ITEM RESPONSE THEORY:
INTRODUCTION AND BIBLIOGRAPHY

Ronald K. HAMBLETON
University of Massachusetts at Amherst, USA.

INTRODUCTION

In a few words, item response theory (IRT) 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 these traits can be described by a monotonically increasing function called and item characteristic function. This function specifies that examinees with higher scores on the traits have higher expected probabilities for answering an item correctly than examinees with lower scores on the traits. In applying item response theory to measurement problems, a common assumption is made that there is one dominant factor or ability which can account for item performance. This so-called «ability» which the test measures could be a broadly or narrowly defined aptitude, achievement, or personality variable.

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 defined for the trait measured by the test. Modifications are made in the interpretations of ICCs when, for example, the underlying trait is an attitudinal variable and the «item response» is a rating from (say) a Likert scale. In addition to the assumption of test unidimensionality, it is common to assume that the item characteristic curves are described by one, two, or three parameters. The specification of the mathematical form of the ICCs and the corresponding number or parameters needed to describe the curves determines the particular item response model. Generating and/or selecting mathematical forms for ICCs are two of the currently important lines of research in the IRT field.

In any successful application of item response theory, item parameter estimates are obtained to describe the test items, and ability estimates are obtained to describe the performance of examinees. Any successful application requires that there be evidence that the chosen item response
model, at least to an adequate degree, fits the test dataset.

Item response theory (IRT) (or latent trait theory, or item characteristic curve theory, as it is sometimes called) has become over the last 20 years a very popular topic in the measurement field. There have been (1) numerous IRT research studies published in the measurement journals, (2) a very large number of conference presentations, and (3) many successful applications of the theory to pressing measurement problems (i.e., test score equating, study of item bias, test development, item banking, and adaptive testing).

Interest in item response theory stems from two desirable features which are obtained when an item response model fits a test dataset: Descriptors of test items (the item statistics) are not dependent upon the particular sample of examinees chosen 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. Neither desirable feature is obtained when the well-known and popular classical test models are used.

There are many well-documented shortcomings of classical testing methods and measurement procedures. The first shortcoming is that the values of such classical item statistics as item difficulty and item discrimination depend on the particular examinee samples in which they are obtained. The average level of ability and the variability of ability scores in a examinee group influence the values of the item statistics, and reliability and validity statistics too, often substantially. One undesirable consequence of sample dependent item statistics is that these 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.

A second shortcoming of classical testing methods and procedures is that comparisons of examinees on an ability scale measured by a set of test items comprising a test are limited to situations where examinees are administered the same (or parallel) test items. Unfortunately, many achievement and aptitude tests are (typically) suitable for middle-ability students only and so these 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, in theory, when the test difficulty is matched to the approximately ability levels of examinees. But, when several forms of a test which vary substantially in difficulty are used, the task of comparing examinees becomes more complex because test scores only cannot be used.

A third shortcoming of classical testing methods and procedures is that they provide 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. Also, when an adaptive test is being administered at a computer terminal, optimal item selection depends on being able to predict how the examinee will perform on various test items.

Item response theory purports to overcome the shortcomings of classical test 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. Item parameter invariance is accomplished by defining the item characteristic curves (from which the item parameters are obtained) in a way that the underlying ability distribution is not a factor in item parameter values or interpretations. Finally, by deriving standard errors associated with individual ability estimates, rather than producing a single estimate of error and applying it to all examinees, another of the criticisms of the classical test model can be overcome.

In summary, item response theory models 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 dataset. Through the parameter estimation process, test items and examinees 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 performance of 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.

Today, item response theory is being used in the United States by most of the large test publishers, credentialing organizations, state departments of education, large school districts, the Armed Services, and industry to (1) construct both norm-referenced and criterion-referenced tests, (2) investigate item bias, (3) equate tests, and (4) report ability scores and diagnostic information. In fact, the various applications have been sufficiently successful that researchers in the IRT field have shifted their attention from a consideration of IRT model advantages and disadvantages in relation to classical test models to consideration of such IRT technical problems as goodness-of-fit investigations, model selection, parameter estimation, and steps for carrying out particular applications. Certainly some issues and technical problems remain to be solved in the IRT field but it would seem that item response model technology is more than adequate at this time to serve a variety of uses.

What follows is an IRT bibliography consisting mainly of important references (up to June of 1989) which have been published in the United States. No attempt was made to catalog the many important IRT articles which have appeared in European journals, or other non-American journals. The bibliography is organized into 13 categories: General Articles/Texts, Models, Parameter Estimation, Model-Fit, Scales, Robustness Studies, Test Development Studies, Adaptive Testing Studies, Item Banking Studies, Equating Studies, Item Bias Studies, Miscellaneous Applications, and Computer Programs.
General Articles/Texts


Models


Parameter Estimation

(1987). Methodology review: Item parameter estimation under the one-, two-,


Model-Fit


--- (1982). Two new test statistics for the Rasch...
model. *Psychometrika*, 47, 123-140.


### Scales


### Robustness Studies


### Test Development Studies


van der Linden, W. J. (1981). A latent trait look at pretest-posttest validation of crite-
Adaptive Testing Studies


Item Banking Studies


Equating Studies


**Item Bias Studies**


**Miscellaneous Applications**


Drasgow, F., Levine, M. V. & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appro-


**Computer Programs**


— (1976). CALFIT: Sample-free item calibration with a Rasch measurement model (Research Memorandum No. 18). Chicago: Statistical Laboratory, Department of Education, University of Chicago.