Journal of Applied Measurement Abstracts

Volume 1, Number 1 (2000)

Attention Deficit Hyperactivity Disorder: Scaling and Standard Setting using Rasch Measurement

Everett V. Smith, Jr.
The University of Illinois at Chicago
Brian D. Johnson
The University of Northern Colorado

This paper explores the dimensionality of responses to the Adult Behavior Checklist—Revised, a screening assessment for Attention Deficit Hyperactivity Disorder (ADHD) in college students, and demonstrates a standard setting process for diagnostic rating scales. Responses from 317 undergraduate college students were used to investigate dimensionality; 8 judges were used for the standard setting procedure. A series of Rasch rating scale analyses support the interpretation of Inattention and Impulsivity/Hyperactivity variables. Principal component analyses of residuals identified the existence of secondary variables that may have clinical implications for the evaluation and treatment of ADHD. For the standard setting process, judges generally displayed less variability than expected by the model. Several judges were in disagreement as to the level of symptomatology needed to indicate potential ADHD. The derived standard for Inattention was found to be more stringent than a previously suggested standard while the derived standard for Hyperactive/Impulsive more lenient. Prevalence rates found a diagnosis of Hyperactive/ Impulsive to be more common than a diagnosis of Inattention. Efficient screening assessments for ADHD in college students are needed in order to determine who may qualify for further evaluation and special services. The standard setting process attempts to identify judge disagreement regarding the level of symptomatology (the standard) needed to qualify for additional evaluation and identify judge response strings not consistent with the item difficulties. Recommendations suggesting additional work on the screening assessment and standard setting process are outlined.

An Investigation of Factors Affecting Test Equating in Latent Trait Theory

Surintorn Suanthong
Randall E. Schumacker
Michael M. Beyerlein
University of North Texas

The study investigated five factors which can affect the equating of scores from two tests onto a common score scale. The five factors were: (a) item distribution type (i.e., normal versus uniform; (b) standard deviation of item difficulty (i.e., .68, .95, .99); (c) number of items or test length (i.e., 50, 100, 200); (d) number of common items (i.e., 10, 20, 30); and (e) sample size (i.e., 100, 300, 500). SIMTEST and BIGSTEPS programs were used for the simulation and equating of 4,860 item data sets, respectively. Results from the five-way fixed effects factorial analysis of variance indicated three statistically significant two-way interaction effects. Simple effects for the interaction between common item length and test length only were interpreted given Type I error rate considerations. The eta-squared values for number of common items and test length were small indicating the effects had little practical importance. The Rasch approach to equating is robust with as few as 10 common items and a test length of 100 items.

An Approach to Studying Scale for Students in Higher Education: A Rasch Measurement Model Analysis

Russell F. Waugh

Teck Kiong Hii
Atique Islam
Edith Cowan University, Western Australia

A questionnaire comprising 80 self-report items was designed to measure student Approaches to Studying in a higher education context. The items were conceptualized and designed from five learning orientations: a Deep Approach, a Surface Approach, a Strategic Approach, Clarity of Direction and Academic Self-Confidence, to include 40 attitude items and 40 corresponding behavior items. The study aimed to create a scale and investigate its psychometric properties using a Rasch measurement model. The convenience sample consisted of 350 students at an Australian university in 1998. The analysis supported the conceptual structure of the Scale as involving studying attitudes and behaviors towards five orientations to learning. Attitudes are mostly easier than behaviors, in line with the theory. Sixty-eight items fit the model and have good psychometric properties. The proportion of observed variance considered true is 92% and the Scale is well-targeted against the students. Some harder items are needed to improve the targeting and some further testing work needs to be done on the Surface Approach. In the Surface Approach and Clarity of Direction in Studying, attitudes make a lesser contribution than behaviors to the variable, Approaches to Studying.

Modeling Effects of Differential Item Functioning in Polytomous Items

Wen-Chung Wang
National Chung Cheng University

Conventional two-group DIF analysis for dichotomous items is extended to factorial DIF analysis for polytomous items where multiple grouping factors with multiple groups in each are jointly analyzed. By adopting the formulation of general linear models, item parameters across all possible groups are treated as a dependent variable and the grouping factors as independent variables. These item parameters are then reparameterized as a set of grand item parameters and sets of DIF parameters representing main and interaction effects of the factors on the items. Results of simulation studies show that the parameters of the proposed modeling could be satisfactorily recovered. A real data set of 10 polytomous items and 1924 subjects was analyzed. Applications and implications of the proposed modeling are addressed.

Rasch Models Overview

Benjamin D. Wright
University of Chicago
Magdalena Mok
The Hong Kong Institute of Education
Macquarie University

This overview of Rasch measurement models begins with a conceptualization of our continuous experiences that are often captured as discrete observations. It goes on to discuss the properties that are require of measures if they are to transcend the occasion in which they were collected, and concludes with a discussion the spiral of inferential development. This is followed by a discussion of the mathematical properties of the Rasch family of models that allow the transformation of discrete deterministic counts into continuous probabilistic abstractions on which science is based. The overview concludes with a discussion of six of the family of Rasch models, Binomial Trials, Poisson Counts, Rating Scale, Partial Credit, and Ranks and the types of data for which these models are appropriate.

Volume 1, Number 2

Development of a Nutrition Self-Efficacy Scale for Prospective Physicians

Jessica A. Schulman
University of Florida
Edward W. Wolfe
Michigan State University

Diet is associated with 5 of the 10 leading causes of death in the U.S., including coronary heart disease, certain types of cancer, atherosclerosis, and type 2 diabetes. Physicians can play a pivotal role in promoting nutritional management of diabetes and other chronic diseases. Therefore, it is important that valid instruments are created so administrators can better assess the educational needs of prospective physicians, their practices, and patient outcomes. Two comparable studies, one year apart, were undertaken to create an instrument that measures nutritional competence and self-efficacy among prospective physicians. This paper: (a) describes the development of a nutrition self-efficacy scale (NSES) and (b) demonstrates reliability and validity of the NSES using Rasch modeling. It concludes with a discussion of potential contributions of this scale for assessing mastery of applied nutrition among prospective physicians.

The Impact of Receiving the Same Items on Consecutive Computer Adaptive Test Administrations

Thomas O'Neill
American Society of Clinical Pathologists
Mary E. Lunz
Measurement Resources, Inc.
Keith Thiede
University of Illinois at Chicago

This study addresses item exposure in a Computerized Adaptive Test (CAT) when the item selection algorithm is permitted to present examinees with questions that they have already been asked in a previous test administration. The results indicate that the combined use of an adaptive algorithm to select items and latent trait theory to estimate person ability provides substantial protection from score contamination. The implications for constraints that prohibit examinees from seeing an item twice are discussed.

A Critique of Rasch Residual Fit Statistics

George Karabatsos
Louisiana State University Health Sciences Center

In test analysis involving the Rasch model, a large degree of importance is placed on the "objective" measurement of individual abilities and item difficulties. The degree to which the objectivity properties are attained, of course, depends on the degree to which the data fit the Rasch model. It is therefore important to utilize fit statistics that accurately and reliably detect the person-item response inconsistencies that threaten the measurement objectivity of persons and items. Given this argument, it is somewhat surprising that there is far more emphasis placed in the objective measurement of person and items than there is in the measurement quality of Rasch fit statistics. This paper provides a critical analysis of the residual fit statistics of the Rasch model, arguably the most often used fit statistics, in an effort to illustrate that the task of Rasch fit analysis is not as simple and straightforward as it appears to be. The faulty statistical properties of the residual fit statistics do not allow either a convenient or a straightforward approach to Rasch fit analysis. For instance, given a residual fit statistic, the use of a single minimum critical value for misfit diagnosis across different testing situations, where the situations vary in sample and test properties, leads to both the overdetection and underdetection of misfit. To improve this situation, it is argued that psychometricians need to implement residual-free Rasch fit statistics that are based on the number of Guttman response errors, or use indices that are statistically optimal in detecting measurement disturbances.

Construct Validity of Scores/Measures from a Developmental Assessment in Mathematics using Classical and Many-Facet Rasch Measurement

Madhabi Banerji
University of South Florida

Data from a developmental assessment comprised of 9 short answer mathematics tasks were validated using classical and three-faceted Rasch measurement methods. Field test data from a mixed age elementary school sample (*N*=280) were analyzed. Descriptive statistics on scores from the overall scale and two subdomains indicated improved performance with age. The data showed better fit with a two-factor model corresponding with the subdomain structure (Bentler's CFI=.94), than a one factor model (CFI=.87). The inter-factor correlation was .76. Convergent validity coefficients of scores with scaled scores of the Stanford Achievement Test mathematics battery ranged from .28 to .47; internal consistency reliability of the total and subdomain scores ranged from .87 to .89, respectively; and median inter-rater reliability was .75. On average, persons, tasks and raters showed acceptable fit with the three-facet Rasch model. Rasch logit difficulties of tasks suggested an ordered scale structure, although tasks tended to have high difficulty levels. The original and calibrated task ordering was consistent at the extreme ends of the scale. Gaps identified on the Rasch item map suggested a need for additional task construction. Conceptual and procedural differences in each technique are considered in deciding future improvements to the scale.

Fit Analysis in Latent Trait Measurement Models

Richard M. Smith
University of Florida

The analysis of fit, whether viewed from the prospective of the fit of the data to the measurement model, or the fit of the measurement model to the data, is an important part of using latent trait models. In the case of the Rasch model, all of the desirable characteristics of the model (interval item and person measures, asymptotic standard errors, parameter invariance across subsets of persons or items, to name a few) are predicated on the requirement that the data fit the model. To the extent that the data do not fit the model, these properties hold to a lesser degree. The analysis of fit is of primary importance if the interpretation of the calibration results is to be useful. This article explores the nature of fit and provides a historical overview of fit indices. It then focuses on a particular family of fit indices that are based on the Pearsonian chi-square approach to fit, in an attempt to show why it is necessary to use a family of standardized fit indices to completely understand the relationship between the data and the model.

Volume 1, Number 3

Development of a Scale for Measuring Invasive Plant Environmentalism

Edward W. Wolfe
Michigan State University
Hallie Dozier
Southeastern Louisiana University

The ecological impact of invasive plant species is a serious concern among environmental scientists and conservationists. Educating the public about invasive plant issues is a major hurdle, given that several invaders come into the environment through ornamental gardening. An important first step in planning an educational program concerning invasive plant issues is to assess public knowledge and attitudes concerning these issues. This paper describes the development of an instrument that measures invasive plant environmentalism. Responses from 237 nursery customers from the southeastern U.S. to a 17-item standardized interview were scaled using the partial

credit model, a member of the family of Rasch (1960) measurement models. Our results indicate that the instrument adequately measures this construct. Substantive interpretations of the results are also discussed.

Factorial Modeling of Differential Distractor Functioning in Multiple-choice Items

Wen-Chung Wang
National Chung Cheng University

A factorial procedure for investigating differential distractor functioning in multiple-choice items is proposed. The procedure adopts the formulation of general linear models and treats grouping factors as independent variables and item parameters across the grouping factors as a dependent variable. Specifically, each distractor in a multiple-choice item is modeled with a distinct distractibility parameter. The distractibility parameters across groups are partitioned into a grand mean distractibility and sets of parameters representing main effects of the individual grouping factors, and interaction effects among them. Results of a simulation study show that the parameters of the proposed modeling were recovered very well. Ten four-choice items in the English test of the 1997 Taiwan Joint College Entrance Examination with seven thousands of examinees in two grouping factors were analyzed.

Measurement of Cognitive Performance in Computer Programming Concept Acquisition: Interactive Effects of Visual Metaphors and the Cognitive Style Construct

Elspeth McKay
RMIT University

An innovative research program was devised to investigate the interactive effect of instructional strategies enhanced with text-plus-textual metaphors or text-plus-graphical metaphors, and cognitive style on the acquisition of programming concepts. The Cognitive Styles Analysis (CSA) program (Riding, 1991) was used to establish the participants' cognitive style. The QUEST Interactive Test Analysis System (Adams and Khoo, 1996) provided the cognitive performance measuring tool, which ensured an absence of error measurement in the programming knowledge testing instruments. Therefore, reliability of the instrumentation was assured through the calibration techniques utilized by the QUEST estimate; providing predictability of the research design. A means analysis of the QUEST data, using the Cohen (1977) approach to size effect and statistical power further quantified the significance of the findings. The experimental methodology adopted for this research links the disciplines of instructional science, cognitive psychology, and objective measurement to provide reliable mechanisms for beneficial use in the evaluation of cognitive performance by the education, training and development sectors. Furthermore, the research outcomes will be of interest to educators, cognitive psychologists, communications engineers, and computer scientists specializing in computer-human interactions.

CAT Administration of Language Placement Examinations

John Stahl
Betty Bergstrom
Computer Adaptive Technologies, Inc.
Richard Gershon
Houghton Mifflin Company

This article describes the development of a computerized adaptive test for Cegep de Jonquiere, a community college located in Quebec, Canada. Computerized language proficiency testing allows the simultaneous presentation of sound stimuli as the question is being presented to the test-taker. With a properly calibrated bank of items, the language proficiency test can be offered in an adaptive framework. By adapting the test to the test-taker's level of ability, an assessment can be made with significantly fewer items. We also describe our initial attempt to detect instances in which "cheating low" is occurring. In the "cheating low" situation, test-takers

deliberately answer questions incorrectly, questions that they are fully capable of answering correctly had they been taking the test honestly.

Metric Development and Score Reporting in Rasch Measurement

Everett V. Smith, Jr.
University of Illinois at Chicago

This article: 1) describes problems in score reporting with the True-Score Model, 2) presents a definition of the Rasch measurement unit, the logit, 3) reviews various transformations of the logit metric, and 4) provides examples of score reporting procedures. Two data sets are used. The first contains dichotomous data drawn from responses to a multiple-response statistics examination taken by Ph.D. students; the second contains polychotomous data from a self-efficacy assessment given to war veterans suffering from Post-Traumatic Stress Disorder.

Volume 1, Number 4

The Social Physique Anxiety Scale: An Example of the Potential Consequence of Negatively Worded Items in Factorial Validity Studies

Robert W. Motl
The University of Georgia
David E. Conroy
The Pennsylvania State University
Patrick M. Horan
The University of Georgia

Social physique anxiety (SPA) based on Hart, Leary, and Rejeski's (1989) Social Physique Anxiety Scale (SPAS) was originally conceptualized to be a unidimensional construct. Empirical evidence on the factorial validity of the SPAS has been contradictory, yielding both one- and two-factor models. The two-factor model, which consists of separate factors associated with positively and negatively worded items, has stimulated an ongoing debate about the dimensionality and content of the SPAS. The present study employed confirmatory factor analysis (CFA) to examine whether the two-factor solution to the 12-item SPAS was substantively meaningful or a methodological artifact. Results of the CFAs, which were performed on responses from four different samples (Eklund, Kelley, and Wilson, 1997; Eklund, Mack, and Hart, 1996), supported the existence of a single substantive SPA factor underlying the responses to the 12-item SPAS. There were, in addition, method effects associated with the negatively worded items that could be modeled to achieve good fit. Therefore, it was concluded that a single substantive factor and a non-substantive method effect primarily related to the negatively worded items best represented the 12-item SPAS.

Moral and Evaluative Reasoning Across the Life-span

Theo Linda Dawson
University of California at Berkeley

In a longitudinal/cross sectional study of moral and evaluative reasoning, Armon interviewed 23 females and 19 males, ages ranging from 5 at the first test time (1977) to 86 at the 4th (1989) test-time. Rasch analysis of Armon's data demonstrated that Armon's and Kohlberg's measures tap a single underlying dimension of reasoning; that individual stages across five items measure the same levels of reasoning, and that development on all items progresses at about the same rate. Participants found it easier to apply already available reasoning structures to new areas than to reason at a new stage, implying that stage transition is step-like.

Methodological Issues in Using the Rasch Model to Select Cross Culturally Equivalent Items in Order to Develop a Quality of Life Index: The Analysis of Four WHOQOL-100 Data Sets (Argentina, France, Hong Kong, United Kingdom)

Alain Leplege
Emmanuel Ecosse
INSERM U292, France
WHOQOL Rasch Project Scientific Committee
Silvia Bonicatto, *FUNDONAR, La Plata, Argentina*
Rex Billington, *WHO, Geneva, Switzerland*
Monica Bullinger, *Hamburg University, Germany*
Guss van Heck, *Department of Psychology, Tilburg University,The Netherlands*
Kwok Fai Leung, *Department of Occupational Therapy, Queen Elizabeth Hospital, Hong Kong*
John Orley, *WHO, Geneva, Switzerland*
Donald Patrick, *Department of Health Services, University of Washington, Seattle, WA*
Mick Power, *Department of Psychiatry, Edimburg, United Kingdom*
Suzy Skevington, *School of Social Sciences, Bath, United Kingdom*.

The Constitution of the World Health Organization (WHO) defines Health as "A state of complete physical, mental, and social well-being not merely the absence of disease . . ." (WHOQOL Group, 1993). It follows that the measurement of health and the effects of health care must include not only an indication of changes in the frequency and severity of diseases but also of changes in well-being, and this can be assessed by measuring the improvement in the quality of life related to health care.

Simultaneous Measurement of Reading Growth, Gender, and Relative-Age Effects: Many-Faceted Rasch Applied to CBM Reading Scores

Peter MacMillan
University of Northern British Columbia

Reading growth, gender effects, relative-age effects, and reading probe difficulty for reading were simultaneously assessed on one linear scale. The reading measure chosen was the Curriculum Based Measurement (CBM), words-read-correctly. A sample of 1619 students in Grades Two through Seven was employed. There is growth in reading within all grades but decreasing growth with increasing grade level. There are consistent gender differences favoring girls but the differences are equivalent to only one month's growth. There is no consistent evidence of a relative-age effect across grades. Two strategies for linking data across grades were compared and found to produce results consistent with each other and individual grade results.

Equating and Item Banking with the Rasch Model

Edward W. Wolfe
Michigan State University

This article describes Rasch measurement procedures for equating multiple test forms or calibrating an item bank. The procedures entail (a) selecting an appropriate data collection design, (b) estimating parameters, (c) transforming the parameters from multiple forms to a common scale, and (d) evaluating the quality of the linkage between these forms. Data collection designs include (a) anchor tests, (b) single group, (c) single data set, and (d) equivalent groups. Estimation procedures may involve (a) separate or (b) simultaneous calibration of data from multiple forms. Transformation is typically accomplished using (a) estimation scaling, but may involve (b)

parameter anchoring or (c) computing equating constants. Link quality is evaluated using four fit indices: (a) item-within-link, (b) item between- link, (c) link-within-bank, and (d) form-within-bank. These procedures are illustrated using an anchor test design.