White Paper

# Rasch Versus Birnbaum: New Arguments in an Old Debate

by
John Richard Bergan, Ph.D.

**Assessment Technology, Incorporated**
6700 E. Speedway Boulevard
Tucson, Arizona  85710
Phone: 520.323.9033 • Fax: 520.323.9139

*Assessment Technology, Incorporated, Publishers*
*Tucson, Arizona, U.S.A.*
*Printed in the United States of America.*
*V4-121013*

# Rasch Versus Birnbaum: New Arguments in an Old Debate

*By John Richard Bergan, Ph.D.*
*Assessment Technology, Incorporated*

## Table of Contents

# I. Acknowledgements

I want to thank Dave Thissen for his insightful comments regarding this manuscript. In addition, I would like to thank John Robert Bergan, Kathryn Bergan, and Sarah Callahan for their many helpful observations regarding the manuscript.

John R. Bergan, Ph.D.
President, Assessment Technology, Incorporated

# II. Introduction

Over 20 years ago Drs. Benjamin Wright and Ronald Hambleton debated the validity and utility of the Rasch model developed by Georg Rasch and the three-parameter logistic model developed by Allan Birnbaum. The debate took place in the context of an extended controversy regarding how student academic achievement and other social science variables should be measured. Although over 20 years have passed since the debate, the controversy has never been resolved. The provision of valid social science measures is a topic of great concern to educators and to those organizations providing assessment services in education and other fields. Each year our company, Assessment Technology Incorporated (ATI), produces tests administered to several million students across the nation. This document outlines ATI's approach to measurement as it relates to the Rasch and Birnbaum debate. In addition, the document relates the ATI approach to the arguments set forth by Dr. Wright in the landmark debate with Dr. Hambleton. The current paper is the second ATI paper addressing the Rasch and Birnbaum models. A link to the first paper follows for the reader's convenience.

http://ati-online.com/pdfs/researchK12/AlternativeIRTModels.pdf

# III. ATI's Approach to the Measurement of Student Achievement

ATI approaches education, in general, and educational measurement, more specifically, from an empirical perspective. In fact, ATI chose to name its flagship application after Galileo to underscore ATI's commitment to a scientific approach to education. This means that ATI uses mathematical models to test hypotheses that can be supported or not supported based on empirical evidence. In ATI's work, one model is not favored over another apriori. For example, ATI does not inherently favor the three-parameter model over the Rasch model or use the three-parameter model simply because it provides a means to describe data. In ATI's work, a mathematical model chosen to represent observable data is seen as a theory to be tested empirically. The testing process generally requires the examination of multiple models representing different hypotheses. For example, analysis of the three-parameter logistic model calls for the examination of multiple models. The selection of a model to represent the data is based on two factors: parsimony and goodness of fit. The law of parsimony holds that given alternate equally acceptable models, the simplest model, i.e., the one with the fewest estimated parameters, should be chosen to represent the data. Goodness of fit is addressed by determining whether or not a given model improves significantly on the fit to the data of an alternative model. Goodness of fit is often assessed using the chi-squared statistic. For example, the chi-squared approach is often used in assessing the fit of loglinear models, latent-class models, structural equation models, Hierarchical Linear Models, and Item Response Theory (IRT) models.

In the case of IRT, the models to be tested hold that observed response patterns to a set of items comprising an assessment can be explained by a latent variable typically referred to as ability or proficiency and one or more item parameters (Bock and Lieberman, 1970). When the number of possible patterns is small, the fit of an IRT model to the data can be assessed directly. When the number of patterns is large, direct assessment of the fit of the model to the data is not practical. However, Bock and Aitkin (1981) demonstrated that the difference between minus twice the log of the likelihood for one model and minus twice the log of the likelihood for a second model is distributed as chi-square. They showed that the difference chi-square can be referred to the chi-square distribution to provide a test of the relative fit of the models under examination to the data.

## IV. Wright's Argument Favoring the Rasch Model

Dr. Wright takes the position that the Rasch model is the only valid approach to measurement. From ATI's point of view, the fundamental flaw in Wright's argument is that he does not offer an empirical test that could provide results that would either support or fail to support his position. Wright's argument lacks the power that scientific hypothesis testing provides for resolving questions based on empirical evidence. In the absence of an empirical test, there is no apparent way to resolve the argument. An unfortunate consequence of this state of affairs is that the Rasch-Birnbaum controversy has lingered for decades without resolution. The paragraphs that follow address critical components in the controversy covered in the Wright-Hambleton debate and discuss the impact on the argument of a commitment to empiricism and advances in technology occurring over the last 20 years.

### A. Specific Objectivity

Wright argues in the debate with Hambleton that the definition of a measure requires that in "science, engineering, business and cooking, you need measures which have this simple essential property: one more is always the same amount, like the inches on this carpenter's ruler I am using for a pointer. To get that result, that kind of numbers, you need to use the additive construction of the Rasch model." The Rasch model requires a kind of objectivity called specific objectivity, which is not provided by other models. In the Rasch model, the probability of a correct response for a given person is a function of the ratio of the ability of the person to the difficulty of the item. When one person's ability is compared to the ability of a second person, the item parameters cancel leaving an item-free comparison of their abilities. From the Rasch perspective, specific objectivity is an essential ingredient of a valid measure. Specific objectivity guarantees that two people will never have the same number-right score and a different ability score.

### B. Pattern Scoring and the Number-Right Score as a Sufficient Statistic

In the Rasch model, the number-right score is a sufficient statistic that contains all the information available in the observable data regarding the unobservable trait (e.g., ability or proficiency) being measured. By contrast, in the Birnbaum model, the number-right score is not a sufficient statistic containing all of the information available regarding the latent trait. In the case of the two-parameter model, the sufficient statistic for the Birnbaum model requires the inclusion of the discrimination parameter for each item. The inclusion of the discrimination parameters is linked to pattern scoring, which creates the possibility that two examinees could have the same number-right score and different ability scores. This state of affairs is at the heart of Wright's objection to the Birnbaum approach.

### C. The Summed Scoring Alternative

Advances in scoring technology have completely overcome Wright's principal objection regarding the relationship between the number-right and the Birnbaum ability score. This scoring technology, called summed scoring, avoids pattern scoring even in cases involving the 3 PL model. In the 1980s, Lord and Wingersky (1984) described a simple recursive algorithm for computing an ability score that was a direct function of the number-right score for dichotomously scored items. In the 1990s Thissen and his colleagues extended the procedure to accommodate polytomous items (Thissen & Wainer, 2001). ATI wrote code to implement summed scoring on a large scale. Thissen has implemented the ATI code in his research and ATI has implemented summed scoring for assessments containing dichotomous and polytomous items in Galileo® Online for over a decade. When summed scoring is used, it can never be the case that two examinees will have the same number right score and different ability scores.

Note that summed scoring does not overcome the objection that the Birnbaum model does not support specific objectivity. The specific objectivity requirement is philosophically based, not empirically based. Thus, there will be those who believe the requirement is essential and those who do not believe it is essential. Supporters of the logistic model assume for each item *i* the existence of an unobserved variable $Y_i$ that is linearly related to the latent variable with some constant variance including measurement error. The idea is that that $Y_i$ is correlated with the latent variable (ability) and that $Y_i$ also includes measurement error. One of the benefits of the logistic model is that it accommodates more parameters than the Rasch model. As a consequence, the logistic model may provide more information about the student's response than the Rasch model. However, it is worth noting that the logistic model does not require the inclusion of more parameters than the Rasch model. The inclusion of specific parameters in the logistic model can be justified through an empirical test.

### D. Complexity of Parameter Estimation

Wright's second major objection to the Birnbaum model involves the complexity of parameter estimation. The inclusion of the discrimination parameters and the pseudo guessing parameters makes the numerical analysis procedures required to estimate the latent trait and each of the included item parameters more complicated than is the case under the Rasch model. Given the computer technology available 20 years ago, Wright was able to argue that parameter estimation for the Birnbaum model was difficult and time consuming, which provided a significant reason for favoring the Rasch approach. However, given advances in computer technology, Wright's argument with respect to computational complexity is no longer compelling. For example, within the last year, ATI has delivered over 4,800,000 scores for tests involving a variety of complex IRT models. That would not have been possible with the technology available 20 years ago. Indeed ATI's test scoring capacity has expanded 20 fold over an eight-year period and continues to expand as technology continues to advance.

### E. The Rasch Model and the One-Parameter Model

Wright's third objection to Birnbaum centered on the distinction between the Rasch model and the one-parameter logistic model. Wright states that even "the arithmetical trick of making parameters 'a' and 'c' disappear, so that the Birnbaum model looks like Rasch doesn't make Birnbaum in spirit, purpose or function equivalent to Rasch." This statement ignores that fact that the one-parameter Birnbaum model overcomes both of the shortcomings identified by Wright. In the case of the one-parameter model, the number-right score is a sufficient statistic containing all of the information available regarding the latent trait being measured. Consequently, pattern scoring is avoided. Moreover, the number of parameters to be estimated is small as is the case for the Rasch model.

### F. Crossing Trace Lines

Wright's fourth objection to Birnbaum involves the fact that in many cases the trace lines for two or more items may cross. Wright regards crossing trace lines as evidence of item bias. The Rasch model does not permit trace lines to cross. If item *I* is more difficult than item *j*, it is assumed to be more difficult across the ability range. That is, the probability of passing item *I* will be greater than or equal to the probability of passing item *j* at any point on the ability distribution.

Trace lines do not cross in the special case in which the skill assessed by one item is prerequisite to the skill assessed by a second item. Trace lines can be expected to cross in the many cases in which the skills measured by two or more items are not ordered in a prerequisite fashion. The lack of a prerequisite relation certainly does not imply item bias. Indeed, it is safe to say that only a minority of narrowly defined skill sets reflect a prerequisite relationship. Although Wright fails to provide a statistical test that could be used to address the prerequisite hypothesis, there does exist a large literature indicating that it is possible to test the assumption of a prerequisite relationship. In ATI's view science can be advanced by empirical tests of prerequisite relations. In fact, much of my own research has been related to the crossing-trace-lines issue.

The notion of skills reflecting an invariant order was advanced in the 1940s by Louis Guttman (1944). Guttman did not provide a statistical test for his assumptions regarding invariant order. However, Leo Goodman (1974a, 1974b) introduced quasi-Independence models and latent-class models that provided a test for the hypothesis that a set of skills formed an ordered sequence. In the 1980s, Bergan and Stone introduced latent-class models that could test the hypothesis that skills were ordered in a prerequisite fashion and latent-class models assessing skills reflecting a Piagetian transition state between non-mastery and skill mastery (Bergan & Stone, 1985). The data from the study was shared with Dr. David Thissen. Thissen and Steinberg (1988) used the data to test the hypothesis that skills were ordered in a prerequisite fashion using a one-parameter item response model.

Failure to recognize and test the assumption that skills are related in a prerequisite fashion can lead to seriously misleading assumptions regarding cognitive functioning. For example, a graduate student at the Massachusetts Institute of Technology asked his 5-year-old daughter to add 75 and 26. She was unable to solve the problem because she had not been taught the carrying operation. However, when asked to add 75 cents and 26 cents, she responded: "Three quarters, four quarters, and a penny makes a dollar one." The child changed the way she represented the problem. She replaced one set of rules with another in order to reach a solution. This is precisely the kind of circumstance that can produce crossing trace lines. The fields of math, science, and English language arts as well as other disciplines are filled with examples in which students apply diverse sets of procedures to solve problems. To assume that all items are ordered in a prerequisite fashion, as the Rasch model does, eliminates useful information about cognitive functioning from assessment results.

### G. Item-Type Limitations

Wright presents a fifth objection to the Birnbaum model that is related to item types. Wright indicates that the Birnbaum model applies almost exclusively to dichotomous items. In fact, logistic IRT models are available for the full range of models including dichotomous as well as polytomous items. Categorical models involving nominal data are also available. As indicated earlier, ATI has been scoring assessments containing both dichotomous and polytomous items in Galileo® for over a decade.

## V. Consequences of Using the Rasch Model When the Model is Not Supported Empirically

While ATI does not find Dr. Wright's arguments compelling, the elegance of the Rasch model is appreciated. The Rasch model provides a highly parsimonious explanation of the data. As a consequence, it is preferred to represent the data in those cases in which alternative less parsimonious models do not improve significantly on the fit of the model to the data. In addition, it affords a useful test of the hypothesis that skills are related in a prerequisite fashion.

What are the consequences of using the Rasch model to represent the data when that decision is not supported empirically? When summed scoring is used, there will be no effect on the measurement of ability. The correlation between a test scored with the Rasch model and a test scored with the three parameter logistic model will be 1.0. However, there will be an effect on the use of the data to guide instruction. IRT models provide estimates of the probability that a student of any given ability will correctly perform items that are indicators of that ability. ATI uses that information to recommend next instructional steps. For example, ATI uses these probabilities to identify intervention groups and recommend what to teach next to reduce student risk of not meeting standards on statewide assessments. Information on the probability of performing items correctly can also be used to evaluate curriculum. For example, suppose that a pretest indicates that on average students receiving instruction in a particular curriculum are estimated to have mastered 80 percent of the skills being taught before instruction begins. Let's assume that after instruction, the estimate only increases to 85 percent. These finding would suggest that the skills targeted for instruction were not well matched to mastery levels of the students. If empirical evidence favors the three-parameter model over the Rasch model, then use of the Rasch model is a case of a model that is misspecified. The mastery probabilities obtained from the model will yield misleading information.

## VI. A Look Ahead: Item Bifactor Analysis

Scientific progress has continued since 1992 when Dr. Wright put forth his arguments favoring the Rasch model over the Birnbaum model. It is possible, in fact likely, that advances begun decades ago and continuing since the 1992 debate will make the interesting controversy between the Rasch and Birnbaum models moot. For example, it has been well-known for decades that achievement tests are often multi-dimensional, not unidimensional. Until recently, multi-dimensional IRT models were not practical to implement. In fact, just a few years ago, Dr. Michael Edwards did a dissertation involving multi-dimensional IRT. Running a single multi-dimensional factor analysis took him as long as 24 hours. However, in recent years, Li Cai and his colleagues at UCLA (Cai, Seung, & Hansen, 2011) have developed new techniques for the implementation of item bifactor analysis that make it possible to run multidimensional IRT models in seconds even with large data sets. Item bifactor analysis is a special case of confirmatory multidimensional item response theory modeling. Item bifactor analysis provides information about the dimensionality of the measuring instrument under examination. It also provides strategies for scaling individual differences and new approaches to computerized adapting testing. ATI has been experimenting with bifactor models for over a year and plan to provide bifactor technology to clients in the near future.

## VII. Text References

Bergan, J. R., & Stone, C.A. (1985) Latent-class models for knowledge domains. *Psychological Bulletin, 98*, 166-184.

Bock, R. D., & Lieberman, M. (1970). Fitting a Response Model for N Dichotomously Scored Items. *Psychometrika*, *35*, 179-197.

Bock, R. D., & Aitken, M. (1981). Marginal Maximum Likelihood Estimation of Parameters: An application of the EM algorithm. *Psychometrika*, *46*, 446-459.

Cai, L, , Seung, J., & Hansen, M. (2011). Generalized Full-Information Item Bifactor Analysis, *Psychological Methods,* 16*, 221-248.*

Goodman, L. A. (1974a). The analysis of systems of quantitative variables when some of the variables are unobservable: Part 1. A modified latent structure approach*. American Journal of Sociology, 79, 179-259.*

Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 167-171

Guttman, L.A. (1944). A basis for scaling qualitative data. *American Sociological Review*, *91*, 139-150.

Wright, B. (1992): IRT in the 1990s: Which Models Work Best? 3PL or Rasch? *Ben Wright's opening remarks in his invited debate with Ron Hambleton, Session 11.05, AERA Annual Meeting 1992.*

Lord, F. M. & Wingersky, M., S. (1984). Comparison of IRT true score and equipercentile “equatings”. *Applied Psychological Measurement,* 8*, 453-461.*

Thissen, D., & Steinberg, L., (1986). Data analysis using item response theory. *Psychology Bulletin, 104,* 385-395.

Wright, B. (1992): IRT in the 1990s: Which Models Work Best? 3PL or Rasch? *Ben Wright's opening remarks in his invited debate with Ron Hambleton, Session 11.05, AERA Annual Meeting 1992.*