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

Psicothema, 1990, vol. 2, nº 1, pp. 97-107
ISSN: 0214 - 9915

model, at least to an adequate degree, fits the test detaset.

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 approximate 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.

ITEM RESPONSE THEORY
BIBLIOGRAPHY

General Articles/Texts

Andrich, D. (1989). Statistical reasoning in psychometric models and educational measurement. *Journal of Educational Measurement, 26*, 81-90.

Baker, F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann.

Birnbaum, A. (1968). 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.

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

— (Ed.) (1983). *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.

— (1989). Principles and selected applications of item response theory. In R. Linn (Ed.), *Educational Measurement* (3rd edition) (pp. 147-200). New York: Macmillan.

— (Ed.) (1989). Applications of item response theory. *International Journal of Educational Research, 13*, 121-231. (6 papers)

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

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Academic Publishers.

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

Hambleton, R. K. & van der Linden, W. (Eds.) (1982). Technical contributions to item response theory. *Applied Psychological Measurement, 6*, 373-492. (7 papers)

Hulin, C. L., Drasgow, F. & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Irwin.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.

— (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989-1.020.

— (1977). Practical applications of item characteristics curve theory. *Journal of Educational Measurement, 14*, 117-138.

— (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27*, 251-280.

McDonald, R. P. (1989). Future directions for item response theory. *International Journal of Educational Research, 13*, 206-231.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

— (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*, 49-57.

Thissen, D. & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*, 385-395.

Traub, R. E. & Lam, R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology, 36*, 19-48.

Traub, R. E. & Wolfe, R. G. (1981). Latent traint theories and the assessment of educational achievement. In D. C. Berliner (Ed.), *Review of Research in Education (Vol. 9)*. Washington: American Educational Research Association.

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

— (Ed.) (1983). *New horizons in testing*. New York: Academic Press.

Weiss, D. J. & Davidson, M. L. (1981). Test theory and methods. *Annual Review of Psychology, 32*, 629-658.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.

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

Models

Andrich, D. (1978). A binomal latent trait model for the study of Likertstyle attitude questionnaires. *British Journal of Mathematical and Statistical Psychology, 31*, 84-98.

— (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

— (1978). Applications of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*, 581-594.

— (1978). Relationship between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2*, 451-462.

de Gruijter, D. N. M. (1986). Small N does not always justify the Rasch model. *Applied Psychological Measurement, 10*, 187 194.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175 186.

— (Ed.) (1985). *Test design: Developments in psychology and psychometrics*. New York: Academic Press

Hambleton, R. K. (1987). The three-parameter logistic model. In D. L. McArthur (Ed.), *Alternative approaches to the asessment of achievement*. Boston, MA: Kluwer Academic Publishers.

Harris, D. (1989). Comparison of 1-, 2-, an 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8*, 35-41.

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph, No. 7). Psychometric Society.

Masters, G. N. & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49*, 269-272.

McDonald, R. P. (1980). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 33*, 205-233.

Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational Statistics, 8*, 271-288.

Perline, R., Wright, B. D. & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*, 237-255.

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

— (1972). *A general model for free-response data* (Psychometric Monograph, No. 18). Psychometric Society.

— (1973). A comment of Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika, 38*, 221-233.

— (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*, 203-219.

— (1974). Normal ogive model on the continuous response level in the multi-dimensional latent space. *Psychometrika, 39*, 111-121.

Whitely, S. E. (1977). Models, meanings and misunderstandings: Some issues in applying Rasch's theory. *Journal of Educational Measurement, 14*, 227-235.

Whitely, S. & Dawis, R. V. (1974). The nature of objectivity with the Rasch model. *Journal of Educational Measurement, 11*, 163-178.

Wright, B. D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.

— (1977). Misunderstanding of the Rasch model. *Journal of Educational Measurement, 14*, 219-226.

Parameter Estimation

Baker, F. B. (1977). Advances in item analysis. *Review of Educational Research, 47*, 151-158.

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

three-parameter logistic models. *Applied Psychological Measurement, 11*, 111-143.

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

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*, 77-90.

Gustafsson, J. E. (1980). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement, 40*, 377-385.

Harwell, M. R., Baker, F. B. & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics, 13*, 243-271.

Jensema, C. J. (1976). A simple technique for estimating latent trait mental test parameters. *Educational and Psychological Measurement, 36*, 705-715.

Lord, F. M. (1970). Estimating item characteristic curves without knowledge of their mathematical form. *Psychometrika, 35*, 43-50.

— (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrica, 39*, 247-264.

— (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement, 21*, 239-243.

— (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157-162.

Ree, M. J. (1979). Estimating item characteristic curves. *Applied Psychological Measurement, 3*, 371-385.

Samejima, F. (1977). A method of estimating item characteristic functions using the maximum likelihood estimate of ability. *Psychometrika, 42*, 163-191.

Schmidt, F. L. (1977). The Urry method of approximating the item parameters of latent trait theory. *Educational and Psychological Measurement, 37*, 613-620.

Swaminathan, H. (1983). Parameter estimation in item response models. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Swaminathan, H. & Gifford, J. A. (1984). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50*, 349-364.

— (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*, 589-601.

Urry, V. W. (1974). Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement, 34*, 253-269.

Wainer, H. & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika, 45*, 373-391.

Wingersky, M. W. & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8*, 347-364.

Wright, B. D. & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1*, 281-295.

— (1977). Conditional versus unconditional procedures for sample-free analysis. *Educational and Psychological Measurement, 37*, 573-586.

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

Model-Fit

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38, 123-140.*

— *(1980). Comparing latent distributions. Psychometrika, 45*, 121-134.

Bejar, I. I. (1988). An approach to assessing unidimensionality revisited. *Applied Psychological Measurement, 12*, 377-379.

Bock, R. D., Gibbons, R. & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Cook, L. L., Dorans, N. J. & Eignor, D. R. (1988). An assessment of the dimensionality of three SAT-Verbal test editions. *Journal of Educational Statistics, 13*, 19-43.

Divgi, D. R. (1986). Does the Rasch model really work for multiple-choice items? Not if you look closely. *Journal of Educational Measurement, 23*, 283-298.

Dorans, N. J. & 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. & Parsons, C. K. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*, 363-373.

Gustafsson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 33*, 205-233.

Hambleton, R. K. (1980). Latent ability scales, interpretations, and uses. In S. Mayo (Ed.), *New directions for testing and measurement: Interpreting test scores (No. 6)*. San Francisco: Jossey-Bass.

Hambleton, R. K. & Murray, L. N. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, BC: Educational Research Institute of Britisth Columbia.

Hambleton, R. K. & Rogers, H. J. (in press). Promising directions for assessing item response model fit to test data. *Applied Psychological Measurement*.

Hambleton, R. K. & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 287-302.

Hambleton, R. K. & Traub, R. E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology, 26*, 195-211.

Hattie, J. A. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49-78.

— (1985). Methodological review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Henning, G. (1989). Does the Rasch model really work for multiple-choice items? Take another look: A response to Divgi. *Journal of Educational Measurement, 26*, 91-97.

Kingston, N. M. & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147-154.

— (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement, 9*, 281-288.

Ludlow, L. H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement, 10*, 217-229.

McDonald, R. P. (May, 1980). Fitting latent trait models. In D. Spearitt (Ed.), *The Improvement of Measurement in Education and Psychology*. Melbourne, Australia: Australian Council of Educational Research.

Rentz, R. R. & Rentz, C. C. (1978). *Does the Rasch model really work? A discussion for practitioners* (Technical Memorandum No. 67). Princeton, NJ: ERIC Clearinghouse on Tests, Measurement and Evaluation, Educational Testing Service.

Rogers, H. J. & Hattie, J. A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11*, 47-57.

Ross, J. (1966). An empirical study of a logistic mental test model. *Psychometrika, 31*, 325-340.

Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological Measurement, 48*, 657-667.

Traub, R. E. (1983). 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.

van den Wollenberg, A. L. (1982). A simple and effective method to test the dimensionality axiom of the Rasch model. *Applied Psychological Measurement, 6*, 83-91.

— (1982). Two new test statistics for the Rasch

model. *Psychometrika, 47*, 123-140.
Wood, R. (1978). Fitting the Rasch model – A heady tale. *British Journal of Mathematical and Statistical Psychology, 31*, 27-32.
Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Scales

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*, 299-325.
Lord, F. M. (1975). The «ability» scale in item characteristic curve theory. *Psychometrika, 44*, 205-217.
Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Robustness Studies

Ansley, T. N. & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from twodimensional data. *Applied Psychological Measurement, 9*, 37-48.
Dinero, T. E. & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement, 1*, 581-592.
Forsyth, R., Saisangjan, U. & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement, 5*, 175-186.
Hambleton, R. K. & Cook, L. L. (1983). The robustness of latent trait models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), *New horizons in testing*. New York: Academic Press.
Linn, R. L. & Harnisch, D. L. (1980). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 17*, 179-194.
Lord, F. M. (1983). Small *N* justifies Rasch methods. In D. J. Weiss (Ed.), *New horizons in testing*. New York: Academic Press.
Slinde, J. A. & Linn, R. L. (1979). The Rasch model, objective measurement, equating and robustness. *Applied Psychological Measurement, 3*, 437-452.
Tinsley, H. E. A. & Dawis, R. V. (1974). An investigation of the Rasch simple logistic model: Sample free item and test calibration. *Educational and Psychological Measurement, 11*, 163-178.
— (1977). Test-free person measurement with the Rasch simple logistic model. *Applied Psychological Measurement, 1*, 483-487.
van de Vijver, F. J. R. (1986). The robustness of Rasch estimates. *Applied Psychological Measurement, 10*, 45-57.
Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17*, 297-311.

Test Development Studies

de Gruijter, D. N. M. & Hambleton, R. K. (1983). Using item response models in criterion-referenced test item selection. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.
Hambleton, R. K. (1983). Applications of item response models to criterion-referenced assessments. *Applied Psychological Measurement, 6*, 33-44.
Hambleton, R. K. & de Gruijter, D. N. M. (1983). Application of item response models to criterion-referenced test item selection. *Journal of Educational Measurement, 20*, 355-367.
Hambleton, R. K. & Rogers, H. J. (1989). Solving criterion-referenced measurement problems with item response models. *International Journal of Educational Research, 13*, 146-161.
Shea, J. A., Norcini, J. J. & Webster, G. (1988). An application of item response theory to certifying examinations in internal medicine. *Evaluation and the Health Professions, 11*, 283-305.
van der Linden, W. J. (1981). A latent trait look at pretest-posttest validation of crite-

rion-referenced test items. *Review of Educational Research, 51*, 379-402.

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

Adaptive Testing Studies

Fischer, G. H. & Pendl, P. (1980). Individualized testing on the basis of the dichotomous Rasch model. In L. J. Th. van der Kamp, W. F. Langerak & D. N. M. de Gruijter (Eds.), *Psychometrics for Educational Debates*. New York: Wiley.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L. & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Hotzman (Ed.), *Computer-assisted instruction, testing and guidance*. New York: Harper and Row.

— (1971). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*, 147-151.

— (1971). A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement, 31*, 805-813.

— (1974). Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce & P. Suppes (Eds.), *Contemporary developments in mathematical psychology, Vol. II*. San Francisco: Freeman.

— (1980). Some how and which for practical tailored testing. In L. J. Th. van der Kamp, W. F. Langerak & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates*. New York: Wiley.

Ree, M. J. (1981). The effects of item calibration sample size and item pool size on adaptive testing. *Applied Psychological Measurement, 5*, 11-19.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*, 233-247.

Urry, V. S. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

van der Linden, W. J. & Zwarts, M. A. (1989). Some procedures for computerized ability testing. *International Journal of Educational Research, 13*, 176-188.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Weiss, D. J. (Ed.) (1978). *Proceedings of the 1977 Computertzed Adaptive Testing Conference*. Minneapolis: University of Minnesota.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Wood, R. L. (1973). Response-contingent testing. *Review of Educational Research, 43*, 529-544.

— (1976). Adaptive testing: A Bayesian procedure for the efficient measurement of ability. *Programmed Learning and Educational Technology, 13*, 34-48.

Item Banking Studies

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

Wood, R. L. (1976). Trait measurement and item banks. In D. N. M. de Gruijter and L. J. Th. van der Kamp (Eds.), *Advances in psychological and educational measurement*. New York: Wiley.

Equating Studies

Cook, L. L. & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.

— (1989). Using item response theory in test score equating. *International Journal of Educational Research, 13*, 162-171.

Cook, L. L., Eignor, D. R. & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25*, 31-45.

Divgi, D. R. (1981). Model free evaluation of equating and scaling. *Applied Psychological Measurement, 5*, 203-208.

Guskey, T. R. (1981). Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. *Applied Psychological Measurement, 5*, 187-201.

Gustafsson, J. E. (1978). The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement, 16*, 153-158.

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

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

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

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

Skaggs, G. & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*, 495-529.

— (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12*, 69-82.

Slinde, J. A. & Linn, R. L. (1977). Vertically equated tests: Fact or phantom? *Journal of Educational Measurement, 14*, 23-32.

— (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement, 15*, 23-35.

— (1978). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement, 16*, 159-165.

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

Item Bias Studies

Hambleton, R. K. & Rogers, H. J. (1986). Evaluation of the plot method for identifying potentially biased test items. In S. H. Irvine, S. Newstead & P. Dann (Eds.), *Computer-based human assessment*. Boston, MA: Kluwer-Nijhoff.

Ironson, G. H. (1983). Using item response theory to measure bias. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 128-144.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

Shepard, L. A., Camilli, G. & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317-375.

Shepard, L. A., Camilli, G. & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*, 83-138.

— (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*, 77-105.

Miscellaneous Applications

Bock, R. D. & Mislevy, R. J. (1988). Comprehensive educational assessment for the States: The duplex design. *Educational Evaluation and Policy Analysis, 10*, 89-105.

Drasgow, F. & Guertler, E. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19-29.

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

priateness indices. *Applied Psychological Measurement, 11*, 59-79.

Embretson, S. (1989). Latent trait models as an information-processing approach to testing. *International Journal of Educational Research, 13*, 190-204.

Harnisch, D. L. & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Levine, M. V. & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35*, 42-56.

Pandey, T. N. & Carlson, D. (1983). Application of item response models to reporting assessment data. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Thissen, D. M. (1976). Information in wrong responses to Raven's Progressive Matrices. *Journal of Educational Measurement, 13*, 201-204.

Computer Programs

Assessment Systems Corporation (1984). User's manual for the MicroCAT testing system. St. Paul, MN: Author.

Hambleton, R. K. & Rovinelli, R. (1973). A FORTRAN IV program for generating examinee response data from logistic test models. *Behavioral Science, 18*, 74.

Mislevy, R. & Bock, R. D. (1986). PC-BILOG. Mooresville, IN: Scientific Software, Inc.

Mislevy, R. J. & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75.

Thissen, D. M. (1983). MULTILOG: A user's guide. Chicago: International Educational Services.

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*. Vancouver, BC: Educational Research Institute of British Columbia.

Wood, R. L. & Lord, F. M. (1976). *A user's guide to LOGIST* (Research Memorandum 76-4). Princeton, NJ: Educational Testing Service.

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

Wright, B. D., Andrich, D. & Edgett, P. (1974). *MESA 21: CALFIT sample free item analysis for small computers*. Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D. & Mead, R. J. (1976). *BICAL: Calibrating rating scales with the Rasch model* (Research Memorandum No. 23). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

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

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*, 275-291.