustness of IRT test development methods in the face of multidimensionality.

Should a test be too multidimensional for the use of a unidimensional IRT model for the test as a whole, that test could be broken down into relatively more unidimensional subtests that could be separately developed (and equated) using IRT methods. By developing subtests that are each highly parallel to the appropriate subtests in another form of the test, the resulting total tests will also be highly parallel.

In addition to satisfying assumptions about dimensionality, it is also necessary to choose, and confirm, the selection of the number of item parameters. Models incorporating different numbers of parameters impose various restrictions on allowable shapes for item response functions. For example, if a one-parameter model is chosen, do all items have the same slope? Is the probability of a correct response a monotonically increasing function of ability? There are two approaches to assessing the fit of the form of the model to the data: graphical approaches such as the analysis of item-ability regressions (see, for example, Kingston & Dorans, 1985) and statistical approaches such as Yen's Q_1 (Yen, 1981, 1984). Graphical approaches have the advantage of allowing one to see readily where and how the model does not fit the data, but they do not provide information regarding whether discrepancies are "real" or due to sampling variability. Existing statistical approaches have not solved this problem: that is, there is no research regarding any statistical approach appropriate for all IRT models in which the probability distribution of the sample statistic is known. Thus, for anything other than simple models, such approaches can only be used as informal guides. Work with likelihood ratio chi-square tests (based on marginal maximum likelihood estimation with no priors on item statistics) may overcome this difficulty.

CHOICE OF IRT SCALING METHOD

True IRT parameters are invariant across samples of items or persons as long as a scale for expressing them has been established. Estimates of these parameters will not have this property for two reasons: 1) the scales may be different (we can fix this) and 2) the errors of estimation may be different (we can not fix this). The difference in scales arises because IRT parameters and their estimates have no natural origin or unit of measurement, as does temperature or length for example. Typical computer programs select

a convenient origin and unit; LOGIST (Wingersky, 1983; Wingersky, Burton, & Lord, 1982) and BILOG (Mislevy & Bock, 1982), for example, essentially choose these values so that the mean and standard deviation of estimated abilities are zero and one respectively. Thus, even in the absence of errors of estimation, all item parameter estimates will be different if estimated once based on a very high ability group and a second time based on a low ability group. If the model holds, however, the relationship between the two arbitrary scales will be linear, and once determined, can be used to place estimates from different groups on the same scale. Note that this can be contrasted with classical item statistics, where the ranking of items by difficulty can be different in different groups, and thus no monotonic transformation can correctly put all item statistics on a single scale.

If a testing program pretests its items on non-equivalent groups, some method must be used to place the IRT parameter estimates on a single scale. A number of methods exist to do this. Although it is possible to use a "common person" design, more typical methods use a "common items" design. By this we mean that there is a set of items administered to two different groups of people who, in turn, may have responded to some non-common items. This communality of items is what allows the construction of a single IRT scale for expressing the results.

One relatively simple method of placing parameter estimates on a single scale is often referred to as concurrent calibration. In concurrent calibration, all items are parameterized in a single calibration run of a program such as LOGIST. All items that were not administered to an examinee are coded as "not reached" and are not used as part of the estimation process for that examinee. Concurrent calibration is a powerful method of maintaining one's IRT scale, but if not used in conjunction with other methods it can be very expensive for a test with an ongoing pretesting program. This is so, because one would have to perform progressively larger and larger LOGIST runs to calibrate one's items.

Several other methods of IRT parameter scaling using results of individual calibrations have been developed. We will not describe these methods here, but references are given for those who are interested. These methods include: 1) fixed b's (see, for example, Hicks, 1983), 2) robust mean and sigma method (Stocking & Lord, 1983), 3) characteristic curve transformation (Stocking & Lord, 1983), and 4) minimum chi-square (Divgi, 1985).

TEST SCORES

The appropriateness of the various IRT tools in test construction is a function of the scaling metric in which the test developer feels most comfortable working. Different IRT test development tools are more or less useful depending upon the metric chosen.

If examinee responses to items in a conventional test are scored either right, wrong, or omitted, there are a number of important aggregates of these item responses that can be called "total test" scores. Perhaps the most familiar of these "observed scores" is the sum of the number of right answers: the number-right score. A variant of this is a scoring mechanism that imposes a "penalty for guessing," generally called a formula score. A less familiar member of the same family is the optimally-weighted sum of item scores: the maximum likelihood estimate of ability, theta-hat. All of these observed scores may be transformed to a different scale by either linear or nonlinear methods for the purposes of score reporting. An observed score obtained in this fashion is referred to as an observed scaled score.

Each of these scoring procedures deals with observed responses provided by examinees to items in a test. Corresponding to each of these observed scores are unobservable and unmeasurable quantities that we wish we could obtain, but cannot. These are (in the same order as above) number-right true score, true formula score, true ability, and true scaled score. Since we cannot obtain these unobservable quantities, the best we can do is to estimate them. While the distinction between observable and unobservable quantities may seem of interest only to theoreticians, such distinctions become important in the selection of appropriate IRT tools for test developers.

IRT TOOLS

There are three important concepts from IRT that are applicable in the context of test development: Information functions, conditional standard error functions, and relative efficiency functions. None of these has a counterpart in conventional item and test analysis.

Information functions are functions of test scores. The word "information" has an intuitive meaning in this context: it is essentially how well we can do when using an observable quantity (for example the observed number-right score) to make inferences about an unobservable quantity (number-right true score). It is obvious that a collection of items that is administered as a test cannot measure with equal precision at all score levels. Information is therefore a function of test score.

If the observable test score is an unbiased estimator of the unobservable quantity it attempts to measure, then conditional standard error functions are computed as the inverse of the square root of information functions. The use of this function puts questions of precision into the metric of the test score.

Tests may be compared with each other in terms of relative efficiency functions. Relative efficiency functions are the simple ratio of two information functions at corresponding values of the test scores. Information functions can change shape drastically depending upon the scoring

metric being used. The conclusions drawn from information functions are only valid for the metric on which they are computed. Relative efficiency functions, however, do not change when the scoring metric is changed. For this reason, relative efficiency functions are very powerful for making comparisons between tests.

IRT TOOLS WHEN THE OBSERVED SCORE IS THETA-HAT

Theta-hat is an optimally weighted sum of item scores. While the computation of optimal weights is complex, the details of the procedure need not concern us here. Although quite useful in theoretical work in IRT, this method of test scoring has not yet achieved wide popularity. The only published conventional test scored using this method is the Comprehensive Test of Basic Skills published by CTB/McGraw Hill (Yen, 1982). Theta-hat is a natural scoring metric for adaptive tests, and as these types of tests become more common, its use is likely to increase.

In this scoring metric, the IRT item parameters (a, b, c) have an interpretation for test developers. More important, the item information function, that is, the amount of information available from a single item response for making inferences about true ability, can be used to indicate where and how well an item is functioning. In addition, the ability level at which the item yields maximum information is easily determined.

But the most important aspect of this metric for test development purposes is the information function for the total test. This information function indicates the amount of information available for estimating true ability, theta, from estimated ability, theta-hat. This function has a unique property -- it is the sum of the independent and additive contributions from each item information function. This means that the measurement properties of the test in this metric can be analyzed in terms of the measurement properties of each item, independent of all other items in the test. In addition, it represents the maximum amount of information available for making inferences about true ability from any method of scoring a test (Lord, 1980, chapter 5.6).

Conditional standard error functions and relative efficiency functions are also tools that are useful in this metric. Relative efficiency functions computed in this metric lead, of course, to the same conclusions in this metric as in any other metric.

IRT TOOLS WHEN THE OBSERVED SCORE IS NUMBER-RIGHT OR FORMULA SCORE

Number-right and formula scores are (non-optimally) weighted sums of item scores. In this metric, the natural interpretation of the IRT item parameters disappears. In addition, item information functions become less

intuitively appealing, since information functions for a total test are no longer simple sums of item information functions.

There are two informations of interest in this context. The first is the information available for estimating true ability from the observed score. This is most useful in comparisons with information functions when the observed score is theta-hat (see, for example, Lord, 1980, page 74). We see from these comparisons that using number-right or formula scores diminishes the precision of estimate somewhat for high ability levels, but has a much greater effect at low ability levels. In general, observed number-right or formula score does not fully utilize all of the data for making inferences about true ability.

Most conventional tests rarely seek to make inferences about true ability, but rather about number-right true score or true formula score. This second information function is the one that is likely to be most useful for test development purposes, along with its corresponding conditional standard error of measurement function.

IRT TOOLS WHEN THE OBSERVED SCORE IS A SCALED SCORE

This metric is perhaps the most common metric for published conventional tests. In this metric, observed scores on a test are transformed to another metric before reporting, by equating the test to some previously scaled test. The IRT tools of item parameters and item information functions are not helpful here.

What is extremely useful, however, is the information function for making inferences about true scaled score from observed scaled score. To compute this function requires an intact, previously equated test form as part of the item pool. As a by-product of the computations to obtain this information, it is also possible to generate an equating table for the candidate test form. This table can be used as a rough guide to what the operational equating might be, or as the final equating for the candidate form, in which case the new test is considered to be "pre-equated."

As with all other metrics considered, conditional standard error functions and relative efficiency functions are also useful in this metric.

CONCLUSION

In order to use IRT for test development there are a number of steps one must take:

1. Collect pretest data,
2. Select an IRT model,
3. Assess the appropriateness of the model,

4. Place all item parameters on a single scale,
5. Choose IRT tools appropriate for your test scoring method,
6. Develop one or more forms of your test, and
7. Assess the success of your test development effort.

Exactly how we accomplish each of these tasks is likely to change as we gain more experience.

REFERENCES

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Bock, R. D. and Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R. D., Gibbons, R. and Muraki, E. (1985). Full information factor analysis. MRC Report 85-1. Chicago: National Opinion Research Center.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. Applied Psychological Measurement, 9, 413-415.

Dorans, N. J. and Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. Journal of Educational Measurement, 22, 249-262.

Drasgow, F. and Parsons, C. K. (1983). Application of unidimensional item response models to multidimensional data. Applied Psychological Measurement, 7, 189-200.

Hambleton, R. K. and Rogers, J. K. (1986). Assessing item response model fit. Paper presented as part of the symposium, "Current issues and developments in item response theory," at the annual meeting of the American Educational Research Association.

Hambleton, R. K. and Swaminathan, H. (1985). Item response theory: Principals and Applications. Boston: Kluwer-Nijhoff.

Hicks, M. M. (1983). True score equating by fixed *b's* scaling: a flexible and stable equating alternative. Applied Psychological Measurement, 7, 255-266.

Kingston, N. M. and Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. Applied Psychological Measurement, 9, 281-288.

Kingston, N. M., Leary, L. F., and Wightman, L. E. (1985). An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test. RR 85-34. Princeton, NJ: Educational Testing Service.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Mislevy R. and Bock R. D. (1982). BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary models. Chicago: International Educational Services.

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

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.

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

Samejima, F. (1972). A general model for free-response data. Psychometric Monograph, No. 19.

Stocking, M. L. and Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement 7, 201-210.

Wilson, D. T., Wood, R. and Gibbons, R. T. (1984). TESTFACT users guide. Moorseville, IN: Scientific Software Inc.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 45-56). Vancouver, BC: Educational Research Institute of British Columbia.

Wingersky, M. S., Barton, M. A., and Lord, F. M. (1982). LOGIST users guide. Princeton, NJ: Educational Testing Service.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.

Yen, W. M. (1982). 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). Vancouver, BC: Educational Research Institute of British Columbia.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.