European Journal of Psychological Assessment, Vol. 13, Issue 1, pp. 21–28

Item Response Theory Models and Testing Practices: Current International Status and Future Directions*,**

Ronald K. Hambleton and Sharon C. Slater

University of Massachusetts at Amherst, USA

Psychological testing has been undergoing major changes. One of the main changes is the transition from the use of classical to modern test models and methods in test development. The purposes of this paper are to describe the shortcomings of classical test models which are overcome with modern test theory, i.e., item response theory, to introduce the basic concepts of item response theory, to describe several important international applications of item response theory models, and finally, to describe some likely IRT directions in the next century.

Keywords: Item response theory, item response models, testing practices

Psychological testing has been undergoing major changes. Demands for tests to measure new and important psychological constructs, increased interest in diagnostic assessment, the influence of cognitive psychology on testing, and the role of computers in test administration, scoring, and score interpretations, are four of many changes taking place. Less well known among psychologists is that the basic psychometric theory for developing psychological tests and evaluating tests and test scores is changing too and these changes are going to make the construction and evaluation of tests and the interpretation of results easier and potentially more valid (Linn, 1990).

Many psychologists have undoubtedly seen references to the Rasch model, the three-parameter logistic model, latent trait theory, item response theory, latent ability, item characteristic curves, computer adaptive testing, etc. in the psychological testing texts, test manuals, and journals they read (see, for example, Anastasi, 1989; McGrew, Werder, & Woodcock, 1991). These are psychometric terms which are associated with modern test theory, known as "item response theory."

Item response theory (IRT) was introduced to the field of measurement in the early 1950s (Lord, 1952), became increasingly popular in the 60s, 70s and 80s (see, Lord, 1980; Wright & Stone, 1979), and today, provides the basis for the development of many important psychological tests (see, Hambleton, Swaminathan, & Rogers, 1991). The purposes of this paper are to address four topics: (1) the reasons for replacing popular classical test models and methods, (2) the basic concepts of IRT, (3) important international IRT applications, and (4) likely directions of IRT modeling and applications in the next century.

Shortcomings of Classical Measurement Models and Methods

Many psychologists will be familiar with "classical test theory." This theory has been used by psychologists for more than 70 years in the design and evaluation of tests (see, for example, Gulliksen, 1950).

Psychologists with training in psychometric methods will know that classical test theory is a theory in which a test score is assumed to consist of two major components, a true score and an error score, and that error scores are assumed to be uncorrelated with true scores, and error scores across parallel forms of a test are also considered to be uncorrelated. The true score of an examinee is defined as the examinee's expected score across infinite replications of parallel-forms of the test of interest. An error score is the difference between the construct

* Laboratory of Psychometric and Evaluative Research Report No. 229. Amherst, MA: University of Massachusetts, School of Education.

** Invited paper presented at the 23rd International Congress of Applied Psychology, Madrid, Spain.

of interest (i. e., true score) and the observable data (i. e., the test score) and every effort is made to minimize factors contributing to error such as improper sampling of content, poorly constructed items, guessing, cheating, misleading responses (e. g., responses reflecting social desirability), and flaws in the administration process such as test speededness. By reducing both random and systematic errors in the testing process, test score and true score are close and reliability and validity are increased.

Classical test theory has produced such well-known results as the Spearman-Brown formula for predicting the reliability of a full-length test from the correlation between scores on two-parallel halves, Kuder-Richardson Formula 20 for estimating the internal consistency of a test, the standard error of measurement used in interpreting test scores, and the attenuation formulas for estimating the correlation between scores on two constructs corrected for the unreliability of the scores used in the estimation of each construct. There is hardly a psychological test for which the corrected split-half reliability estimate, the Kuder-Richardson Formula 20, and the standard error of measurement are not reported.

Despite the usefulness of classical test theory and models in test development, shortcomings in the basic theory have been recognized for a long time (Gulliksen, 1950; Lord & Novick, 1968; Rasch, 1960). One shortcoming is that classical item statistics – item difficulty and item discrimination – depend on the particular examinee samples from which they were obtained. A consequence of this dependence on a specific sample of examinees is that these item statistics are only useful when constructing tests for examinee populations that are similar to the sample of examinees from which the item statistics were obtained. Unfortunately, one cannot always be sure that the population of examinees for whom a test in intended is similar to the sample or samples of examinees used in obtaining, for example, pilot test item statistics. Test statistics, too, such as those addressing test score reliability and validity are also examinee sample dependent.

A second well-known shortcoming of classical test theory is that comparisons of examinees on the test score scale are limited to situations where examinees are administered the same (or parallel) tests. The seriousness of this shortcoming becomes evident when one recognizes that examinees often take non-parallel forms of a test or it may be desirable to administer them different non-parallel forms of a test. When several forms of a test that vary in difficulty are used, examinee scores across nonparallel forms are *not* comparable unless one makes use of equating procedures, which are often quite complex. In some cases, it may not even be possible to equate scores from one form of a test to another.

Currently, there are many situations where the use of non-equivalent tests are of interest. A computer-adaptive test (CAT) may be the best example (Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, & Thissen, 1990). A CAT is a test administered by a computer, where the items administered to an examinee are dependent on the examinee's performance on previous items: all examinees may begin the test with several medium difficulty items; perform well and the computer selects harder items; perform poorly and the computer selects easier items. Advantages of a CAT include reduced testing time (on average, perhaps about 50% of the testing time can be eliminated), increased test security, flexibility in test scheduling, quicker and enhanced score reporting, and improved precision in estimating psychological constructs. Because examinees are administered unique and non-parallel forms of a test, valid comparisions of test scores are impossible. What is needed, if the goal is to tailor or adapt the administration of tests to examinees, is an approach to ability estimation which is not test dependent.

Basic IRT Concepts

Item response theory (IRT) purports to overcome the shortcomings of classical test theory by providing a reporting scale on which examinee ability (the construct measured by the test) is independent of the particular items which make up the test constructed from the bank of items which measures the construct. In addition, item statistics can be obtained which are independent of the particular sample of examinees from the population of examinees for whom the test is intended.

What began in the 1940s and 1950s as a goal of psychometricians to provide item-free ability estimates and person-free item statistics, became reality beginning in the 1960s and 1970s (Lord & Novick, 1968; Rasch, 1960). By the early 1970s, the theory was being advanced nicely, computer software was available, and applications of IRT were beginning to appear. Today, IRT is well developed and being used by test publishers, large testing agencies, test developers, and researchers around the world to address technical problems such as the design of tests, the study of item bias, equating test scores, and com-

puter-adaptive testing (see, for example, Hambleton, Swaminathan, & Rogers, 1991).

IRT, in its basic form, postulates that (1) underlying examinee performance on an test is a single ability or trait, and (2) the relationship between the probability that an examinee will provide a correct answer (or agree to a statement, in the case of a personality or attitude survey) and the examinee's ability can be described by a *monotonically* increasing curve. This is not a controversial point. Certainly we would expect examinees with more ability to have a higher probability of providing a correct answer than those with less ability. Or, in the case of (say) an instrument measuring student attitudes towards school, we would expect those persons with very positive attitudes toward school (the construct measured by the instrument) to agree with a positive statement more frequently than those persons with less positive attitudes. This curve representing the relationship between probability of a correct response and ability is called an "item characteristic curve" (ICC). It is "S-shaped" and indicates the probability of examinees at various ability levels answering an item correctly. An ICC is estimated for each item in the item bank from which a test is constructed.

The useful features about ICCs are that, in principle, they reflect where on the ability scale the corresponding items do their best discrimination *and* how well they discriminate. The fact that items and examinees are reported on the same scale opens up possibilities for matching items and examinees for more effective measurements.

For each item, a set of descriptors (i. e., the item statistics) can be used to describe the ICC. The particular values of the item parameters for any item determine the exact shape of the ICC. With highly discriminating items, the kind of item every test developer wants, the ICCs are very steep; for easy items, the ICCs are shifted to the left end of the ability scale (where lower performing examinees have a moderate to high probability of correctly responding), and for hard items, the ICCs are shifted to the right end of the ability scale where probabilities of successful performance will be low for low performing examinees and much higher for higher performing examinees (Hambleton, Swaminathan, & Rogers, 1991).

ICCs for dichotomously scored items (e. g., correct/incorrect or true/false) are typically described by one, two, or three parameters. The number of parameters identifies the IRT model. With the popular Rasch model, or one-parameter model, items are described by a single item parameter, called the "item difficulty statistic." With more general models, items may be described by two or more item parameters. For example, in the two-parameter logistic model, items are described by both "item difficulty" and "item discrimination" parameters (Hambleton, Swaminathan, & Rogers, 1991).

The number of parameters in an IRT model often depends on a researcher's philosophical orientation to measurement or sometimes on the number of item parameters the researcher needs to adequately fit his/her test data (Hambleton, 1994 a). Both the Rasch model and the three-parameter logistic model are popular in education and psychology and are used regularly in test development (see, for example, Woodcock, 1978; Yen, 1983). And, as will be described later, one of the most important IRT directions, is the development of new models to handle data which is polytomously scored and may be multidimensional in character (van der Linden & Hambleton, 1996).

The key features of IRT models that make them so appealing in measurement work are (1) the invariance property of item and person parameters, (2) the capability of providing a measure of precision for each ability score, (3) reporting items and examinees on a common reporting scale, and (4) the presence of item information functions (which indicate the contributions of items to measurement precision along the ability continuum and serve as the "building blocks" of test development). When an IRT model fits the test data to which it is applied, these four features (and others) permit many successful applications of IRT to testing problems.

The ways IRT can be applied to these and other testing problems, and examples of how IRT is being used, are described in the next section. Examples of applications of IRT are readily found in the United States. Many large testing corporations in America currently use IRT, for example, Educational Testing Service (ETS), American College Testing (ACT), California Test Bureau/McGraw-Hill, National Board of Medical Examiners (NBME), Psychological Corporation, and the Law School Admissions Council (LSAC). However, the focus of this paper is on the international impact of IRT, therefore, the examples which follow will be from applications outside of the United States. Applications of IRT models have been found in the following countries: Australia, Belgium, Canada, China, Germany, Great Britain, Indonesia, Israel, Korea, Japan, The Netherlands, Spain, Sweden, Taiwan, Turkey, and the United States.

International Applications of Item Response Theory Models

In this section, seven applications of IRT models will be described. Both the application itself and the special property or properties of IRT models used in the application will be highlighted.

Test Development

The properties of IRT create a number of benefits in the area of test design. One is in the improvement of item banking made possible because of the invariance of item parameters. Item invariance essentially means that the item parameters do not depend on the particular sample of examinees used to calibrate the items. Given that the examinee sample is large and heterogenous, and given model fit to the test data, the item parameters calibrated will hold across other samples of examinees from the population of examinees for whom the test was constructed. This allows items pre-tested at different times and with different examinee samples to be included in the same item bank.

A group in Israel, Ben-Simon, Tractinsky, and Cohen (1989) developed an item bank of 545 English as a Foreign Language (EFL) items using the three-parameter logistic model. They even developed their own software program, *NITEST*, for estimating IRT parameters (Cohen & Bodner, 1989). Other work in Turkey has used the Rasch model to re-evaluate a depression measure, the *Depressive Attribution Style Questionnaire* (Aydin & Berberoglu, 1990).

The Australian Council for Educational Research is using IRT applications for several aspects of their Basic Skills Testing Program, including item-calibration, banking, and test design. Their work is based on the Rasch model (Masters, Lokan, Doig, Khoo, Lindsey, Robinson, & Zammit, 1990).From item banks like these, IRT can be used to design tests with certain specifications. Through the use of item and test information functions, it is possible to custom design tests by specifying a *target test information function* (TIF) and then adding items, with known item information functions, to the test in order to achieve the specified target. Item information functions (IIF) have the property of additivity, so you simply add the IIFs from enough items until the target TIF is obtained. Because IRT is a theory based on item level data (as opposed to test level data, like classical test theory) it is possible to see the effects of adding or omitting any particular item (see, for example, van der Linden & Bookooi-Timminga, 1989). The concepts of item and test information are new to test development and very useful.

A target TIF can be selected either to maximize test information at a passing score or to yield high information across any range of ability scores. There are two computer software programs available commercially that perform this "optimal test design" or "computerized test assembly" (*ConTEST*, Timminga & van der Linden, 1996; *Optimal Test Design*, Verschoor, 1991). These enable the user to enter test specifications, such as, the desired TIF and content considerations, and the computer then constructs the first draft of a desired test with these constraints in mind (van der Linden & Bookooi-Timminga, 1989).

In China, Yu (1991) used IRT to develop an attitude scale to measure a student's attitudes regarding the teaching profession. There has also been IRT test development work done in Japan. Watanabe and Takahashi (1994) used IRT to develop a new job interest index. Items were calibrated using the two-parameter logistic model. This paper also references many other applications of IRT used in Japan (e. g., Takahashi, 1994).

Computer Adaptive Testing

An application that would be next to impossible without IRT is computer adaptive testing (CAT). CAT uses item information functions (and the associated standard errors) to construct tests tailored to an examinee's ability level. In CAT, the computer constructs the test "on the fly," while the examinee is taking the test. According to a number of prespecified rules, the computer selects and administers items to an examinee that yield the maximum information from the examinee. A simplified explanation is that an examinee is administered the most discriminating item that he or she has about a 50% probability of answering correctly. In other words, the most informative items for an examinee are administered after reassessing the ability estimate. This estimate is based on the responses to the previous items in the CAT. This IRT application capitalizes on the invariance property of ability estimates and the presence of items and ability estimates on a common reporting scale.

In essence, every examinee sees a unique set of items tailored to his/her estimated ability level. Even though examinees are administered different items, their scores (i. e., ability estimates) may be compared because of the invariance of the ability

estimates (i.e., non-dependence on the particular choice of items). For more information on CAT, both theory and applications, readers are encouraged to read Wainer et al. (1990).

In Israel, Cohen, Ben-Simon, and Tractinsky (1989) used their item bank, calibrated with the three-parameter logistic model, to develop a CAT for their *English as a Foreign Language Exam* using a software package called *MicroCAT*. They also developed a software package for research on CAT and IRT called *NITECAT* (Cohen, Bodner, & Ronen, 1989). In China, Liu (1990) developed a CAT version of the *Raven Progressive Matrices Test* (RCAT).

Test Score Equating

The property of model parameter invariance makes it possible to separate the difficulty of a given test from the ability of the sample of examinees taking that test. This property of IRT may appear to make test equating unnecessary for test designed with IRT. If ability estimates do not depend on the set of items administered, it should not be necessary to equate those sets of items. Theoretically, this is true. However, in practice, different tests (and the ability estimates derived from those tests) will be on different scales and therefore need to be placed on the same scale in order to be compared. This is often called scaling instead of equating, and is one of the most popular IRT applications.

In IRT, the *b*-parameters (i.e., the item difficulty statistics) are linearly related (unlike p-values in classical testing) so the scaling is relatively easy, especially if items are already calibrated using IRT. If the items are not already calibrated with IRT, the classical test theory procedures (e.g., equipercentile) are comparable to IRT equating procedures in the case of horizontal equating, as is used for parallel or near parallel test forms. However, in the case of vertical equating, where tests of varying difficulty need to be placed on the same scale, IRT equating procedures are superior.

Glas (1992) in the Netherlands stated that the Rasch model with a multivariate distribution of ability is used for vertical equating, to equate examinations over consecutive years by the Netherlands Department of Education. In China, Gui and Li (1991) reported that IRT has been used to equate English tests in the higher education entrance examination of Guangdong Province since 1988.

There are a number of other advantages to equating using IRT. Multiple tests are able to be equated easily. Also, changes in the test are easy to make even after equating has been done. This 're-equating' simply involves removing the desired items, and revising the test characteristic curve (the test characteristic curve is the sum of the item characteristic curves in the test), without the need to recalibrate the test. When item parameter estimates are known, pre-equating is also possible (i.e., tests can be equated before they are administered). (See Cook & Eignor, 1991, for more details on IRT equating).

Differential Item Functioning

An application that is naturally suited to IRT procedures is the identification of potentially biased items, or DIF (differential item functioning) items. If reference and focal groups differ in their mean performance on an item, it doesn't automatically mean that the item is biased. Legitimate differences may exist in performance due to differences in ability between the two groups. It is difficult to separate a group's ability from their performance on a set of test items, but this is necessary to determine if item bias exists.

It is said that "an item shows DIF if individuals having the same ability, but from different groups do not have the same probability of getting the item right" (Hambleton, Swaminathan, & Rogers, 1991). Because IRT can provide ability estimates independent of the set of test items administered, IRT provides a natural framework for studying DIF. If the ICCs for an item are different for two groups (e.g., males and females), then DIF is present. It is possible to compare the item parameters or to compute the area between the two ICCs to determine if there is DIF. These methods are not clearly better than classical testing methods, like the Mantel-Haensel procedure, but IRT DIF methods are being used successfully (see Camilli & Shepard, 1994; Holland & Wainer, 1993 for more details on DIF).

In Turkey, Berberoglu (1989) used the one-parameter logistic model to determine DIF for selection and placement tests into higher education. Two British researchers used IRT to detect bias in a translated locus of control scale (Ertubey & Russell, 1994). They used an IRT software package (MULTILOG) for handling polytomously-scored data to fit these data. Instead of looking for DIF between two different groups of examinees, they used the same group, bilingual Turkish teenage students, to determine if DIF existed between the English and Turkish versions of the locus of control scale.

Score Reporting

Reporting can also be improved upon by using IRT methods. There are two features that IRT offers to reporting of scores: (1) the calculation of a more accurate standard error of measurement (for each ability score), and (2) the possibility to predict performance of examinees on items not administered in a test (but calibrated on the same scale as the items which are administered).

The standard error of ability estimation is computed for *each* score level. This obviously provides a more accurate estimate of error than the popular group standard error of measurement used in classical test theory. It is possible to calculate more accurate estimates of error using classical test theory methods (Feldt, Steffen, & Gupta, 1985), however, these methods are not commonly used. In IRT, it is routine to report the more accurate standard errors.

When an examinee's ability has been estimated from a test, it is possible to map this ability estimate onto ICCs of items not administered to the examinee. This allows richer information about what a person does and does not know or can and cannot do. Of course the validity of these inferences is dependent on model fit to the test data. The information about items seen and not seen by an examinee is able to be incorporated allowing more detail to be included in the score report.

An excellent example of IRT reporting is in the work done by the Basic Skills Testing Program of the Australian Council for Educational Research (Masters et. al, 1990). The reports generated from this testing program include much more detailed information, listing some of the skills students with any particular score are generally able to do along with the test score.

Alternative Forms of Testing (Performance Assessment)

One of the complexities of using performance assessments is in determining how to score them. The responses of performance assessments are usually not dichotomous multiple-choice responses, and need to be scored in different ways. IRT offers some solutions to this testing problem.

There are a number of polytomous IRT models that are much better suited to open-ended questions, such as essays, or multiple-step problems. Some of the models being used in these situations are Bock's nominal-response model, Samejima's graded-response model, and Master's partial-credit model (see, van der Linden & Hambleton, 1996). These models are helpful once the scoring has been done by the raters. However, IRT obviously cannot help in the judgmental process of scoring performance assessments.

The work already mentioned by Ertubey and Russell (1994) employs a polytomous IRT model, Samejima's graded-response model. Also, in Australia, Masters and Wright (1996) have applied the partial-credit model to assess student essay writing and to identify differences among judges in their interpretations and uses of grades.

Test Adaptations

A couple of the applications already mentioned are quite useful when attempting to translate tests from one language to another. The process is actually more complicated than just translating the text. Cultural issues and values must also be taken into account. For this reason, some prefer to call the process test adaptation (see, for example, Hambleton, 1994 b).

To perform a test adaptation, both equating and DIF play a role. Especially in the case where international comparisons are to be made, it is essential that the tests in both languages are equated. Also, it is important to determine that the test or specific items are not biased against any of the cultural groups being tested. DIF studies can help to guard against this. This is the application seen in the study by Ertubey and Russell (1994) mentioned under the heading of differential item functioning. (See Hambleton, 1993, 1994 b; van de Vijver & Hambleton, 1996; Woodcock & Munoz-Sandoval, 1993, for more on test adaptations).

Future Directions

There is little doubt that IRT models are providing technical solutions to many important and challenging technical problems in assessment. The examples given in this paper are of course not an exhaustive list of the current international uses of IRT in assessment practices. The fact that such a list would probably be impossible to compile today attests to the widespread use of IRT. On the other hand, IRT is not some kind of magic psychometric wand that can correct for major flaws in item writing, test construction, and test administration. But in the hands of competent assessment specialists, IRT models

provide useful solutions to challenging technical problems and open up new directions for test development (e. g., automated test construction), test administration (e. g., computer-adaptive testing), and test score reporting (e. g., enhanced score reporting).

Much research remains to be done in the area of IRT. Two of those areas include the development of polytomous and multidimensional models. Polytomous IRT models are needed to analyze the new wave of assessment formats and scoring paradigms associated with performance assessment in education. They are also needed to handle new applications of IRT models to the analysis of personality data. For a description of many of these new models, approaches to parameter estimation and model fit, readers are referred to van der Linden and Hambleton (1996).

Multidimensional models are needed to better fit current educational and psychological data which are multidimensional in structure. Such models seem preferrable to attempting to decompose educational and psychological data into unidimensional components and estimating examinee scores on each component. Lost in this type of analysis is the complexity of the original structure among the underlying components or traits which are measured by the test.

Author's Address:
Professor Ronald K. Hambleton
University of Massachusetts
Hills South, Room 152
Amherst, MA 01003 , USA
Telephone: 413–545–0262, Fax: 413–545–4181
e-mail: RKH@EDUC.UMASS.EDU

References

Anastasi, A. (1989). *Psychological testing* (6th ed.). New York: Macmillan.

Aydin, G., & Berberoglu, G. (1990). *The construct validity of a Turkish depressive attribution style questionnaire (DASQ) for university students*. Unpublished manuscript.

Ben-Simon, A., Tractinsky, N., & Cohen, Y. (1989). *Item-banking of EFL items using the 3-p logistic model* (CAT Project Report No. 4). Jerusalem, Israel: National Institute for Testing and Evaluation.

Berberoglu, G. (1989, August). *The use of one parameter logistic model in the university entrance examinations in Turkey*. Paper presented at the 11th EAIR Forum, University of Trier, Germany.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Cohen, Y., Ben-Simon, A., & Tractinsky, N. (1989). *Computerized adaptive test of English proficiency* (CAT Project Report No. 6). Jerusalem, Israel: National Institute for Testing and Evaluation.

Cohen, Y., & Bodner, G. (1989). *A manual for NITEST – a program for estimating IRT parameters* (CAT Project Report No. 1). Jerusalem, Israel: National Institute for Testing and Evaluation.

Cohen, Y., Bodner, G., & Ronen, T. (1989). *A manual for NITECAT a software package for research on CAT/IRT version 1* (CAT Project Report No. 2). Jerusalem, Israel: National Institute for Testing and Evaluation.

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice*, *10*(3), 37–45.

Ertubey, C., & Russell, R. J. H. (1994, August). *Using item response theory to detect bias in a translated locus of control scale*. Paper presented at the meeting of the British Psychological Society, Brighton.

Feldt, L. S., Steffen, M., & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement*, *9*(4), 351–361.

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement: Theory into practice: Vol. 1*. Norwood, NJ: Ablex Publishing Corporation.

Gui, S., & Li, W. (1991). Applications of item response theory in equating. *China Examinations*, *3*, 25–29.

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

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment*, *9*(1), 57–68.

Hambleton, R. K. (1994 a). Item response theory: a broad psychometric framework for measurement advances. *Psicothema*, *6*(3), 535–556.

Hambleton, R. K. (1994 b). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, *10*, 229–240.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Linn, R. L. (1990). Has item response theory increased the validity of achievement test scores? *Applied Measurement in Education*, *3*(2), 115–141.

Liu, J. (1990). Testing the reliability of Raven computerized adaptive tests. *Information on Psychological Sciences*, 34–39.

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

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Masters, G. N., Lokan, J., Doig, B. A., Khoo, S. T., Lindsey, J., Robinson, L., & Zammit, S. (1990). *Profiles of learning: The Basic Skills Testing Program in New South*

Wales, 1989. Hawthorn, Australia: Australian Council for Educational Research.

Masters, G. N., & Wright, B. (1996). Partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer-Verlag Publishers.

McGrew, K. S., Werder, J. K., & Woodcock, R. W. (1991). *Woodcock-Johnson Technical Manual*. Allen, TX: DLM.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Neilsen and Lydiche.

Takahashi, K. (1994). Empirical test of the stage model of organizational socialization: Development of integrated model and verification of its validity. *The Japanese Journal of Administrative Behavior*, *9*(1).

Timminga, E., & van der Linden, W. (1996). *ConTEST Technical Manual*. Groningen, the Netherlands: ProGamma.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, *1*, 89–99.

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika*, *54*(2), 237–247.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1996). *Handbook of modern item response theory*. New York: Springer-Verlag Publishers.

Verschoor, A. (1991). *OTD: Optimal test design* (Manual). Arnhem, The Netherlands: CITO.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (Eds.). (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.

Watanabe, N., & Takahashi, K. (1994, July). *A development of a job interest index by item response theory: Based on Japanese samples*. Paper presented at the 23rd International Congress of Applied Psychology, Madrid, Spain.

Woodcock, R. W. (1978). *Development and standardization of the Woodcock-Johnson psycho-educational battery*. New York: Teaching Resources.

Woodcock, R. W., & Munoz-Sandoval, A. F. (1993). An IRT approach to cross-language test equating and interpretation. *European Jounral of Psychological Assessment*, *9*, 233–241.

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

Yen, W. M. (1983). Use of the three-parameter logistic 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.

Yu, J. (1991). Developing a professional attitude scale for teachers school students. *Proceedings of International Academic Symposium on Psychological Measurement*, Peking, China.