Journal of Applied Measurement Abstracts

Volume 4, Number 1 (2003)

The Effect of Missing Data on Estimating a Respondent's Location using Ratings Data

R. J. De Ayala
University of Nebraska-Lincoln

In social science research there are a number of instruments that utilize a rating scale such as a Likert response scale. For a number of reasons a respondent's response vector may not contain responses to each item. This study investigated the effect on a respondent's location estimate when a respondent is presented an item, has ample time to answer the item, but decides to not respond to the item. For these situations different strategies have been developed for handling missing data. In this study, four different approaches for handling missing data were investigated for their capability to mitigate against the effect of omitted responses on person location estimation. These methods included ignoring the omitted response, selecting the "midpoint" response category, hot-decking, and a Likelihood-based approach. A Monte Carlo study was performed and the effect of different levels of omissions on the simulees' location estimates was determined. Results showed that the hot-decking procedure performed the best of methods examined. Implications for practitioners were discussed.

Rasch Simultaneous Vertical Equating for Measuring Reading Growth

Ong Kim Lee
National Institute of Education
Nanyang Technological University

The longitudinal measurement of ability growth requires that the measures that are taken at the various time points be obtained using the same yardstick. Tests given for this purpose, therefore, need to be equated. The popular practice still in use for the purpose of equating tests, is the grade-equivalent. This paper compares the observation of children's growth in reading using grade-equivalents with that using Rasch Simultaneous Vertical Equating procedure. It is found that grade equivalents differ much more between two different test forms compared to ability measures obtained using Rasch Simultaneous Vertical Equating. It is also found that the spread of students' grade-equivalents, increased over the years as they grow while the standard deviations of their Rasch measures remain relatively constant over the same period of time. Student responses to the ITBS Form 7 and the CPS90 and CPS 91 were used. A total of 5,623 students were tracked over eight years.

An Examination of Exposure Control and Content Balancing Restrictions on Item Selection in CATs using the Partial Credit Model

Laurie Laughlin Davis
Pearson Educational Measurement
Dena A. Pastor
James Madison University
Barbara G. Dodd
The University of Texas at Austin
Claire Chiang
Brown University
Steven J. Fitzpatrick
Pearson Educational Measurement

The purpose of the present investigation was to systematically examine the effectiveness of the Sympson-Hetter technique and rotated content balancing relative to no exposure control and no content rotation conditions in a computerized adaptive testing system (CAT) based on the partial credit model. A series of simulated fixed and variable length CATs were run using two data sets generated to multiple content areas for three sizes of item pools. The 2 (exposure control) X 2 (content rotation) X 2 (test length) X 3 (item pool size) X 2 (data sets) yielded a total of 48 conditions. Results show that while both procedures can be used with no deleterious effect on measurement precision, the gains in exposure control, pool utilization, and item overlap appear quite modest. Difficulties involved with setting the exposure control parameters in small item pools make questionable the utility of the Sympson-Hetter technique with similar item pools.

Aerobic Exercise Equipment Preferences Among Older Adults: A Preliminary Investigation

Marilyn A. Looney
Northern Illinois University
James H. Rimmer
University of Illinois at Chicago

The purpose of this study was to develop an instrument that would measure the aerobic exercise equipment preferences of a frail older population and to see, despite a small sample size, if a many-facet Rasch analysis would provide useful information concerning the construct validity of the instrument and the equipment preferences of the sample. Sixteen ambulatory seniors (M = 82.0 yr + 6.6; 4 males & 12 females), who resided in a local retirement community and were involved in a structured fitness program, evaluated the following exercise equipment: Schwinn Air-Dyne, Nu-Step Recumbent Stepper; Monark bicycle ergometer; Stairmaster; and PTS Turbo Recumbent Bicycle. Participants used the equipment for 5 min. and then completed the survey via a structured interview technique. Test-retest reliability coefficients indicated the participants' responses were stable across days for each piece of exercise equipment (proportions of agreement > .83; km > .77). A many-facet (equipment, items, participants) Rasch analysis verified that 12 closed format items defined a linear construct of equipment preference (separation = 1.8; reliability = .77). The pieces of equipment were placed on the linear continuum according to their equipment preference measures (separation = 3.21; reliability = .91) derived from the participants' response patterns to the items (separation = 1.43; reliability = .67). Although the MNSQ infit and outfit statistics were acceptable for each facet, the items did not target the equipment very well. Suggested changes to the instrument include converting questions to statements to use with Likert response categories; converting negative wording to positive phrasing, and adding items related to seat comfort, foot pedal placement, and visibility of display panel. The Nu-Step and Schwinn were the most preferred pieces of equipment while the Stairmaster was the least preferred . This preliminary investigation illustrates how useful information can be obtained from a many-facet Rasch analysis to guide instrument revision and better understand exercise equipment preferences among a frail older population.

Measurement Precision of the Clinician Administered PTSD Scale (CAPS): A RASCH Model Analysis

Elizabeth J. Betemps
University of Cincinnati
Richard M. Smith
Data Recognition Corporation
Dewleen G. Baker
University of Cincinnati Medical Center
Barbara A. Rounds-Kugler
Cincinnati Veterans Affairs Medical Center

The Clinician Administered PTSD Scale (CAPS), originally developed as a diagnostic tool, is frequently used to evaluate treatment responses. Defining a case and measuring symptom changes are different processes that require different attributes for the instrument. Measuring symptom changes requires precision in measurement. Using the Rasch rating scale model, we evaluated this instrument for construct validity in a veteran sample. The distribution of the veteran measures did not align with the distribution of the item measures in the CAPS instrument. Separate analysis of the CAPS Frequency subscale and Intensity subscale were conducted. The Frequency subscale produced measures that encompassed the level of severity found in the veteran sample. Items from this instrument can be used to develop an equal interval scale to provide precise measurements for treatment evaluations and to identify clinical cut points for diagnostic purposes.

A Comparative Evaluation of Methods of Adjusting GPA for Differences in Grade Assignment Practices

Pui-Wa Lei
Pennsylvania State University
Dina Bassiri
E. Matthew Schulz
ACT, Inc.

Numerous methods have been proposed for constructing an adjusted grade point average (adjusted-GPA) that controls for differences in grading standards across college courses and departments. Compared to the raw GPA, adjusted-GPA measures are generally more predictable from preadmissions variables, such as standardized tests and high school achievement. Relative rankings of students on adjusted-GPA measures are also more consistent with their relative standings within courses. This study compared the performance of 4 polytomous IRT and 3 linear models for constructing adjusted-GPA measures. Unlike previous studies, the regression weights of predictor variables and the course parameter estimates used to compute adjusted-GPA were cross validated. Adjusted-GPA retained noticeable advantages over raw GPA on cross-validation. The largest advantages were seen in the multiple correlation of adjusted-GPA with preadmission variables, when adjusted- GPA was constructed with the rating scale and partial credit IRT models. The cross-validity of adjusted-GPA was the weakest with the graded response model.

An Introduction to Multidimensional Measurement using Rasch Models

Derek C. Briggs
Mark Wilson
University of California, Berkeley

The act of constructing a measure requires a number of important assumptions. Principle among these assumptions is that the construct is unidimensional. In practice there are many instances when the assumption of unidimensionality does not hold, and where the application of a multidimensional measurement model is both technically appropriate and substantively advantageous. In this paper we illustrate the usefulness of a multidimensional approach to measurement with the Multidimensional Random Coefficient Multinomial Logit (MRCML) model, an extension of the unidimensional Rasch model. An empirical example is taken from a collection of embedded assessments administered to 541 students enrolled in middle school science classes with a hands-on science curriculum. Student achievement on these assessments are multidimensional in nature, but can also be treated as consecutive unidimensional estimates, or as is most common, as a composite unidimensional estimate. Structural parameters are estimated for each model using ConQuest, and model fit is compared. Student achievement in science is also compared across models. The multidimensional approach has the best fit to the data, and provides more reliable estimates of student achievement than under the consecutive unidimensional approach. Finally, at an interpretational level, the multidimensional approach may well provide richer information to the classroom teacher about the nature of student achievement.

Volume 4, Number 2

Maximum Information Approach to Scale Description for Affective Measures Based on the Rasch Model

Huynh Huynh
J. Patrick Meyer
University of South Carolina

Using the Rasch model for ordered categories, this paper provides a method for qualitative interpretations of data from an affective measure such as an attitude scale or survey instrument. The Bock procedure is first used to partition the total item information to each response category. The location of each response category is determined by maximizing the information associated with this category. The response categories that cluster around a given point on the latent trait are then used to provide a qualitative description of this point. The description is explicit in terms of the behaviors or activities that are likely to be displayed by a respondent at this point. An illustration is provided using an alumni survey used at a large university in the southeast.

Measuring Client Satisfaction with Public Education I: Meeting Competing Demands in Establishing State-wide Benchmarks

John A. King
Trevor G. Bond
James Cook University

By its very nature, a large-scale evaluation of client satisfaction with public education using a quantitative approach, places almost impossibly competing demands on the research methodology. This paper reports on the use of a suite of Rasch measurement techniques to meet the competing demands in establishing statewide benchmarks relating to the School Opinion Survey carried out over 1200 government schools in one state of Australia. Although the evaluation had to establish system-wide representative parent and student benchmarks, meaningful quantitative estimates of client satisfaction had to be provided at the smallest public schools. The final 20-item School Opinion Survey Parent and Student Forms were designed following feedback from the administration of trial forms. Instrument development was monitored by the results of Rasch modeling. The Rasch modeling property of specific objectivity was empirically verified when calculation of identical benchmark estimates resulted from the construction of simulated population proportional samples using sample-population size weightings.

Developing an Initial Physical Function Item Bank from Existing Sources

Rita K.Bode 1, 2, 4
David Cella 2, 3
Jin-shei Lai 2, 3
Allen W. Heinemann 1, 2, 4
1 *Rehabilitation Institute of Chicago*
2 *Institute of Health Services Research & Policy Studies, Feinberg School of Medicine, Northwestern University*
3 *Evanston Northwestern Healthcare*
4 *Department of Physical Medicine and Rehabilitation, Feinberg School of Medicine, Northwestern University*

The objective of this article is to illustrate incremental item banking using health-related quality of life data collected from two samples of patients receiving cancer treatment. The kinds of decisions one faces in establishing an item bank for computerized adaptive testing are also illustrated. Pre-calibration procedures include: identifying common items across databases; creating a new database with data from each pool; reversescoring "negative" items; identifying rating scales used in items; identifying pivot points in each rating scale; pivot anchoring items at comparable rating scale categories; and identifying items in each instrument that measure the construct of interest. A series of calibrations were conducted in which a small proportion of new items were added to the common core and misfitting items were identified and deleted until an initial item bank has been developed.

Breakthrough Measuring Neighborhoods

Nikolaus Bezruczko

An empirical strategy is presented for transforming ordinal counts and percentages to interval scale measures by recoding them as ordered categories and estimating Rasch model rating scale parameters. This strategy is demonstrated for a neighborhood construct socioeconomic disadvantage operationally defined by eight characteristics of Chicago neighborhoods (N = 77). Results show surprisingly sound model fit and satisfactory scale invariance between 1980 and 1990 census. A striking finding obscured by traditional methods is many Chicago neighborhoods are four times more disadvantaged than official U.S. poverty threshold. Intramodel construct validation confirms this scale structure is consistent with sociological expectations about property values, income, and race. A general benefit of this approach over conventional categorical socioeconomic indices is neighborhood measurement on a linear scale.

Rasch Fit Statistics as a Test of the Invariance of Item Parameter Estimates

Richard M. Smith
Data Recognition Corporation
Kyunghee K. Suh
American Institutes for Research

The invariance of the estimated parameters across variation in the incidental parameters of a sample is one of the most important properties of Rasch measurement models. This is the property that allows the equating of test forms and the use of computer adaptive testing. It necessarily follows that in Rasch models if the data fit the model, than the estimation of the parameter of interest must be invariant across sub-samples of the items or persons. This study investigates the degree to which the INFIT and OUTFIT item fit statistics in WINSTEPS detect violations of the invariance property of Rasch measurement models. The test in this study is a 80 item multiple-choice test used to assess mathematics competency. The WINSTEPS analysis of the dichotomous results, based on a sample of 2000 from a very large number of students who took the exam, indicated that only 7 of the 80 items misfit using the 1.3 mean square criteria advocated by Linacre and Wright. Subsequent calibration of separate samples of 1,000 students from the upper and lower third of the person raw score distribution, followed by a t-test comparison of the item calibrations, indicated that the item difficulties for 60 of the 80 items were more than 2 standard errors apart. The separate calibration t-values ranged from +21.00 to –7.00 with the t-test value of 41 of the 80 comparisons either larger than +5 or smaller than -5. Clearly these data do not exhibit the invariance of the item parameters expected if the data fit the model. Yet the INFIT and OUTFIT mean squares are completely insensitive to the lack of invariance in the item parameters. If the OUTFIT ZSTD from WINSTEPS was used with a critical value of $| t | > 2.0$, then 56 of the 60 items identified by the separate calibration t-test would be identified as misfitting. A fourth measure of misfit, the between ability-group item fit statistic identified 69 items as misfitting when a critical value of $t > 2.0$ was used. Clearly relying solely on the INFIT and OUTFIT mean squares in WINSETPS to assess the fit of the data to the model would cause one to miss one of the most important threats to the usefulness of the measurement model.

Measuring Attitudes and Behaviors to Studying and Learning for University Students:

A Rasch Measurement Model Analysis

Russell F. Waugh
Edith Cowan University

A Studying and Learning Scale was created using a model of Motivation (sets of ordered stem-items based on Striving for Excellence, Desire to Learn and Personal Incentives), with each item answered from three selfreported perspectives (an Ideal Self-view, a Capability Self-view, and a Studying and Learning Self-view). The response categories were the number of subjects studied. The stem-item sample was 23, each answered in three aspects, so each stem-item had three 'difficulties', making an effective item sample of 69. The person convenience sample was 372 students in education at an Australian university. The 69 items fit a Rasch measurement model and formed a scale in which the 'difficulties' of the items were ordered from 'easy' to 'hard' and the student measures of Studying and Learning were ordered from 'low' to 'high'. The person separation reliability was high at 0.94. The response categories were answered consistently and logically and the results supported many (but not all) of the conceptually ordered-by-difficulty item patterns. Students found it 'easy' to form a high view of How they would like to be, much 'harder' to form a high view of What they think they are capable of doing and even 'harder' to perform, at a high level, their Studying and Learning behavior for all stem-items, in accordance with the model.

Rasch Techniques for Detecting Bias in Performance Assessments: An Example Comparing the Performance of Native and Non-native Speakers on a Test of Academic English

Catherine Elder
University of Auckland
Tim McNamara
University of Melbourne
Peter Congdon
Victorian Curriculum and Assessment Authority

The use of common tasks and rating procedures when assessing the communicative skills of students from highly diverse linguistic and cultural backgrounds poses particular measurement challenges, which have thus far received little research attention. If assessment tasks or criteria are found to function differentially for particular subpopulations within a test candidature with the same or a similar level of criterion ability, then the test is open to charges of bias in favor of one or other group. While there have been numerous studies involving dichotomous language test items (see e.g. Chen and Henning, 1985 and more recently Elder, 1996) few studies have considered the issue of bias in relation to performance based tasks which are assessed subjectively, via analytic and holistic rating scales. The paper demonstrates how Rasch analytic procedures can be applied to the investigation of item bias or differential item functioning (DIF) in both dichotomous and scalar items on a test of English for academic purposes. The data were gathered from a pilot English language test administered to a representative sample of undergraduate students (N= 139) enrolled in their first year of study at an English-medium university. The sample included native speakers of English who had completed up to 12 years of secondary schooling in their first language (L1) and immigrant students, mainly from Asian language backgrounds, with varying degrees of prior English language instruction and exposure. The purpose of the test was to diagnose the academic English needs of incoming undergraduates so that additional support could be offered to those deemed at risk of failure in their university study. Some of the tasks included in the assessment procedure involved objectively-scored items (measuring vocabulary knowledge, text-editing skills and reading and listening comprehension) whereas others (i.e. a report and an argumentative writing task) were subjectively-scored. The study models a methodology for estimating bias with both dichotomous and scalar items using the programs Quest (Adams and Khoo, 1993) for the former and ConQuest (Wu, Adams and Wilson, 1998) for the latter. It also offers answers to the practical questions of whether a common set of assessment criteria can, in an academic context such as this one, be meaningfully applied to all subgroups within the candidature and whether analytic criteria are more susceptible to biased ratings than holistic ones. Implications for test fairness and test validity are discussed.

Volume 4, Number 3

Conditional Pairwise Estimation in the Rasch Model for Ordered Response Categories using Principal Components

David Andrich
Guanzhong Luo
Murdoch University

In the Rasch model for items with more than two ordered response categories, the thresholds that define the successive categories are an integral part of the structure of each item in that the probability of the response in any category is a function of all thresholds, not just the thresholds between any two categories. This paper describes a method of estimation for the Rasch model that takes advantage of this structure. In particular, instead of estimating the thresholds directly, it estimates the principal components of the thresholds, from which threshold estimates are then recovered. The principal components are estimated using a pairwise maximum likelihood algorithm which specializes to the well-known algorithm for dichotomous items. The method of estimation has three advantageous properties. First, by considering items in all possible pairs, sufficiency in the Rasch model is exploited with the person parameter conditioned out in estimating the item parameters, and by analogy to the pairwise algorithm for dichotomous items, the estimates appear to be consistent, though unlike for the dichotomous case, no formal proof has yet been provided. Second, the estimates of each item parameter are a function of frequencies in all categories of the item rather than just a function of frequencies of two adjacent categories. This stabilizes estimates in the presence of low frequency data. Third, the procedure accounts readily for missing data. All of these properties are important when the model is used for constructing variables from large scale data sets which must account for structurally missing data. A simulation study shows that the quality of the estimates is excellent.

Reliability and True-Score Measures of Binary Items as a Function of Their Rasch Difficulty Parameter

Dimiter M. Dimitrov
George Mason University

This article provides formulas for expected true-score measures and reliability of binary items as a function of their Rasch difficulty when the trait (ability) distribution is normal or logistic. The proposed formulas have theoretical value and can be useful in test development, score analysis, and simulation studies. Once the items are calibrated with the dichotomous Rasch model, one can estimate (without further data collection) the expected values for true-score measures (e.g., domain score, true score variance, and error variance for the number-right score) and reliability for both norm-referenced and criterion-referenced interpretations. Thus, given a bank of Rasch calibrated items, one can develop a test with desirable values of population true-score measures and reliability or compare such measures for subsets of items that are grouped by substantive characteristics (e.g., content areas or strands of learning outcomes). An illustrative example for using the proposed formulas is also provided.

Using Logistic Regression to Detect Item-Level Non-Response Bias in Surveys

Edward W. Wolfe
Michigan State University

This article describes a procedure for evaluating item-level non-response bias in questionnaire items. Specifically, logistic regression is used to determine whether non-responses are random or systematic in nature for one question from the National Educational Longitudinal Study of 1994 concerning drug use behaviors. It turns out that, indeed, non-responses are systematic with males and lower achieving students being more likely to contribute to non-response along with two-way interactions between ethnicity and SES and ethnicity and geographic region. In addition, the magnitude of the potential bias is estimated, which demonstrates that the parameter estimates

obtained by assuming that the data are missing at random may be extremely biased, given this frame of reference. Finally, several steps are suggested for evaluating the threat of non-response bias in survey research.

Rasch Measurement in the Assessment of Amytrophic Lateral Sclerosis Patients

Josephine M. Norquist 1
Ray Fitzpatrick 1
Crispin Jenkinson 2, 3
1 *Department of Public Health,*
Institute of Health Sciences, University of Oxford
2 *Department of Public Health,*
Health Services Research Unit, University of Oxford
3 *Picker Institute Europe, Oxford*

This paper examines the sensitivity to change over time of the Amyotrophic Lateral Sclerosis Assessment Questionnaire (ALSAQ-40). Individuals' health status change was assessed by means of the Rasch-based Reliable Change Index (RCI) for ALSAQ-40 questionnaires completed on two occasions, three months apart. In addition, at follow-up respondents indicated how much change they had experienced since baseline via dimension-specific self- reported transition questions. 764 individuals returned questionnaires at baseline and follow-up. For all dimensions, of respondents defined by the RCI as worse, a majority rated themselves as worse. However, on two dimensions over 60% of the respondents who rated themselves as being worse were defined as unchanged by the RCI. As with effect size smaller RCI cut-off points might be needed for subjects with ALS. This study confirms that the ALSAQ-40 is a valid and responsive disease specific health related quality of life instrument for use in studies of patients with ALS or other motor neuron diseases.

Measuring Client Satisfaction with Public Education II: Comparing Schools with State Benchmarks

Trevor G. Bond
John A. King
James Cook University

Because the results of the client satisfaction evaluation trials conducted in the state's public schools revealed that levels of client satisfaction differed in significant and meaningful ways between parents and students as well as between types of schools, this research consultancy provided school versus benchmark comparison reports based on groups of generally comparable school types. The state benchmark for each set of comparable schools estimated how easy it was, on average, for the members of the benchmark group (parents or students) to endorse each of the 20 School Opinion Survey Likert-scale items (King and Bond, 2003). The school satisfaction level for each item was calculated by a similar process to estimate how easy it was, on average, for the members of the school sample (parents or students) to endorse that item. Reports to individual schools used easy to interpret Parent/Student Satisfaction Graphs which plotted the Rasch-modeled differences between the state benchmark level and the school level for each item. In very small schools where the small amount of data did not allow for item-by-item graphs to be constructed, overall satisfaction graphs provided one global comparison with the appropriate benchmark to be reported.

The Recovery of the Density Scale using a Stochastic Quasi-Realization of Additive Conjoint Measurement

Timothy W. Pelton
University of Victoria
C. Victor Bunderson
Brigham Young University

This paper attempts to illuminate some of the practical limitations that the Rasch model (and by extension, Item Response Theory models) may have by focusing on the recovery of the density scale. Five simulation trials were conducted—the first four to recover the density scale with different deviations from the assumptions implicit in the use of the Rasch model and the fifth trial with an almost ideal data set. Results demonstrate that when error distributions are insufficient the results may be ordinal at best, and when error distributions are non-symmetrical, the positions of items may be biased with respect to the positions of persons. Results also confirm that errors of estimation, and test and sample information functions are sample dependent.

Substantive Scale Construction

Mark H. Stone
Adler School of Professional Psychology

Variable construction requires careful attention to substantive issues; a theory guiding its development, a hierarchy of illustrative items constructed to define the variable, the subsequent production of item difficulties and person measures, and the analysis of fit. Rasch measurement practitioners should give careful attention to these matters so practical suggestions are given for designing variables based on theory, item construction, and Rasch models for the analysis of data. Variable maps are emphasized to guide variable construction and interpret the results.

Volume 4, Number 4

Measurement: A Beginner's Guide

Joel Michell
University of Sydney

This paper provides an introduction to measurement theory for psychometricians. The central concept in measurement theory is that of a continuous quantitative attribute and explaining what measurement is requires showing how this central concept leads on to those of ratio and real number and distinguishing measurements from measures. These distinctions made, the logic of quantification is described with particular emphasis upon the scientific task of quantification, as opposed to the instrumental task. The position presented is that measurement is the estimation of the magnitude of a quantitative attribute relative to a unit and that quantification is always contingent upon first attempting the scientific task of acquiring evidence that the relevant attribute is quantitative in structure. This position means that the definition of measurement usually given in psychology is incorrect and that psychologists' claims about being able to already measure psychological attributes must be seriously questioned. Just how the scientific task of investigating whether psychological attributes are quantitative may be undertaken in psychology is then considered and the corollary that psychological attributes may not actually be quantitative is raised.

Rasch Modeling and the Measurement of Social Participation

Claire Dumont
Institut de réadaptation en déficience physique de Québec
Richard Bertrand
Laval University
Patrick Fougeyrollas
Institut de réadaptation en déficience physique de Québec
Marie Gervais
Société de l'assurance automobile du Québec and Laval University

Social participation is the main outcome of physical rehabilitation programs. The aim of this study is to improve the measurement of social participation, using an instrument called the Assessment of Life Habits Scale and the Rasch model. The interval level measurement, the dimensionality and the generalizability of the item hierarchy were verified. The data from a large sample of people with spinal cord injury was analyzed and specific results

were compared with expert opinions. The main properties of the instrument were satisfactory and the agreement with expert opinion was high. Principal component analysis showed multidimensionality. The item difficulty hierarchy obtained with spinal cord injury experts was different from the one obtained with traumatic brain injury experts, indicating a different difficulty level of items in relation to each population characteristics. We conclude that the instrument is appropriate for the measurement of social participation and suggest ways to improve the instrument.

Measuring Client Satisfaction with Public Education III: Group Effects in Client Satisfaction

Trevor G. Bond
John A. King
James Cook University

A contracted research project to evaluate client satisfaction with public education (King and Bond, 2003; Bond and King, 2003) required that the satisfaction of certain groups of clients received particular attention. A number of target groups were specifically identified by the state education department as those requiring separate satisfaction analyses. ConQuest software (Wu, Adams and Wilson, 1997) provided convenient techniques for estimating group mean effects for identified sub-groups of the parent and student samples. This allowed for the analysis of directly comparable satisfaction estimates for groups such as these, as well as for school size, school type, school location and parents according to the year level of the child's school class. Rasch analyses of group mean effects revealed generally, that for students, differences in overall satisfaction levels between identified sub-samples and the whole student sample even if statistically significant, were, at most, substantively marginal. However, for identified groups of parents and caregivers, the results were not so equivocal: one group of indigenous Australians, and parents of children with disabilities showed marked positive group mean effects, while another group of indigenous Australians reported lower satisfaction levels than did the whole parent sample. Moreover, analyses based on groupings of schools according to their relative complexity, appeared to reiterate the finding mentioned en passant in a previous paper (King and Bond, 2003), that client satisfaction decreased as school size increased.

Toward A Hierarchical Goal Theory Model of School Motivation

Dennis M. McInerney
Herbert W. Marsh
University of Western Sydney
Alexander Seeshing Yeung
Hong Kong Institute of Education

Instead of concentrating on mastery and performance goal orientations, recent research on school motivation has suggested a multidimensional structure of achievement goal orientations. Students in Australian high schools (N =774) responded to 35 survey items on 10 goal orientation constructs (effort, task, sense of purpose, praise, competition, power, token, social concern, social dependence, and affiliation) and 14 items on general mastery, general performance, general social, and global motivation constructs. Confirmatory factor analysis results supported a hierarchical, multidimensional school motivation construct. The hierarchical, multidimensional model has provided a strong theoretical structure for further school motivation research.

Examining Reliability and Validity of Job Analysis Survey Data

Ning Wang
Widener University

Historically, job analysis has played a fundamental role for developing and validating licensure and certification examinations. Still, research on what constitutes reliable and valid job analysis data is lacking. This paper illustrates several ways to examine the reliability and validity of job analysis survey results. Generalizability theory and the many-facets Rasch model are applied to investigate consistency and generalizability in task importance

measures, to suggest reliable sample size, and to justify the number and use of rating scales. By using random samples from job analysis data for two professions with divergent job activities, this study finds that a representative sample as small as 400 respondents produces reliable estimates of task importance to the same degree of generalizability as obtained from a larger sample of job analysis respondents. Analyses of rating scales suggest that the effectiveness of using different numbers and types of rating scales depends on the nature of a profession.

Measuring Coping at a University Using a Rasch Model

Russell F. Waugh
Edith Cowan University

Coping with academic difficulties at a university was based on six aspects: Motivation and Planning, Friends and Planning, Studying and Planning, Emotions, Spiritual Help and Coping by Doing Nothing. Stem-items for each aspect were conceptually ordered by difficulty. Each of the stem-items was answered from three perspectives, Good Coping Strategies, Actual Coping Strategies, and Stress Reduction Strategies. The three response categories were No, not on any occasion this semester; Yes, on 1 to 3 occasions this semester, and Yes, on 4 or more occasions this semester. The convenience sample was 337 students studying education at an Australian university and data were analyzed with a Rasch measurement model. A scale was created in which the difficulties of the items were ordered from easy to hard and the student measures of Coping were ordered from low to high. Coping by Doing Nothing and Using Spiritual Help stem-items didn't fit the measurement model and were deleted. This left an effective item sample of 21 (7 stem-items times 3). The proportion of observed student variance considered true was 0.88. The results supported the theory behind the construct of Coping as using Motivation and Planning, Friends and Planning, Studying and Planning, and Emotions, in which Expected Good Coping Strategies are easier than Actual Coping Strategies which, in turn, are easier than Stress Reducing Coping Strategies.

Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I

Carol M. Myford
University of Illinois at Chicago
Edward W. Wolfe
Michigan State University

The purpose of this two-part paper is to introduce researchers to the many-facet Rasch measurement (MFRM) approach for detecting and measuring rater effects. The researcher will learn how to use the Facets (Linacre, 2001) computer program to study five effects: leniency/severity, central tendency, randomness, halo, and differential leniency/severity. Part 1 of the paper provides critical background and context for studying MFRM. We present a catalog of rater effects, introducing effects that researchers have studied over the last three quarters of a century in order to help readers gain a historical perspective on how those effects have been conceptualized. We define each effect and describe various ways the effect has been portrayed in the research literature. We then explain how researchers theorize that the effect impacts the quality of ratings, pinpoint various indices they have used to measure it, and describe various strategies that have been proposed to try to minimize its impact on the measurement of ratees. The second half of Part 1 provides conceptual and mathematical explanations of many-facet Rasch measurement, focusing on how researchers can use MFRM to study rater effects. First, we present the many-facet version of Andrich's (1978) rating scale model and identify questions about a rating operation that researchers can address using this model. We then introduce three hybrid MFRM models, explain the conceptual distinctions among them, describe how they differ from the rating scale model, and identify questions about a rating operation that researchers can address using these hybrid models.