Journal of Applied Measurement Abstracts

Volume 6, Number 1 (2005)

Rasch Analysis of Inattentive, Hyperactive and Impulsive Behavior in Young Children and the Link with Academic Achievement

Christine Merrell
Peter Tymms
University of Durham

The Attention Deficit Hyperactivity Disorder (ADHD) criteria from the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders were used to assess a large sample of children at the end of their first year at school in England. These data were explored using Rasch measurement and the measures for the items together with their frequencies are reported. The data were further analyzed in three ways:
a) The results were compared with a previous similar analysis of college students.
b) A principal components analysis of the item residuals from the Rasch analysis was conducted.
c) The measures were linked to reading and mathematics attainment assessed at three different time points.
The exploration supported previous work and theoretical positions, and in doing so raised issues about the appropriateness of the use of the criteria across all ages. It also suggested that one of the currently recognized ADHD sub-types could be further sub-divided into verbal and physical hyperactivity. The links to academic achievement raised questions about the integrity of the currently recognized ADHD sub-types and the paper calls for further investigations.

Measuring Statistical Literacy

Rosemary Callingham
University of New England
Jane M. Watson
University of Tasmania

This study considers the measurement of Statistical Literacy understanding that goes beyond the basic chance and data skills and knowledge in the mathematics curriculum. This understanding requires application of mathematical skills in a range of contextual situations and draws on aspects of statistics, such as variation and inference, which may not be explicit in the school curriculum. The study reports the outcomes from tests of Statistical Literacy given to 673 students from Grades 5 to 10. It confirms the nature and structure of a previously identified construct of Statistical Literacy and proposes three subgroups of items that address aspects of Statistical Literacy that might usefully be measured by classroom teachers.

Expected Linking Error Resulting from Item Parameter Drift among the Common Items on Rasch Calibrated Tests

G. Edward Miller
Texas Education Agency
Paul Randall Gesn
PRO-ED, Inc.
Ourania Rotou
Educational Testing Service

In state assessment programs that employ Rasch-based common item linking procedures, the linking constant is usually estimated with only those common items not identified as exhibiting item difficulty parameter drift. Since state assessments typically contain a fixed number of items, an item classified as exhibiting parameter drift during the linking process remains on the exam as a scorable item even if it is removed from the common item set. Under the assumption that item parameter drift has occurred for one or more of the common items, the expected effect of including or excluding the "affected" item(s) in the estimation of the linking constant is derived in this article. If the item parameter drift is due solely to factors not associated with a change in examinee achievement, no linking error will (be expected to) occur given that the linking constant is estimated only with the items not identified as "affected"; linking error will (be expected to) occur if the linking constant is estimated with all common items. However, if the item parameter drift is due solely to change in examinee achievement, the opposite is true: no linking error will (be expected to) occur if the linking constant is estimated with all common items; linking error will (be expected to) occur if the linking constant is estimated only with the items not identified as "affected".

Measuring College Sailing Teams Ability: An Application of the Many-Facet Rasch Model to Ordinal Data

William Steve Lang
Judy R. Wilkerson
University of South Florida St. Petersburg

For those who look at typical approaches to sports ranking, sailing seems an almost impossible challenge, thereby making the evidence supporting Rasch measurement in this study even more intriguing. This article describes our application of MFRM and the results from our analysis of one year's data from the North American college sailing competitions. We discuss the following issues for inclusion in the development of a Rasch model useful to college sailing team ability:
1. The level of data quality (as described by Stevens, 1946)
2. The connectedness of the contests
3. Empty cells (missing data)
4. Interpreting outliers, trends, or unusual results
5. Judges' or polls' bias
Our results indicate the utility and precision of MFRM as a tool generally appropriate for ordinal ranking applications and sailing ability specifically.

On the Lack of Comonotonicity between Likert Scores and Rasch-Based Measures

Lucio Bertoli-Barsotti
University of Bergamo, Italy

The rating scale model (RSM) and the partial credit model (PCM) are fairly well-known examples of Rasch models for polytomously scored items. In addition to a number of threshold parameters, both the models contain two scalar parameters characterizing item and person location on a common interval-level scale. The rank order of items and persons defined by the Likert summative scores (i.e. the raw total scores) is compared with that obtained from the Rasch-based measures (i.e. the maximum likelihood estimates of person and item parameters). It is proved that: 1) the property of comonotonicity between Likert summative scores and Rasch-based measures holds for both the person and item parameters of the RSM; 2) the property of comonotonicity between Likert summative scores and Rasch-based measures holds for the PCM only with reference to the person parameters; 3) violations of comonotonicity are possible, for particular datasets, for the item parameters of the PCM.

An Analysis of Dimensionality Using Factor Analysis (True-Score Theory) and Rasch Measurement: What Is the Difference? Which Method Is Better?

Russell F. Waugh
Elaine S. Chapman
The University of Western Australia

One often hears the question asked, "For questionnaire data measuring a variable, what difference does it make to use factor analysis/principal components analysis (true-score theory) or Rasch measurement in testing for dimensionality?" This paper reports both factor analysis and Rasch measurement analysis for two sets of data. One set of data measures social anxiety for primary school students (N=436, I=10) and the second measures attitude to mathematics for primary-aged students (N=774, I=10). For both sets of data, the factor analysis suggests that the scores are reliable, and that inferences can be made that are valid for measuring school anxiety and attitude to mathematics. For both sets of data analyzed with Rasch measurement techniques, the reliability of the measures, the dimensionality of the measures, and the initial conceptualization of the items, are called into question. It suggests that one cannot make valid inferences from the measures that were initially set up for true-score theory. The Rasch analysis suggests that items intended to measure a variable should be initially developed on a conceptualized scale from easy to hard, and that students should answer the items from this perspective, so that the Rasch analysis of the data tests this conceptualization, and a linear scale can be created based on a mathematical measurement model with consistent units (logits).

Does Data Rounding-Off Influence Reproducibility Index Estimates?

Bruno Giraudeau
INSERM CIC 202, Faculté de Médecine
INSERM ERIT-M 0321, Université Paris 7
Philippe Ravaud
INSERM EMI 0357, Université Paris 7
Jean Yves Mary
INSERM ERIT-M 0321, Université Paris 7

The intraclass correlation coefficient (ICC) is the reproducibility index commonly used to assess the reproducibility level of continuous data. Such data are collected with finite precision (i.e., a finite number of decimal digits corresponding to the level of data rounding-off error). We therefore investigated the consequences of having finite precision for data on ICC estimates, thus comparing ICC estimates in presence of rounding-off to those associated with data with infinite precision. As long as the rounding-off level is less than 0.3sX (where sX is the standard deviation of the outcome measured), there is no practical consequence on the ICC estimate. Therefore, in most practical situations the consequences of rounding-off data on the estimation of reproducibility levels can be ignored.

Computer Adaptive Testing

Richard C. Gershon
Northwestern University

The creation of item response theory (IRT) and Rasch models, inexpensive accessibility to high speed desktop computers, and the growth of the Internet, has led to the creation and growth of computerized adaptive testing or CAT. This form of assessment is applicable for both high stakes tests such as certification or licensure exams, as well as health related quality of life surveys. This article discusses the historical background of CAT including its many advantages over conventional (typically paper and pencil) alternatives. The process of CAT is then described including descriptions of the specific differences of using CAT based upon 1-, 2- and 3-parameter IRT and various Rasch models. Numerous specific topics describing CAT in practice are described including: initial item selection, content balancing, test difficulty, test length and stopping rules. The article concludes with the author's reflections regarding the future of CAT.

Volume 6, Number 2

Estimating Parameters in the Rasch Model in the Presence of Null Categories

Guanzhong Luo
Hong Kong Examinations and Assessment Authority
David Andrich
Murdoch University, Australia

A category with a frequency of zero is called a null category. When null categories are present in polytomous responses, then in the Rasch model for such responses, the thresholds that define the categories are inestimable with the commonly used joint maximum likelihood, marginal maximum likelihood, or standard conditional maximum likelihood estimation algorithms. The reason for this situation is that in principle, these estimation algorithms involve frequencies of each category. Andrich and Luo (2003) describe an alogorithm in which the thresholds are reparameterized into their principal components and in which the estimate of any threshold is based on a function of the frequencies of all categories of the item rather than the frequency of a particular category. This algorithm works in the presence of null categories. However, in situations where the null categories are at the extremes of a set of categories, the estimates themselves can become too extreme. This paper describes a procedure in which the solution algorithm described by Andrich and Luo is further adapted in the presence of null categories by using their expected frequencies. The procedure is demonstrated with simulated and real data.

Effect of Item Redundancy on Rasch Item and Person Estimates

Everett V. Smith Jr.
University of Illinois at Chicago

One of the assumptions of many latent trait models is local independence. This assumption specifies that, after controlling for the underlying trait, item responses are independent. Given the lack of studies of model robust-ness against such violations, it appears that this assumption is frequently taken for granted. Therefore, this study investigated the robustness of Rasch item and person estimates with simulated data under varying number of items, sample sizes, and levels of item redundancy. Item and person reliabilities, the standard deviations of the person and item estimates, the root mean squared differences and mean signed differences among person and item estimates, the correlations between person estimates, and the percentage of person estimates shifting by more than .50 logits were used to evaluate the impact of item redundancy. Both norm and criterion-reference interpretations may be influenced by the imputation of redundancy into the data. However, it appears that the amount of redundancy needs to be considerable before such interpretations would be adversely impacted. Suggestions for further simulation research are provided.

Measuring Progress Toward Smoking Cessation

Melinda F. Davis
Lee B. Sechrest
Dan Shapiro
University of Arizona

Measuring the effect of behavioral interventions is often limited to a single outcome variable for ease of analysis. In the case of low probability outcomes, this narrow focus may often result in Type II errors, reducing the likelihood of detecting an effect of an intervention. The development and use of a scale to measure progress toward the ultimate desired change in behavior might result in greater sensitivity to subtle, but important, effects of interventions. That possibility is illustrated by the development and exploratory testing of a scale meant to measure penetration into the process of change with respect to smoking cessation. The scale consists of a set of outcome

indicators that are intended to represent the sequential steps that smokers go through in moving toward and ultimately giving up smoking. Rasch analyses indicate that the scale is coherent and merits further development. It seems likely that similar scales might be developed to assess progress toward change for many other behaviors that seem to require a gradual process of change that can be indexed by items representing discrete steps along the way.

Daredevil Barnstorming to the Tipping Point: New Aspirations for the Human Sciences

William P. Fisher, Jr.
MetaMetrics, Inc.

Aviation history provides an apt metaphor for the state of Rasch measurement practice, and its potential future. Flying was initially widely believed to be nothing but a spectacular and dangerous fad. Few saw in it any potential for the huge industry that it is today. The current state of Rasch measurement practice is quite akin to daredevil barnstorming in that the field is focused on isolated demonstrations of disconnected technical effects. Only when the analogues of air traffic control, airports, support staff, training programs, text-books, and partner industries (hotels, restaurants) are in place will Rasch measurement come into its own as the technical medium of a widespread industry. The point at which current practice tips into a new paradigm depends on the realization of operationally validated theory in a supportive social context. The paper closes with speculations on what crossing Rasch measurement's tipping point might entail.

Comparing Rasch Analyses Probability Estimates to Sensitivity, Specificity and Likelihood Ratios when Examining the Utility of Medical Diagnostic Tests

Daniel Cipriani
The Medical College of Ohio
Christine Fox
The University of Toledo
Sadik Khuder
The Medical College of Ohio
Nancy Boudreau
Bowling Green State University

Introduction: Medical diagnostic tests are evaluated based on measures of sensitivity (Sn), specificity (Sp), and likelihood ratios (LR). These procedures are limited in the event of a biased gold standard or missing data. Interpretations of these measures are frequently inappropriate. Purpose: The Rasch measurement model (RMM) was examined as a method to provide evidence of diagnostic test utility in order to overcome the limitations of Sn, Sp, and LR. Methods: Patients suspected of a knee ligament tear (n = 825) were studied, by evaluating four diagnostic tests. The RMM probability estimates for each test were compared to estimates of Sn, Sp, and LR. Results: The RMM provided probability estimates for the diagnosis that were comparable to likelihood ratios. These probability estimates correlated with the estimates of Sn, Sp, and LR. The RMM estimates were not affected by missing data. Discussion: The RMM may provide an alternative means to study the utility of medical diagnostic tests to estimate the probability of disease presence/absence.

A Rank-Ordering Method for Equating Tests by Expert Judgment

Tom Bramley
University of Cambridge

This paper describes a new method of comparing the raw mark scales on two tests using expert judgment. The two tests do not need to have any common items, nor to be taken by common groups of candidates. This study used

scripts (i.e. the complete work of a candidate on the test) from England's National Curriculum Test for Reading at Key Stage 3 (14-year olds) in 2003 and 2004. Each member of a panel of 12 experts was given four packs each containing ten scripts—five scripts from each year's test. Marks and annotations from these scripts had been removed. Their task was to put the ten scripts into a single rank order, based on a holistic judgment of the level of performance exhibited in each. Because the design of the study linked scripts across judges and packs it was possible to construct a single latent trait of judged quality of performance. This was done using two different analytical methods: the Rasch formulation of Thurstone paired comparisons, and the Rasch partial credit model. Relating the two raw mark scales to the single latent scale allowed the two years' tests to be equated. The merits of using this standard-maintaining method as opposed to a standard-setting method in this particular context are discussed.

Using the Rasch Model to Validate Stages of Understanding the Energy Concept

Xiufeng Liu
Sarah Collard
SUNY at Buffalo

In recent years, there have been efforts to bridge science education with developmental psychology to develop theories on students developing understanding of science concepts from elementary to high school and be-yond. The present study intends to test one such theory on students developing understanding of the energy concept. The theory states that students develop understanding of the energy concept by going through the following qualitatively distinct stages: (a) energy as activity/work; (b) energy as sources/forms, (c) energy transfer, (d) energy degradation, and (e) energy conservation. Three classes, one each from 4th grade, 8th grade, and high school physics class (grades 10, 11, and 12), completed a performance assessment. Students' performances were scored based on three traits of energy understanding: attention capacity, qualitative relations, and quantitative relations; each of the traits was defined into five hierarchical levels consistent with the five stages of understanding the energy concept. The Many-Facet Rasch Measurement (MFRM) model was used to analyze the effects of rater scoring severity, students' stages of energy understanding (theta), and difficulties of energy understanding traits. Results show that there was a discontinuity among the stages of understanding the energy concept, supporting the theory on students developing the understanding of the energy concept.

Volume 6, Number 3

The Multilevel Measurement Model: Introduction to the Special Issue

S. Natasha Beretvas
University of Texas at Austin
Akihito Kamata
Florida State University

An introduction to the special issue on the multilevel measurement model (MMM) is provided. The two- and three-level multilevel models for continuous outcomes are reviewed. The extension to the hierarchical generalized linear model and its use as a multilevel measurement model for dichotomous measurement indicators is demonstrated. The six articles in the special issue are described.

Demonstration of Software Programs for Estimating Multilevel Measurement Model Parameters

J. Kyle Roberts
Baylor College of Medicine
Rich Herrington

University of North Texas

A brief overview of the relevant parameterizations of the Rasch measurement model will first be provided. Next, the use of five different hierarchical linear modeling software programs (SAS, MLWIN, S-PLUS, R and HLM) using a heuristic data set will be demonstrated. For each program, researchers will be offered: 1) instructions describing the way to set up data sets, as well as 2) directions for running each program, 3) guidelines to assist with the appropriate interpretation of the output. Last, testing of the local independence assumption will be discussed. All data and code for each of the models run in this analysis may be downloaded from the website: http://www.hlm-online.com/papers/mmmpaper.htm/ .

Mixed Model Estimation Methods for the Rasch Model

Frank Rijmen
Francis Tuerlinckx
Michel Meulders
Dirk J.M. Smits
Katalin Balázs
K.U. Leuven, Belgium

Mixed models take the dependency between observations based on the same person into account by introducing one or more random effects. After introducing the mixed model framework, it is explained, by taking the Rasch model as a generic example, how item response models can be conceptualized as generalized linear and nonlinear mixed models. Common estimation methods for generalized linear and nonlinear models are dis-cussed. In a simulation study, the performance of four estimation methods is assessed for the Rasch model under different conditions regarding the number of items and persons, and the degree of inter-individual differences. The estimation methods included in the study are: an approximation of the integral over the random effect by means of Gaussian quadrature; direct maximization with a sixth-order Laplace approximation to the integrand; a linearized approximation of the nonlinear model employing PQL2; and finally a Bayesian MCMC method. It is concluded that the estimation methods perform almost equally well, except for a slightly worse recovery of the variance parameter for PQL2 and MCMC.

Some Links between Classical and Modern Test Theory via the Two-Level Hierarchical Generalized Linear Model

Yasuo Miyazaki
Virginia Polytechnic Institute and State University

This article considers some links between classical test theory (CTT) and modern test theory (MTT) such as item response theory (IRT) and the Rasch model in the context of the two-level hierarchical generalized linear model (HGLM). Conceptualizing items as nested within subjects, both the CTT model and the MTT model can be reformulated as an HGLM where item difficulty parameters are represented by fixed effects and subjects' abilities are represented by random effects. In this HGLM framework, the CTT and MTT models differ only in the level 1 sampling model and the associated link function. This article also contrasts the Rasch and two-parameter IRT models by considering the property of specific objectivity in the context of CTT. It is found that the essentially tau-equivalent model exhibits specific objectivity if the data fit the model, but the congeneric measures model does not. Data from English composition scores on essay writing used by Jöreskog (1971) are reanalyzed for illustration.

Modeling Local Item Dependence with the Hierarchical Generalized Linear Model

Hong Jiao
Shudong Wang

Harcourt Assessment, Inc.
Akihito Kamata
Florida State University

Local item dependence (LID) can emerge when the test items are nested within common stimuli or item groups. This study proposes a three-level hierarchical generalized linear model (HGLM) to model LID when LID is due to such contextual effects. The proposed three-level HGLM was examined by analyzing simulated data sets and was compared with the Rasch-equivalent two-level HGLM that ignores such a nested structure of test items. The results demonstrated that the proposed model could capture LID and estimate its magnitude. Also, the two-level HGLM resulted in larger mean absolute differences between the true and the estimated item difficulties than those from the proposed three-level HGLM. Furthermore, it was demonstrated that the proposed three-level HGLM estimated the ability distribution variance unaffected by the LID magnitude, while the two-level HGLM with no LID consideration increasingly underestimated the ability variance as the LID magnitude increased.

The Cross-Classified Multilevel Measurement Model: An Explanation and Demonstration

S. Natasha Beretvas
University of Texas at Austin
Jason L. Meyers
Pearson Educational Measurement
Rolando A. Rodriguez
University of Texas at Austin

The link between the hierarchical generalized linear model (HGLM) and the Rasch model's parameterization has already been demonstrated by several researchers. Extensions have been described that include higher clustering levels to model more appropriately the contextual effects that are frequently encountered in educational research. However, pure hierarchies are relatively rare and instead cross-classified data structures are more frequently encountered. Cross-classified random effect modeling (CCREM) is still not commonly used. Use of CCREM in combination with the multilevel measurement model (MMM) has been recently introduced and is described further in the current study. Specifically, the link between the MMM and the CCREM MMM (termed "CCMMM" model) is provided. A dataset was simulated to demonstrate interpretation of the CCMMM model's parameters and to compare results under a CCMMM versus HGLM analysis. An Appendix is provided to demonstrate SAS GLIMMIX code used to estimate HGLM and CCMMM models' parameters.

Test Equating in the Presence of DIF Items

Kwang-lee Chu
Harcourt Assessment Inc.
Akihito Kamata
Florida State University

This paper proposes a multilevel measurement model that controls for DIF effects in test equating. The accuracy and stability of item and ability parameter estimates under the proposed multilevel measurement model were examined using randomly simulated data. Estimates from the proposed model were compared with those resulting from two multiple-group concurrent equating designs, including 1) a design that replaced DIF-items with items with no DIF; and 2) a design that retained DIF items with no attempt to control for DIF. In most of the investigated conditions, the results indicated that the proposed multilevel measurement model performed better than the two comparison models.

Volume 6, Number 4

Unique Properties of Rasch Model Item Information Functions

Randall D. Penfield
University of Miami

The Rasch family of models displays several well-documented properties that distinguish them from the general item response theory (IRT) family of measurement models. This paper describes an additional unique property of Rasch models, referred to as the property of item information constancy. This property asserts that the area under the information function for Rasch models is always equal to the number of response categories minus one, regardless of the values of the item location parameters. The implication of the property of item information constancy is that, for a given number of response categories, all items following a Rasch model contribute equally to the height of the test information function across the entire latent continuum.

Evaluation of the Diabetes Self-Care Scale

Nantawadee P. Lee
NorthShore Regional Medical Center
William P. Fisher, Jr.
MetaMetrics, Inc.

The purpose of this study was to evaluate the psychometric properties of the Diabetes Self-Care Scale (DSCS). A convenience sample of 175 adults with diabetes who met the inclusion criteria from a local hospital in southern Louisiana participated in this study. Data from a pilot study with 50 respondents were also used to calibrate the instrument. The analysis tested the hypothesis that the difference between any respondent's agreement with an item and the difficulty presented by that item is equal to the natural logarithm of the odds of a response. Results indicated that respondent separation reliability is acceptable (.80) and item separation reliability is high (.99). All 35 items measure a single construct of diabetes self-care. Construct validity was supported by the meaningfulness of the item endorsability order and by the consistency of that order across respondents. Seven recommendations for modification of the instrument, future research, and practice are proposed.

Rasch Analysis Examining Processing Mechanisms of the Object Location Memory Test Revised

Tracy L. Kline
Karen M. Schmidt
University of Virginia

This study's objective was the construction and examination of the Object Location Memory Revised (OLM- R), an instrument designed to measure spatial memory. The OLM-R measured a participant's (N = 111) ability to inspect a spatial array and recall image identities and positions after a distractor task. Rasch methodology and regression analyses were employed to explore the influence a priori design factors have on performance and item difficulty. Rasch analyses revealed that, while the OLM-R has a misfitting item, overall the instrument shows good measurement properties. Specific analyses examining complexity factors indicated that Object Manipulation (e.g. moving, replacing, or unchanging) and the Number of Items in an array were leading influences of OLM-R performance.

Using the Rasch Model to Develop a Measure of Second Language Learners' Willingness to Communicate within a Language Classroom

Christopher Weaver
Tokyo University of Agriculture and Technology

The purpose of this investigation was to use Rasch measurement to study the psychometric properties of a 34 item questionnaire designed to measure second language learners' willingness to communicate (WTC) in English inside their language class. 490 Japanese university students' responses to the questionnaire were subjected to a number of different analyses. The first involved a comparison of the category threshold estimates produced by the Rating Scale and Partial Credit models. The questionnaire's items were then evaluated according to how well they defined the willingness to communicate construct. The potential dimensionality of using items that involved different speaking and writing tasks/situations in order to gain a more comprehensive understanding of students' willingness to communicate was also investigated. Next there was an examination of the questionnaire's four-point scale to ensure that it captured meaningful differences in students' WTC. Finally, the questionnaire items were compared using differential item functioning to determine if second year students were more willing than first year students in any of the different speaking and writing tasks/situations. This investigation closes with some suggestions on how the WTC questionnaire can inform second language instruction and curriculum design.

Knowledge and Understanding of Asia: Using a Common Item Pool to Obtain National Estimates

Patrick Griffin
Kerry Woods
The University of Melbourne

A series of tests were developed to assess the proficiency of Australian Year 5 and Year 8 students in Asian Studies. This paper presents results of analyses that involved calibrating items distributed over 14 overlapping subtests, developed to cater for state and territory curricula and two year-levels. This allowed for state and year-level preferences to be selected from a common pool of 105 items. The project used common item anchoring to map all students and items onto a single, underpinning scale that was identified and interpreted using concurrent equating procedures and a skills audit of items.

Measuring and Comparing Higher Education Quality between Countries Worldwide

Anatoli A. Maslak
Slaviansk-on-Kuban State Pedagogical Institute, Russia
George Karabatsos
College of Education, University of Illinois-Chicago, USA
Tatijana S. Anisimova
Sergei A. Osipov
Slaviansk-on-Kuban State Pedagogical Institute, Russia

The purpose of this investigation is to establish a unidimensional interval scale for measuring each country on the quality of higher education, based on indicators (items) characterizing various aspects of a country's quality. The data from these indicators are publicly available through the United Nations Educational, Scientific, and Cultural Organization (UNESCO), for all countries worldwide. Currently, a country's quality of higher education is summarized by simple descriptive statistics of these many indicators, and there seems to be a need to combine these results into a single measure of quality. A single measure of quality can facilitate a country's education quality to be monitored over time, enable a comparison of education quality between different countries, and help decision making relative to improving a country's quality of education. The results show that, with the Rasch partial credit model, it is possible to represent each country's quality of higher education by a single measure. Possible implications and extensions of this new scale are discussed.

The Relationship between the Rating Scale and Partial Credit Models and the Implication of Disordered Thresholds of the Rasch Models for Polytomous Responses

Guanzhong Luo
Hong Kong Examinations and Assessment Authority

There is a perception in the literature that the Rating Scale Model (RSM) and Partial Credit Model (PCM) are two different types of Rasch models. This paper clarifies the relationship between the RSM and PCM from the perspectives of literature history and mathematical logic. It is shown that not only are the RSM and the PCM identical, but the two approaches used to introduce them are statistically equivalent. Then the implication of disordered thresholds is discussed. In addition, the difference between the structural thresholds and the Thurstone thresholds are clarified.
