Advances in Item Response Theory and Applications: An Introduction

Ronald K. Hambleton
University of Massachusetts, Amherst

Wim J. van der Linden
Twente University of Technology

Test theories can be divided roughly into two categories. The first is classical test theory, which dates back to Spearman's conception of the observed test score as a composite of true and error components, and which was introduced to psychologists at the beginning of this century. Important milestones in its long and venerable tradition are Gulliksen's *Theory of Mental Tests* (1950) and Lord and Novick's *Statistical Theories of Mental Test Scores* (1968).

The second is item response theory, or latent trait theory, as it has been called until recently. At the present time, item response theory (IRT) is having a major impact on the field of testing. Models derived from IRT are being used to develop tests, to equate scores from nonparallel tests, to investigate item bias, and to report scores, as well as to address many other pressing measurement problems (see, e.g., Hambleton, 1983; Lord, 1980). IRT differs from classical test theory in that it assumes a different relation of the test score to the variable measured by the test. Although there are parallels between models from IRT and psychophysical models formulated around the turn of the century, only in the last 10 years has IRT had any impact on psychometricians and test users. Work by Rasch (1980/1960), Fischer (1974), Birnbaum (1968), Wright and Panchapakesan (1969), Bock (1972), and Lord (1974) has been especially influential in this turnabout; and Lazarsfeld's pioneering work on latent structure analysis in sociology (Lazarsfeld, 1950; Lazarsfeld & Henry, 1968) has also provided impetus.

One objective of this introduction is to review the conceptual differences between classical test theory and IRT. A second objective is to introduce the goals of this special issue on item response theory and the seven papers. Some basic problems with classical test theory are reviewed in the next section. Then, IRT approaches to educational and psychological measurement are presented and compared to classical test theory. The final two sections present the goals for this special issue and an outline of the seven invited papers.

Some Problems with Classical Test Theory

In this section the classical test model for a population of persons and a collection of tests, as presented by Lord and Novick (1968, chap. 2), is considered. The classical test model is a weak model.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 6, No. 4, Fall 1982, pp. 373–378

0416-6216/82/040373-06$1.30

It is a model for the propensity distribution of an observed test score; that is, it conceives this score for a fixed person and test as a random variable across replications. The model consists of definitions for true score and error score, and one primary assumption—that the correlations between errors on distinct measurements are zero. Definitions cannot be true or false. The only part of the model that entails the necessity of an empirical test is the assumption of uncorrelated errors. Studies have been conducted to determine the consequences of violations of this assumption. For some cases it is possible to test the presence of uncorrelated errors with the aid of the theory of linear structural models with unmeasured variables (see, e.g., Jöreskog, 1971).

In spite of the fact that the classical test model is a weak model, it is not without problems. These problems do not pertain to the model as such, which is largely a tautology, but to practical difficulties and interpretation problems arising when the model is applied to behavioral measurements.

The first problem with the classical test model is the test-dependent score. The starting point for the model is the random variable X_{ga} representing the observed score of person a on test g. The true score $\tau_{ga} \equiv EX_{ga}$ is defined as the *person* parameter which should be the focus of the measurement process. However, as is demonstrated by its double indices, τ_{ga} can be considered as a *test* parameter as well. That τ_{ga} is person dependent is obvious; tests are constructed to estimate these differences. However, τ_{ga} is also test dependent and that feature (or limitation) of the score engenders problems. Every test entails its own true score, even when it is known that a group of tests measure the same ability or achievement variable. True scores are thus as typical of tests as they are of persons. What is desired, however, is a test-free person score, a score that does not depend on the specific items chosen for the test, that does not reflect the difficulty of the items but only the position of the person on the variable of interest. The absence of test-free scores has not only led to serious problems that could not be solved in classical test theory but has also hindered behavioral measurement from achieving its potential. For example, because of the test-dependent characteristic of scores, it is necessary to administer the same set of test items to examinees to enable proper comparisons to be made. However, it is well known that measurement precision is enhanced when test items are selected to match the specific ability (or other trait) levels of examinees.

The second problem with the classical test model is associated with the first. Not only are examinee scores from classical test theory test dependent, but item and test parameters derived from classical test theory are also sample dependent. That the classical reliability coefficient changes as a function of the true score variance in the selected sample of examinees, even when the size of the measurement errors remains constant, is one indication of this dependency. Comparable dependencies hold for such classical parameters as item p values, item-test correlations, validity coefficients, and so forth. Sample-dependent item and test statistics have created many practical problems that cannot be solved within the framework of classical test theory.

The third problem involves the status of the observed test score and the error of measurement. It is certainly correct to consider these as stochastic across replications. Behavioral measurements seem to be processes with outcomes susceptible to all kinds of random influences, and it seems unlikely that replications of measurement should yield exactly the same outcomes. Although the random character of observed and error scores can thus be defended theoretically, practical problems are involved, since replications of measurements cannot be realized experimentally. Individuals are never exactly the same at a second administration of the test—they forget, learn new skills, are motivated or frustrated by the first administration, or their situation changes. The construction of parallel tests is a requirement that can never exactly be met. Nevertheless, classical test theory leans heavily on the first two moments of the propensity distribution, and its most important test and item parameters are defined using these moments. As a consequence, for example, the reliability coefficient cannot be estimated satisfactorily. The psychometrician must be content with either (1) lower bound estimates or (2) estimates with unknown biases due to the testing design used. Comparable difficulties hold for estimating other item and test parameters.

It can be concluded that classical test theory, while syntactically correct, yields several serious problems. It provides no test-free scores, has sample-dependent test and item parameters, and leans heavily on the availability of parallel measurements which are difficult to obtain in practice.

Item Response Theory

Item response models differ from the classical test model in many respects. The first conspicuous difference is that the modeling starts prior to the scoring of the test. Unlike classical test theory, IRT does not provide models *of* test scores but consists of models *providing* test scores. In IRT the quantity of interest is the person parameter or latent variable as specified in the model. The statistical task is to estimate the parameters of the model and to assess the fit of the model to the item responses.

The second difference is associated with the first. In IRT modeling is not aimed at the level of the test as in classical test theory; instead, it is aimed at the item level. This is one of the principal advantages of IRT. It is recognized that the data gathered in educational and psychological testing are always qualitative responses to test items. Quantitative information is obtained from qualitative data via the use of measurement models formulated to this end. Specifically, IRT achieves this by modeling in which *quantitative* item and ability parameters are used to explain qualitative item responses. The ability parameter represents the position of the person on the variable to be measured; the item parameters represent the properties of the item affecting the examinee responses. IRT does not take for granted a (pseudo-) quantitative test score but considers the information included in the item responses as merely qualitative and uses quantitative parameters to explain them. The scale for the parameters is given by the structure of the chosen model.

The third difference is that item response models are stochastic. Item responses are considered to be outcomes of stochastic processes which can be characterized by certain parameters but in which, nevertheless, all kinds of random disturbances play their roles. As a consequence, item response models are not models for the explanation of item responses but of the *probabilities* of these responses. It is these probabilities that are written as a function of a person parameter and one or more item parameters. To be sure, classical test theory considers its variables as stochastic as well, but it is more correct to qualify its models as error-component models; they do not go any further than the inclusion of a random error component with an unspecified distribution. Item response models do go further, specifying a distributional form with a parameter structure for the item responses. It is for this reason that item response models are termed stochastic.

The final difference to be mentioned here is that IRT replaces measurement by statistical estimation. Measuring the ability of an examinee takes the form of using his/her responses to estimate his/her parameter from the model. By doing so, IRT replaces the mostly gratuitous scoring rules that prevail in the practice of testing by methods of estimation anchored well in statistical theory.

Earlier in this introduction, three problems inherent in the classical test theory model were mentioned: (1) examinee test scores are test dependent, (2) item and test parameters are sample dependent, and (3) the availability of parallel measurements is required. IRT provides workable solutions to these three problems.

In IRT the role played by the true score τ_{ga} in classical test theory is replaced by the latent parameter θ_a. The latter is not indexed to the test. Although the probability of success is a function of the examinee and the item, the parametric structure in the model used to explain this probability has different and exclusive parameters from the ability of the person to be measured and from the properties of the items that influence this probability. If the items vary in difficulty, parameters can be added to the model to account for that. The same can be done if differences in discriminating power between the items are a non-negligible factor or if guessing is possible. By explicitly accounting for these differential properties

of items, their effects are "removed" from the success probabilities and only the effect of an item-free person parameter remains. It is in this basic fashion that any well-formulated explanatory theory functions. Classical test theory has neglected these possibilities; as a consequence, for example, differences in item difficulty or in the possibility of guessing create different examinee true scores. The fact that IRT offers descriptors of examinees that do not depend upon the particular selection of test items has opened up the possibility for new testing technology (e.g., adaptive testing).

Likewise, the item parameters in item response models are sample independent. The presence of a person parameter in the model is used to remove the effect of the ability of the person on his/her probability of success (assuming that the dimensionality of the parameter is complete, of course). This has opened up the possibility of creating banks of "calibrated" items from which can be selected items that are optimal for measuring certain levels of ability or for use in certain testing strategies.

Finally, although IRT considers item responses (and thus test scores) as stochastic, replications of measurements are not required to analyze their accuracy. This is because person parameters are not defined as expected values of such replications but, independently of these, as model parameters. Ability parameters can be estimated from one test administration using statistical estimation procedures with known properties. In IRT the difficult problems inherent in obtaining replications of behavioral measurements have been solved by replacing classical reliability analyses by statistical analyses of the qualities of model parameter estimates.

In summary, then, there is now substantial theoretical as well as empirical evidence to indicate that IRT is useful in overcoming several of the shortcomings of classical test theory models. Perhaps it is not surprising, therefore, to observe the large number of recent IRT applications to testing problems.

Purposes of This Issue

In view of the substantial interest internationally in IRT, this issue presents a number of technical contributions on IRT to highlight the present work being done in several countries. Specifically, the purposes of this issue are (1) to draw attention to a number of important technical advances from several of the leading contributors to the IRT literature and (2) to highlight some of the important IRT research being conducted outside the United States. In addressing these two purposes, the goal was to contribute to the growth of the IRT field by introducing many new IRT models and applications, and by suggesting several promising directions for additional research and development.

Introduction to the Papers

The seven papers in this issue can be organized around three broad topics: New models (Roderick McDonald; Gerhard Fischer and Anton Formann; Robert Mokken and Charles Lewis), parameter estimation (Darrell Bock and Robert Mislevy; Erling Andersen), and applications (Frederic Lord; David Weiss). Roderick McDonald provides a general framework for organizing many of the presently available item response models, including those that are multidimensional. In addition, his framework provides a basis for generating many more models for comparing and contrasting present models in the areas of parameter estimation, and for testing the fit of models and data. Gerhard Fischer and Anton Formann describe a special class of linearly restricted logistic models and their applications. These models have been studied in Europe for at least 10 years but remain relatively unknown in the United States. The models, which are an extension of the Rasch model, are being used by psychologists in Europe to investigate the cognitive processes underlying test item performance.

Robert Mokken and Charles Lewis describe a new class of item response models in which the mathematical form of the item characteristic curve (ICC) is not specified. The only assumption about the items made with these models is that the ICCs are a montonic function of ability. Lord (1970) is the only other researcher to date to have formulated this type of model. These nonparametric models, as Mokken and Lewis call them (because they make no assumptions about the mathematical form of ICCs or the ability distribution), include the logistic test models as special cases. Mokken and Lewis consider their new models in test development and ability estimation.

Two papers describe current developments in parameter estimation. The first, by Darrell Bock and Robert Mislevy, describes their latest work with estimated a posteriori (EAP) estimators. Specifically, their work is directed toward those testing situations where efficiency of computation is essential, for example, as it might be when estimating abilities on a microcomputer. Their estimation methods require fewer calculations than several other prominent estimation methods, and it appears that their method can be applied easily to multiple category scoring models where there are substantial complexities in estimation. In the second paper, Erling Andersen draws attention to connections among IRT, the theory of exponential family distributions, contingency table analysis, and latent structure analysis. He considers the substantial implications of these associations for parameter estimation in IRT.

In the final section of this issue, two of the most frequent and important applications of IRT are considered. Frederic Lord derives a formula for the standard error associated with the highly recommended method of true-score equating by IRT. He then uses his new formula to compare several methods of equating and to consider the quality of equated scores using anchor tests of variable lengths. David Weiss provides a historical view of adaptive testing and describes several of his recent research studies. The studies he describes present a number of guidelines for successfully implementing adaptive testing when applied to both problems of measurement and classification.

References

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 MA: Addison-Wesley, 1968.

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

Fischer, G. H. *Einführung in die Theorie psychologischer Tests*. Bern: Huber, 1974.

Gulliksen, H. *Theory of mental tests*. New York: Wiley, 1950.

Hambleton, R. K. (Ed.) *Applications of item response theory*. Vancouver: Educational Research Institute of British Columbia, 1983.

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. *Psychometrika*, 1971, *36*, 109–133.

Lazarsfeld, P. F. The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, *Measurement and prediction*. Princeton: Princeton University Press, 1950.

Lazarsfeld, P. F., & Henry, N. W. *Latent structure analysis*. New York: Houghton-Mifflin, 1968.

Lord, F. M. Estimating item characteristic curves without knowledge of their mathematical form. *Psychometrika*, 1970, *35*, 42–50.

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

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.

Rasch, G. *Probabilistic models for some intelligence and attainment tests (expanded edition)*. Chicago: University of Chicago Press, 1980. (Originally published Copenhagen: Danmarks Paedagogiske Institut, 1960)

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

Editor's Note

Both authors of this paper contributed equally to the planning, preparation, and editing of this issue.

Authors' Addresses

Send requests for reprints or further information to Ronald K. Hambleton, University of Massachusetts, Laboratory of Psychometric and Evaluative Research, Hills South, Room 152, Amherst MA 01002, USA; or Wim J. van der Linden, Afdeling Toegepaste Onderwijskunde, Technische Hogeschool Twente, Postbus 217, 7500 AE Enschede, The Netherlands.