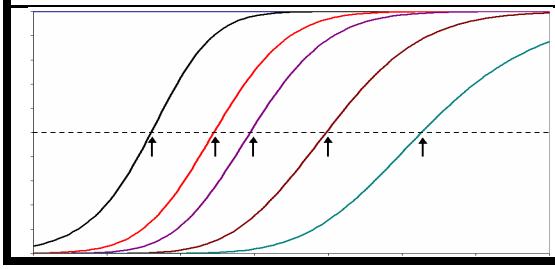
Rasch-Thurstone Thresholds

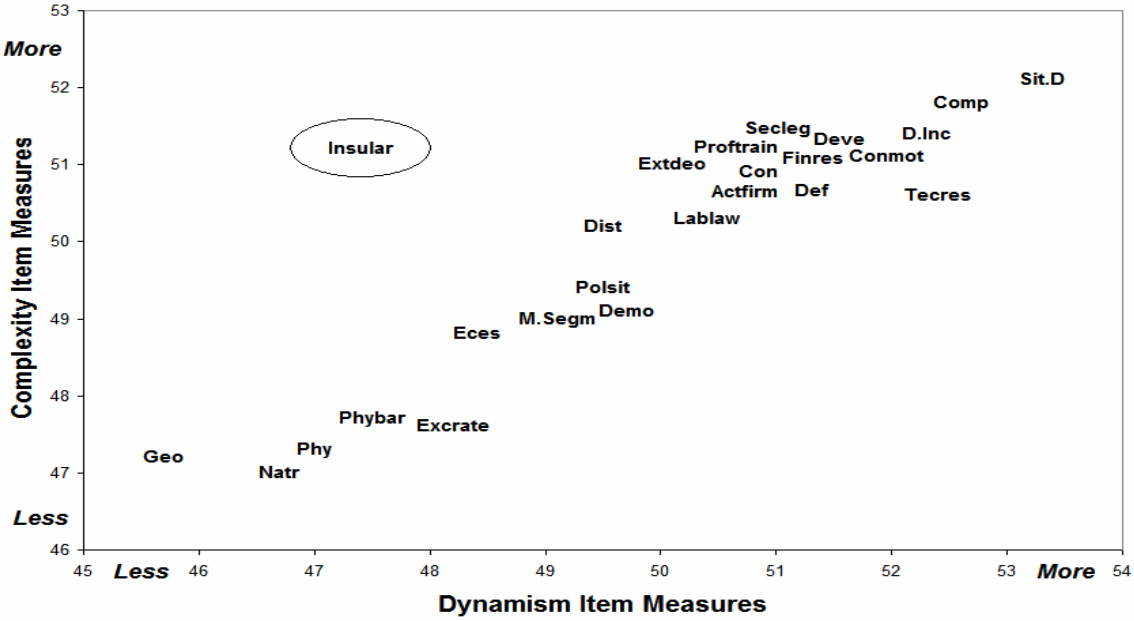


RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

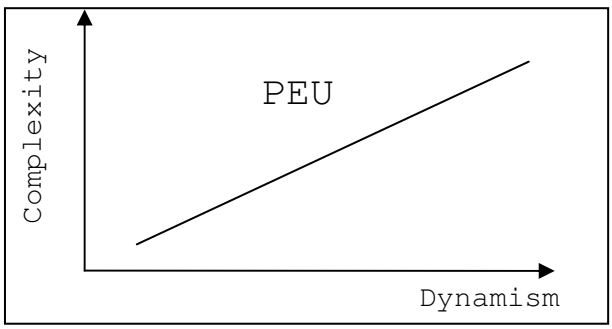
Vol. 19 No. 4 Spring 2006 ISSN 1051-0796

Mapping Perceived Environmental Uncertainty



Organizations are in continuous interaction with their environment - an environment that is a primary source of opportunities and threats. Detailed information about that environment is required for making sound decisions, but that information is never complete. There is always uncertainty. Perceived Environmental Uncertainty (PEU) quantifies the lack of information about the world around a firm as perceived by its decision-makers.

The lack of information can be because the decision-related items, *i,* are difficult to understand (complexity, C_i) and because their nature is changing (dynamism, D_i).



This suggests a conceptual plot PEU. Any point on the plot gives the PEU experienced by a decision-maker based on the perceived complexity and dynamism of that environmental item.

The core idea is to measure environmental complexity and dynamism by means of latent variables (empirically defined by operational items) without introducing any external weights as was done in previous research (e.g., Daft *et al.*, 1988). Accordingly, we obtained a sample of 338 surveys answered by business managers in the Canary Islands (Spain). We asked them about their perceptions (on a 5-point scale) of complexity and dynamism for 25 items relevant to the local economy.

Cross-plotting the Rasch item calibrations for the complexity and dynamism constructs yields an instructive

Table of Contents

map of PEU, shown above. There is a strong correlation ($R^2 = 0.85$) between the two constructs. This confirms the findings of previous research (Daft *et al.*, 1988). According to the scatterplot, dynamism and complexity contribute roughly equally to PEU. This finding differs from Duncan (1972). He found that dynamism contributes more to uncertainty.

The "Insularity" item is an outlier (top left). It is perceived to be more complex than dynamic. This could be because of a special meaning of the term "insularity" for people living on islands. In one sense it is a constant factor in managing their businesses.

One way in which this work advances this field of research is that the relationship between complexity and dynamism derives from the decision-makers themselves and is not forced into the data by some arbitrary weighting system.

Another way is that it assists organizations faced by the problem of allocating resources for the improvement of decision-making. This methodology identifies which environmental items are perceived to be more uncertain and whether that uncertainty is more due to complexity or dynamism. This information can then be used to optimize decision support systems.

From Yanes-Estévez V., Oreja-Rodríguez J.R. & Alvarez P. (2004) Computing the perceived environmental uncertainty function by Rasch model, WSEAS Transactions on Business and Economics, 4, 1, 281-285

Rasch Workshops

May 2006 - March 2007 10 two & three day Rasch courses, Leeds, UK
home.btconnect.com/Psylab_at_Leeds/

June 23, 2006, Friday Workshop: Latent variables, Russia
www.rasch.org/russia.htm

June 25, 2006, Sunday, Hong Kong (pre-PROMS)
Introduction to *Winsteps,* conducted by Mike Linacre
June 26, 2006, Monday, Hong Kong (pre-PROMS)
Introduction to *Facets,* conducted by Mike Linacre
www.promshk.org

July 3-7, 2006, Mon.-Fri. Measurement in the Psychosocial Sciences: from raw scores to Rasch measures, Andrew Stephanou, Brisbane, Australia
www.acspri.org.au

July 21 - Aug.18, 2006 Practical Rasch Measurement, Online (J. Linacre)
www.statistics.com/content/courses/rasch

July 24 - Oct. 30, 2006 Rasch Unit of Study, Online (D. Andrich)
www.education.murdoch.edu.au/educ_RaschCourse2006.html

August 7-8, 2006, Mon.-Tue. Introduction to Winsteps, Chicago IL www.winsteps.com/workshop.htm

Election of Rasch SIG Officers

Rasch Measurement SIG officers elected to hold office from April 2006 to April 2008 are:

Chair: **Thomas O'Neil**

Secretary: **Ed Wolfe**

Ballots mailed electronically to active members: 170
Ballots returned: 32. Ballots spoiled: 0

(Signed) Steven Stemler, Rasch SIG Election Officer, February 23, 2006

As of 3-25-2006, the SIG has 186 members.

Daft R.L., Sormunen, J., Parks, D. (1988) Chief executive scanning, environmental characteristics, and company performance: an empirical study. Strategic Management Journal, 9, 123-139.

Duncan, R. (1972) Characteristics of organizational environment and perceived environment uncertainty, Administrative Science Quarterly, 17, 313 – 327.

MOMS

Midwest Objective Measurement Seminar

Chicago Circle Center (CCC), UIC
750 South Halsted Street, Room 605
Friday, April 28, 2006, 9:00 – 4:45

Measurement of Participation
Allen Heinemann, Rehabilitation Institute of Chicago

Equating Versions of the FIM
Anne Deutsch,, Rehabilitation Institute of Chicago

Special Presentation
David Andrich, Murdoch University, Australia

A Domain Level Model
Matthew Schultz, ACT

Hierarchical Rater Modeling
Timothy Muckle, Promissor

A Comparison of Pre-Equating with Post-Administration One-Step Equating For Initial Computerized Testing
Linjun Shen,, NBOME

Using Person Fit Statistics to Screen for Atypical Persons At-risk for Suicide
Kendon J. Conrad, Nikolaus Bezruczko, HyeJung Park, Michael Dennis, UIC

The New Lexile Framework for Writing
Jackson Stenner, MetaMetrics

Stability of Item Calibrations
Ross Brown, Measurement Research Associates, Inc.

Complex Adaptive Functionality via Measurement
William P. Fisher, Jr., Avatar International

The Second Pacific Rim Objective Measurement Symposium
PROMS HK 2006
The Hong Kong Institute of Education
Tai Po, HONG KONG
Tuesday 27th - Thursday 29th June, 2006

With 60 paper proposals accepted and invited keynote presentations, the program will be both varied and interesting. Rasch measurement workshops (including *Facets* And *Winsteps* workshops) will precede the conference (Sunday 25, Monday 26).

The theme of the conference is

"Rasch Measurement:
A Tool for Scientific Progress for the Asia Pacific"

It will focus on recent advances in objective measurement as a tool for scientific progress in education, health and the social sciences. It will provide an international forum for discourse on the latest research in using Rasch measurement as well as opportunities to learn more about Rasch measurement.

The program will include a Teachers' Day (Thursday 29th June) and pre-conference workshops (Sunday 25th - Monday 26th June). For details about registration and accommodation, please visit our website:

www.promshk.org

We look forward to welcoming you in this important and exciting event! We would be delighted if you could forward this invitation to your colleagues.

Optometry and Vision Science
is soliciting papers for a Feature Issue

Vision-related Quality of Life

Deadline for submission October 1, 2006.

A feature issue provides the opportunity for your work to be published alongside similar subject matter. Past feature issues of OVS have included a number of important and highly-cited papers. For example, the 20 or so papers published in the 1999 feature issues on myopia have been cited well over 300 times.

Patient-centered assessments of vision have become standard supplements to vision tests in clinical trials, with these instruments gaining importance as main outcome measures. With a developing research focus, these instruments are evolving from relatively simple measures to increasingly discriminatory, reliable and valid ones. It is intended for the feature issue to include a broad spectrum of topics associated with vision-related quality of life, including:

• the development and validation of questionnaires that quantify vision-related quality of life, visual disability and/or visual symptoms

• studies which demonstrate benefits of contemporary methodologies for questionnaire design and development e.g. Rasch analysis, item banking, computer adaptive testing

• a comparison or assessment of questionnaires that quantify vision-related quality of life, visual disability and/or visual symptoms

• the relationship between questionnaire scores and clinical vision tests or task performance

• the use of quality of life measures as outcome measures in clinical research and clinical trials

• the use of quality of life measures as clinical tools for the practitioner.

Manuscripts must be submitted online at ovs.edmgr.com and should be prepared according to the instructions to authors at this web site. Indicate that your paper is being submitted for this feature issue.

Manuscripts will be subjected to peer review under the editorial leadership of David Elliott, with Trudy Mallinson, and Konrad Pesudovs serving as Co-Editors.

IOMW2006

13th International Objective Measurement Workshop
University of California, Berkeley, California, USA
April 5–7, 2006
http://bearcenter.berkeley.edu/IOMW2006/

Wednesday, April 5, 2006.

8:45 a.m. Welcome - Mark Wilson

9:00 a.m. Opening Presentation - Ray Adams - Reliability and Item Response Modeling: Myths, Observations and Applications

9:45 Symposium 1: Multidimensionality: Theory and Practice. Chair: Linda Morell

David Andrich - On characterizing the thickness of an educational measurement

Steffen Brandt - Modeling Tests With Sub-Dimensions

Tim Gaffney - On the Factor Structure of Standardized Educational Achievement Tests

Hiroyuki Yamada, Karen Draney, Tzur Karelitz, Stephen Moore, and Mark Wilson - Comparison of dimension-aligning techniques in a multidimensional IRT context

11:15 a.m. Roundtables 1: Differential Item Functioning and Bias.

Donna J. Butterbaugh, Richard M. Smith, Vincent A. Maurelli - Examining Type I and Type II Error Rates in Small Sample DIF Statistics

Xiaoting Huang, Kelly Lei Wang - Validity Equivalence Between the Chinese and English Versions of the IEA Child Cognitive Developmental Status Test

Ou Lydia Liu - Gender Similarities or Differences: Analysis based on PISA mathematics 2003

Maria Veronica Santelices - Differential Item Functioning in the SAT Reasoning Test

Yiyu Xie, Mark Wilson - Imperial vs. Metric Study (IMS)

11:15 a.m. Roundtables 2: Item Bundles and Testlets.

Steffen Brandt - Exploring Bundle Dependencies for the Embedded Attitudinal Items in PISA 2006

Cherdsak Iramaneerat, Carol M. Myford, Rachel Yudkowsky - Item dependency in an objective structured clinical examination

Insu Paek, Haniza Yon, Mark Wilson - Random Parameter Structure and Testlet Model: Extension of the Rasch Testlet Model

1:15 p.m. Roundtables 3 Objective Measurement Across the Disciplines I.

Pedro Alvarez, Francisco J. Moral, Jose Canito - Rasch model and geostatistics techniques for atmospheric pollution

Tsair-Wei Chien, Mike Linacre, Wen-Chung Wang, Ou Lydia Liu - Rasch Analysis of Teams Abilities and Home Court Advantages for 2005-2006 NBA Ranks

Curt Hagquist - The Psychometric Properties of the Strengths and Difficulties Questionnaire - an Analysis of Swedish Data based on the Rasch Model

Rense Lange, James Houran - Perceived importance of employees' traits and abilities for performance in hospitality jobs

Nathan J. Markward, William P. Fisher, Bronya J. B. Keats - A Measurement Theoretic Version of the Ohta-Kimura Stepwise-Mutation (Ladder) Model

A. Jackson Stenner, Mark H. Stone - Does the Reader Comprehend the Text Because the Reader is Able or Because the Text is Easy?

1:15 p.m. Roundtables 4: Equating Test Forms I.

Mark H. Moulton, Howard A. Silsdorf - Multidimensional Equating: Linking Multidimensional Test Forms by Constructing an Objective n-Space

Richard Patz, Venessa Lall, Christopher Domaleski - Estimating the Rasch Model with Block-Diagonal Item Response Matrix: An Exploration of Winsteps Software with Implications for Equivalent-Groups Equating

1:15 p.m. Roundtables 5: Precision, Error, and Fit I.

Clemens Draxler - Sequential Tests for the Rasch Model

Anatoli A. Maslak - A Simulation Study of Rasch Measurement Precision for Dichotomous Items

Matthew Stearns, Richard Smith - On the Estimation of Classification Consistency Indexes for Complex Assessments

2:15 p.m. Symposium 2: A Matter of Judgment: Standard Setting and Rater Consistency. Chair: Diane Allen

Sun-Geun Baek, In-Hee Choi - The Mastery Level Judgment Consistency Rate of a Rasch Model Based Standard Setting Method for Classroom Achievement Tests

Trevor G. Bond, Noor Lide Abu Kassim - Use of the Many Facet Rasch Model in Resolving Standard Setting Issues

William P. Fisher, Jr., Batya Elbaum, Lisa Persinger, Alan Coulter - Survey-Based Service Quality Standards under IDEA: An Open Source Platform for Metrological Uniformity

Peter D. MacMillan - Rater stability and applicant pool quality across successive applicant pools: A many-faceted Rasch rating scale analysis

3:45 p.m. Symposium 3: Assessment for e-Learning: Case studies of an emerging field. Chair: Kathleen Scalise. Discussant: Cathleen A. Kennedy.

Diana J. Bernbaum - NetPASS: Construct and content validity in e-learning products

Mike Timms - Quantum Tutors: Matching instructional goals with assessment in e-learning

S. Veeragoudar Harrell - Cognitive Diagnostics Using Rasch family models to map student understanding

Kristen Burmester - ALEKS: Making e-learning assessment reports useful in classroom instruction

Thursday, April 6, 2006

9:00 a.m. Plenary Session - Mark Wilson

9:45 Symposium 4: Person and Population Models. Chair: Karen Draney

John Gargani - Using the Rasch Model to Evaluate Programs and Program Theories: An Example from the Evaluation of an Integrated Science and Literacy Curriculum

Robert W. Massof - The Theoretical Difference Between IRT and Rasch Models (It's Not What You Think!)

George Jay Unick - An Analysis of the Latent Structure of the DSM IV Criteria for Major Depressive Disorder

Yiyu Xie - To Guess or Not to Guess? It is a student's choice

11:15 a.m. Symposium 5: Measuring Growth. Chair: Cathleen A. Kennedy

Derek Briggs, Ed Wiley, Jon Weeks - Vertical Scaling in Value-Added Models for Student Learning

Theo Dawson-Tunik - Cognitive change is stage-like: The cumulative evidence from a decade of Rasch modeling

Shudong Wang, Hong Jiao, Michael J. Young, Ying Jin - The Effects of Linking Designs in Vertical Scaling on the Growth Patterns of Student Achievement

1:15 p.m. Roundtables 6: Objective Measurement Across the Disciplines II.

Pedro Alvarez, M.A. Blanco - Measuring Sensorial Perception

Robert Frederick Cavanagh, Joseph Romanoski - Rasch and structural equation modeling analyses of teacher observations of school principal leadership

Sergij Gabrscek - Why not do it differently: Analysis of examination results in Slovenia

Hung-Jung Lin, Tsair-Wei Chien, Wen-Chung Wang - Rasch analysis assists a hospital with salary allocation for physicians in emergency department

Nathan J. Markward - Separating the Parameters of Genealogy and Mutation: Violations of Local Independence as Deviations from Genetic Equilibrium

Kavita L. Seeratan - Evaluative Implementations: Meaning Equivalence Instructional and Assessment Methodology for Deep Understanding

Sharon G. Solloway, W.P. Fisher, Jr. - Mindfulness Practice: A Rasch Variable Construct Innovation

1:15 p.m. Roundtables 7: Equating Test Forms II.

Anli Lin, Don Meagher, Eugene Bowles, Christina P. Stellato - Creating Equivalent Groups for Equating with Bootstrap and Matched Samples

Xiaohui Zheng - A Comparison Study of IRT Item Parameter Scaling Methods in Common-Item Nonequivalent Groups Equating

1:15 p.m. Roundtables 8: Precision, Error, and Fit II.

Kirk A. Becker, George Karabatsos - Determining confidence intervals for IRT statistics through parametric bootstrapping

Timothy Muckle, Betty Bergstrom, Kirk Becker, John Stahl - Impact of Altering Randomization Intervals on Precision of Measurement and Item Exposure

2:15 p.m. Symposium 6: Structured Construct Development. Chair: Cheryl Schwab

Diane D. Allen - Using Item Response Modeling Methods to Develop Theory Related to Human Performance

Brent Duckor - Measuring Measuring: An item response theory approach

Jennifer Randall Thomas, George Engelhard, Jr. - Using Guttmann's Facet Theory to Develop an Instrument that Examines the Grading Practices of Teachers

Christopher Weaver - Optimizing the compatibility between rating scales and product measures of second language competence

3:45 p.m. Symposium 7: Current Critical Issues Related to Science Assessments. Chairs: Hong Jiao, Shudong Wang. Discussant: Richard Patz

Shudong Wang, Hong Jiao, Michael J. Young, Lihua Yao - The effect of construct shift in science achievement across grades on a science vertical scale

Jiahe Qian - Linking 2005 NAEP science assessments through bridge samples

Hong Jiao, Shudong Wang, Zarko Vukmirovic - Investigation of local item dependence in scenario-based science assessment

Nathaniel J.S. Brown, Cathleen A. Kennedy, Karen Draney, Mark Wilson - Assessing a learning progression in science: Solving psychometric issues

Friday, April 7, 2006.

Post-Conference Workshops

8:00 a.m. Workshops I

John M. Linacre, William Bonk - Modeling many-facets data using Facets: An introduction

David Andrich - Interactive analysis of data using RUMM2020

Cathleen A. Kennedy - GradeMap

10:00 a.m. Workshops II

John M. Linacre, Mark H. Moulton - Winsteps applied to messy data

Karen Draney, Hiro Yamada, Ray Adams - ConQuest

Rasch Measurement Transactions
P.O. Box 811322, Chicago IL 60681-1322
www.rasch.org/rmt/
Editor: John Michael Linacre

SIG Chair: Thomas O'Neill, Secretary: Ed Wolfe
Program Chair: Trevor Bond
SIG website: www.raschsig.org
www.raschsig.org/news.html

AERA-NCME Rasch-related Papers

San Francisco, April 7-11, 2006

Friday, April 7, 2006

Diverse Topics in Survey Research
SIG-Survey Research in Education
12:00m - 12:40pm Moscone Center West 3005

Applying the Rasch Rating-Scale Model in Survey Research: A Study of Black University Students' Perceptions of Marriage. *Kelly D. Bradley (University of Kentucky), William E. Harris (University of Kentucky)*

Measuring and Modeling in Learning Environment Research. SIG-Learning Environments
12:00pm - 2:00pm Moscone Center South 222

Parental Involvement and Classroom Learning Culture: A LISREL Analysis Using Rasch Model Instrumentation. *Graham B. Dellar (Curtin University of Technology), Robert Frederick Cavanagh (Curtin University of Technology), Joseph Thomas Romanoski (Curtin University of Technology)*

Diverse Explorations and Interpretations of Constructivist Theory, Research, and Practices . SIG-Constructivist Theory, Research and Practice
2:15pm - 3:45pm Moscone Center South 236

Dialogical Constructivism: Measurement Technology Can Mediate the Construction of Shared Meanings. *William P. Fisher (Avatar International, Inc.), Jackson A. Stenner (MetaMetrics, Inc.)*

Applying the Rasch Model in Educational Settings
SIG-Rasch Measurement
4:05pm - 5:35pm Moscone Center West 2009
Chair: Gilles Raiche (U. du Quebec a Montreal)

Deriving Proficiency Scales From Performance Indicators Using the Rasch Model. *Michel D. Laurier (Université de Montréal), Jean-Guy Blais (Université de Montréal), Christian Rousseau (Quebec Ministry of Education)*

Investigating the Fit and Functioning of a High-School Algebra Assessment for English Language Learners Using the Dichotomous Rasch Model. *Shannon O. Sampson (University of Kentucky), Kelly D. Bradley (University of Kentucky)*

Using Paired Comparison Matrices to Analyze Connectivity of Assessment Data. *Mary Garner (Kennesaw State University)*

Using Rasch Modeling to Compare Scientific Conceptions of High-School and University Students. *Debra L. Panizzon (University of New England), Trevor G. Bond (Hong Kong Institute of Education)*

Rasch Measurement SIG Business Meeting
SIG-Rasch Measurement
6:15pm - 8:15pm Moscone Center West 2009
Chair: Randall E. Schumacker (U. of North Texas)
Secretary: Steven Stemler (Wesleyan University)

Invited Address: The Ties That Bind. *Richard Smith (Journal of Applied Measurement)*

Saturday, April 8, 2006

Rasch Roundtables: Applied to Education and Caring Professions. SIG-Rasch Measurement
8:15am - 8:55am Moscone Center West 3005

An Investigation Into Variability of Tasks and Teacher-Judges in Second-Language Oral Performance Assessment: A Many-Faceted Rasch Measurement Analysis. *Youn-Hee Kim (McGill University)*

Assessment of the Rescoring Procedure for Discrepant Raters Using a Many-Faceted Rasch Rating Scale Analysis. *Peter D. MacMillan (University of Northern British Columbia), Colin Chasteauneuf (University of Northern British Columbia)*

Employing the Many-Facet Rasch Model to Investigate the Domains of the MSLSS. *Kelly D. Bradley (University of Kentucky), Richard Gilman (University of Kentucky), Jessica Dawn Cunningham (University of Kentucky)*

Mostly the Teacher Is the Test! *Martin Kennings Caust (James Cook University)*

Rasch Analysis of Functional Caregiving: A New Construct for Mothers' Caregiving. *Nikolaus Bezruczko*

Assessing Measurement Model Fit
Division D-Measurement and Research Methodology. Section 1: Educational Measurement, Psychometrics and Assessment
8:15am - 10:15am Renaissance Parc 55 / Aragon
Chair: Thomas R. O'Neill (National Council of State Boards of Nursing)
Discussant: Ronald T. Mead (DRC)

An Investigation of the Power and Type I Error Rates for Winsteps' Fit Statistics. *Huiqin Hu (Data Recognition Corporation), Guangrong Dai (Central Michigan University)*

Explaining "Examinee Misfit" via Case Studies. *Alexandra Petridou (The University of Manchester), Julian S. Williams (The University of Manchester)*

Rasch Roundtables: A Focus on Methods
SIG-Rasch Measurement
12:25pm - 1:05pm Moscone Center West 3006

Interpreting Reliability Using Rasch Measurement Models. *Randall E. Schumacker (University of North Texas)*

Parallel Analysis for Dichotomous Data in the Context of Rasch Residual Principal Components Analysis. *Gilles Raiche (Universite du Quebec a Montreal), Jean-Guy Blais (Université de Montréal)*

A Comparison of the Rasch Model and Three-Parameter Model on Ability Estimation Based on Summed Scores. *Angela T. Austin (CTB/McGraw-Hill), Seung W. Choi (CTB/McGraw-Hill)*

Effects of Sample Size on the Multilevel Measurement Model DIF Detection. *Kwang-Lee Chu (Harcourt Assessment, Inc.), C. Allen Lau (Harcourt Assessment, Inc.)*

Use of Rasch Step and Scale Statistics in Identification of Category Order. *Amjed Al-Owidha (University of Denver), Kathy E. Green (University of Denver), Jane Kroger (Universitetet i Tromsø)*

Rasch Measurement Investigations of DIF and Scaling
SIG-Rasch Measurement
4:05pm - 5:35pm Moscone Center West 2009
Chair: Steven Stemler (Wesleyan University)

CAT Ability Estimation With Varying Item Exposure Controls and Types of DIF Using the Partial Credit Model. *Candace Macken-Ruiz (The University of Texas), Linda L. Hargrove (University of Texas at Austin), Barbara G. Dodd (University of Texas At Austin)*

Evaluation of a Categorization Procedure on the DIF Results Produced by Winsteps. *Yuming Liu (Riverside Publishing Company)*

Understanding the Utility of Magnitude Estimation Scaling. *Svetlana A. Beltyukova (The University of Toledo), Christine M. Fox (The University of Toledo), Gregory E. Stone (The University of Toledo)*

Uniform Differential Item Functioning on the Basis of Birth Cohort in a Verbal Ability Test Used in National Public Opinion Surveys: A Rasch Analysis. *Randall MacIntosh (California State University-Sacramento)*

IRT Parameter Estimation
Division D-Measurement and Research Methodology.
Section 1: Educational Measurement, Psychometrics and Assessment
4:05pm - 6:05pm Renaissance Parc 55 / Cervantes

Rasch and 3PL Ability Estimates When the True Model is Multidimensional. *Tammiee S. Dickenson (University of South Carolina-Columbia), Brian T. Habing (University of South Carolina)*

Psychometric Research with Policy Implications for Certification & Licensure
Invited Symposium. NCME
4:05 p.m.-6:05 p.m., Nikko Ballroom I, E1
Organizer/Moderator: Thomas O'Neill, National Council of State Boards of Nursing
Discussants: G. Gage Kingsbury, NWEA; Barbara S. Plake, University of Nebraska, Lincoln

Impact of English as a second language (ESL) status on NCLEX performance. *Thomas O'Neill, Weiwei Liu, Michelle Reynolds, National Council of State Boards of Nursing*

Assessing scientific content in an integrated, clinically relevant context. *Gene A. Kramer, Laura M. Neumann, American Dental Association*

Modeling item difficulty for performance assessments that include critical steps. *Betty Bergstrom, Kirk Becker, Jim Masters, Timothy Muckle, Promissor*

The implications of DIF and bias in certification testing. *Russell W. Smith, Thomson Prometric*

A comparison of traditional and IRT based item quality criteria. *Brian Bontempo, Mountain Measurement, Jerry Gorham, Pearson VUE*

Graduate Student Poster Session: NCME
4:05 p.m.-6:05 p.m., Nikko Ballroom III, E3

Comparing BILOG-MG and WINSTEPS in item parameter recovery of Rasch model. *Chen-Miao Chen, Young-Sun Lee, Columbia University*

Sunday, April 9

Assessment in the Professions: From Testlets to Performance Exams
Division I-Education in the Professions
10:35am - 12:05pm Moscone Center North 111

Rater Effects in Clinical Performance Ratings of Surgery Residents. *Cherdsak Iramaneerat (University of Illinois at Chicago), Carol M. Myford (University of Illinois at Chicago)*

Monday, April 10, 2006

The Rasch Model Applied to Polytomous Data
SIG-Rasch Measurement
8:15am - 10:15am Moscone Center West 2004
Chair: Gregory E. Stone (The University of Toledo)

Calibrating a Single-Prompt Writing Test: An Investigation of Rasch Polytomous Model Behavior. *Michelle LD Barrett (CTB/McGraw-Hill), Seung W. Choi (CTB/McGraw-Hill), Bruce F. Randel (McREL)*

Is More Less? Impact of Number of Response Categories in Self-Reported Pain. *Karon F. Cook (University of Washington)*

Obtaining a Common Scale for a Mixed-Item Format Test Under the Partial-Credit Model. *Daeryong Seo*

(Harcourt Assessment, Inc.), Se-Kang Kim (Harcourt Assessment, Inc.), Michael J. Young (Harcourt Educational Measurement), Husein M. Taherbhai (Harcourt Assessment, Inc.)

Rasch Measurement in Developing Faculty Ratings of Students Applying to Graduate School. *Sooyeon Kim (ETS), Patrick Kyllonen (ETS)*

Constructing Data, Modeling Worlds. Division C-Learning and Instruction. Section 3: Mathematics
10:35am - 12:05pm Moscone Center West 2006

Measuring Statistical Reasoning: Development of an Assessment System for Data Modeling. *Xiaohui Zheng (University of California-Berkeley), Kristen Orourke Burmester (University of California-Berkeley), Tzur Karelitz (University of California-Berkeley), Mark R. Wilson (University of California-Berkeley)*

Applications of Rasch Measurement Methods
SIG-Rasch Measurement
4:05pm - 5:35pm Moscone Center West 2005
Chair: Kelly D. Bradley (University of Kentucky)

A Comparison of Separate Versus Concurrent Methods After Centering Person Scale Under the Rasch Model. *Daeryong Seo (Harcout Assessment, Inc.), Michael J. Young (Harcourt Educational Measurement), Se-Kang Kim (Harcourt Assessment, Inc.), Tim O'Neil (Harcourt Assessment, Inc.)*

Objective Standard Setting for Judge-Mediated Examinations. *Gregory E. Stone (The University of Toledo), Svetlana A. Beltyukova (The University of Toledo), Christine M. Fox (The University of Toledo)*

Teaching for Social Justice: An Application of Rasch Measurement Principles. *Larry H. Ludlow (Boston College), Sarah Enterline (Boston College), Marilyn Cochran-Smith (Boston College)*

Using Rasch Measurement Model to Validate the Korean Version of the Marlowe-Crowne Social Desirability Scale. *Hyun Soo Seol (Chung-Ang University)*

Technical Issues in Test Equating
Division D-Measurement and Research Methodology. Section 1: Educational Measurement, Psychometrics and Assessment
4:05pm - 6:05pm Renaissance Parc 55 / Aragon

An Evaluation of Procedures for the Screening of "Affected" Common Items in Rasch-Based Equating. *Nathan L. Wall (Harcourt Assessment, Inc.), Qing Yi (Harcourt Assessment, Inc.), Michael J. Young (Harcourt Educational Measurement)*

Comparing Screening Approaches to Investigate Stability of Common Items for Test Equating. *Alvaro J. Arce-Ferrer (Harcourt Assessment, Inc.), C. Allen Lau (Harcourt Assessment, Inc.)*

Item Position and Item Difficulty Change in an IRT-Based Common Item Equating Design. *Jason L. Meyers (Pearson Educational Measurement), Walter D. Way (Pearson Educational Measurement), Edward Miller (Texas Education Agency)*

Tuesday, April 11

Measuring Trends in International Comparative Research: Results from the First Two Cycles of the OECD/PISA Study
SIG-Large Scale Assessment
8:15am - 10:15am Moscone Center West 2009
Research: Results From the First Two Cycles of the OECD/PISA Study
Chair: Geoff Masters (Australian Council for Educational Research)
Discussants: Mark R. Wilson (University of California-Berkeley).
David P. Baker (The Pennsylvania State University)

The Impact of Item Choice on the Measurement of Trends in Educational Achievement. *Dominique Marie Lafontaine (Université de Liège), Christian Monseur (Université de Liège)*

The Estimation of Equating Error in Studies of Educational Trends. *Christian Monseur (Université de Liège), Alla Berezner (Australian Council for Educational Research)*

Use of Different Models for Estimating Trends. *Raymond J. Adams (Australian Council for Educational Research), Eveline Gebhardt (Australian Council for Educational Research)*

Examining the Student Effect of Effort in PISA: The Impact on the Validity of Trends. *Jayne Butler (University of Melbourne), Raymond J. Adams (Australian Council for Educational Research)*

Measuring the Socioeconomic Background of Students and Its Effect on Achievement in PISA 2000 and PISA 2003 . *Wolfram H. Schulz (Australian Council for Educational Research)*

Rasch-related Conferences

May 19-21, 2006, Fri.-Sun. Quality of Life Research in Asia HKSoQOL, Hong Kong
www.hksoqol.org/conference

June 24 - July 2, 2006 Summer school: Rasch measurement, Russia
www.rasch.org/russia.htm

June 27-29, 2006, Tues.-Thur. Pacific Rim Objective Measurement Symposium PROMS, Hong Kong
www.promshk.org

April 9-13, 2007, Mon.-Fri. AERA Annual Meeting
Chicago, Illinois
www.aera.net

Demarcating Category Intervals

There are three widely-used methods for identifying the sections of a latent variable line that correspond to each category of a rating scale (or other ordered polytomy) for an item. For well-behaved rating scales, the three methods produce similar results, so the choice of method depends on what best communicates the findings to the audience. Each method has advantages and disadvantages.

Imagine a 7 category rating scale, scored 1-7: very strongly disagree VSD, strongly disagree SD, disagree D, neutral N, agree A, strongly agree SA, and very strongly agree VSA. This item is in an instrument administered to 900 persons. For this item, the observed frequencies of each category are: VSD 100, SD 0, D 100, N 50, A 200, SA 350, VSA 200. Notice that SD, strongly disagree, is not observed for this item in this data set. When this data set is analyzed using Masters' Rasch Partial Credit Model (PCM), a response structure for our item is estimated.

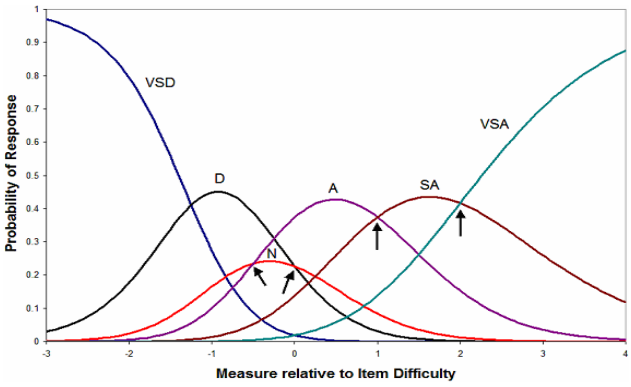


Figure 1. Modeled category probability curves

The PCM category probability curves are shown in Figure 1. There are 6 curves visible. Starting from the left, VSD is most likely to be observed for low-measure respondents. Then as respondent measures increase, the probability of observing D increases. SD is never observed in this data set, so is modeled to have a zero probability and is not shown in this Figure. As the respondent measure increases, the probability of category N increases, but it is never the most probably category at any point on the latent variable. With increasing measure, category A becomes most probable, then SA, and finally VSA.

The points at which adjacent categories are equally probable (indicated by arrows) are the Rasch-Andrich thresholds or step calibrations. They are the response-structure parameters of the PCM model. Since category SD is not observed, the equal-probability point between category VSD and SD is modeled to be at plus infinity, and between SD and D is modeled to be at minus infinity. If plotable, they would also be indicated by arrows.

Thus there are 6 Rasch-Andrich adjacent-category equal-probability points. In this example, they are located on the latent variable at {+∞,-∞, 0, -0.5, +1, +2} logits relative to the overall item difficulty. The overall item difficulty is defined as the point at which the lowest and highest categories, VSD and VSA, are equally probable. According to this definition, the sum of the Rasch-Andrich thresholds is zero. Thus, for estimation purposes, the set of Rasch-Andrich thresholds can be approximated by {37.5, -40, 0, -0.5, +1, +2} logits.

The Rasch-Andrich thresholds (arrowed) between categories A and SA, and categories SA and VSA, are in ascending order with the categories along the latent variable. This is termed "ordered". The thresholds between categories D and N and categories N and A are in reverse order on the latent variable to the substantive advance of the categories, D then N then A. This is termed "disordered". So, overall, the Rasch-Andrich thresholds are disordered for this item.

Demarcation by Modal Categories

One approach is to partition the variable according to which category is the most probable to be observed according to the Rasch model. In this example, the intervals correspond to the tops of the "hills" in the Figure. They are VSD: -∞ to -1.25; D: -1.25 to -0.25; A: -0.25 to +1; SA: +1 to +2; VSA: +2 to +∞. This can be summarized as {-1.25, n, -0.25, n, +1.00, +2.00) where "n" means "Non-modal category", i.e., categories SD and N which do not appear. When the Rasch-Andrich thresholds are ordered, then the modal values coincide with those threshold values, and all categories appear on the latent variable.

Demarcation by Median Categories

Another approach is to partition the variable according to which category is in the middle of the probability range. In other words, there is a less than .5 probability that the respondent would choose a category below this category and there is also less than a .5 probability that the respondent would choose a category above this category. L. L. Thurstone reported item locations using medians, so these locations are termed Rasch-Thurstone thresholds.

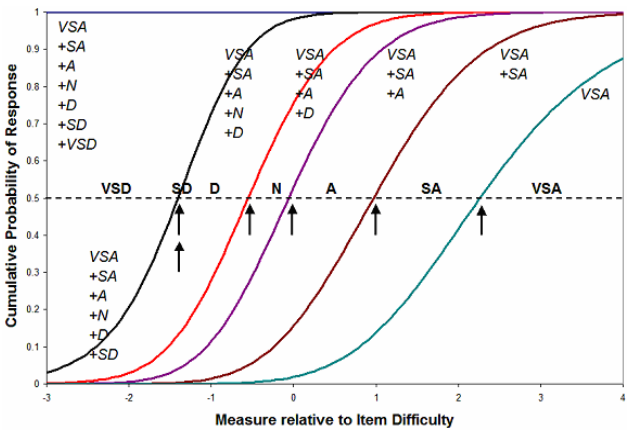


Figure 2. Modeled cumulative category probability curves

In Figure 2, these are indicated by arrows. They indicate the locations where the cumulative probability curves intersect the .5 probability line. The rightmost curve is the probability of observing category VSA. The next left curve is the accumulated probability of observing VSA or SA. The next left curve is the accumulated probability of VSA or SA or A. And so on. The left most curve would be the probability of observing any of the categories. This always has the value of 1.0. The interval between curves along the .5 probability line corresponds to the median category. These are indicated by arrows.

Category N, which always has a low probability of being observed is shown with a narrow interval. The VAS+SA+A+N+D and VAS+SA+A+N+D+SD curves coincide because SD has zero probability of being observed in this data set. Consequently category SD is regarded as a point located where these curves intersect the .5 probability line.

The Rasch-Thurstone thresholds are useful for indicating the cut-points for the probability that the response will be at or above a certain category or below a certain category. In this example they are {-1.40, -1.40, -.44, -.06, .97, 2.26}.

Demarcation by Mean Categories

A third approach is to identify the intervals by means of the expected average value of the responses at each point on the latent variable. These expected values are the sum of the category values multiplied by their probabilities. These form a monotonic ascending ogive from the lowest category value to the highest category values, 0 to 6 in this example. In this conceptualization, the interval or zone on the latent variable corresponding to, say, category 2, contains the measures corresponding to expected scores on the item from 1.5 to 2.5.

These Rasch-half-point thresholds are indicated by arrows in Figure 3. Each arrowed location can be thought of as the measure on the latent variable at which the average response rating of 1,000 persons is expected to approximate 0.5 score-points away from an integer category value. So these intervals are the zones on the latent variable in which the rounded expected scores are

the category values. In this example, they are {-1.94, -1.21, -0.63, +.06, +1.00, +2.53}.

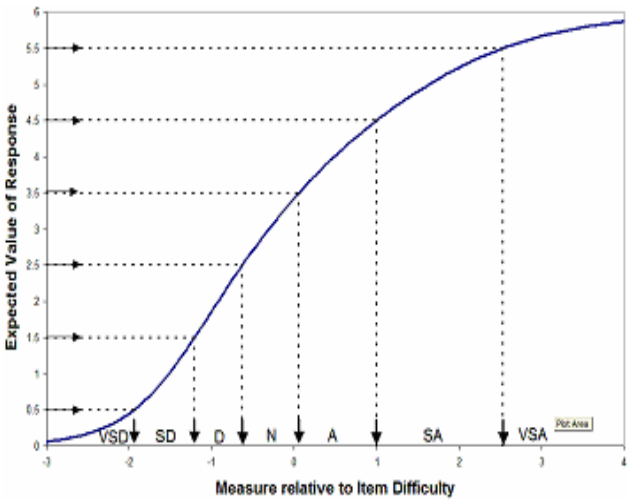


Figure 3. Model expected score ogive for item

This ogive is fundamental to the estimation of Rasch measures. It is seen that, though category SD (=1) is never observed, an average rating corresponding to its category value of 1 is expected. If this makes no sense for a particular instrument, then the unobserved category must be dropped from the analysis by renumbering the observed categories to exclude the unobserved category. In this case, the categories would be renumbered 0=VSD, 1=SD, 2=N, 3=A, 4=SA, 5=VSA.

Communicating with Category Demarcations

The three demarcations communicate the same measurement information, but in ways which answer different questions.

For a person at a particular location on the latent variable:

1. What is the person's most probable response category ? *The modal category.*

2. At or above what response category is the person's probable response? *The median category.* Also, at or below, etc., what response category?

3. What category value is nearest to the person's average expected response? *The mean category.*

Practical note: The modal-category interpretation (particularly when the Rasch-Andrich thresholds are ordered) provides a direct connection between the categories most likely to be observed and their measures. The median-category interpretation provides convenient above-or-below locations for cut-points on items when printed on item maps. The mean-category interpretation is useful for summarizing expected sample behavior. It has the advantage of relating the rating scale to the latent variable as one plotted line without explicit mention of the underlying probabilities.

John Michael Linacre

Multiple Significance Tests

In Rasch analysis, multiple significance tests predominate. Commonly, there is a significance test associated with every item, every person, every rating scale category, every differential-item-functioning (DIF) effect, and more.

1. Conceptualized as single significance tests.

Significance tests are usually reported as the probabilities of single tests. On seeing that the response string associated with item 23 has a very low probability of being generated in accord with a Rasch model, we are prone to say to ourselves, "The purpose of this experiment was to test a hypothesis regarding the fit of the response string for Item 23. Consequently, the single-test probability is the relevant one."

2. Conceptualized as multiple independent tests of the same process.

Consider a 100 item test with responses that accord with the Rasch model. Then the expectation is that 5 or so item response strings have a probability of $p \leq .05$ of according with the Rasch model. So how unlikely must a response string be for it to be significantly unexpected? A technique attributed to Carlo Bonferroni employs the following logic for testing "the universal null hypothesis":

α is the Type I error for a single test (incorrectly rejecting a true null hypothesis). This is .05 for a single test of $p \leq 0.05$. So, when the data fit the model, the probability of a correct finding for one test is $(1-\alpha)$, and for two independent tests $(1-\alpha)^2$, and for n tests, $(1-\alpha)^n$. Consequently the Type I error for n independent tests is $1-(1-\alpha)^n$. Thus, if we intended the Type I error for the multiple test to be α, then the level for each single test is $\alpha' = 1 - (1-\alpha)^{1/n} \approx \alpha/n$. So that for a finding of $p \leq .05$ to be found for 100 items, then at least one item would need to be reported with $p \leq .0005$ on a single item test for the hypothesis that "the entire set of items fits the Rasch model" to be rejected .

An obvious problem with adopting this technique routinely in Rasch work is that a set of items may be accepted that includes obviously bad items. For a finding of $p \leq .05$ to be found for 100 items, then at least one item would need to be reported with $p \leq .0005$ ($t \geq 3.4$) on a single item test. This degree of misfit generally requires a sample size of about 1,000 to be observable. 20 items reported with $.005 < p < .01$ ($2.8 > t > 2.6$) would not be deemed as sufficient to reject the null hypothesis that the data fit the Rasch model. It can be seen that the Bonferroni logic considers Type I error, but ignores Type II error (incorrectly rejecting a true alternative hypothesis), especially for individual "bad" items.

So, when does Bonferroni correction work? Apparently in decision-making situations in which a production batch is to be accepted or rejected based on testing of the quality of a sample from the batch. *T. V. Perneger (1998) What's wrong with Bonferroni adjustments? British Medical Journal, 316, 1236-1238.*

3. Multiple tests conceptualized as accumulating individual tests.

Benjamini and Hochberg (1995) suggest that an incremental application of Bonferroni correction overcomes its drawbacks. Here is their procedure:

i) Perform the n single significance tests.

ii) Number them in ascending order by probability $P(i)$ where $i = 1,n$ in order.

iii) Identify k the largest value of i for which $P(i) \leq \alpha * i / n$

iv) Reject the null hypothesis for $i = 1, k$

In our example of a 100 item test with 20 bad items with $.005 < p < .01$, the threshold values for cut-off with α of $p \leq .05$ would be: 0.0005 for the 1st item, .005 for the 10th item, .01 for the 20th item, .015 for the 30th item. So that k in our example would be at least 20 and perhaps more. All our bad items have been flagged for rejection.

There are other techniques for multiple significance tests. Please contact Rasch Measurement Transactions if you have found any to be useful.

Benjamini Y. & Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B, 57,1, 289-300.

Fred Wolfe, Randy MacIntosh, Svend Kreiner, Rense Lange, Roger Graves, John Michael Linacre contributed to the Rasch Listserv discussion on this topic.

Ben Wright at 80

Benjamin D. Wright celebrated his 80th birthday on March 30, 2006. At that time, he had recently fallen and sustained a hip injury. He is recovering in the Warren N. Barr Pavilion of the Illinois Masonic Medical Center. He welcomes visitors and telephone calls.

If you set an item on fire,
it will not retain its difficulty.

Benjamin D. Wright
May, 1976

Long on the wall of Ben's office
in Judd Hall, University of Chicago

In 1962, Georg Rasch remarked during his Inauguration Lecture as Professor in *Theoretical Statistics* at the University of Copenhagen that he would rather have had a chair in *Heretical Statistics.*

Courtesy of Helmuth Nyborg