Journal of Applied Measurement Abstracts

Volume 3, Number 1 (2002)

A Confirmatory Study of Rasch-Based Optimal Categorization of a Rating Scale

Weimo Zhu
University of Illinois at Urbana-Champaign

The purpose of this study was to determine if the characteristics of the optimal categorization identified by the Rasch analysis in a previous study can be maintained when the revised scale is applied to the same population. Based on the results of the previous Rasch analysis, a 23-item exercise barrier scale was modified from its original five-category structure (*Very often* = 1, *Often* = 2, *Sometimes* = 3, *Rarely* = 4, and *Never* = 5) to a three-category structure (*Very often* = 1, *Sometimes* = 2, and *Never* = 3). The modified scale was then mailed to the original sample (*N* = 381), of which 206 returned the survey; a return rate 57.5%. The data was again analyzed using the Rasch Rating Scale model. Overall, the Rasch model fit data well and similar change patterns were observed in two category statistics provided by the Rasch analysis. The order of item severity was also well kept and the correlation of item severities generated from two studies was very high, with *r* = .98. In addition, similar results were also found in respondents' ability estimations, and the correlation between the two studies was moderately high, with *r* = .68. These results verified that the characteristics of the optimal categorization identified by the Rasch post-hoc analysis can be maintained after the original scale was modified based on such an analysis.

Dimensionality and Construct Validity of School Development Expectation Scale for Secondary Students

Mo Ching Magdalena Mok
The Hong Kong Institute of Education
Marcellin Flynn
The Australian Catholic University

This study aims to establish the dimensionality and construct validity of the School Development Expectation Scale for use with Year 12 students. The scale is made up of five expectation subscales in the vocational, academic, personal, social and religious development domains. The validation sample comprised 8,310 Year 12 students from 70 schools. Traditional confirmatory factor analysis supported the theory postulated five factor model. The scale was then calibrated using Rasch rating scale model. Recommendations were made regarding how to further refine the scale.

Item Grouping Effects on Invariance of Attitude Items

Catherine Frantom
University of Missouri, Columbia
Kathy E. Green
University of Denver
Tony C.M. Lam
University of Toronto

The purpose of this study was to evaluate the effects of item grouping on local independence and item invariance, the characteristics of items scaled under the Rasch model that make them sample-free. Item fit and calibration for attitude items presented in a grouped versus random order were examined. It was hypothesized that grouping items to facilitate interpretation central to a construct may result in a failure of invariance. Data were 107 responses to a 40-item mail survey of teachers' opinions about the Ontario Ministry's grade 9 literacy test. Effects of grouping

and item phrasing on invariance were found. Results, however, generally support the use of grouping of items to provide a higher person separation, and potentially higher quality data.

Level of Activity in Profound/Severe Mental Retardation (LAPMER): A Rasch-derived Scale of Disability

Luigi Tesio
Salvatore Maugeri Foundation, Pavia
Maria Rosa Valsecchi
Marina Sala
Paolo Guzzon
Fondazione Istituto Sacra Famiglia, Cesano Boscone-Milan
Mario Alberto Battaglia
Institute of Hygiene, University of Siena, Italy

Classification of Mental Retardation (MR) into severe and profound is based on IQ threshold (<35 and 20% respectively) and on quite generic descriptions of deficits in adaptive behavior. The LAPMER scale (after Level of Activity in Profound/severe Mental Retardation) was developed as a measure of severity through observed behavior in adult patients. The Rasch analysis (RA, in its rating scale model) was adopted as a guide for selection of items, conceptualization of item levels, and validation of the overall instrument. The RA provides estimates on a continuum measure corresponding to the discrete cumulative score. A model prescribes the expected scores on each subject-item interaction. Discrepancies between observed and expected scores allow diagnostic procedures on coherence (fit) of both subjects and items. The final version included 8 items: Feeding, Sphincters, Communication, Manipulation, Dressing, Locomotion, Spatial Orientation and Praxiae, scored 0/1 or 0/1/2 (cumulative range for the total set of items was 0-13) the higher the score, the better the performance. The test can be administered in 15 minutes through observation or inquiry from proxies and personnel. A psychologist rated 146 permanent hosts of a large Institute for mentally retarded adults (51 profound and 95 severe, 91 male, age 18-63, median 36). Median score was 6/13, IQR 1-9, range 0- 12, 19% of cases scored 0. Cronbach's α for internal consistency was 0.90. Fifty-seven patients were also independently scored by another psychologist. Between-rater Cohen's κ reliability index ranged from 0.77-0.96 across items. Median raw scores were 1 and 8 in profound and severe cases, respectively (p<0.001). Rasch person reliability coefficient, a 0 to 1 index of internal consistency analogous to Crohnbach's α, was 0.92. For each item the standardized differences between observed and model-expected scores (residuals) were χ^2 tested (α level 0.05) across sub-groups of patients. These were: profound vs. severe cases, and classes of motor impairment (tetra-,hemi-,para-plegic and unimpaired), matched for overall ability measure. For 6 items some residuals were found to be statistically significant. Absolute differences ranged from 0 to 0.7 raw score points, with no systematic patterns. Gender, age group and rater did not bias the measure. Residuals did not correlate meaningfully across pairs of items (r<|0.5|), further supporting the unidimensionality of the measure. The scale seems a valid tool for classification of adult severe and profound MR cases.

Optimizing Rating Scale Category Effectiveness

John M. Linacre
University of Chicago

Rating scales are employed as a means of extracting more information out of an item than would be obtained from a mere "yes/no", "right/wrong" or other dichotomy. But does this additional information increase measurement accuracy and precision? Eight guidelines are suggested to aid the analyst in optimizing the manner in which rating scales categories cooperate in order to improve the utility of the resultant measures. Though these guidelines are presented within the context of Rasch analysis, they reflect aspects of rating scale functioning which impact all methods of analysis. The guidelines feature rating-scale based data such as category frequency, ordering, rating-to-measure inferential coherence, and the quality of the scale from measurement and statistical perspectives. The

manner in which the guidelines prompt recategorization or reconceptualization of the rating scale is indicated. Utilization of the guidelines is illustrated through their application to two published data sets.

Volume 3, Number 2

An Eigenvector Method for Estimating Item Parameters of the Dichotomous and Polytomous Rasch Models

Mary Garner
Kennesaw State University
George Engelhard, Jr.
Emory University

The purpose of this paper is to describe a technique for obtaining item parameters of the Rasch model, a technique in which the item parameters are extracted from the eigenvector of a matrix derived from comparisons between pairs of items. The technique can be applied to both dichotomous and polytomous data. In application to a previously published data set, it is shown that the technique provides item parameter estimates comparable to those produced by joint maximum likelihood estimation, and for the most difficult items, the technique appears to produce superior estimates. This method has several advantages. It easily accommodates missing data, and makes transparent the basis for item parameter estimation in the presence of missing data. Furthermore, the method provides a link to other methods in the social sciences and, in particular, provides the framework for application of graph theory to the analysis of assessment networks. Finally, it exploits several characteristics that are unique to the Rasch model.

Two Strategies for Fitting Real Data to Rasch Polytomous Models

Antonio J. Rojas Tejada
University of Almería
Andrés González Gómez
José L. Padilla García
Cristino Pérez Meléndez
University of Granada

A comparative study of the results provided by two strategies for fitting data to Latent Trait Theory Models has been performed. The first, called Total-Persons-Items (TPI), is structured in three phases: 1) assessment of item fit, 2) assessment of person fit; and finally, 3) overall fit of data to the models (items and persons). The second strategy, the Total-Items-Persons (TIP), changes the order of the phases: 1) assessment of person fit, 2) assessment of item fit and, 3) overall fit of data to the models. To verify the results of these two strategies, a set of 30 items, designed to measure religious attitude, was administered to a sample of 821 persons. The latent trait theory models used were the partial credit model and the rating scale model. The results underline an important difference between the two procedures: the TPI maximizes the number of persons with good fit and the TIP maximizes the number of items with good fit. Moreover, a procedure for controlling the sensitivity of fit to sample size is proposed.

A Comparison of Three Developmental Stage Scoring Systems

Theo Linda Dawson
University of California at Berkeley

In social psychological research the stage metaphor has fallen into disfavor due to concerns about bias, reliability, and validity. To address some of these issues, I employ a multidimensional partial credit analysis comparing moral

judgment interviews scored with the Standard Issue Scoring System (SISS) (Colby and Kohlberg, 1987b), evaluative reasoning interviews scored with the Good Life Scoring System (GLSS) (Armon, 1984b), and Good Education interviews scored with the Hierarchical Complexity Scoring System (HCSS) (Commons, Danaher, Miller, and Dawson, 2000). A total of 209 participants between the ages of 5 and 86 were interviewed. The multidimensional model reveals that even though the scoring systems rely upon different criteria and the data were collected using different methods and scored by different teams of raters, the SISS, GLSS, and HCSS all appear to measure the same latent variable. The HCSS exhibits more internal consistency than the SISS and GLSS, and solves some methodological problems introduced by the content dependency of the SISS and GLSS. These results and their implications are elaborated.

Development of a Functional Movement Scale for Infants

Suzann K. Campbell
University of Illinois at Chicago
Benjamin D. Wright
J. Michael Linacre
University of Chicago

The increasing survival rate of infants with a complicated birth and perinatal history generated the need for a test of functional motor performance with the capability of identifying children under four months of age with delayed development which could be addressed with physical therapy. This paper describes a Rasch analysis of the psychometric qualities of the Test of Infant Motor Performance (TIMP) for the purpose of reducing the length of the test while maintaining its precision as a measurement device. Following analysis of fit statistics, item-to-total correlations, redundancy of item difficulty measures, and consideration of clinically-relevant features of test items from analysis of 1732 tests, the TIMP was reduced from 59 to 42 items forming a functional motor scale for prematurely born infants. The resulting person separation index was 4.85 and the item separation index was 23.79.

Detecting and Evaluating the Impact of Multidimensionality using Item Fit Statistics and Principal Component Analysis of Residuals

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

The purpose of this research is twofold. First is to extend the work of Smith (1992, 1996) and Smith and Miao (1991, 1994) in comparing item fit statistics and principal component analysis as tools for assessing the unidimensionality requirement of Rasch models. Second is to demonstrate methods to explore how violations of the unidimensionality requirement influence person measurement. For the first study, rating scale data were simulated to represent varying degrees of multidimensionality and the proportion of items contributing to each component. The second study used responses to a 24 item Attention Deficit Hyperactivity Disorder scale obtained from 317 college undergraduates. The simulation study reveals both an iterative item fit approach and principal component analysis of standardized residuals are effective in detecting items simulated to contribute to multidimensionality. The methods presented in Study 2 demonstrate the potential impact of multidimensionality on norm and criterion-reference person measure interpretations. The results provide researchers with quantitative information to help assist with the qualitative judgment as to whether the impact of multidimensionality is severe enough to warrant removing items from the analysis.

Volume 3, Number 3

Test Scores, Measurement, and the Use of Analysis of Variance: An Historical Overview

Joseph Romanoski

Graham Douglas
University of Western Australia

In order to establish a firmer statistical foundation from which to draw inferences from factorial design study data, transformations of raw scores are occasionally employed in order to make their distributions more generally normal or to provide linearity. To date, few studies have been conducted to determine whether or not raw scores—transformed or otherwise—constitute measures for the purposes of statistical analysis. In this article, the historical development of the understanding of the term "measurement" by researchers in the social sciences is traced, and the development and use of One and Two-way ANOVA by researchers in the social sciences are presented and evaluated.

Using Rasch Measurement to Investigate the Cross-form Equivalence and Clinical Utility of Spanish and English Versions of a Diabetes Questionnaire: A Pilot Study

Ben Gerber
Everett V. Smith, Jr.
Mariela Girotti
Lourdes Pelaez
Kimberly Lawless
Louanne Smolin
Irwin Brodsky
Arnold Eiser
University of Illinois at Chicago

The purpose of this research was to use Rasch measurement to study the psychometric properties of data obtained from a newly developed Diabetes Questionnaire designed to measure diabetes knowledge, attitudes, and self-care. Specifically, a methodology using principles of Rasch measurement for investigating the cross-form equivalence of English and Spanish versions of the Diabetes Questionnaire was employed. A total of fifty diabetes patients responded to the questionnaire, with 26 participants completing the English version. Analyses detected problems with the attitude items. We attributed the scaling problems to the use of negatively worded items with participants having generally low educational backgrounds. Analysis of the knowledge and self-care items yielded unidimensional variables with clinically meaningful item hierarchies that may have relevance to treatment protocols. Furthermore, the knowledge and the self-care items from the two versions of the Diabetes Questionnaire met our criteria for establishing cross-form equivalence and thus allow quantitative comparisons of person measures across versions. Limitations of the study and suggested refinements of the Diabetes Questionnaire are discussed.

Moving the Cut Score on Rasch Scored Tests

G. Edward Miller
Texas Education Agency
S. Natasha Beretvas
University of Texas at Austin

Empirically based item selection guidelines are presented for moving the cut score on equated tests consisting of n dichotomous items calibrated assuming the Rasch model. The cut score on a test form B, c_B, may be made higher than test form A's cut score, c_A, in the following ways: (1) select items for test form B such that the variance of test form B's item difficulties, σ^2_B, will be equal to test form A's, σ^2_A, but test form B's mean item difficulty, μ_B, will be less that of test form A, μ_A; (2) given $c_A > n/2$, select items for test form B such that $\mu_B < \mu_A$, and $\sigma^2_B < \sigma^2_A$; (3) given $c_A < n/2$, select items for test form B such that $\mu_B < \mu_A$ and $\sigma^2_B > \sigma^2_A$. To make c_B lower than c_A, the direction of the changes listed above for the two tests' item difficulties' σ^2 and μ should be reversed. Derivations of lemmas that underlie the guidelines are provided as well as a simulated example.

Examining Item Difficulty and Response Time on Perceptual Ability Test Items

Chien-Lin Yang
American Dental Association
Thomas R. O'Neill
National Council of State Boards of Nursing
Gene A. Kramer
American Dental Association

This study examined item calibration stability in relation to response time and the levels of item difficulty between different response time groups on a sample of 389 examinees responding to six different subtest items of the Perceptual Ability Test (PAT). The results indicated that no Differential Item Functioning (DIF) was found and a significant correlation coefficient of item difficulty was formed between slow and fast responders. Three distinct levels of difficulty emerged among the six subtests across groups. Slow responders spent significantly more time than fast responders on the four most difficult subtests. A positive significant relationship was found between item difficulty and response time across groups on the overall perceptual ability test items. Overall, this study found that: 1) the same underlying construct is being measured across groups, 2) the PAT scores were equally useful across groups, 3) different sources of item difficulty may exist among the six subtests, and 4) more difficult test items may require more time to answer.

When Raters Disagree, Then What: Examining a Third-rating Discrepancy Resolution Procedure and Its Utility for Identifying Unusual Patterns of Ratings

Carol M. Myford
Educational Testing Service
Edward W. Wolfe
Michigan State University

The purpose of this study was to examine a procedure for identifying and resolving discrepancies in ratings. We sought to determine to what extent the third-rater adjudication procedure employed in scoring the Test of Spoken English (TSE) successfully identified all anomalous ratings. We analyzed data from the April 1997 TSE scoring session using FACETS, a rating scale analysis computer program. The results suggest that, while it is important for an assessment program to identify cases in which there is obvious disagreement in the ratings assigned and have a policy to resolve those disagreements, implementing a discrepancy resolution procedure is not sufficient in and of itself for quality control monitoring. Often times, there are other anomalous ratings that discrepancy resolution procedures may miss. Fit analysis can provide a valuable adjunct to a discrepancy resolution procedure, flagging suspect rating profiles in need of expert review before a final score report is issued.

Understanding Resistance to the Data-model Relationship in Rasch's Paradigm: A Reflection for the Next Generation

David Andrich
Murdoch University

The case for the Rasch models, that the comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; … and vice versa (Rasch, 1961), does not depend on the models accounting for any data set. This has two distinctive consequences on the data-model relationship for the Rasch models. First, and this was recognized by Rasch, when there are deviations of one sort or another, it turns upside down the question of whether it is the model or the test that has gone wrong (Rasch, 1960). Second, because the invariance of comparisons among stimuli, and vice versa, is built into the model rather than being merely a

requirement of data, further implications of this requirement can be derived mathematically. These implications, too, inevitably turn some questions, and their solutions, upside down. It is argued that having to look at these implications upside down produces substantial psychological and intellectual resistance amongst those schooled in looking at them in the traditional way. It is also argued that in turning the question upside down, Rasch had an insight that goes beyond the mathematical derivations, and that to sustain this insight requires a paradigm shift (Kuhn, 1970) in the data-model relationship. Using an illustrative example, it is suggested that to maintain this paradigm shift, even by those who research the Rasch models, requires the same uncompromising consistency and passion that Rasch displayed in maintaining faith in his insight.

Volume 3, Number 4

A Multi-factor Rasch Scale for Artistic Judgment

Nikolaus Bezruczko

Measurement properties are reported for a combined scale of abstract and figurative artistic judgment aptitude items. Abstract items are synthetic, rule-based images from Visual Designs Test which implements a statistical algorithm to control design complexity and redundancy, and figurative items are canvas paintings in five styles, Fauvism, Post- Impressionism, Surrealism, Renaissance, and Baroque especially created for this research. The paintings integrate syntactic structure from VDT Abstract designs with thematic content for each style at four levels of complexity while controlling redundancy. Trained test administrators collected preference for synthetic abstract designs and authentic figurative art from 462 examinees in Johnson O'Connor Research Foundation testing offices in Boston, New York, Chicago, and Dallas. The Rasch model replicated measurement properties for VDT Abstract items and identified an item hierarchy that was statistically invariant between genders and generally stable across age for new, authentic figurative items. Further examination of the figurative item hierarchy revealed that complexity interacts with style and meaning. Sound measurement properties for a combined VDT Abstract and Figurative scale shows promise for a comprehensive artistic judgment construct.

Establishing Longitudinal Factorial Construct Validity of the Quality of School Life Scale for Secondary Students

Magdalena Mo Ching Mok
The Hong Kong Institute of Education
Marcellin Flynn
Australian Catholic University

The purpose of this study was to establish the longitudinal factorial construct validity of the Quality of School Life (QSL) scale, which was initially designed to measure the wellbeing of students in Australian high schools. The items comprising the scale were based on theoretical models and existing measurement instruments concerning the domains of schooling and the quality of life experienced by adults. Three latent structure models, two reported earlier in the literature and one new in this context, were reviewed and their merits compared on the basis of two sets of empirical data comprising, respectively, 5932 secondary students in the 1993 cohort and 8269 secondary students in the 1999 cohort using confirmatory factor analysis procedures. Results indicate that there is strong evidence in support of the factorial construct validity of the Quality of School Life scale.

Rasch-transformed Raw Scores and Two-way ANOVA: A Simulation Analysis

Joseph Romanoski
Graham Douglas
University of Western Australia

This article demonstrates, through the exposition of underestimated variable effects or spurious interaction effects, the inherent inadequacy of untransformed 0-1 raw scores for analysis via Two-Way analysis of variance (ANOVA). The pioneering work in this field, conducted in the 1990's by Dr. Susan Embretson of the University of Kansas, USA, is highlighted, and the eminent suitability of Rasch transformations of 0-1 raw scores for analysis via Two-Way ANOVA is also demonstrated. In this study Monte Carlo techniques or simulations are utilized to determine the precise psychometric conditions under which differences between raw scores and Rasch transformations of those raw scores are detectable via Two-Way ANOVA. This study partially replicates Dr. Embretson's studies, and also defines the extent of underestimation and spuriousness which ensue when uniform or skewed distributions of item difficulties are used instead of normal distributions, and misfitting raw data are utilized instead of fitting data.

Development of Measurability and Importance Scales for the NATA Athletic Training Educational Competencies

Edward W. Wolfe
Sally Nogle
Michigan State University

A recent mandate issued by the National Athletic Trainers' Association Board of Certification (NATABOC) stated that beginning in 2004 a student must graduate from an accredited Commission on Accreditation of Allied Health Education Programs (CAAHEP) in order to qualify for the NATABOC exam. The content of this exam is based on the National Athletic Trainers' Association (NATA) Athletic Training Educational Competencies. These 542 competencies in 12 different domains were developed through role delineation studies with the most recent edition published in 1999. Therefore, these competencies must be included in each athletic training curriculum program across the country in order to prepare their students for certification and to achieve program accreditation. Concern over the large number of competencies to be attained within the educational time frame created the need to develop an instrument to examine this issue. In response, instruments were developed to examine one domain of the NATA Educational Competencies. Specifically, the General Medical Conditions and Disabilities competencies were assessed for their perceived importance and measurability by certified athletic trainers and sports medicine physicians. This article reports the results of a validation study of an instrument designed to measure the perceived measurability and importance of the NATA Athletic Training Educational Competencies. Generally, the results are encouraging. The data supports six constructs, and each of these constructs exhibits high reliabilities. Relative competency calibrations within and between scales were consistent with theory. And, ratings assigned by different groups of trainers were comparable.

Measuring Leader Perceptions of School Readiness for Reforms: Use of an Iterative Model Combining Classical and Rasch Methods

Madhabi Chatterji
Teachers College, Columbia University

This study examines validity of data generated by the School Readiness for Reforms: Leader Questionnaire (SRR-LQ) using an iterative procedure that combines classical and Rasch rating scale analysis. Following content-validation and pilot-testing, principal axis factor extraction and promax rotation of factors yielded a five factor structure consistent with the content-validated subscales of the original instrument. Factors were identified based on inspection of pattern and structure coefficients. The rotated factor pattern, interfactor correlations, convergent validity coefficients, and Cronbach's alpha reliability estimates supported the hypothesized construct properties. To further examine unidimensionality and efficacy of the rating scale structures, item-level data from each factor-defined subscale were subjected to analysis with the Rasch rating scale model. Data-to-model fit statistics and separation reliability for items and persons met acceptable criteria. Rating scale results suggested consistency of expected and observed step difficulties in rating categories, and correspondence of step calibrations with increases in the underlying variables. The combined approach yielded more comprehensive diagnostic information on the quality of the five SRR-LQ subscales; further research is continuing.

Construction of Measures from Many-facet Data

John M. Linacre
Benjamin D. Wright
University of Chicago

An extension to the Rasch model for fundamental measurement is described in which there is parameterization not only for examinee ability and item difficulty but also for judge severity. Variants of this model are discussed and judging plans reviewed. Its use and characteristics are explained by an application of the model to an empirical testing situation. A comparison with Generalizability Theory using a common data set is presented as a contrast in approaches to resolving judge indeterminacy.