Psicothema, 1994. Vol 6, n.º 3, pp. 535-556
ISSN 0214 - 9915 CODEN PSOTEG

# ITEM RESPONSE THEORY: A BROAD PSYCHOMETRIC FRAMEWORK FOR MEASUREMENT ADVANCES[1, 2]

Ronald K. Hambleton
University of Massachusetts at Amherst

Measurement theory and practice have changed considerably in the last 25 years. For many assessment specialists to-day, item response theory (IRT) has replaced classical measurement theory as a framework for test development, scale construction, score reporting, and test evaluation. The most popular of the item response models for multiple-choice tests are the one-parameter (i.e., the Rasch) and three-parameter models. Some researchers have been quite adamant about using only the one-parameter model and have been rather critical of applications of multi-parameter models such as the three-parameter model. In this paper, nine arguments are offered for continuing research and applying multi-parameter IRT models. Also, the position is taken that both single *and* multi-parameter IRT models (and many others) have potentially important roles to play in the advancement of measurement practice and judgments about which models to use in particular situations should depend on model fit to the test data.

Measurement theory and practice have changed considerably since the seminal publications of Lord (1952, 1953a, 1953b), Birnbaum (1957, 1958a, 1958b), Lord and Novick (1968), Rasch (1960), and Fischer (1974). Since the publication of Lord and Novick's *Statistical Theories of Mental Test Scores* in 1968, large numbers of item response models have been proposed, estimation and goodness-of-fit procedures developed, and small- and large-scale applications, almost too numerous and varied to count, have followed. A quick check of the *Journal of Educational Measurement* and *Applied Psychological Measurement* since 1977, located over 200 papers! And these papers represent only a fraction of the publications that are available to interested readers. For useful summaries of current as well as new item response theory (IRT) models, readers are referred to Hambleton, Swaminathan, and Rogers (1991), Lord (1980), Thissen and Steinberg (1986), McDonald (1982, 1989), Mellenbergh (1994), Masters (1982), Masters and Wright (1984), Goldstein and Wood (1989), and van der Linden and Hambleton (in press). Unidimensional and multi-dimensional models to handle dichotomous as well as polytomous data with various properties (e.g., nominal, ordinal, interval) from the cognitive, affective, and psychomotor domains, can now be found in the measurement literature (see van der Linden & Hambleton, in press).

Some of the largest and most influential assessment instruments in the country are influenced in some way or other by item response models. Arguably the most important data to address the third goal of President Bush's Education 2000 is provided by the

Correspondence: Ronald K. Hambleton
University of Massachusetts at Amherst
School of Education
Amherst, MA 01003 USA

[1] Paper presented at the meeting of NCME, San Francisco, 1992.

[2] *Laboratory of Psychometric and Evaluative Research Report No. 235.* Amherst, MA: University of Massachusetts, School of Education.

*National Assessment of Educational Progress,* an assessment system for grades 4, 8, and 12 and based upon the three-parameter logistic model. Two of the major standardized achievement tests, *Comprehensive Tests of Basic Skills,* and the *Metropolitan Achievement Tests,* are developed and scaled with the three-parameter and one-parameter logistic models (or Rasch model, as it is often called), respectively. Major national selection tests such as the *Scholastic Aptitude Test, Graduate Management Admissions Test, Law School Admission Test,* and the *Graduate Record Exam* utilize the three-parameter model in item selection, scoring, equating, detecting differentially functioning test items, and other ways. The *Armed Services Vocational Aptitude Battery* (ASVAB) was calibrated with the three-parameter logistic model. And, numerous other achievement, aptitude, and personality tests could be added to this list. Test applications based upon IRT principles and applications impact on millions of students in the U.S. and in other countries each year (Hambleton, 1989).

Working within the item response theory field is a group of researchers led by Professor Benjamin Wright and several prominent former students from the University of Chicago, who reject many of the current IRT models and their applications. Their rejection appears to be based on theoretical as well as empirical grounds (e.g., Wright, 1968, 1977, 1984; Wright & Stone, 1979). Their special interest is in the Rasch model (and extensions) and its applications to test data (Rasch, 1960). Many European scholars, in addition to Wright and his co-workers, too (e.g., Fischer, 1974; Gustaffson, 1980a, 1980b), have been responsible for important Rasch model developments.

Whereas many IRT researchers have taken the position that the selection of psychometric models should be based in the final analysis on their psychological meaningfulness, as well as goodness of fit evidence and utility, and have adopted multi-parameter IRT models in their work, Professor Wright and his co-workers argue for «fundamental measurement.» His models preclude item discrimination and pseudo-guessing parameters, for example, because these models violate the assumptions of simple ordering of persons and/or items required of conjoint measurement models and hence such models do *not* lead to «fundamental measurement.» On the other hand, multi-parameter models have found wide use in many testing applications. What seems clear is that researchers fall into two categories: those who are basically model builders and are willing to use many IRT models in their psychometric work when they fit, and those who believe in fundamental measurement and restrict their attention, therefore, to Rasch models and extensions which do meet the strict criteria or assumptions of conjoint measurement.

In the remainder of this paper, nine arguments will be offered to support the strong current interest in and use of IRT models that include more than a single model parameter to account for item difficulty:

1. Nearly a century of testing experience.
2. Reasonableness of fitting IRT models to data.
3. Central importance of the property of «parameter invariance.»
4. Usefulness of item discrimination as an IRT model parameter.
5. Usefulness of the «pseudo-guessing» parameter as an IRT model parameter.
6. Reasonableness of IRT multi-parameter software packages.
7. Successes in applications of multi-parameter IRT models.
8. Shortcomings in the arguments to support the Rasch model.
9. Promising future of multi-parameter IRT models.

The remainder of this paper will be organized around the nine arguments. Concluding remarks will summarize the arguments and point to future IRT directions and development.

## 1. *Nearly a Century of Experience*

Item response theory, as clearly outlined in Lord and Novick (1968) and Lord (1980), was developed over at least a 50-year period and evolved from classical measurement theory. Basic concerns for item characteristic curves and invariant item and ability parameters can be traced to work by Tucker (1946) and Gulliksen (1950). What troubled Tucker, Gulliksen, Lawley, Lord, and other psychometricians about their models and methods in the 1940s was that they produced item and ability parameters which were sample dependent. For example, Figure 1 highlights the problem (see Lord, 1953b). The same group of examinees would have relatively *low* true scores on a difficult test and relatively *high* true scores on an easy test measuring the same ability. However, examinees approach both tests with the same ability. Lord felt it would be useful to find psychometric models which could incorporate the characteristics of the test items into the ability parameter estimation process so that the abilities and a single ability distribution could be estimated. Popular item statistics such as item difficulty (e.g., p-value) and item discriminating power (e.g., the point-biserial correlation), too, were sample dependent, which limited their usefulness in test design. In fact, interest in the biserial correlation as a measure of item discrimination increased in the 1950s because of evidence that it tended to be more sample independent than other item discrimination indices.

The goal of these early psychometricians was to obtain *invariant* parameters: descriptors of items which would not depend upon the particular examinee sample, and descriptors of examinees (i.e., ability scores) which would be independent of the sample of items from the larger domain of items measuring the ability of interest. Lord's contributions in 1952 and 1953 (1952, 1953a, 1953b) were especially important. In these papers, he developed the normal ogive model, offered approaches for model parameter estimation and model fit, and formally connected IRT models to well-known classical measurement models. McDonald (1989, p. 209) went so far as to say «...just about anything done in the field of binary item response theory can be thought of as a footnote to the seminal research of Frederic M. Lord» (p. 209). This statement captures the sentiments of many psychometricians.

Lord offered a two-parameter normal-ogive unidimensional model in 1952: Unidimensional because this was then, and remains today, a reasonable assumption for many sets of test data, normal-ogive ICCs (rather than logistic) because normal ogives were popular (at the time) «S»-shaped curves bounded by 0 and 1 and had been used successfully by other psychometricians (Tucker, 1946), *and* two item parameters because of the long tradition of utilizing two item statistics, difficulty and discrimination, in test development and test analysis work. On this last point, both item statistics had simple and important relationships to test score characteristics. Thus, the use of a second item parameter was «not an obsession for complexity...» as one critic reported, but rather a response to well-established and validated psychometric procedures. An item discrimination parameter was placed in the model by Lord because of empirical evidence of its importance and utility. To assume all discriminating powers equal, when it was well known then, as it is today, that items typically show variability in their discriminating power, would have been unlikely.

Later, Birnbaum (1957, 1958a, 1958b) substituted the logistic function for the normal-ogive, and added an additional parameter to the model to account for non-zero performance of low-ability examinees. In 1960, Georg Rasch published his own version of a one-parameter IRT model for use with achievement tests, which, in fact, in a different (but equivalent) form, had been known by Lord in 1952. But Lord rejected

the model because he felt it was unsuitable for use with multiple-choice items. Rasch, who was a statistician and not a psychometrician himself or he would have known of the earlier work by Lord and others which appeared in *Psychometrika* and *Educational and Psychological Measurement,* desired separability of item and person parameters, and he achieved it with his one-parameter model. In contrast, many psychometricians of the day valued (at least) two item parameters in their psychometric models, and were prepared to give up sufficient statistics (which were available with the Rasch model) to improve the fit of their models. Also, as Lord (1980) noted later, sufficient statistics are *not* guaranteed by the Rasch model. They are only present when this model *fits* the data.

Were it not for the lack of computer power in the 1950s, applications of item response models would have proceeded more quickly. Conditions were more favorable in the late 1960s and serious research began in many places. Publication of Lord and Novick's (1968) *Statistical Theories of Mental Test Scores* was influential, and Wright (1968) was influential, too. His publications (Wright, 1968; Wright & Panchapakesan, 1969), his level of energy, and his stimulating Rasch model training programs at the annual meetings of AERA attracted many researchers to the IRT area, though not always to his philosophical viewpoint.

Rasch model advocates have adopted «specific objectivity» and «sufficient statistics» as fundamental and essential in their work. To many psychometricians, their position represents a narrow basis on which to build a measurement model. It must be kept in mind that most (perhaps all) psychometricians would be willing to use the Rasch model when it can be demonstrated that it fits their data. In this way, they regard the Rasch model as one of many logistic models that lead to invariant model parameters. When it fits, the Rasch model can be used. In contrast, Rasch model advocates adhere to one model (or, more correctly, to one family of models) and they will redefine the ability measured by the test (explicitly, and, more often, implicitly, to the detriment of construct validity of the ability of interest) by deleting misfitting items.

## 2. *Models or Data: Which Should Come First?*

Few psychometricians would dispute the point that models are valuable in advancing psychometric work such as constructing tests and scaling and equating scores. And, many psychometricians would be quick to argue that meaningful model building follows from careful data analysis. Lord, Bock, Samejima, etc., are model builders. Lord, for example, delayed his IRT research program for 15 years (from 1952 to 1967), until he was able to derive parameter estimates for a model (the three-parameter model) that he felt he needed to fit multiple-choice test data. Justification for his position comes from several empirical studies (perhaps the best evidence is reported in Lord, 1970).

Scientists develop models to explain or fit their data, not usually the reverse. Otherwise, we may still be thinking that the sun travels around the earth and the earth is the center of the universe. Scientists in the social sciences build their models and theories from carefully analyzing their data. Awkward data may force changes in a model; but these data are not discarded for a more appropriate set.

But, as John Tukey was quoted (by Howard Wainer) as stating, «All models are wrong, but they may be useful!» Consider this quote:

> That the model is not true is certainly correct, no models are - not even the Newtonian laws. When you construct a model, you leave out all the details which you, with the knowledge at your disposal, consider inessential. . . . *Models should not be true, but it is important that they*

*are applicable,* and whether they are applicable for any given purpose must of course be investigated.

> . . . we may tentatively accept the model described, investigate how far our data agree with it, and perhaps find discrepancies which may lead us to certain revisions of the model.

This is a quote with which that many psychometricians could agree. It describes how we go about our psychometric work. Perhaps this quote came from Fred Lord or Darrell Bock, two of the leading model-builders in psychometric methods. Actually, the quote came from Georg Rasch, in his book, *Probabilistic Models for Some Intelligence and Attainment Tests.* Georg Rasch was obviously not in support of models being more important than the data. Those who take this position advocating the Rasch model seem to be in opposition to Rasch's advice about conducting psychometric research.

Does a model first, data second approach have any merit? «Specific objectivity» is a useful feature of a measuring instrument. «Sufficient statistics» are valuable, too. A psychometric model with these properties could be designed and was, in the form of the Rasch model. Is it of any value? If the model fits data, it could be. Few would dispute this position. As Rasch himself notes, empirical evidence should be compiled to check the fit of the model, and «perhaps find discrepancies which may lead us to certain revisions of the model» (Rasch, 1960). Of course, one alternative is to discard items and/or persons which are not consistent with the model. To many, this position is not defensible. Curriculum specialists would find even less use for measurement specialists if we asked them to narrow their test content, or delete some of their most discriminating items, so as to capitalize on the features of a simple psychometric model.

A simple example of the point that sufficient statistics are not essential in testing is easy to find. The mean of a set of test scores is a sufficient statistic which is used to estimate the population mean in a normal distribution. However, the sample mean is *not* a sufficient statistic if the distribution of scores is non-normal. In fact, considerable misinformation may be conveyed about the distribution of scores were it to be used. One popular response of statisticians or data analysts is to report the median, which is *not* a sufficient statistic, but which conveys more valuable data about the test score distribution than the mean with non-normal distributions. At the very least, statistics which are *not* sufficient can and do often provide valuable information.

Whether sufficiency is essential or not, the key point with respect to ability estimation is that sufficient statistics are only available to the extent that the Rasch model fits the data. When the basic model assumptions are violated, test score is *not* a sufficient statistic for the estimation of ability. Lord (1980) and others have made these points often. It is not a matter of model robustness either. Either sufficient statistics are present or they are not.

Sufficient statistics are valued by statisticians but they are not the only basis on which to produce a test theory. Lord's (1980) view was that the concept of test information should be central. He felt maximizing the information provided by a test was a desirable goal, and scoring weights for items should be chosen to maximize test information. Readers are referred to Lord (1980) to see the derivation of optimal scoring weights for the one–, two–, and three-parameter logistic models. The loss of information in ability estimation with the one-parameter model can be substantial when items vary substantially in their discrimination power, and performance of low-ability examinees is not zero (see, for example, Lord, 1980).



### 3. *Parameter Invariance is the Cornerstone of IRT*

The property of invariance of ability and item parameters is the *cornerstone* of IRT. It is the major distinction between IRT and classical test theory (see, for example, Hambleton & Jones, 1993). Figures 1 and 2 highlight ability parameter invariance (over tests of differing difficulty) and item parameter invariance (over two examinee samples of differing ability). The property implies that the parameters that characterize an item do not depend on the ability distribution of the examinees and the parameter that characterizes an examinee does not depend on the set of test items.

The invariance property is a characteristic of *all* item response models. Of course, the property is only present when the IRT model fits the test data, and when model parameters are estimated properly. As Lord (1980) reminds us, invariance is a property of the model parameters *not* the estimates. Thus, care must be taken in practice to estimate model parameters well. For example, a homogeneous low-performing group would be a poor choice of examinee sample to calibrate item statistics on a set of relatively difficult items.

Some will argue that item parameters cannot be invariant if the choice of examinee sample needs to be considered. But this argument is incorrect. An example may be helpful. In the linear regression model, the regression line for predicting a variable Y from a variable X is obtained as the line joining the means of the Y variable for each value of the X variable. When the regression model holds, the same regression line will be

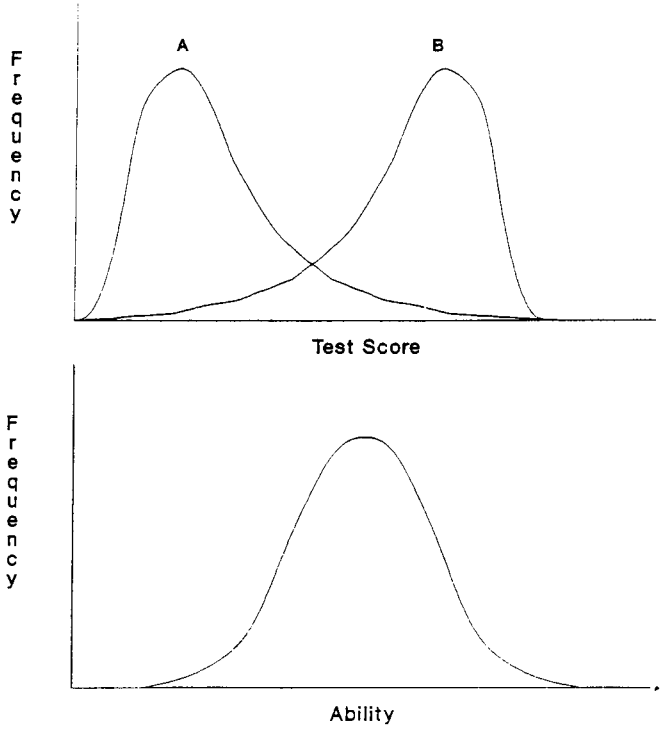


*Figure 1.*–Test Score and Ability Distributions for the Examinee Group

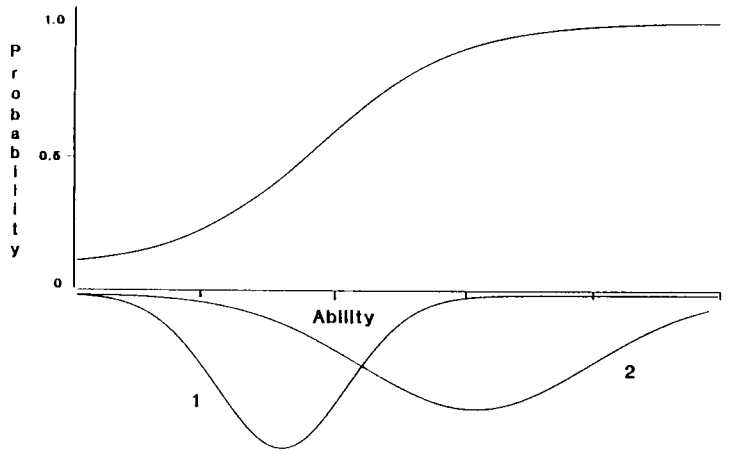


*Figure 2.*–An Item Characteristic Curve and Distributions of Ability for Two Groups of Examinees

obtained within any restricted range of the X variable, that is, in any subpopulation on X, meaning that the slope and intercept of the line will be the same in any subpopulation on X. A derived index such as the correlation coefficient, which is not a parameter that characterizes the regression line, is *not* invariant across subpopulations. The difference between the slope parameter and the correlation coefficient is that the slope parameter does not depend on the characteristics of the subpopulation, such as its variability, whereas the correlation coefficient does. Note, however, that the proper *estimation* of the regression line does require a heterogeneous sample. A homogeneous sample of examinees will provide unstable estimates of the model parameters. The same concepts also apply in item response models, which can be regarded as nonlinear regression models.

It is important to determine whether invariance holds, since every application of item response theory capitalizes on this property. Although invariance is clearly an all-or-none property in the population and can never be observed in the strict sense, we can assess the «degree» to which it holds when samples of test data are used. For example, if two samples of different ability are drawn from the population and item parameters are estimated in each sample, the congruence between the two sets of estimates of each item parameter can be taken as an indication of the degree to which invariance holds. The degree of congruence can be assessed by examining the correlation between the two sets of estimates of each item parameter or by studying the corresponding scatterplot. Figure 3 shows a plot of the difficulty values for 75 items based on two samples from a population of examinees. Suppose that the samples differed with respect to ability. Since the difficulty estimates based on the two samples lie on a straight line, with some scatter, it can be concluded that the invariance property of item parameters holds. Similar checks can be carried out on other item parameters in the model. Some degree of scatter can be expected because of the use of samples; a large amount of scatter would indicate a lack of invariance that might be caused either by model-data misfit or poor item parameter estimation (which, unfortunately, are confounded).

The assessment of invariance described above is clearly subjective but is used because no objective criteria are currently available. Such investigations of the degree to which invariance holds are, as seen above, investigations of the fit of the model to the

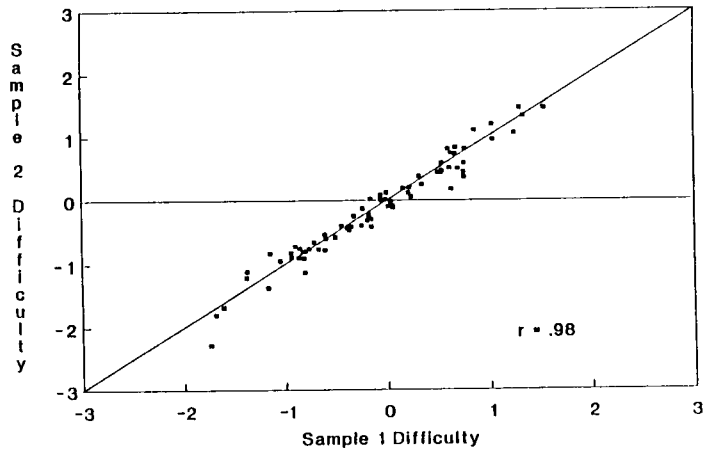


*Figure 3.*–Plot of 3P Item Difficulty Values Based on Two Groups of Examinees

data, since invariance and model-data fit are equivalent concepts (see, for example, Hambleton, Swaminathan, & Rogers, 1991).

A similar example can be given to highlight ability invariance (see Hambleton, Swaminathan, & Rogers, 1991). Invariance of item and ability parameters is a feature of *all* item response models, one–, two–, and three-parameter logistic models and others, and can be obtained in the model parameter *estimates* when the model fits the data, and appropriate data collection designs are used. The less complex IRT models typically require simpler designs. The parallel to fitting linear and non-linear regression models is obvious.

### 4. *Item Discrimination*

One of the strongest criticisms of the one-parameter logistic model is that it does not account for variability in item discriminating power. As Traub (1983) so cogently states,

> . . . these [Rasch] assumptions fly in the face of common sense and a wealth of empirical evidence accumulated over the last 80 years...The fact that otherwise acceptable achievement items differ in the degree to which they correlate with the underlying trait has been observed so very often that we should expect this kind of variation for any set of achievement items we choose to study. (p. 64)

Table 1 contains the means, standard deviations, and low and high values of the item biserial correlations on the nine subtests in the *Armed Services Vocational Aptitude Battery* (ASVAB). This is an excellent example of data for which an item discrimination parameter is needed to fit the test data.

Lord (1952) certainly was not prepared to proceed with his IRT research program until he could satisfactorily estimate *both* item difficulty and discrimination (see Bejar, 1983, pp. 12-13). Many other researchers, too, doubt the wisdom of discarding the item discrimination parameter. Hundreds of applications of multi-parameter IRT models attest to this point.

Two common arguments for using an item discrimination parameter are (1) you don't want to eliminate some of your best items, and (2) a concern that the ability of interest may be changed if non-fitting items are removed from the test or item pool. Figures 4 and 5 highlight the first concern using over 250 1977 NAEP Mathematics Assessment items (Hambleton & Rogers, 1990). Items with low and high biserial correlations are *not* fit well by a one-parameter

*Table 1*
Item Biserial Correlations in the Armed Services Vocational Aptitude Battery Subtests[1]
(Total Calibration Sample, N = 3000 +)

| Descriptive Statistic | Content Area | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | GS | AR | WK | PC | AI | SI | MK | MC | EI |
| Mean | .56 | .63 | .66 | .65 | .57 | .53 | .64 | .53 | .49 |
| Standard Deviation | .15 | .12 | .22 | .14 | .17 | .15 | .13 | .13 | .13 |
| Low Value | .07 | –.05 | .07 | .19 | .09 | .08 | .25 | .06 | .11 |
| High Value | .82 | .86 | 1.00 | .99 | .82 | .77 | .93 | .79 | .80 |
| Number of Items | 228 | 245 | 258 | 230 | 240 | 228 | 230 | 230 | 226 |

[1] From a report by Prestwood, Vale, Massey, & Welsh (1985).

model (see Figure 4). The fits are considerably better with the two-parameter model and the curvilinear pattern vanishes (see Figure 5). Users of the one-parameter model would be forced to eliminate some of the best items in the test or simply proceed with all of the items and a model that doesn't fit their data. But, in this latter case, the presence of the invariance property would depend upon model robustness. Neither option seems satisfactory. The choice of the better fitting two-parameter model is the decision many researchers would make.

Table 2 highlights the second concern. The example is from a one– and three-parameter model analysis of the 75-item *Maryland Functional Reading Test.* Item fit statistics (above and below 1.0) are reported

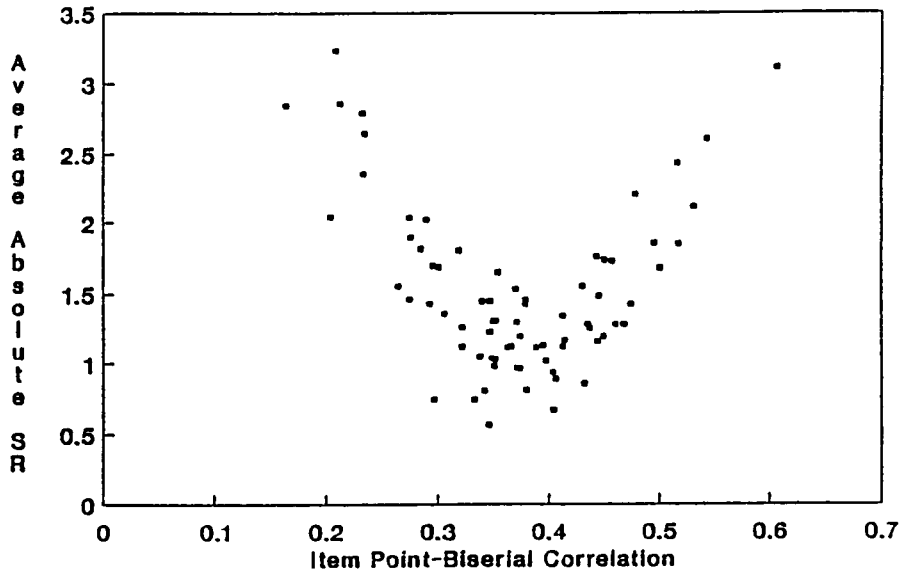


*Figure 4.*–Plot of 1P Average Absolute SRs Against Point-Biserial Correlations



for each of five content areas measured in the test. Items measuring «Main Idea» are fitted less well than items measuring the other four content areas. Deleting items in the «main idea» content area would definitely distort the ability measured by the full test.

In a recent paper, Masters (1988) alerted the measurement field to the fact that *not* all items with high discriminating power are necessarily useful. This is an interesting point, and it is possible that some items may be showing high discrimination for inappropriate reasons. Proper empirical and judgmental item bias investigations can help reduce the problem of misinterpreting high item discriminating power. In fact, when item bias studies are conducted, certain items show up as discriminating in the majority group and non-discriminating in the minority group. The result is known as non-uniform bias and can be detected with two- and three-parameter model item bias studies (see, for example, Hambleton & Rogers, 1989). Also, results reported in the Masters' (1988) paper again highlight the well-known point that valid test development procedures require more than a simplistic look at item statistics. Careful, thoughtful, systematic test development work is needed to insure that test scores are valid.

Some have argued that item discrimination might be modelled by multidimensional IRT models. Still, the advantages and disadvantages of a one-parameter multidimensional Rasch model versus a one-dimensional, two-parameter model would need to be determined. Certainly, in 1994, the answer seems easy. Two-parameter logistic models can handle the variability in item discrimination well, and the application of any multidimensional model, even a simple multidimensional model, is fraught with problems at this time (e.g., parameter estimation, in-

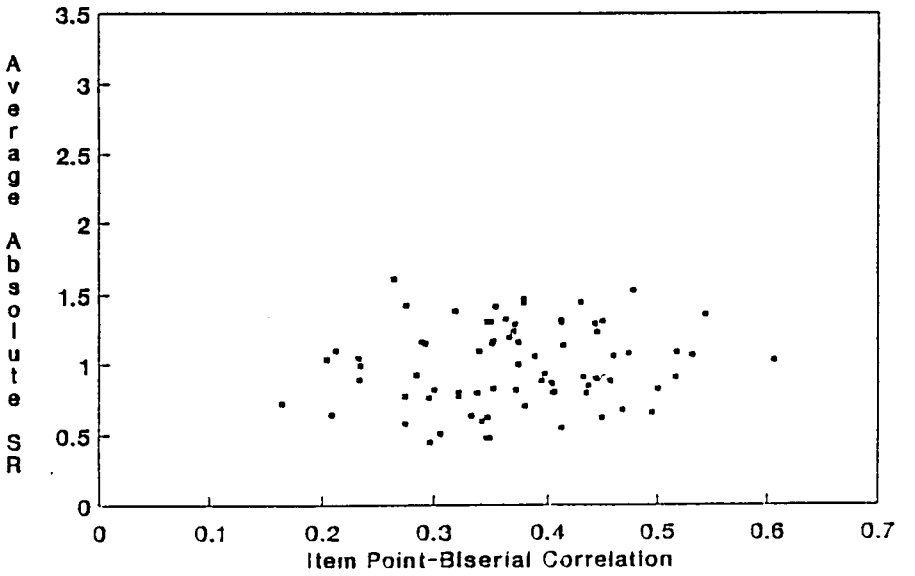


*Figure 5.*–Plot of 2P Average Absolute SRs Against Point-Biserial Correlations

*Table 2*
Association Between Absolute-Valued Standardized Residuals (SRs) and Item Content on the Maryland Functional Reading Test[1]

| | | % of Standardized Residuals[2] | | | |
|---|---|---|---|---|---|
| | | One-Parameter | | Three-Parameter | |
| Content Category | Number of Items | $\|SR\| \leq 1$ (n = 16) | $\|SR\| > 1$ (n = 59) | $\|SR\| \leq 1$ (n = 56) | $\|SR\| > 1$ (n = 19) |
| Following Directions | 17 | 41.2% | 58.8% | 88.2% | 11.8% |
| Locating Information | 17 | 23.5 | 76.5 | 82.4 | 17.6 |
| Main Idea | 12 | 0.0 | 100.0 | 41.7 | 58.3 |
| Using Details | 17 | 11.8 | 88.2 | 76.5 | 23.5 |
| Understanding Forms | 12 | 25.0 | 75.0 | 75.0 | 25.0 |

[1] From a paper by Hambleton, Murray, and Williams (1983).
[2] Twelve (absolute-valued) standardized residuals (SRs) were available on each test item. The average of these 12 SRs for an item was used in the calculations above.

terpretations). Some recent research by Glas (1992) is responsive to concerns about this point. His solution involves fitting a multi-dimensional Rasch model in which the items measuring each dimension in the solution space have their own level of discriminating power. The model fitting begins by first forming subtests of items with similar discriminating powers. But the procedure is very complicated, not ready for wide use at this time, and, interestingly, requires multiple item discrimination values, one for each subtest.

Finally, recently measurement with the Rasch model has been enriched by contributions from several Dutch researchers who have extended the Rasch model by adding an item discrimination parameter (Verstralen & Verhelst, 1991a, 1991b). What makes this work novel is that they have developed a method for using «imputed values.» They found that the addition of an item discrimination parameter improved model fit, preserved the property of specific objectivity, while not complicating model parameter estimation. The important point at this juncture is that item discrimination was found to be a useful addition to the Rasch model. The topic of model parameter estimation will be considered in Section 6.

### 5. *Pseudo-Guessing Parameter*

The inclusion of a pseudo-guessing parameter in an IRT model has caused considerable controversy. For one, as Mellenbergh (1994) has noted, the three-parameter logistic model does not conveniently fit into a general framework of psychometric models. Another problem has to do with the psychological interpretation of the parameter. And, finally, parameter estimation has been difficult.

With respect to interpretation, the parameter is not strictly a «guessing parameter.» Lord (1974) determined this in his evaluative study of item parameter estimates from LOGIST. He coined the term «pseudo-



chance level parameter.» For others, the parameter is simply needed to account for the non-zero item performance of low ability candidates. But numerous data analyses (see, for example, Hambleton & Cook, 1983; Hambleton & Rogers, 1990; Hambleton, Swaminathan, & Rogers, 1991) have highlighted the utility of the *c*-parameter in improving model fit. Hambleton, Swaminathan, and Rogers fitted the one-parameter and three-parameter logistic models to NAEP math items and then considered the (absolute-valued) standardized residuals (i.e., misfit statistics) for items sorted by format (open-ended and multiple-choice) and difficulty. The results, report in Table 3, are clear and compelling. When the fit is good, mean residuals around a value of .80 are expected. The findings are that the three-parameter model generally fits the data well and better than the one-parameter model for the four combinations of item format and level of difficulty. The one-parameter model fit is especially poor when guessing arises with the hard multiple-choice items.

With respect to estimation, Thissen and Wainer (1982) noted very large standard errors for the *c*-parameter estimates.

Fortunately, proper choice of examinee sample and/or the use of Bayesian priors in the estimation procedure (for example, such priors are used in BILOG) can be very helpful in reducing parameter estimation errors (see, for example, de Gruijter, 1984; Swaminathan & Gifford, 1986). Researchers have often settled for a common *c*-parameter value across items. This model is sometimes called a modified three-parameter model, and has resulted in improved model fit over the two-parameter model.

An interesting problem arises with small sample sizes. Do you fit a model with a single difficulty parameter? Lord (1983) showed that when samples sizes are less than 200, a Rasch model with equal scoring weights is better than a two-parameter model with relatively poor estimates of the item discrimination parameters. On the other hand, in a very nice simulation pertaining to optimal item selection in test development, de Gruijter (1986a) showed that even with small samples, if guessing occurs, a model that handles guessing

$$P_i(\theta) = c + (1-c)\left[1 + e^{-Da(\theta - b_i)}\right]^{-1}$$

is preferable to the Rasch model. The poor estimation of *c*-parameters in small samples

*Table 3*
Analysis of the (Absolute-Valued) Standardized Residuals[1]

| Item Difficulty Level | Item Format | Number of Items | 1-p Results $\bar{X}$ | 1-p Results SD | 3-p Results $\bar{X}$ | 3-p Results SD |
|---|---|---|---|---|---|---|
| Hard ($p < .5$) | MC | 70 | 2.73 | 1.55 | .82 | .23 |
| | OE | 54 | 1.64 | .81 | .86 | .28 |
| Easy ($p \geq .5$) | MC | 70 | 1.79 | 1.10 | .90 | .64 |
| | OE | 66 | 1.67 | .72 | .97 | .38 |

[1] Math Booklets No. 1 and 2, 260 Items, 9- and 13-Year-Olds, 1977-78

is addressed in two ways: (1) only a single *c*-parameter is estimated, and (2) a Bayesian prior on the *c*-parameter estimation process can be used.

### 6. *IRT Computer Software Packages*

Solving hundreds and often thousands of non-linear equations simultaneously is a complex numerical analysis problem. This is the problem that BILOG (Mislevy & Bock, 1986), LOGIST (Wingersky, Barton, & Lord, 1982), MULTILOG (Thissen, 1986), and MicroCAT (Assessment Systems Corporation, 1988) have been designed to address. In solving these equations to obtain model parameter estimates and standard errors, it is common for numerical analysts to place constraints on some or all of the parameters, to incorporate prior beliefs about the parameters (e.g., the c-parameter cannot be less than zero, and is unlikely to exceed .30), and to use sensible initial values of the parameters (see, for example, Mislevy & Stocking, 1989).

Several facts about these computer programs are not disputed. First, they take considerably more time to provide two– and three-parameter model estimates than programs providing Rasch model estimates, such as BIGSCALE (Wright, Schultz, & Linacre, 1989) and RIDA (Glas, 1990). Second, LOGIST, BILOG, and others sometimes encounter convergence problems with some items and/or persons (see, for example, Yen, Burket, & Sykes, 1991). These events will occur with more frequency when tests are short, and examinee sample sizes are small, though some estimation methods, such as Bayesian, seem to reduce the number of problems (Swaminathan & Gifford, 1985, 1986). That the problems of convergence, large standard errors, and so on, are generally known to users is due to the substantial amount of research that has been conducted.

Considerably less technical information is available on BICAL, BIGSCALE, MICROSCALE, and other software packages which are used to obtain Rasch model parameter estimates. Even in this considerably simpler case, bias in the unconditional maximum likelihood parameter estimates has been observed (de Gruijter, 1986b, 1990; Divgi, 1986; van den Wollenberg, Wierda, & Jansen, 1988) and the goodness of fit statistics associated with the Rasch model have been challenged (see, for example, Divgi, 1986; Rogers & Hattie, 1987). Rogers and Hattie concluded, «the results reported in this study suggest that the mean-square residual and total-t person and item fit statistics will contribute very little to an investigation of one-parameter model fit and, indeed, if relied on solely, will provide incorrect information about the appropriateness of the model» (p. 56).

Divgi (1986) reported that the percent of items misfitting the Rasch model in the famous *Anchor Test Study* of the middle 1970s (Loret, Seder, Bianchini, & Vale, 1974) increased from 17 %, as reported with the improper statistical tests, to 68 % with what he claimed were the correct statistical tests. Unfortunately, however, the impression remains among many practitioners (based upon the Rentz & Rentz [1979] study) that the Rasch model fits all types of data well. To quote Divgi (1986, p. 284): «It seems safe to conclude that the studies that found the Rasch model satisfactory for multiple-choice tests did so because their methods of analysis were not powerful enough.» Fortunately, improved statistical tests for the Rasch model appear to be on the way (Smith, 1988, 1991).

Our point is not to argue that the two studies by Rogers and Hattie and Divgi invalidate most of the conclusions about Rasch model fit using BICAL. In fact, papers by Henning (1989) and Smith (1988, 1991) have been helpful in understanding and explaining some of the technical problems with various goodness-of-fit measures.



Rather, the point is that the Rasch model estimation and goodness-of-fit procedures, particularly as implemented, in BICAL and its successors, are *not* without controversy themselves. Rasch model research perhaps would be enhanced if more up-to-date information were available about the main programs in use in the U.S. More documentation is available on several programs in use in Europe (see, for example, Glas, 1990; Verhelst et al., 1991).

In a rapidly developing field such as IRT, it is not surprising to see constant updating of computer software packages to respond to new requests (e.g., standard errors), to technical advances (e.g., Bayesian priors), and to correct errors (e.g., goodness of fit statistics). In fact, such developments are constructive and necessary, but these developments make the task of providing timely comments on software packages difficult. Perhaps it is sufficient to report that the main software packages (LOGIST, BILOG, MULTILOG) for obtaining parameters for the two-parameter and three-parameter logistic models and the graded response model work quite well under reasonable conditions of test length and sample size (see, for example, Mislevy & Stocking, 1989; Reise & Yu, 1990; Yen, 1987). Precise numbers needed are impossible to specify because they interact with desired degree of parameter precision, estimation procedures, and examinee ability distribution. For example, a smaller heterogeneous sample is generally preferable to a larger, more homogeneous sample in item parameter estimation with the two- and three-parameter models.

What does it mean to say that a software program works? One requirement is that the estimation process recover known model parameters in simulation studies without bias or standard errors that are inconsistent with the sample sizes. In the most comprehensive study to date, Mislevy and Stocking (1989), have provided the clearest study on the relative strengths and weaknesses of LOGIST and BILOG and more generally technical problems that arise with joint and marginal maximum likelihood estimation and Bayesian estimation, and how they are addressed in the programs. Basically, the recovery of true ability and item parameters was done well by both programs with moderately long tests (n = 45) and large sample sizes. LOGIST was not very successful with a short test (n = 15), a finding which has been reported at other times also (see, for example, Lord, 1974). BILOG, with its use of Bayesian priors is more successful than LOGIST with the shorter test lengths. Other comparative studies by Yen (1987), Hulin, Lissak, and Drasgow (1982), and, more recently, Wingersky (1992), have shown the utility of the LOGIST program under many conditions that occur in practice. Still, it must be said that there is plenty of room for improvement in parameter estimation, and neither program, but especially LOGIST, is intended for the IRT newcomer or for small-scale testing applications.

Mislevy and Stocking (1989) have suggested that many of the problems observed in LOGIST and BILOG are common to other computer programs as well that attempt to solve many non-linear equations simultaneously. de Gruijter (1990), Divgi (1986), and others, for example, have noted that unconditional parameter estimation methods popular with the Rasch model are biased.

Yen, Burket, and Sykes (1991) have recently addressed the problem of non-unique solutions to the likelihood equations for the three-parameter logistic model and offered ways in which the problem can be addressed in practice. Some researchers will no doubt point again to flaws in the three-parameter model. Recall, though, as noted by Yen et al. (1991), that the problems are due to correct guessing on the part of some examinees, *not* due to the three-parameter model. Use of the Rasch model will eliminate the multiple maxima problem but can worsen mo-

del fit. Choose your poison! Fortunately, there are possible ways to identify and handle multiple maxima when they occur. And, other IRT models which provide more psychologically satisfying responses to the guessing problem are under development (see, for example, Goldstein & Wood, 1989).

Some, such as Wright (1984), have argued that the technical problems described above are insurmountable and are best overcome with the use of a «better behaved» Rasch model. But two points might be noted. First, these problems are being identified because of the careful and serious way in which researchers working with these multi-parameter models and procedures are proceeding. Problems such as the handling of omits and «not reached» items which have seriously concerned Lord since as early as 1952 or the serious problems of «item order» and «context effects» on item parameter estimation (Yen, 1980; Zwick, 1991) are either unknown to many users of the Rasch model or unappreciated. These problems and many others are every bit as serious for the validity of Rasch model applications as they are for applications of the multi-parameter models. Second, for many psychometricians, the goal is to find IRT models or other models that can be used to represent their data. Simple but non-fitting models are of little interest.

In sum, many research studies (Harwell & Janosky, 1991; Hulin, Lissak, & Drasgow, 1982; Lord, 1974; Ree, 1979; Skaggs & Stevenson, 1989; Swaminathan & Gifford, 1983; Vale & Gialluca, 1988; and others) have shown that logistic models can recover the true parameters in simulation studies when proper estimation designs (i.e., suitable samples in size and location on the ability scale, test lengths, and priors) are used. Generally, item difficulty parameters can be properly estimated with smaller sample sizes than other item parameters.

### 7. *Successful Applications of Multi-Parameter IRT Models*

A complete accounting of the successful applications of multi-parameter IRT models is beyond the scope of this paper. This topic itself could serve as the basis for a long paper. Readers are referred to Hambleton and Swaminathan (1985), Hambleton, Swaminathan, and Rogers (1991), Lord (1980), and the plethora of articles in the *Journal of Educational Measurement, Applied Psychological Measurement,* and *Applied Measurement in Education,* and other measurement journals.

In view of the serious doubts expressed by Wright (1984) about the utility of multi-parameter IRT models, it seems only necessary to highlight a few prominent examples to counter his objections. Perhaps the most visible IRT multi-parameter application for assessments of interest to policy makers is the *National Assessment of Educational Progress.* Since 1984, scale construction and score reporting have been done with the three-parameter logistic model. This year, too, multi-parameter model extensions to handle polychotomously scored items will be introduced. CTB/McGraw-Hill/Macmillan has been using the three-parameter logistic model in its test development, scale construction, and score reporting of standardized achievement tests. Perhaps the «cadillac» of state assessment systems was in California and this system, until it was discontinued last year, used multi-parameter IRT models and carefully monitored for technical quality by Darrell Bock from the University of Chicago. The *Law School Admissions Test,* the *Graduate Management Admissions Test,* the *Scholastic Aptitude Test,* the *Graduate Record Exam,* and the *Tests of English as a Foreign Language* are a number of other high profile, national or international tests that use multi-parameter IRT models in test design, item bias analyses, score equating, etc.

## 8. *Shortcomings in the Arguments to Support the Rasch Model*

Controversy has followed the Rasch model since its introduction to U.S. measurement specialists in a paper by Wright (1968). Goldstein (1981), Whitely and Dawis (1974), Divgi (1986), McDonald (1985, 1989), are some of the best known measurement experts to challenge the validity of the Rasch model. Shortcomings in the arguments advanced by Wright (1968, 1977, 1984) for the Rasch model include:

1. Placing too much importance on sufficient statistics;
2. Failing to satisfactorily account for variability in item discriminating power;
3. Failing to satisfactorily account for the non-zero item performance of low-performing examinees;
4. Failing to successfully attend to problems such as «omits,» «not reached,» item context effects, item order effects, etc., which impact on model utility and validity;
5. Providing incorrect results on model fit because of the use of inappropriate statistics.
6. Failing to acknowledge (or recognize) that the properties of item and ability invariance are characteristics of *all* IRT models which fit the data to which they are applied;
7. Recommending the deletion of misfitting items (rather than adding model parameters) which can distort the trait underlying the test;
8. Failing to recognize serious problems with the Rasch model in important applications such as vertical equating (see, for example, Schulz, Perlman, Rice, & Wright, 1992); and
9. Overstating problems with current multi-parameter IRT software packages.

Each of the nine points has been addressed directly or indirectly in one or more of the previous sections. It is important to add, however, that, though there are many shortcomings in the arguments used to support the Rasch model, the model (and extensions) itself has an important role to play in many testing applications. A review of Wilson (1992) and the *JEM, APM,* and *AME,* to name just four major references, will turn up many important technical advances and applications of the Rasch model, including use in (1) national standardized achievement tests, (2) item banking projects, and (3) many state testing and credentialing exam programs. Perhaps the best argument to support its use is that it sometimes provides a close fit to actual test data. Also, the model may have special utility in situations where samples are modest in size and the need for high precision in ability estimates is not great (such as in some school applications).

## 9. *Promising Future of IRT Models*

While IRT provides solutions to many testing problems that previously were unsolved or solved in a less than satisfactory way, it is not a magic wand which can be waved to overcome deficiencies such as poorly written test items and poor test designs. In the hands of careful test developers, however, IRT models, the Rasch model *and* multi-parameter models, and IRT methods can become powerful tools in the design and construction of sound educational and psychological instruments, and in reporting and interpreting test results. But it is highly unlikely that any single family of IRT models will be able to meet the challenges and demands on measurement practices in the coming decade.

Research on IRT models and their applications is being carried out at a phenomenal rate (see Thissen & Steinberg, 1986, and Mellenbergh, 1994, for taxonomies of models; and van der Linden & Hambleton, in press). Entire issues of several journals have been devoted to developments in IRT. For the future, two directions for research appear to be especially important: polytomous unidimensional response models and both

dichotomous and polytomous multidimensional response models. Research in both directions is well under way (Masters & Wright, 1984; McDonald, 1989; van der Linden & Hambleton, in press). With the growing interest in «authentic measurement,» special attention must be given to IRT models that can handle polytomous scoring, since authentic measurement is linked to performance testing and non-dichotomous scoring of examinee performance.

Multidimensional IRT models were introduced originally by Lord and Novick (1968), Samejima (1974), and, more recently, by Embretson (1984), Fischer and Seliger (in press), and McDonald (1989). Multidimensional models offer the prospect of better fitting current test data and providing multidimensional representations of both items and examinee abilities. It remains to be seen whether parameters for these multidimensional models can be properly estimated, and whether multidimensional representations of items and examinees are useful to practitioners.

Goldstein and Wood (1989) have argued for more IRT model-building in the future, but feel that more attention should be given to placing IRT models within an explicit linear modeling framework. Advantages, according to Goldstein and Wood, include model parameters that are simpler to understand, easier to estimate, and that have well-known statistical properties.

Three other areas are likely to draw special attention from educators and psychologists in the coming years. First, large-scale state, national, and international assessments are attracting considerable attention, and will continue to do so for the foreseeable future (see, for example, the Third International Mathematics and Science Study involving over 60 countries). Item response models are being used at the all-important reporting stages in these assessments. It will be interesting to see what technical controversies arise from this type of application (see, for example, Zwick, 1991). One feature that plays an important role in reporting is the ICC. Are ICCs invariant to the nature and amounts of instruction? The assumption is that ICCs are invariant, but substantially more research is needed to establish this point.

Second, cognitive psychologists such as Embretson (1984) are interested in using IRT models to link examinee task performance to their ability through complex models that attempt to estimate parameters for the cognitive components that are needed to complete the tasks. This line of research is also consistent with Goldstein and Wood's goal to see the construction of more meaningful psychological models to help explain examinee test performance. See, for example, recent work by Mislevy and Verhelst (1990), and Sheehan and Mislevy (1990), which is along these general lines.

Third, educators and psychologists are making the argument for considerably more use of test scores than simply rank ordering of examinees on their abilities or determining whether they have met a particular achievement level or standard. Diagnostic information is becoming increasingly important to users of test scores. *Inappropriateness measurement*, developed by M. Levine and F. Drasgow (see, for example, Drasgow, Levine, & McLaughlin, 1987), which utilizes IRT models, provides a framework for identifying aberrant responses of examinees and special groups of examinees on individual and groups of items. Such information can be helpful in successful diagnostic work. More use of IRT models in providing diagnostic information can be anticipated in the coming years.

### *Conclusions*

There has been a considerable amount of debate in the measurement literature about the merits of various IRT models, estimation procedures and designs, and the specific steps for equating tests, identifying

DIF, and constructing tests. There are even researchers who have some serious reservations about the whole IRT direction in measurement (see, for example, Hoover, 1992).

In 1994, it is nonsensical to argue that the Rasch model (and its extensions) are the *only* models providing a sound technical basis for constructing tests and evaluating test scores. The testing field has many outstanding examples of successful applications of multi-parameter IRT models. And, technical advances on many fronts (see, for example, Suen & Lee, 1992) will make future applications even more successful. The argument in this paper is that there is room for many IRT models, but only models which fit data, and have parameters which can be estimated well will be of interest. One important example of model fit was presented earlier in the paper and the results were poor for the one-parameter model. But, our point was not to reject the Rasch model for every application. Our only purpose was to highlight the need for, and the acceptability of, other models. In fact, we laud the excellent work of many one-parameter model researchers and acknowledge the role their research has played in the technical foundations of IRT. We note, too, in passing, that many of the improvements in the Rasch model methods and procedures have been due to constructive suggestions from IRT researchers not exclusively interested in the Rasch model. These are constructive actions on the part of many researchers to make IRT models more useful in practice.

In a way, the long-standing debate between one-parameter and three-parameter model advocates has been counter-productive. There is a need for both psychometric models. In fact, there is substantial evidence to suggest that both models have been used very successfully in many types of testing applications. On the negative side, the focus of attention on the one– and three-parameter logistic models has meant that there is less familiarity on the part of practitioners with many new promising directions for psychometric model-building. Goldstein and Wood (1989), Garcia-Perez and Frary (1991), Mellenbergh (1994), and McDonald (1982, 1989) have all suggested new models or classes of models and, of course, extensions of logistic models to handle polychotomous as well as multi-dimensional data are well under way (see, for example, Embretson, in press; Fischer&Seliger, in pres Reckase, in press).

The main points of this paper are that both the Rasch model and the more general (i.e., multi-parameter) logistic models have important roles to play in the field of testing. But model fit is essential, and it is far better to find models that fit the test data than to discard data simply to fit the Rasch model. Educational and psychological testing practices are changing. New item formats and scoring schema are being introduced to measure higher-order thinking skills. These new measures should not be narrowed to enable a simple (or even extended) Rasch model to fit the resulting data. Curriculum specialists would be rightly shocked. At the same time, to argue that variations in item discrimination is really «multidimensionality» in the test data and should be handled by multidimensional Rasch models is to disregard 80 years of measurement experience. To advocate the use of multidimensional IRT models that are still in their early developmental stage seems highly inappropriate and fails to recognize the success of unidimensional multi-parameter models in fitting educational and psychological data.

*References*


Assessment Systems Corporation. (1988). *User's manual for the MicroCAT testing system* (Version 3). St. Paul, MN: Author.

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 1-23). Vancouver, BC: Educational Research Institute of British Columbia.

Birnbaum, A. (1957). *Efficient design and use of tests of ability for various decision-making problems* (Series Report No. 58-16, Project No. 7755-23). Randolph Air Force Base, TX: USAF School of Aviation Medicine.

Birnbaum, A. (1958a). *On the estimation of mental ability* (Series Report No. 15, Project No. 7755-23). Randolph Air Force Base, TX: USAF School of Aviation Medicine.

Birnbaum, A. (1958b). *Further considerations of efficiency in tests of a mental ability* (Series Report No. 17, Project No. 7755-23). Randolph Air Force Base, TX: USAF School of Aviation Medicine.

De Gruijter, D. N. M. (1984). A comment on «Some standard errors in item response theory.» *Psychometrika, 49,* 269-272.

De Gruijter, D. N. M. (1986a). Small N does not always justify Rasch model. *Applied Psychological Measurement, 10(2),* 187-194.

De Gruijter, D. N. M. (1986b). The bias of an approximation to the unconditional maximum likelihood procedure for the Rasch model. *Kwantitatieve Methoden, 20,* 49-55.

De Gruijter, D. N. M. (1990). A note on the bias of UCON item parameter estimation in the Rasch model. *Journal of Educational Measurement, 27(3),* 285-288.

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.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11(1),* 59-79.

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

Embretson, S. E. (in press). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory.* New York: Springer-Verlag.

Fischer, G. H. (1974). *Ein fuhrung in die theorie psychologischer tests* (Introduction to the theory of psychological tests). Bern: Huber.

Fischer, G. H., & Seliger, E. (in press). Multidimensional linear logistic models for change. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory.* New York: Springer-Verlag.

Garcia-Perez, M. A., & Frary, R. B. (1991). Finite state polynomic item characteristic curves. *British Journal of Mathematical and Statistical Psychology, 44,* 45-73.

Glas, C. A. W. (1990). *RIDA: Rasch incomplete design analysis.* Arnhem, The Netherlands: National Institute for Educational Measurement.

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

Goldstein, H. (1981). Limitations of the Rasch model for educational assessment. In C. Lacey & D. Lawton (Eds.), *Issues in evaluation and accountability* (pp. 172-188). London: Methuen.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42,* 139-167.

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

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

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

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York: Macmillan.

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. Weiss (Ed.), *New horizons in testing* (pp. 31-49). New York: Academic Press.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12(3),* 38-47.

Hambleton, R. K., Murray, L. N., & Williams, P. (1983). *Fitting item response models to the Maryland Functional Reading Tests* (Laboratory of Psychometric and Evaluative Research Report No. 139). Amherst, MA: University of Massachusetts, School of Education.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2(4),* 313-334.

Hambleton, R. K., & Rogers, H. J. (1990). Using item response models in educational assessments. In W. H. Schreiber & K. Ingekamp (Eds.), *International developments in large-scale assessment* (pp. 155-184). Windsor, UK: NFER-Nelson.

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

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

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15,* 279-291.

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(1),* 91-97.

Hoover, H. D. (1992, April). *Some shortcomings of item response models.* Paper presented at the meeting of AERA, San Francisco.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6,* 249-260.

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

Lord, F. M. (1953a). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18,* 57-75.

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

Lord, F. M. (1970). Item characteristic curves estimated without knowledge of their mathematical form - a confrontation of Birnbaum's logistic model. *Psychometrika, 35,* 43-50.

Lord, F. M. (1974). Estimation of latent-ability and item parameters when there are omitted responses. *Psychometrika, 39,* 29-51.

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

Lord, F. M. (1983). Small N justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 51-61). New York: Academic Press.

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

Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. A. (1974). *Anchor test study* (Final Report). Princeton, NJ: Educational Testing Service.

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

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25,* 15-29.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49,* 529-544.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6(4),* 379-396.

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Lawrence Erlbaum Associates.

McDonald, R. P. (1989). Future directions for

item response theory. *International Journal of Educational Research, 13(2),* 205-220.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115(2),* 300-307.

Mislevy, R. J., & Bock, R. D. (1986). *BILOG: Maximum likelihood item analysis and test scoring with logistic models.* Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13(1),* 57-75.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55(2),* 195-215.

Prestwood, J. S., Vale, C. D., Massey, R. H., & Welsh, J. R. (1985). *Armed Services Vocational Aptitude Battery: Development of an adaptive item pool* (AFHRL-TR-85-19). Brooks Air Force Base, TX: Manpower and Personnel Division.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks Paedagogiske Institut.

Reckase, M. (in press). A linear logistic model for cichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory.* New York: Springer-Verlag.

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

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27(2),* 133-144.

Rentz, R. R., & Rentz, C. C. (1979). Does the Rasch model really work? *Measurement in Education, 10(2),* 1-8.

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(1),* 47-57.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39,* 111-121.

Schulz, E. M., Perlman, C., Rice, W. K., & Wright, B. D. (1992). Vertically equating reading tests: An example from Chicago public schools. In M. Wilson (Ed.), *Objective measurement: Theory into practice, Volume 1* (pp. 138-156). Norwood, NJ: Ablex Publishing Co.

Sheehan, K., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement, 27*(3), 255-272.

Skaggs, G., & Stevenson, J. (1989). A comparison of pseudo-Bayesian and joint maximum likelihood procedures for estimating parameters in the three-parameter IRT model. *Applied Psychological Measurement, 13,* 391-402.

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

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement, 51,* 541-565.

Suen, H. K., & Lee, P. S. C. (1992). Constraint optimization: A perspective of IRT parameter estimation. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (pp. 289-300). Norwood, NJ: Ablex Publishing.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. Weiss (Ed.), *New horizons in testing* (pp. 13-30). New York: Academic Press.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50,* 349-364.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51,* 589-601.

Thissen, D. M. (1986). *MULTILOG: Item analysis and scoring with multiple category response models* (Version 5). Mooresville, IN: Scientific Software.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567-577.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397-412.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11,* 1-13.

Vale, C. D., & Gialluca, K. A. (1988). Evaluation of the efficiency of item calibration. *Applied Psychological Measurement, 12(1),* 53-67.

Van den Wollenberg, A. L., Wierda, F. W., & Jansen, P. G. W. (1988). Consistency of Rasch model parameter estimtion: A simulation study. *Applied Psychological Measurement, 12,* 307-313.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (in press). *Handbook of item response theory.* New York: Springer-Verlag.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1991). *OPLM* (The One Parameter Logistic Model). Arnhem, The Netherlands: CITO.

Verstralen, H. H. F. M., & Verhelst, N. D. (1991a). *The sample strategy of a test information function in computerized test design* (Measurement and Research Department Reports, 91-6). Arnhem, The Netherlands: CITO.

Verstralen, H. H. F. M., & Verhelst, N. D. (1991b). *Decision accuracy in IRT model* (Measurement and Research Department Reports 91-7). Arnhem, The Netherlands: CITO.

Whitely, S. E., & Dawis, R. V. (1974). The nature of objectivity with the Rasch model. *Journal of Educational Measurement, 11(3),* 163-178.

Wilson, M. (Ed.). (1992). *Objective measurement: Theory into practice.* Norwood, NJ: Ablex Publishing Co.

Wingersky, M. S. (1992). *Significant improvements to LOGIST* (Research Bulletin 92-22). Princeton, NJ: Educational Testing Service.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide.* Princeton, NJ: Educational Testing Service.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems* (pp. 85-101). Princeton, NJ: Educational Test Service.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14(2),* 97-116.

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review, 1,* 281-288.

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

Wright, B. D., Schulz, M., & Linacre, J. M. (1989). BIGSCALE: Rasch analysis computer program. Chicago: MESA Press.

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

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.

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

Yen, W. M., Burket, G. R., & Sykes, R. C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model. *Psychometrika, 56(1),* 39-54.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10,* 10-16.

*Accepted: June, 30, 1994*