Journal of Applied Measurement Abstracts

Volume 2, Number 1 (2001)

Historical View of the Influences of Measurement and Reading Theories on the Assessment of Reading

George Engelhard, Jr.
Emory University

The purpose of this study is to briefly explore the interactions among measurement theories, reading theories, and measurement practices from an historical perspective. The assessment of reading provides a useful framework for examining how theories influence, and in some cases fail to influence, the practice of reading assessment as operationalized in reading tests. The first section describes a conceptual framework for examining the assessment of reading. Next, I describe the major research traditions in measurement theory that have dominated measurement practice during the 20th century. In the next section, I briefly introduce major reading theories. Next, I bring together the previous two sections in order to examine the adequacy of the proposed conceptual framework for examining the assessment of reading. This section includes criticism of measurement theory by selected reading theorists. It also provides a brief history of the use of Rasch measurement theory to calibrate reading tests. Finally, the main points of the study are summarized and discussed. It should be recognized that this study represents a preliminary analysis of these issues.

The Assessment of Unidimensionality of Normal and Lognormal Data: A Look at Two Nonparametric Procedures

Anne E. Seraphine
James J. Algina
M. David Miller
University of Florida

In this Monte Carlo study, the Type I error rate and the power of the Stout T procedure (DIMTEST) and the Holland-Rosenbaum procedure (HR) were examined for normal and lognormal data sets. Both procedures are based on a nonparametric item response model, where the key assumption is the item response function is monotonically nondecreasing. The two procedures performed adequately under certain conditions for the both the normal and lognormal data sets. Of the two, however, the Stout T procedure showed adequate power under more conditions than the Holland-Rosenbaum procedure.

Rasch Measurement in the Assessment of Growth Hormone Deficiency in Adult Patients

Luis Prieto
Universitat Ramon Llull
Hospital de la Santa Creu i Sant Pau
World Health Organization
Montse Roset
Health Outcomes Research Europe
Xavier Badia
Hospital de la Santa Creu i Sant Pau

The Assessment of Growth Hormone Deficiency in Adults (AGHDA) questionnaire was previously designed, translated and validated in several European countries to evaluate the impact of the disease on Quality of Life. This study aimed to test the metric properties of the Spanish version by means of Rasch analysis. A sample of 356 consecutive adult patients with untreated GHD was included in the study. Patients responded to the questionnaire at baseline and 12 months apart. Answers were analyzed following the dichotomous logistic response model. Parameter estimates, model-data fit and separation statistics were computed. The invariance of the item parameters across time was tested in the follow-up. Rasch results were additionally employed to ascertain score changes through the calculation of the Reliable Change Index (RCI). Items varied in severity from 8.3-16.8 units (*SE*= 0.4-0.5) and fit to define a unidimensional variable. The item separation index (SI) (5.2) indicates a good and reliable (0.9) separation of the items along the variable they define. Moreover, results showed the AGHDA conforms to the model expectation of item parameter invariance between administrations. The substantial (2.3) and reliable (0.8) person SI also suggests the sample was well targeted by the questionnaire. According to the RCI, 84% of the patients did not show a significant transition in their measures. Results denote the items of the AGHDA succeeded in defining a scale characterized by the interval-level of its measures, suggesting the questionnaire could be a useful complement of the clinical evaluation of GHD patients at both group and individual level.

An Investigation of Gender Differences in the Components Influencing the Difficulty of Spatial Ability Items

Gene A. Kramer
American Dental Association
Richard M. Smith
University of Florida

This study examines the role that gender differences play in the determination of the components influencing the difficulty of spatial ability items. Considerable research has examined the role of gender differences in spatial abilities, with sometimes contradictory findings. In general, the findings show that males tend to outperform females on spatial ability items. Other research has focused on determining the components of items that contribute to their difficulty. This research has usually been based on mixed-gender populations, however. The present study attempts to determine if gender influences the extent to which different components contribute to the difficulty of items. The results indicate that component difficulties show little variation across gender. This finding supports the notion that any differences in raw scores observed for males and females are not due to differences in the manner in which males and females process spatial information or solve spatial ability items.

Partial Credit Model and Pivot Anchoring

Rita K. Bode
Rehabilitation Institute of Chicago
Northwestern University Medical School

This article contains information on the Rasch measurement partial credit model: what it is, how it differs from other Rasch models, when to use it, and how to use it. The calibration of instruments with increasingly complex items is described, starting with dichotomous items and moving on to polychotomous items using a single rating scale, and mixed polychotomous items using multiple rating scales, and instruments in which each item has its own rating scale. It also introduces a procedure for aligning rating scale categories to be used when more than one rating scale is used in a single instrument. Pivot anchoring is defined and an illustration of its use with the mental health scale of the SF-36 that contains positive and negative items is provided. It finally describes the effect of pivot anchoring on step calibrations, the item hierarchy, and person measures.

Volume 2, Number 2

Using the Rasch Measurement Model to Investigate the Construct of Motor Ability in Young Children

Beth Hands
Dawne Larkin
The University of Western Australia

This paper reports the use of a Rasch measurement model, the Extended Logistic Model of Rasch (Andrich, 1988), to explore the construct of a general motor ability in young children. Data were collected from 332 five and six year old children performing 24 motor skills, including run, hop, balance and ball skills. The data were categorized based on threshold estimates provided by the measurement model. Gender differences in performances on items were hypothesized to contribute to initial item and person misfit for the total sample. The data for boys and for girls were separated and independently analyzed resulting in improved item and person fit. Two different, unidimensional scales for boys and for girls were created.

The Alternate Forms Reliability of the New Tasks Added to the Assessment of Motor and Process Skills

Stephanie Ellison
Anne G. Fisher
Leslie Duran
Colorado State University

The purpose of this study was to evaluate the alternate forms reliability of new tasks vs. old tasks of the Assessment of Motor and Process Skills (AMPS). The participants in this study were 44 persons taken from the AMPS database who had completed two old tasks and two new tasks within a 4-day period. Paired *t*-tests revealed no significant difference between the means of ADL ability measures based on the performance of new vs. old tasks. The Pearson product moment correlations between the ADL ability measures based on the performance of new vs. old tasks was $r = .92$, $p < .001$ for motor ability measures and $r = .77$, $p < .001$ for process ability measures. We found that 100% of the ADL motor ability measures had standardized differences less than 2.00 ($p < .05$) and 97% of the ADL process ability measures had standardized differences less than 2.00 ($p < .05$). Considered together, the results support good alternate forms reliability of the ADL motor and ADL process ability measures. This study supported the finding that the 20 newly calibrated IADL and PADL tasks can be used reliably in clinical practice. When the AMPS is used to evaluate change, we can have 80 to 93% confidence that paired ability measures that change by more than +0.5 logits are the result of actual changes in ability.

Cross-Cultural Validation of the Inventory of School Motivation (ISM): Motivation Orientations of Navajo and Anglo Students

Dennis M. McInerney
University of Western Sydney
Alexander Seeshing Yeung
Hong Kong Institute of Education
Valentina McInerney
University of Western Sydney

The Motivation Orientation scales of the Inventory of School Motivation (ISM) were validated across Navajo ($n = 760$) and Anglo ($n = 1012$) students in the U.S. Confirmatory factor analysis (CFA) supported the 8-factor structure of motivation orientations for the total sample and the Navajo and Anglo subsamples, although Navajo students did not distinguish well between the Effort and Task constructs. However, of 39 survey items, only 30 items were invariant across groups in factor loadings on respective a priori constructs. The findings show that even though the ISM Motivation Orientation scales are applicable to students of different cultural backgrounds, meaningful cross-cultural comparisons should use the 30 items that mean the same to both cultural groups; whereas studies that do not involve cross-cultural comparisons may use the complete version of the scales.

All Prompts Are Created Equal, But Some Prompts Are More Equal than Others

Mark D. Shermis
Jeffrey Lee Rasmussen
D. W. Rajecki
Jennifer Olson
Cliford Marsiglio
Indiana University, Purdue University Indianapolis

Scores assigned to college placement essays by a computer program (PEG) showed high agreement with the evaluations of human readers (r = .82). Further, both types of graders tended to assign higher or lower scores to essays written about particular topics. Content analyses by a second program (MCCA) indicated that themes in essays varied in terms of emphasis on “analytic,” “emotional,” or “practical” dimensions. Human and machine readers tended to give higher scores for analytic and practical themes, and lower scores for those involving emotion. The ranks of mean prompt-related grades were concordant with the ranks of mean analytic and practical content across topics. Such findings call for the refined standardization of prompts for future testing, and the need for care in the evaluation of existing essays.

Measurement Issues in Screening Outstanding Teachers

Wen-Chung Wang
National Chung Cheng University
Ying-Yao Cheng
National Sun Yat-Sen University

Measurement issues that arise in a two-stage teacher evaluation for an outstanding faculty award are addressed. A teacher evaluation inventory with ten Likert-type items was developed. Thirty college teachers were rated by 293 students on the new inventory. The facet Rasch technique was applied to analyze the test data, and the items fit the Rasch models fairly well. The separation reliability for the teachers is .98, indicating that the items together with these students can differentiate the teachers extremely well. A cut score was set to .80 logits so that only those teachers with efficacy estimates above that level are eligible to apply for the award. A short version containing only half of the items was also developed.

Objective Standard Setting (or Truth in Advertising)

Gregory Ethan Stone
MetriKs Consulting Ltd.
Dental Assisting National Board, Inc.

Over the past 40 years criterion referencing has become the major method for setting passing standards on most high-stakes examinations. This paper serves three purposes. First it reveals the limitations and methodological weaknesses of most popular standard setting models, supported by a large volume of prior investigation. Second, it presents a radically different approach to the question of setting standards. The new model called Objective Standard Setting was developed in the early 1990s and has been successfully practiced in a variety of settings since that time. While presented in educational forums for many years, this paper represents the first published account of its methods. Thirdly, suggestions for new considerations of validity in standard setting are addressed. The report concludes with the suggestion that however well-intentioned popular standard setting efforts may be, psychometric experts must more carefully and fully understand the models behind their practices and the validity, meaning and implications of their thought processes.

Volume 2, Number 3

Congruence Between a Theoretical Continuum of Masculinity and the Rasch Model: Examining The Conformity to Masculine Norms Inventory

Larry H. Ludlow
James R. Mahalik
Boston College

The purpose of this study was to examine the psychometric structure of the Conformity to Masculine Norms Inventory (CMNI) in relation to the Rasch model. The CMNI was specifically constructed to measure a set of unidimensional constructs. As such, the items were intended to define a uniform spread of locations along each construct. The CMNI measures conformity to twelve masculine norms: winning, emotional control, risk-taking, violence, dominance, playboy, self-reliance, primacy of work, power over women, disdain for homosexuals, physical toughness, and pursuit of status. Three hundred forty-eight men participated in the study. In addition to examining global Rasch characteristics and the unidimensionality of each of the 12 scales, a detailed Rasch rating scale analysis is provided for the Violence Scale with unusual response patterns discussed in terms of their clinical usefulness. The results across all 12 scales reveal an excellent congruence between the theoretically derived construct of conformity to masculine norms and the theoretically defined objectives of the Rasch rating scale model.

Detecting Unexpected Variables in the MMPI–2 Social Introversion Scale

Chih-Hung Chang
Evanston Northwestern Healthcare
Northwestern University
Benjamin D. Wright
University of Chicago

The standard scoring structure of the revised Minnesota Multiphasic Personality Inventory (MMPI-2) Social Introversion (Si) scale was reexamined with Rasch Measurement. The 69-item Si scale split into two distinct dimensions when their standardized residuals were factor analyzed. Items keyed “true” to Si defined one dimension and items keyed “false” defined another. Relationships between Lexile values (an index of reading difficulty and comprehension) and item difficulties were also explored. The article shows how to use Rasch Measurement to understand and improve personality assessment.

A Cross-cultural Procedure to Assess Reliability and Measurement Invariance

Anil Mathur
Frank G. Zarb School of Business
Hofstra University
Benny Barak
Yong Zhang
Keun S. Lee
Hofstra University

Steenkamp and Baumgartner (1998) developed a procedure to assess measurement invariance across cultures. The study presented here applied their procedure to a scale to measure cognitive age (Barak, 1979; 1987; 1998; Barak and Schiffman, 1981) and relied on data collected in three Non-Western societies: India (*N*=195), China (*N*=250), and Korea (*N*=251). The results from a series of confirmatory factor analyses indicate that the technique provides a valuable tool to assess measurement invariance across cultures. The results further showed the cognitive age scale to be applicable in the three diverse cultures surveyed.

Detecting Differential Rater Functioning over Time (DRIFT) Using a Rasch Multi-Faceted Rating Scale Model

Edward W. Wolfe
Michigan State University
Bradley C. Moulder
University of Florida
Carol M. Myford
Educational Testing Service

This paper describes a class of rater effects that depict rater-by-time interactions. We refer to this class of rater effects as DRIFT—differential rater functioning over time. This article describes several types of DRIFT (primacy/recency, differential centrality/extremism, and practice/fatigue) and Rasch measurement procedures designed to identify these types of DRIFT in rating data. These procedures are applied to simulated data and are shown to be useful in classifying raters as being aberrant or non-aberrant for primacy, recency, and differential centrality and extremism, particularly for moderate or larger effect sizes. Rates of correct classification for practice and fatigue were lower and statistical power exceeded .50 only with very large effect sizes. Type I error rates (i.e., incorrect nomination) were near expected levels in all cases.

Evidence for the Reliability of Measures and Validity of Measure Interpretation: A Rasch Measurement Perspective

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

In an era of high stakes testing and evaluation in education, psychology, and health care, there is need for rigorous methods and standards for obtaining evidence of the reliability of measures and validity of inferences. Messick (1989, 1995), the Standard for Educational and Psychological Testing (American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, 1999), and the Medical Outcomes Trust (1995), among others, have described methods that may be used to gather evidence for reliability and validity, but ignored the potential role Rasch measurement may contribute to this process. This article will outline methods in Rasch measurement that are used to gather evidence for reliability and validity and attempt to articulate how these methods may be linked with the current views of reliability and validity.

Volume 2, Number 4

Toward Establishing a Unified Metric for Performance and Learning Goal Orientations

Everett V. Smith Jr.
University of Illinois at Chicago
Caroline Dupeyrat
Université Toulouse 2

Findings in the goal orientation literature are convoluted due to sample and scale dependent interpretations and potentially the use of inappropriate measurement models for ordinal data. The purpose of this investigation was to explore the possibility of developing a common metric in order to enhance communication among goal orientation researchers using different scales. A simultaneous calibration of four commonly used goal orientation scales was undertaken to investigate if item responses meet the unidimensionality requirement of Rasch models and hence provided a common metric. Participants were 305 third year psychology students at a large University in France. Results indicated that 30 of 36 learning goal and 28 of 33 performance goals items from the four scales defined unidimensional constructs. Correlations between person measures from the four scales and the simultaneous calibration for both learning goal and performance goals support the interpretation of unidimensional variables. Raw score to interval measure conversion tables provide goal orientation researchers using different scales a common metric for interpretation and statistical analyses. Limitations and extensions of this research are discussed.

Controlling for Rater Effects when Comparing Survey Items with Incomplete Likert Data

E. Matthew Schulz
Anji Sun
ACT, Inc.

The rating scale model (Andrich, 1978) was applied to data from a survey that directed students to rate their satisfaction with college services on a five point Likert scale. Because students used different services, and students were directed to rate only the services they used, the items were differentially exposed to a person factor that we call “pleasability.” Differential exposure to pleasability makes items’ average rating a biased measure of their performance. In contrast, item parameter estimates in the rating scale model corrected for differential exposure to pleasability. Compared to items’ average ratings, item parameter estimates in the rating scale model did a better job of predicting which item received the higher rating when any two items were rated by the same rater.

Polytomous Modeling of Cognitive Errors in Computer Adaptive Testing

LihShing Wang
University of Cincinnati
Chun-Shan Li
National Chung Cheng University

In the past two decades of psychometric research, an array of extended item response models has been proposed to capture the complex nature of human cognition. While the literature abounds in model fit analysis, the debate on model selection in different testing conditions continues. This study examines the problems of model selection in computer adaptive testing (CAT) of cognitive errors by comparing the relative measurement efficiency of polytomous modeling over dichotomous modeling under different scoring schemes and termination criteria. Monte Carlo simulation was adopted as the inquiry paradigm to generate 1000 subjects and 100 items in the calibration sample and 200 simulees in the CAT sample. The results suggest that polytomous CAT yields marginal gains over dichotomous CAT when termination criteria are more stringent (shorter test length or smaller standard error of ability estimate). When the conventional dichotomous scoring scheme is adopted, in which all partially correct answers are scored as incorrect, polytomous CAT cannot prevent the non-uniform gain in test information as was observed in paper and pencil testing.

Comparing Holistic and Analytic Scoring for Performance Assessment with Many-Facet Rasch Model

Eunlim Chi
Kyunghee University

This paper compares holistic and analytic scoring methods to explore how the alternative scorings can make differences for performance assessment using many-faceted Rasch model. The model is especially pertinent for analyzing performance assessment since the model can include several facets simultaneously. Forty three students’ reports for social studies were scored by four raters with the holistic method and the analytic method. The result demonstrated that scoring rubrics could be improved by investigating rating scale categories. Also, the comparison of student scores between the two scoring methods revealed that the selection of scoring methods might not be significant for the relative comparison of students but it could have serious implication for the assessment of students' absolute abilities. For rater severity, analytic scoring provided more consistency than holistic scoring. These findings can be used to select and improve scoring methods for performance assessment.

The Rasch Model, Additive Conjoint Measurement, and New Models of Probabilistic Measurement Theory

George Karabatsos
LSU Health Sciences Center

This research describes some of the similarities and differences between additive conjoint measurement (a type of fundamental measurement) and the Rasch model. It seems that there are many similarities between the two frameworks, however, their differences are nontrivial. For instance, while conjoint measurement specifies measurement scales using a data-free, non-numerical axiomatic frame of reference, the Rasch model specifies measurement scales using a numerical frame of reference that is, by definition, data dependent. In order to circumvent difficulties that can be realistically imposed by this data dependence, this research formalizes new non-parametric item response models. These models are probabilistic measurement theory models in the sense that they explicitly integrate the axiomatic ideas of measurement theory with the statistical ideas of orderrestricted inference and Markov Chain Monte Carlo. The specifications of these models are rather flexible, as they can represent any one of several models used in psychometrics, such as Mokken's (1971) monotone homogeneity model, Scheiblechner's (1995) isotonic ordinal probabilistic model, or the Rasch (1960) model. The proposed non-parametric item response models are applied to analyze both real and simulated data sets.