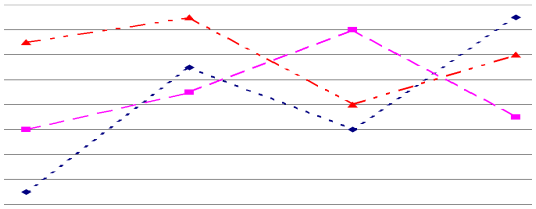
Measurement Profiles



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 21 No. 3 Winter 2007 ISSN 1051-0796

Local Dependency and Rasch Measures

Local independence of items is an assumption in Rasch model and all IRT models. That is, the items in a test should not be related to each other. Sharing a common passage, which is prevalent in reading comprehension tests and cloze tests cab be a potential source of local item dependence (LID). It is argued in the literature that LID results in biased parameter estimation and affects the unidimensionality of the test. In this study the effects of the violation of the local independence assumption on the person measures are studied.

The items that are put to Rasch analysis are required to be independent of each other. That is, a correct or wrong reply to one item should not lead to a correct or wrong reply to another item. This means that there should not be any correlation between two items after the effect of the underlying trait is conditioned out, i.e., the correlation of residuals should be zero. The items should only be correlated through the latent trait that the test is measuring (Lord and Novick, 1968). If there are significant correlations among the items after the contribution of the latent trait is removed, i.e., among the residuals, then the items are locally dependent or there is a subsidiary dimension in the measurement which is not accounted for by the main Rasch dimension (Lee, 2004). In other words, performance on the items depends to some extent on a trait other than the Rasch dimension which is a violation of the assumptions of local independence and unidimensionality. If the assumption of local item independence is violated, any statistical analysis based on it would be misleading. Specifically, estimates of the latent variables and item parameters will generally be biased because of model misspecification, which in turn leads to incorrect decisions on subsequent statistical analysis, such as testing group differences and correlations between latent variables. In addition, it is not clear what constructs the item responses reflect, and consequently, it is not clear how to combine those responses into a single test score, whether IRT is being used or not (Wang et al., 2005, p.6).

When a set of items are locally dependent they can be bundled into polytomous super-items, that is, the set of items which are related to a common stimulus are considered as one polytomous item to partial out the influence of local item dependence (LID) among items within each super-item. Polytomous Rasch models or IRT models such as Andrich's rating scale model or Masters' partial credit model, etc. are then applied to analyze the testlets. The drawback to bundling dichotomies into polytomies is a loss of statistical and diagnostic information.

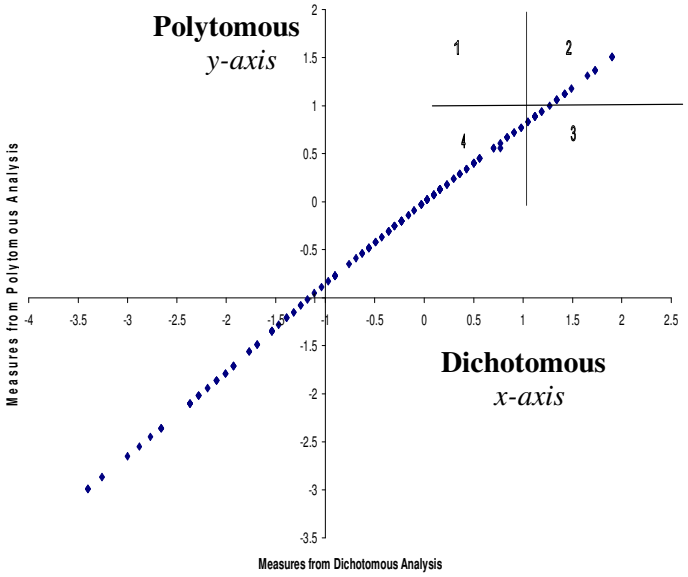


Figure: Plot of person measures from the two analyses. The "+" indicates a hypothetical cut-off score.

The problem of LID is not new and has also been addressed in the classical test theory. Dependency among items can inflate reliability and give a fake impression of the precision and quality of the test. It is argued in the literature that if the local independence assumption does

Table of Contents

not hold, the local dependence itself acts as a dimension. If the effect of LID is substantial it is difficult to say what dimension the main Rasch dimension is. Even if the effect is small, the derived measures will be contaminated, i.e., the measures partially reflect the LID dimension to the extent that LID exists. In fact, LID is a form of violating the unidimensionality principle. LID also results in artificially small standard errors of estimates (SEE) and the overestimation of reliability.

Case Study

In this section the effects of the violation of the assumption of local item independence on the person ability measures in a C-Test are investigated and the impact of LID on decision-making in a hypothetical assessment is studied.

A four-passage C-Test, each passage containing twenty-five blanks, was administered to 160 persons. The C-Test is a variation of the cloze test where the second half of every second word is deleted. Test-takers have to reconstruct the broken words. The C-Test was chosen to conduct this study because the format of the C-Test should be conducive to local dependency and the level of local dependency is presumably high in the context of a C-Test. The data were analyzed twice, once using Rasch's (1960) dichotomous model, treating each gap as an independent dichotomous item and once treating each passage as a polytomous item or testlet (with 25 categories) using Master's (1982) partial credit model. For each person two measures were obtained, one based on the dichotomous analysis and one based on the polytomous analysis.

The measures from the two analyses are cross-plotted in the Figure. The range of the ability measures is wider for the dichotomous measures (5.3 logits) than the polytomous measures (4.5 logits).

As far as criterion-referenced decision-making is concerned we do make somewhat different decisions depending on which analysis we use. In the Figure, a hypothetical cut-score at +1 logit is imposed. For persons who fall in areas 2 and 4 we will be making the same decisions. Test-takers who fall in areas 1 and 3 would have opposite decisions depending on the analysis. Here, no one falls in area 1 but four test-takers fall in area 3. That is, if we base our decision-making on the dichotomous analysis these four people pass and if we decide on the basis of polytomous analysis these four test-takers fail. Depending on the manner in which the +1 logit cut-score was determined, four people may be mistakenly passed or failed depending on the analytical approach.

Conclusion

When the data are expressed in dichotomous form, the local dependence makes the data too predictable. The practical effect is to increase the range of the measures. When the data are summarized into polytomous items, the local dependence is lessened, so making the data less predictable and the range of the abilities narrower.

In the case study, the relationship between the two sets of ability measures is almost linear. Consequently, when the ability measures are rescaled into a more convenient unit for communication to stake-holders, the logit-differences due to local dependence may vanish. Nevertheless, the artificially high reliability and the impact on examinees near a cut-score remain.

Purya Baghaei

Lee, Y. (2004) Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing*, 21:1, 74-100.

Lord, F. M. and Novick, M. R. (1968) *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.

Wang, W. & Wilson, M. (2005) Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, 29: 4, 296-318.

PROMS 2008 - Tokyo
July 31, August 1-3, 2008

The Pacific Rim Objective Measurement Symposium, PROMS, will be held at Ochanomizu University in the Kanda district of Tokyo, Friday - Sunday, August 1-3, with Rasch workshops scheduled for Thursday, July 31.

The deadline for proposals is February 28th. All abstracts will be submitted and reviewed through our website, which can be accessed at:

http://www.proms-tokyo.org/

The PROMS 2008 website was created using Open Conference Systems, an open source web publishing tool created by the Public Knowledge Project. From the web site, each PROMS 2008 participant can create a user account, submit presentation proposals for review, volunteer to be a reviewer, manage and review proposals online, as well as pay registration fees via credit card. Please have a look around the website and create a user account. If you can volunteer some time to review and comment on a few proposals, please tick the "reviewer" role when you create your user accounts The website will be updated regularly as we get further information on invited speakers, scheduled events, and symposium sponsors, so check back regularly.

We look forward to your participation in the symposium, and a chance to show you a bit of Tokyo.

The PROMS 2008 organizing Team:

Jim Sick, J.F. Oberlin University, Tokyo
Edward Schaefer, Ochanomizu University, Tokyo
Christopher Weaver, Jissen Women's University
Jeff Durand, Kanda University of International Studies

Please direct PROMS 2008 correspondence to:

proms -x- windshimmer.com

Preliminary IOMW 2008 Conference Program

New York University, New York, NY
March 22-23, 2008

Saturday, March 22

Registration and Breakfast (8:00 - 8:45)

Session 1 (8:45 - 10:15)

The Rasch Model as a Power Series Distribution for Equating Tests Which Measure the Same Proficiency. *David Andrich, The University of Western Australia*

Historical View of Theories of Measurement and Language Proficiency within the Context of Second Language Testing. *George Engelhard, Jr., Emory University*

Generally Objective Measurement of Human Temperature and Reading Ability: Some Corollaries. *Jack Stenner, MetaMetrics*

A Study of the Influence of Labels Associated with Anchor Points of Likert-Type Response Scales in Survey Questionnaires. *Jean-Guy Blais, University of Montreal, Julie Grondin, University of Montreal*

Session 2a (10:30-12:00)

Symposium Title: Partial Credit Model Analyses of Psychological, Social, and Cultural Factors and Relationships with Individual Experiences of Chronic Pain: A Symposium

Symposium Organizer: *Karen M. Schmidt, University of Virginia*

Never Getting a Break: Persistent High Pain Intensity Relationships with Personality in Chronic Pain Sufferers. *Karen M. Schmidt*

Pain Intensity, Catastrophizing, and Affect in Chronic Pain Sufferers. *Monica K. Erbacher*

Influence of Significant Other Response on Pain Intensity and Psychological Distress in Chronic Pain Sufferers. *Katie J. Ameringer*

Rasch Partial Credit Model (PCM) and Differential Item Functioning (DIF) Analysis of the Impact of Culture on Locus of Control in Chronic Pain Sufferers. *Juliana R. Schroeder*

Pain Coping and Significant Other Relationships for Chronic Pain Sufferers. *David J. Lick*

Session 2b (10:30-12:00)

Using Rasch Analysis to Construct a Trust in Medical Technology Instrument. *Enid Montague, Virginia Tech, Edward W. Wolfe, Virginia Tech, Brian M. Kleiner, Virginia Tech, Woodrow Winchester II, Virginia Tech*

Assessing Student Perceptions of High School Science Classroom Environments: A Validation Study. *Christine D. Luketic, Virginia Tech, Edward W. Wolfe, Virginia Tech, Kusum Singh, Virginia Tech, Erin Dolan, Virginia Tech*

Measuring Positiveness Towards Educational Policy. *Jinnie Choi, University of California, Berkeley*

A Multilevel Item Response Theory Analysis of Health-Related Quality of Life: An Illustration with 11,158 Healthy and Chronically Ill Children using the PedsQL™ Emotional Functioning Scale. *Prathiba Natesan, University of Miami, Christine Limbers, Texas A&M University, James W. Varni, Texas A&M University*

IOMW 2008

March 22-23, 2008 - New York

Data Recognition Corporation, New York University, and JAM Press are pleased to announce that International Objective Measurement Workshop, IOMW 2008, will be held in New York City at New York University on March 22 and 23, 2008, just prior to the AERA annual meeting. This is the fourteenth meeting of IOMW, a series of biannual meetings that originated in 1981. The first IOMW was organized by Ben Wright and held at the University of Chicago. IOMW is an opportunity to meet Rasch colleagues and hear about exciting new developments in objective measurement, theory and practice.

Conference registration is now open. The registration fee is $40 USD (early) and $50 USD (late). A printable pdf is on http://www.jampress.org/ that can be completed and returned by e-mail or surface mail. The final date for early registration is March 14, 2008. Onsite registration will be accepted. Checks, VISA, and MasterCard payments are accepted for registration fees.

For further information, see http://www.jampress.org/

Richard M. Smith, Editor
Journal of Applied Measurement

Session 3a (1:30 – 3:00)

Measuring Student Proficiency with the Constructing Measures Framework. *Brent Duckor, University of California, Berkeley, Mark Wilson, University of California, Berkeley*

A Comparison of Structural Equation and Multidimensional Rasch Modeling Approaches to Confirmatory Factor Analysis. *Edward W. Wolfe, Virginia Tech, Kusum Singh, Virginia Tech*

Using the Standardized Letters of Recommendation in Selection: Results from a Multidimensional Rasch Model. *Ou Lydia Liu, Educational Testing Service, Jennifer Minsky, Educational Testing Service, Guangming Ling, Educational Testing Service, Patrick Kyllonen, Educational Testing Service*

A Summary Index of Multidimensionality in Scales Composed of Subscales: Applications to Traditional and Rasch Measurement Theory. *Barry Sheridan, RUMM Laboratory, David Andrich, University of Western Australia*

Session 3b (1:30 – 3:00)

ConstructMap: A Software Demonstration. *Andy Maul, Berkeley Evaluation and Assessment Research (BEAR) Center, Cathleen Kennedy, Berkeley Evaluation and Assessment Research (BEAR) Center*

Session 4a (3:30 – 5:00)

A Bootstrap Approach to Evaluating Person and Item Fit to the Rasch Model. *Edward W. Wolfe, Virginia Tech*

Are QOL and Spirituality Separate Constructs? A Lesson from Breast Cancer Patients. *Nikolaus Bezruczko, Measurement and Evaluating Consulting, Kyle Perkins, Florida International University, David Cella, Northwestern University*

Measures That Count, Measures That Matter. *William P. Fisher, Jr., Avatar International, Brent Duckor, University of California, Berkeley*

Exploration of a Taxonomic Framework to a New Instrument Development and Item Types: Dimensional Disaster or Informed Instrument? *Judy R. Wilkerson, Florida Gulf Coast University, W. Steve Lang, University of South Florida St. Petersburg*

Session 4b (3:30 – 5:00)

ConQuest: A Software Demonstration. *Mark Wilson, Berkeley Evaluation and Assessment Research (BEAR) Center*

Sunday, March 23

Breakfast (8:00 - 8:45)

Session 5 (8:45 - 10:15)

A History of Benjamin Wright in New York City. *Nikolaus Bezruczko, Measurement and Evaluating Consulting*

Functional Assessment in a Wellness Program for the Frail Elderly. *Dr. Carl V. Granger, SUNY at Buffalo, Department of Rehabilitation Medicine, Uniform Data System for Medical Rehabilitation*

Validity and Objectivity in Health Related Scales: A Second Look at SF36. *Svend Kreiner, University of Copenhagen*

Measuring Mental Health Problems among Adolescents– the Youth Self-Report Examined with the Rasch Model. *Curt Hagquist, Karlstad University, David Andrich, University of Western Australia, Sven R. Silburn, Curtin University of Technology, Stephen R. Zubrick, Curtin University of Technology*

Session 6a (10:30 - 12:00)

Solving Incomplete Paired Comparison Matrices. *Ronald Mead, Data Recognition Corporation*

Measurement of Student Nurse Performance in the Safe Administration of Medication. *Deborah Ryan, Emory University*

Measurement of Visual Disability in Low Vision Patients: Does DIF for Health Status Matter? *Lohrasb Ahmadian, Johns Hopkins University*

Proficiency at Scoring and Preventing Touchdowns: Pairwise Comparisons. *Ronald Mead, Data Recognition Corporation, Christie Plackner, Data Recognition Corporation, Vincent Primoli, Data Recognition Corporation*

Session 6b (10:30 - 12:00)

Using the Rating Scale Model to Examine the Angoff Ratings of Standard-Setting Panelists. *Jade Caines, Emory University, George Engelhard, Emory University*

Examining the Bookmark Ratings of Standard-Setting Panelists: An Approach Based on the Multifaceted Rasch Measurement Model. *Rubye Sullivan, Emory University, Jade Caines, Emory University, Courtney Tucker, Emory University, George Engelhard, Jr., Emory University*

The Construction of the Malaysian Educators Selection Inventory (MedSI): A Large Scale Assessment Initiative. *Joharry Othman, International Islamic University*

Construct Development for Linear Measurement of Accessibility to Education in Regions of the Russian Federation. *Anatoli Maslak, Slavyansk-on-Kuban State Pedagogical Institute, Tatyana Anisimova Slavyansk-on-Kuban State Pedagogical Institute, Nikolaus Bezruczko, Measurement and Evaluating Consulting*

Session 7a (1:30 - 3:00)

Defining a Measurement Scale for Curriculum Evaluation. *Ronald Mead, Data Recognition Corporation, Julie Korts, Data Recognition Corporation, Kyoungwon (Kei) Lee, Data Recognition Corporation*

Inferring an Experimentally Independent Response Space for the Rasch model for Ordered Categories from its Experimentally Dependent Subspace. *David Andrich, The University of Western Australia*

Optimizing Response Categories in a Measure of Health Care Quality Perceptions. *William Fisher, Jr., Avatar International, Geoffrey A. Nagle, Tulane University, Clayton Williams, Louisiana Public Health Institute*

Is the Partial Credit Model a Rasch Model? *Robert W. Massof, Johns Hopkins University*

Session 7b (1:30 - 3:00)

Development of the "Chinese Character Difficulty Scale" for Primary students. *Magdalena Mo Ching MOK, The Hong Kong Institute of Education, Dr Yee Man Cheung, The Hong Kong Institute of Education*

Assessing Students' Perceived Quality of University Courses using a Multilevel Two-Dimensional Item Response. *Isabella Sulis, University of Cagliari, Vincenza Capursi, S. Vianelli*

Constructing the Variable "Explanation in Mathematics", *Brian Doig, Deakin University, Susie Groves, Deakin University*

Testing the Assumption of Sample Invariance of Item Difficulty Parameters in the Rasch Rating Scale Model, *Joseph A. Curtin, Brigham Young University*

Session 8a (3:30 - 5:00)

Calibrating Instruments for Improving What We Do: Establishing Rasch Measurements for Self-Theories of Intelligence, *Sharon Solloway, Bloomsburg University*

Constructing a Variable: Hotel Satisfaction, *Trevor Bond, Hong Kong Institute of Education*

The Effect of Assessment Context on Construct Definition in Direct Writing Assessment. *Sharon E. Osborn Popp, Arizona State University*

Déjà vu: The Rasch Measurement Model and Google's PageRank Algorithm. *Mary Garner, Kennesaw State University*

Session 8b (3:30 - 5:00)

The Conjoint Additivity of the Lexile Framework for Reading. *Andrew Kyngdon, MetaMetrics*

A Developmental Framework for the Measurement of Writer Ability. *Harold Burdick, MetaMetrics, Jack Stenner, MetaMetrics, Donald Burdick, MetaMetrics, Carl Swartz, MetaMetrics*

Using Confirmatory Factor Analysis and Rasch Measurement Theory to Assess Measurement Invariance in a High Stakes Reading Assessment, *Jennifer Randall, University of Massachusetts, Amherst, George Engelhard, Jr., Emory University*

An Examination of Fairness of Math Word Items. *Xuejun (Ina) Shen, Stanford University, Xiaohui Zheng, University of California, Berkeley, Edward Haertel, Stanford University*

AERA 2008 Rasch-Related Activities

Monday, March 24

Applications of the Rasch Model to the Analysis of Ratings. SIG - Rasch Measurement
Mon, Mar 24 - **12:00pm - 1:30pm Hilton New York / Harlem Suite, 4th Fl.**
Comparing Expert and Nonexpert Raters in Essay Scoring With Many-Facet Rasch Model. *Eunlim Chi (Kyung Hee University)*
Longitudinal Reliability of Decisions Derived From Objective Standard-Setting for Judge-Mediated Examinations. *Gregory E. Stone (University of Toledo), Svetlana A. Beltyukova (University of Toledo), Christine M. Fox (University of Toledo), Douglas Edward Stone (University of South Florida)*
Quality of Uni-Level Writing Tasks Linked to the CEFR: Results from Generalizability and Facets Analyses. *Claudia Harsch (Humboldt University - Berlin), Andre A. Rupp (Institute for Educational Progress), Raphaela Oehler (Humboldt University - Berlin), Guido Martin (IEA Data Processing Center - Hamburg)*
Rater Effects in the Consensual Assessment of Creative Products. *Kun-Shia Liu (National Sun Yat-Sen University), Ying-Yao Cheng (National Sun Yat-Sen University), Wen-Chung Wang (National Chung Cheng University)*
Chair: *William S. Lang (University of South Florida - St. Petersburg)*
Discussant: *Karen L. Draney (University of California - Berkeley)*

Psychometric Issues with Performance Assessment. Division D - Measurement and Research Methodology - Section 1: Educational Measurement, Psychometrics and Assessment
Mon, Mar 24 - **4:05pm - 5:35pm New York Marriott Marquis Times Square / Barrymore Room, 9th Fl.**
Comparing the Effectiveness of Two Models for Equating a Large-Scale Standardized Performance Assessment. *Irina Grabovsky (National Board of Medical Examiners), Raja G. Subhiyah (National Board of Medical Examiners), Kimberly A. Swygert (National Board of Medical Examiners), Kevin Balog (National Board of Medical Examiners)*

Tuesday, March 25

PDC16: An Introduction to Latent Class Models, Mixture Rasch Models, and Diagnostic Mixture Models. Professional Development Training
Tue, Mar 25 - **8:00am - 12:00pm Crowne Plaza Hotel Times Square / Broadway Ballroom, Act IV, 4th Fl.**
Director: *Matthias Von Davier (ETS)*

Approaches to Enacting and Evaluating Science Curriculum. Division C - Learning and Instruction - Section 4: Science
Tue, Mar 25 - **10:35am - 12:05pm Hilton New York / Green Room, 4th Fl.**
Fidelity of Implementation to Instructional Strategies as a Moderator of Science Curriculum Unit Effectiveness. *Carol L. O'Donnell (George Washington University), Sharon J. Lynch (George Washington University)*

Spirituality and Education: Paper Discussion. SIG - Spirituality and Education
Tue, Mar 25 - **12:25pm - 1:05pm Hilton New York / Trianon Ballroom/Petit Trianon, 3rd Fl.**
Getting Mindfulness Into the Curriculum at a Public University: The Continuing Story of the Impact of a Project Measuring Mindfulness and Mindset Orientations. *Sharon G. Solloway (Bloomsburg University of Pennsylvania), William P. Fisher (Avatar International, Inc.)*

Paper Discussions: Psychometrics. Division D - Measurement and Research Methodology - Section D: Measurement and Research Methodology
Tue, Mar 25 - **12:25pm - 1:05pm New York Marriott Marquis Times Square / Broadway Ballroom, Broadway North, 6th Fl.**
Biased Sample Calibration of Structural Parameters for the Rasch And Two-Parameter Logistic IRT Models. *Insu Paek (ETS)*
Comparing Three Calibration Methods in a Mixed-Format Large-Scale Assessment Using Empirical Data. *Kevin Fatica (CTB/McGraw-Hill), Kooghyang Ro Um (Pearson Educational Measurement), Dong-In Kim (CTB/McGraw-Hill LLC), Leonardo S. Sotaridona (CTB/McGraw-Hill LLC)*
Modeling Growth: A Longitudinal Study Based on a Vertical Scaled English-Language Proficiency Test. *Zhen Wang (Harcourt), Husein Taherbhai, Husein (The Federation of the State Boards of Physical Therapy), Ming Xu (The New York State Education Department), Sz-Shyan Wu (New York State Education Department)*

Applications in Rasch Measurement. SIG - Rasch Measurement
Tue, Mar 25 - **1:15pm - 1:55pm Sheraton New York Hotel & Towers / Metropolitan Ballroom, Metropolitan East, 2nd Fl.**
A Multifacet Rasch Analysis of a Career Commitment Essay-Scoring Process. *Susan M Gracia (Rhode Island College)*
Applying Multidimensional Partial Credit Model in a Longitudinal Design of Diagnostic Assessment. *Feifei Ye (University of Pittsburgh), Wenyi You (Pearson Educational Measurement)*
Applying the Rasch Model to Develop a Tacit Knowledge Measure of Effective Superintendent Leadership. *Christian E. Mueller (University of Memphis), Kelly D. Bradley (University of Kentucky)*
Construct Development for Linear Measurement of Accessibility to Education in Regions of the Russian Federation. *Anatoly Andreyevich Maslak (Slavyanskon-Kuban State Pedagogical Institute), T. S. Anisimova (Slavyansk-on-Kuban State Pedagogical Institute), Nikolaus Bezruczko (Measurement and Evaluation Consulting)*
Exploring the Structure of Achievement Goal Orientations Using Multidimensional Rasch Models. *Daeryong Seo (Harcourt Assessment, Inc.), Husein Taherbhai, Husein (The Federation of the State Boards of Physical Therapy), Yu Sun (Harcourt Assessment, Inc.)*
Revision of the Assessment Practice Inventory (API): A Combined Exploratory Factor Analysis and Polytomous IRT Approach. *Judith A. Burry-Stock (University of Alabama)*

Issues in Large-Scale Assessment. SIG - Large Scale Assessment
Tue, Mar 25 - **4:05pm - 5:35pm New York Marriott Marquis Times Square / Shubert Complex, Royale Room, 6th Fl.**
Examining Population Invariance in Equating and Linking Functions in Statewide Large-Scale Assessment Programs. *Yi Du (Data Recognition Corp.)*

Investigations in Computer-Based and Computer-Adaptive Testing. Division D - Measurement and Research Methodology - Section 1: Educational Measurement, Psychometrics and Assessment
Tue, Mar 25 - **4:05pm - 6:05pm New York Marriott Marquis Times Square / Barrymore Room, 9th Fl.**
Item-Selection Strategies in Computerized Adaptive Testing Under the Rasch Testlet Model. *Ching-Lin Shih (National Taichung University), Wen-Chung Wang (National Chung Cheng University)*

Rasch Measurement SIG Business Meeting
Tue, Mar 25 - **6:15pm - 7:45pm New York Marriott Marquis Times Square / Jolson Room, 9th Fl.**
Chair: *Thomas R. O'Neill (National Council of State Boards of Nursing)*
Participant: *Edward W. Wolfe (Virginia Tech University)*
Invited Speaker: Explanatory item response models: a matrix representational framework *Paul De Boeck (Katholieke Universiteit Leuven)* http://www.kuleuven.be/cv/u0002630e.htm

Wednesday, March 26

Innovations in Rasch Measurement. SIG - Rasch Measurement
Wed, Mar 26 - **8:15am - 8:55am Sheraton New York Hotel & Towers / Metropolitan Ballroom, Metropolitan East, 2nd Fl.**
A Comparison of the Test Design Variations in Panel Structures of the Computerized Adaptive Sequential Testing System Under the Partial Credit Model. *Jiseon Kim (University of Texas - Austin), Cheryl H. H. Tseng (University of Texas - Austin), Hyewon Chung (University of Texas - Austin), Barbara G. Dodd (University of Texas - Austin)*
Assessing Invariance in Polytomous Items Following the Partial Credit Model Within the Rasch Framework. *Nicholas D. Myers (University of Miami), Randall D. Penfield (University of Miami), Edward W. Wolfe (Virginia Tech University)*
DIF Analysis of the English and Chinese Versions of the Fagerstrom Test for Nicotine Dependence. *Hiroyuki Yamada (University of California - Berkeley), Janice Y. Tsoh (University of California - San Francisco), Scott Acton (Rochester Institute of Technology)*
Effects of Missing Data Proportion on Parameter Recovery Under the Facets Model. *Yi-Hung Lin (National Chung Cheng University - Taiwan), Wen-Chung Wang (National Chung Cheng University)*
Sensitivity of the Least Squares Distance Method of Cognitive Analysis to Misspecifications in the Incidence Q-matrix: A Simulation Study. *Dimiter M. Dimitrov (George Mason University), Sonia Janeth Romero (Universidad Autonoma de Madrid, Spain), Vicente Ponsoda (Universidad Autonoma De Madrid)*
Traditional Rasch Methods for Selecting Stable Linking Items. *Anita Rawls (University of South Carolina), Huynh Huynh (University of South Carolina)*
Treatments of Missing Data and Parameter Estimation in WINSTEPS. *Sungworn Ngudgratoke (Michigan State University), Laddawan Petchroj (Sukhothai Thammathirat Open University), Nalinee Na Nakorn (Sukhothai Thammathirat Open University), Ratchaneekool Pinyopanuwat (Sukhothai Thammathirat Open University), Wanna Denkajornkiat (Chulalongkorn University)*

Technical Issues in Rasch Measurement. SIG - Rasch Measurement
Wed, Mar 26 - **10:35am - 12:05pm Sheraton New York Hotel & Towers / Lenox Ballroom, 2nd Fl.**
Is Absolute Magnitude Estimation Scaling a Viable Alternative to Categorical Rating Scaling in Social Science? An Experimental Study. *Kristin Lea Kelly (University of Toledo), Toni Ann Sondergeld (University of Toledo), Svetlana A. Beltyukova (University of Toledo), Christine M. Fox (University of Toledo)*
Investigating the Equivalence of Test Items: An Extension of Rasch Item Information Properties. *Adam Edward Wyse (Michigan State University), Raymond Mapuranga (ETS)*
Item Exposure Constraints for Mixed-Format Test With the Partial Credit Model. *Tsung-Han Ho (University of Texas - Austin), Barbara G. Dodd (University of Texas - Austin)*
Stability of the Philadelphia Geriatric Center Morale Scale: A Multidimensional Item Response Theory Analysis. *Lin Ma (University of Denver), Kathy E. Green (University of Denver), Enid O. Cox (University of Denver)*
The Rasch Model across More Than One Frame of Reference. *David Andrich (University of Western Australia), Stephen M Humphry (University of Western Australia)*
Chair: *Kelly D. Bradley (University of Kentucky)*
Discussant: *G. Gage Kingsbury (Northwest Education Association)*

New Technologies and New Approaches to Learning. Division C - Learning and Instruction - Section 7: Technology Research
Wed, Mar 26 - **12:25pm - 1:55pm Sheraton New York Hotel & Towers / Carnegie Suite East, 3rd Fl.**
Assessment in E-Learning: Analyzing Architectural Sketching With a Digital Pen. *Nargas Oskui (University of Oregon), Kathleen Scalise (University of Oregon)*

Directions in Reading and Literacy Research III. SIG - Research in Reading and Literacy
Wed, Mar 26 - **1:15pm - 1:55pm Hilton New York / Trianon Ballroom/Petit Trianon, 3rd Fl.**
Mind Your Ps and Qs: Approaching Literacy Instruction From an Item Response Theory Base. *Luke S. Duesbery (San Diego State University), Julie Alonzo (University of Oregon), Leanne Rae Bettesworth (University of Oregon)*

Applied Item Analyses. Division D - Measurement and Research Methodology - Section 1: Educational Measurement, Psychometrics and Assessment
Wed, Mar 26 - **2:15pm - 3:45pm New York Marriott Marquis Times Square / Marquis Ballroom, Salon A, 9th Fl.**
An Investigation of the Changes in Item Parameter Estimates for Items Re-Field Tested. *Jason L. Meyers (Pearson Educational Measurement), Xiaojing Kong (James Madison University)*

Thursday, March 27

Measuring Change Over Time. Division D - Measurement and Research Methodology - Section 1: Educational Measurement, Psychometrics and Assessment
Thu, Mar 27 - **8:15am - 9:45am New York Marriott Marquis Times Square / Barrymore Room, 9th Fl.**
A Polytomous Rasch Model for Nonlinear Individual Change Over Time in Reviewing a Developmentally Appropriate Assessment. *Sunhee Kim (Rutgers University), Gregory Camilli (Rutgers University)*

Contemporary Challenges With Equating and Linking. Division D - Measurement and Research Methodology - Section 1: Educational Measurement, Psychometrics and Assessment
Thu, Mar 27 - **10:35am - 12:05pm New York Marriott Marquis Times Square / Barrymore Room, 9th Fl.**
Handling the Assumption of Randomly Equivalent Groups in Equating: A Comparison of Rasch and 3PL Software. *David Chayer (Data Recognition Corporation), Larissa Smith (Data Recognition Corporation)*

Applications of Survey Research Methods in Education. SIG - Survey Research in Education
Thu, Mar 27 - **12:25pm - 1:05pm Sheraton New York Hotel & Towers / Metropolitan Ballroom, Metropolitan East, 2nd Fl.**
The Usefulness of Investing in More Fully Mathematical Survey Research: Applying Probabilistic Models to Develop Curricula. *Sharon G. Solloway (Bloomsburg University of Pennsylvania), William P. Fisher (Avatar International, Inc.)*

Scaling and Measurement Issues in Survey Research. SIG - Survey Research in Education
Thu, Mar 27 - **4:05pm - 5:35pm Hilton New York / Concourse D, Concourse Level**
Using Measurement Principles to Construct and Restructure a Teacher Perception Survey. *Jessica D. Cunningham (University of Kentucky), Kelly D. Bradley (University of Kentucky)*

Applied Issues in Rasch Measurement. SIG - Rasch Measurement
Thu, Mar 27 - **4:05pm - 6:05pm Sheraton New York Hotel & Towers / Executive Conference Center, Conference Room C, Lower Lobby**
Exploring Patterns of Classroom Contribution to Person Aberrance. *Alexandra Petridou (University of Manchester), Julian S. Williams (University of Manchester)*
Analysis of Cognitive Attributes for Mathematics Items in the Framework of Rasch Measurement. *Dimiter M. Dimitrov (George Mason University), Encho N. Gerganov (New Bulgarian University - Bulgaria), Maurice Greenberg (New Bulgarian University - Bulgaria), Dimitar V. Atanasov (New Bulgarian University - Bulgaria)*
New Uses of Rasch Scaling for Achievement Progress Monitoring Tests. *Gale H. Roid (Southern Methodist University), Mark F Ledbetter (Riverside Publishing)*
Reliability and Validity of Student Evaluations: A Rasch Analysis. *Zongmin Kang (University of Toledo), Gregory E. Stone (University of Toledo)*
When Are Preschool Teacher Ratings Valid? *Nikolaus Bezruczko (Measurement and Evaluation Consulting)*
Chair: *Gregory E. Stone (University of Toledo)*
Discussant: *Kathy E. Green (University of Denver)*

Friday, March 28

Applications of the Rasch Model in Teaching and Learning. SIG - Rasch Measurement
Fri, Mar 28 - **8:15am - 9:45am Sheraton New York Hotel & Towers / Executive Conference Center, Conference Room B, Lower Lobby**
An Alternative Approach to Establishing Cut Points for Classroom-Based Assessments. *Diana Bernbaum Wilmot (University of California - Berkeley), Cathleen A. Kennedy (University of California - Berkeley)*
Assessing the Fidelity of Structural and Instructional Model Implementation in New Small Schools: The Application of IRT Techniques. *Denis W. Jarvinen (Strategic Measurement and Evaluation, Inc.), Edward W. Wolfe (Virginia Tech University), Bill Conrad (Oakland Unified School District), Jean Y. Wing (Oakland Unified School District)*
Measuring Teacher Dispositions With Different Item Structures: An Application of the Rasch Model. *William S. Lang (University of South Florida - St. Petersburg), Judy Wilkerson (Florida Gulf Coast University)*
The Investigation of a Measure of Transformative Experience: Assessing In-Class and Out-of-Class Engagement. *Kristin Lea Kelly (University of Toledo), Victoria Caterina Stewart (University of Toledo), Kevin J. Pugh (University of Northern Colorado), Christine L. Manzey (University of Toledo)*
Chair: *Alan C. Bugbee (American Society for Clinical Pathology)*
Discussant: *Jon S. Twing (Pearson)*

Assess This: Issues in Course Evaluation. Division J - Postsecondary Education - Section 2: Faculty, Teaching and Learning
Fri, Mar 28 - **8:15am - 10:15am Crowne Plaza Hotel Times Square / Room 405, 4th Fl.**
Patterns of Student Evaluations: A Many-Faceted Rasch Analysis. *Zongmin Kang (University of Toledo), Gregory E. Stone (University of Toledo)*

Measuring Teachers in the Classroom. Division D - Measurement and Research Methodology - Section 1: Educational Measurement, Psychometrics and Assessment
Fri, Mar 28 - **12:25pm - 1:55pm Crowne Plaza Hotel Times Square / Room 504, 5th Fl.**
Measuring Teacher-Centrism. *Maria Pampaka (University of Manchester), Julian S. Williams (University of Manchester), Pauline S Davis (University of Manchester), Geoff Wake (University of Manchester)*

Reading Rasch Closely: The History and Future of Measurement. SIG - Rasch Measurement
Fri, Mar 28 - **12:25pm - 1:55pm Hilton New York / Nassau Suite B, 2nd Fl.**
The Lexile Framework as a Close Reading of Rasch's Probabilistic Models. *Jackson A. Stenner (Metametrics, Inc.)*
Compelling Historical and Practical Reasons for Strong Theory in Scientific Research. *David Andrich (University of Western Australia)*
Rasch, Frisch, Two Fishers, and the Prehistory of the Separability Theorem. *William P. Fisher (Avatar International, Inc.)*

Professional Identity and Attitudes. Division I - Education in the Professions
Fri, Mar 28 - **2:15pm - 3:45pm Crowne Plaza Hotel Times Square / Room 1505, 15th Fl.**
Validation of a New Health Professions Teamwork Attitudes Instrument. *David W. Hollar (University of North Carolina - Chapel Hill), Cherri D. Hobgood (University of North Carolina - Chapel Hill), Beverly Foster (University of North Carolina), Marco Aleman (University of North Carolina), Susan Sawning (University of North Carolina)*

NCME 2008 Rasch-Related Activities

Item Response Theory: Parameter Estimation Techniques. Training Session.
Sunday, 8:00 a.m. – 5:00 p.m., New York, Crowne Plaza Hotel, Room 403-404, EE
Presenter: Seock-Ho Kim, University of Georgia

Invited Address by the Recipient of NCME's 2007 Award for Career Contributions to Educational Measurement
Wednesday, 10:35 a.m. – 12:05 p.m., New York, Crowne Plaza Hotel, Times Square Ballroom C, F1
Schrödinger's Cat, Rasch's P and the Most Dangerous Equation
Presenter: *Howard Wainer, National Board of Medical Examiners*

Differential item Functioning Detection: New Procedures and Comparisons.
Wednesday, 12:25 p.m. – 1:55 p.m., New York, Crowne Plaza Hotel, Room 403 – 404, G3
A Range-Null Hypothesis Approach for Testing DIF under the Rasch Model. *Craig Wells, University of Massachusetts, Amherst, Allan Cohen, University of Georgia, Athens, Jeffrey Patton, University of Massachusetts, Amherst*

Test-Taking Effort and Response Time.
Wednesday, 12:25 p.m. – 1:55 p.m., New York, Crowne Plaza Hotel, Room 405 – 406, G4
Development and Applications of Detection Indices for Measuring Guessing Behaviors and Test-Taking Effort in Computerized Adaptive Testing (CAT). *Shu-Ren Chang, Rockford PS, IL, Barbara Plake, University of Nebraska, Lincoln, Shu-Mei Lien, University of Nebraska, Lincoln*
A Mixture Rasch Model with Item Response Time Components. *Patrick Meyer, James Madison University*
An Investigation of How Restrictive Time Limits Affect the Fundamental Assumptions of Item Response Theory (IRT) Models. *Aaron Douglas, Mathematica Policy Research, Inc.*

Modifications and Applications of Standard-Setting Procedures.
Thursday, 12:25 p.m. – 1:55 p.m., New York, Crowne Plaza Hotel, Room 501 – 502, L5
Standard Setting for the Rasch Poisson Count Model. *Rianne Janssen, University of Leuven, Ernesto San Martin, Pontificia Universidad Catolica De Chile*

Dimensionality: Assumptions, Error, and Effect Sizes.
Thursday, 2:15 p.m. – 3:45 p.m., New York, Crowne Plaza Hotel , Times Square Ballroom C, M1
Formalizing the Distinction Between Dimension and Response Violations of Local Independence in the Unidimensional Rasch Model. *Ida Marais, University of Western Australia, David Andrich, University of Western Australia*

Third United Kingdom Rasch Day
Tuesday, February 5, 2008

Thank you to everyone who participated in the Third UK Rasch Day hosted by Assessment and Qualifications Alliance (AQA) in Manchester, England. Thirty four researchers attended from fields as diverse as optometry and high-stakes national assessment.

The presentations, many on test-equating, can be downloaded from: http://www.rasch.org.uk/ . Please use the comment form on the website to suggest ways to make the Day yet more productive or well-known.

You can contact any of the participants through the website to pursue mutual interests or to seek advice.

Chris Wheadon, Peter Tymms and Tom Bramley.

Rasch-related Coming Events

March 2008 - Dec. 2009 3-day Rasch courses (A. Tennant, RUMM), Leeds, UK
http://home.btconnect.com/Psylab_at_Leeds/Courses.htm

March 20-21, 2008, Thur.-Fri. An Introduction to Rasch Measurement: Theory and Applications (Smith & Smith), New York, http://www.jampress.org/

March 22-23, 2008, Sat.-Sun. IOMW 2008, New York, http://www.jampress.org/

March 24-28, 2008, Mon.-Fri. AERA Annual Meeting, New York, http://www.aera.net/

March 27-28, 2008, Thur.-Fri. Introduction to IRT/Rasch workshop (K. Conrad, Winsteps), Chicago, http://www.winsteps.com/workshop.htm

May 2-30, 2008, Fri.-Fri. Many-Facet Rasch Measurement online course, (M. Linacre, Facets), http://www.statistics.com/courses/facets

May 15-18, 2008, Thur.-Sun. 2008 Asian Chinese Quality of Life Conference (T. Bond), China http://www.hksoqol.org/conf2008

June 16-19, 2008, Mon.-Thur. MetaMetrics 2008 Lexile National Conference & Quantile Symposium, San Antonio TX http://www.lexile.com/

July 28 - Nov. 22, 2008 Introduction to Rasch Measurement and Traditional Test Theory online course (D. Andrich, RUMM2020), http://www.education.uwa.edu.au

Aug. 1-3, 2008, Fri.-Sun. 2008 Pacific Rim Objective Measurement Symposium (PROMS), Japan http://www.proms-tokyo.org/

Sept. 11-13, 2008, Thurs.-Sat. International Conference on Outcomes Measurement (ICOM) , Washington D.C. http://icom-2008.org/

Third International Rasch Measurement Conference

University of Western Australia, Perth, Western Australia

January 22 - 24, 2008

Chair: David Andrich

Welcome Addresses:

Prof. Bill Louden, Dean, Graduate School of Education, University of Western Australia
Prof. Alan Robson, Vice-Chancellor, University of Western Australia

Keynote Addresses:

Model Synergy: Conceptual, Measurement and Structural. An integrated framework for understanding the consequences of disease and injury. *Alan Tennant*

Substantive Theory, General Objectivity and an Individual-Centered Psychometrics. *Jack Stenner*

The role of the Unit in Rasch Models. *Stephen M. Humphry*

Paper Presentations:

The Lexile Framework for Writing. *A Kyngdon*

ACTIVLIM: A Rasch-built measure of activity limitations in patients with neuromuscular disorders. *L Vandervelde, P Van den Bergh, N Goemans, J-L Thonnard*

Assessing and scoring a structured interview. *A Mercer*

Using the Rasch model to Develop a Test of Reading and Writing for the Deaf. *A Hameed, H Bano*

Identifying the problem based learning experience of preregistration occupational therapy and physiotherapy students using Rasch Analysis. *A Slade, SV Smith*

Coursework assessment in high stakes examinations: authenticity, creativity, reliability. *R Kimbell, A Pollitt*

Using Rasch Measurement to construct a Diagnostic Reading Assessment Battery. *KH Koh*

Using a Rasch analysis to assess the internal construct validity of the London Handicap Scale (LHS). *B Bhakta, M Horton, N Reay, A Tennant*

Quality supervision of the PhD program at the International Islamic University, Malaysia: A Rasch measurement analysis. *M Ibrahim*

Analyzing the differential item functioning (DIF) on pupil's gender for SACMEQ II Mathematics test. *M Saito*

Using Rasch analysis to assess the internal construct validity of the Fatigue Severity Scale (FSS) in a Systematic Sclerosis population. *M Horton, J Chan, N Reay, R Kent, A Tennant*

Mental Self-Government: Development of the Additional Democratic Learning Style Scale using Rasch Measurement Models. *T Nielsen, S Kreiner, I Styles*

A Rasch family with a structural parameter for different model components of a test and its use in the Hong Kong Certificate of Education Examinations. *G Luo, PW Hill*

Analysis of the Stress-Energy Questionnaire using modern test theory – Part I – a comparison with classic test theory. *Å Lundgren-Nilsson, A Pousette, A Ekman, P Larsman*

Analysis of the Stress-Energy Questionnaire using classic test theory – Part II – a comparison with modern test theory. *Å Lundgren-Nilsson, A Pousette, A Ekman, P Larsman*

Performance and certainty of response: Exploring the relationship for first year chemistry students at three tertiary institutions in South Africa. *E Venter*

Using the Rasch-Kuhnian paradigm to conduct experimental work on writing assessment. *SM Humphry, S Heldsinger*

Some plausible polynomial conjoint systems test performance. *A Kyngdon*

Differential item functioning and the implications for assessment development in a value added monitoring system. *V Scherman*

Do patients assess themselves differently to clinicians? An investigation of levels of agreement between raters using Lysholm Knee Scores. *H Smith, A Tennant*

LIBIRT: A new program for item response theory. *P Valois, B Abdous, S Germain*

Does 3 go into 2? The transition from a 3 tier to a 2 tier scheme of assessment. *D Fowles, Q He*

Resolving differential item functioning using the Rasch model – An example based on adolescent health data. *C Hagquist*

Development of a measure for Computer Graphic Drawing Ability using the Rasch model. *LC Mooi*

The Development of an "HIV and AIDS Knowledge Test" for use in Sub-Saharan Africa: Instrument Design and Preliminary Results. *S Dolata*

Computer adaptive testing with partial credit items. *H Albeck, S Kreiner*

An Analysis of a Large-Scale Placement Test for Establishing an Item Bank. *Y Nakamura*

Analysis of multi-dimensionality in Knee injury and Osteoarthritis Outcome Score (KOOS) – analyses in RUMM and DIGRAM. *J Brodersen, S Kreiner*

Investigating the Threshold Ordering of the Audience Engrossment Scale using the Polytomous Rasch Model. *J Scott, T Salzberger*

Evaluating the efficacy of link items in the construction of a numeracy achievement scale from kindergarten to grade 6. *J Looveer, J Mulligan*

Rasch analysis of the health utility scale EQ5-D in a population of young people with physical and complex disability. *B Bhakta, A Tennant*

Measuring course experience in higher education in Hong Kong: Rasch modeling of affective variables with ordered response categories. *BJ Webster, M Prosser*

Identifying Students' Understanding of Fractions using the dichotomous and Partial Credit Models. *M Wong, D Evans*

Rasch It Not Rush It! Using Rasch analysis to identify the psychometric properties of a student OSCE. *A Slade, J Cronin-Davis, M Molineux*

Measurement of student-teacher adaptive strategies in a cross-cultural setting using principles of the Rasch measurement model. *S Holland, D Andrich*

Applying the RM in RM: Apply the Rasch Model in a Reading Motivation Questionnaire. *S-Y Lin, M-N Yu*

The Rasch Cutoff Point for Diagnosis Usage of the Taiwanese Depression Scale. *M-N Yu, Y-J Liu, R-H Li*

The development of a described student engagement scale. *C Darr, H Ferral, A Stephanou*

Testing the Correlation between Students' Achievement and Their Mathematical Belief: Using the TIMSS 2003 data to explore Taiwan's Eighth Graders. *F-C Chang*

Rasch Factorial Analysis and Scoring with application to Health related Quality of Life. *M Mesbah*

Construction of a civic disposition inventory using a Rasch model analysis. *G-H Tor*

Identifying and confirming partial credit form distracters in multiple choice items: a routine application of the Rasch model. *D Andrich, I Styles*

Rasch analysis of Cochin Hand Functional Disability Scale. *A-M Keenan, PG Conaghan, A Tennant*

Reducing the item number to obtain same-length self-assessment scales: A systematic approach using results of graphical log-linear Rasch modeling. *T Neilsen, S Kreiner*

An investigation into 'ratio and proportional reasoning': using Rasch measurement techniques to establish anchor points. *C Long, H Wendt*

Equating Health related Quality of Life Scales. *A Benmelik, M Mesbah*

Fuzzy partial credit scaling: combining the Rasch model with fuzzy theory as a scoring scheme. *S-C Yu, M-N Yu*

Application of Pairwise Comparison Methodology in Ranking Students' Exemplars in Two Courses to Develop Grade Descriptors. *JN Njiru*

The development of a graphical fit analysis tool for Rasch measurement. *H Ferral, A Stephanou*

The measurement of conceptual understanding in physics. *A Stephanou*

Assessing practical performance in an applied information technology course using digital representations and paired comparison making. *CP Newhouse*

Using Rasch analysis to develop an extended matching question (EMQ) item bank for undergraduate medical education. *M Horton, B Bhakta, A Tennant*

Methodological Aspects of Differential Item Functioning in the Rasch Model. *J Brodersen, S Kreiner*

Application of Collaborative and Active Learning in an Electromagnetic Theory Course: A Wisdom. *RA Rashid, R Abdullah, M Mosodi, A Zaharim*

Rasch model diagnostics and Scale analysis: Measuring Nursing educators' beliefs about diversity in personal and professional contexts. *MP Bourke, W Boone*

Documenting validity evidence for the Rasch and other IRT models using empirical item characteristic curves. *J Tognolini*

The use of the Rasch Model in an item analysis of an item-bank of fifth grade mathematics from textbook suppliers. *HF Lin*

Creation of a new psoriasis quality of life measure from five preexistent instruments using Rasch analysis. *F Sompogna, I Styles, S Tabolli, D Abeni*

Units in Measurement: Putting the Quantity Back into Quantitative Science. *SM Humphry*

Computerized adaptive testing for mathematics in primary schools in Thailand. *C Chusathuchon, RF Waugh*

The Responsiveness Paradox. *J Hobart, SJ Cano*

The trade-off between consistency and precision of measurement (including an interactive demonstration). *G Cooper*

A Rasch measurement of the use of Cohesive devices used in writing English as a Foreign Language by secondary students in Hong Kong. *LFM Ho, RF Waugh*

Attenuating the attenuation paradox. *A Pace, SJ Cano, LE Barrett, JP Zajicek, JC Hobart*

Rasch-Boltzmann Machines. *J Linacre*

Rasch Analysis of academic motivational scales. *GTH Wong, RF Waugh*

Strengthening the structure of item response categories: are we too quick to demolish instead of rebuild? *LE Barrett, SJ Cano, A Pace, JP Zajicek, JC Hobart*

The Power series distribution in the theory of Rasch models. *S Kreiner, D Andrich, G Leunbach*

Teachers' Views of teacher-student relationships in the primary school. *N Leitao, RF Waugh*

Confirming Constructs: the importance of quality as well as quantity. *SJ Cano, A Thompson, JP Zajicek, JC Hobart*

Using the polytomous Rasch model as a power series distribution to equate tests. *D Andrich, S Kreiner, G Leunbach*

Psychometric properties of self-regulated learning in an ICT-rich university environment scale. *JN Njiru*

Inferring an experimentally independent response space for the Rasch model for ordered categories from its experimentally dependent subspace. *D Andrich, G Luo*

Using Rasch Modeling to Develop a Measure of 12-step Counseling Practices. *R Claus, H Gotham*

A Rasch Measurement of Methods of Fostering Creativity for students studying mechatronics' at a polytechnic in Singapore. *KCL Teo, RF Waugh*

Local Polynomial smoothing Kernel Score Equating Methods for health related Quality of Life. *B Abdous, M Mesbah, K El Fassi*

Maintaining performance standards: aligning raw score scales on different tests via a latent trait created by rank-ordering examples of examinees' work. *T Bramley, B Black*

University students' receptivity to peers with disabilities. *M Biswas, RF Waugh*

Local item dependence and scoring options for sentence-level sequencing items. *K Yoshizawa*

Equating the 2006-07 Cooperative Scholarship Mathematics and Humanities Test using the Rasch Model. *J Harding, A Inglis, A Raivars, D Urbach, D Weeding*

Student receptivity to project work at a junior college in Singapore. *RF Waugh, KC Choe*

Response dependence and the measurement of change. *I Marais*

The development of a Malaysian critical thinking instrument (MaCTi) prototype: Conceptualization and Psychometric Properties. *AM Mahdzir*

Vanishing Tricks and Intellectualist Condescension: Measurement, Metrology, and the Advancement of Science

What exactly does it mean for data to fit a Rasch model? Satisfaction of Rasch's separability theorem provides access to sufficient statistics, invariant metrics, separable parameters, etc., but isn't there a more tangible and practical sense of these technical accomplishments?

I think there is and that many of us who value Rasch's models and who use them routinely may not have grasped the full meaning of one of the primary concrete consequences of fit to a Rasch model. To begin to trace out this full meaning, let's start with a statement from Jane Loevinger's 1965 review of Rasch's book:
"Rasch must be credited with an outstanding contribution to one of the two central psychometric problems, the achievement of non-arbitrary measures" (Loevinger, 1965, p. 151).

"Non-arbitrary measures": measures that are not arbitrary, that are not capriciously based in fleeting preferences or whims, or left to individual judgment. And indeed, "non-arbitrary" is the right word to choose for describing the way individual instruments function across multiple samples, and the way multiple instruments can converge on a common construct.

But how far does this non-arbitrariness go? Consider Rasch's (1960, pp. 110-5) own sense of the results he obtained from reading test data. He observes that the multiplicative form of the model he employed has the same structure as that used by Maxwell in the study of mass, force, and acceleration, meaning that the model actually states a law concerning the relation of reading ability, text difficulty, and comprehension rate.

Georg Rasch then claims that
"Where this law can be applied it provides a principle of measurement on a ratio scale of both stimulus parameters and object parameters, the conceptual status of which is comparable to that of measuring mass and force. Thus, ... the reading accuracy of a child ... can be measured with the same kind of objectivity as we may tell its weight ...."

This bold statement has recently been further substantiated by Burdick, Stone, and Stenner (2006), who draw an analogy between the Rasch reading law and the combined gas law's prediction of how temperature and pressure relate to a constant volume. Burdick, et al. close with the statement that "the implications of this kind of law-making for construct validity should be evident."

Yes, the implications for construct validity and for construct theory should indeed be evident. However, if it were evident, would it not be universally apparent that, insofar as a reading test measures reading ability and calibrates the reading difficulties of texts, it must follow the Rasch Reading Law? And does it not also then follow, that if the test follows this law, whatever reading test is used, and no matter what range of numbers is used as the metric, insofar as the test really measures reading ability, it will measure in Lexiles?

Given ongoing research into the measurement of reading and writing performance, these implications are apparently not self-evident. The implications seem to remain so far unperceived, unapprehended, that no one has even been disturbed by what some might find to be the grandiosity of the claim. Is no one provoked enough to challenge the hegemony of this Rasch Reading Law and show readers and tests for which it does not apply?

Laws for Measurement

Perhaps no one is so provoked, and that may be because one implication of laws for the measurement of valid constructs seems quite lost, and not just on the Rasch measurement audience, but on researchers in general, as well as the general public. One of the first to bring out this lost implication was Thomas Kuhn (1977, p. 219), who observed that

"The road from scientific law to scientific measurement can rarely be traveled in the reverse direction. To discover quantitative regularity one must normally know what regularity one is seeking and one's instruments must be designed accordingly; even then nature may not yield consistent or generalizable results without a struggle."

A few pages before this passage, Kuhn points out that examples of productive measurement in the Scientific Revolution are found only in longstanding areas of research, such as optics, mechanics, and astronomy. Measurement in areas involving heat, electricity, magnetism, and chemistry did not come into their own as sciences until the 19th century, because of the extensive qualitative understandings that had to be developed before quantification could be achieved.

Building on decades of others' qualitative understandings applied to reading measurement, Rasch made a point of ensuring that his measurement research and model formulation would arrive at predetermined ends capable of supporting the kinds of mathematical conclusions and scientific generalizations that he wanted to be able to support. In so doing, he arrived at a formulation that is nothing less than a law of laws, a model of models, and a very broad basis for generalization.

That is an incredible accomplishment. But, contrary to what seems assumed in common practice, Rasch models do not automatically articulate a construct theory for whatever kind of data happen to be analyzed using Rasch software. That is, data that fit a Rasch model certainly provide evidence supporting the existence of a lawlike structure relative to the construct measured. But the law itself goes unstated as long as no one explicitly says it out loud. And it furthermore goes untested as long as no one uses it to generate new items that exhibit the properties predicted by the theory.

So Rasch used an intuited or implicit sense of what makes a scientific law valuable in formulating a model, and then observed that his reading test data behaved in conformity with that law. He made bold assertions about being able to measure reading ability with the same kind of objectivity that has been long since established to hold in measuring weight.

But Rasch did not articulate a predictive theory of the reading construct. Neither did the authors of the Anchor Test Study (Jaeger, 1973) conducted 20 years later. Instead, what we had were decades of routinely repeated expressions of the reading construct, with the Anchor Test Study showing conclusively that the major reading tests were measuring the same thing, and that they could do so in the same universal, uniform metric.

The non-arbitrariness of the repeated emergence of the same construct over tests and samples culminated in the articulation of a reading construct specification equation, after the model devised by Stenner and Smith (1982), and Stenner, Smith, and Burdick (1983). The simple, elegant, and parsimonious description of the structure of texts and tests has a predictive structure that has been studied by dozens of psychometricians in multiple state departments of education, and book and test publishing companies, and that has been validated in reading tests taken by millions of students.

It is safe to say that no other construct measured in education or in the social sciences has either the empirical evidence or the theoretical stature enjoyed by the Rasch Reading Law (Burdick, Stone, and Stenner, 2006).

And so, researchers are avidly exploiting this boon for the advancement of science, aren't they? Given the assumption that researchers espouse and embody ostensibly mathematical research values, that is what one would expect. But how many research publications take advantage of, or seek to expose the flaws in, the Rasch Reading Law? How many grant applications are focused on determining how listening, reading, oral, and written comprehension relate to one another relative to the established lawlike relation between an individual's abilities and the difficulty of what is spoken or written?

Just about none. I'm no expert, and my search has been cursory, but it isn't happening. Reading research has been atheoretical for most of its history, reading theory traditionally has had little influence on reading tests, and measurement theory, usually in the form of classical test theory, has had too much influence on reading tests (Engelhard, 2001). Why should this be so?

Incommensurable Beliefs

Two factors come to bear. First, working from research into the history of science, Galison (1999) offers the possibility that experimentalists focused on data, technicians focused on instrumentation, and theoreticians focused on ideas each function within separate, distinct communities, with incommensurable beliefs and behaviors. In this scenario, no field of research is driven exclusively or even primarily by just empirical data (privileged by the positivists) or by theoretical expectations (privileged by the post-positivists).

Instead, each of these subcultures has its own criteria, standards, and methods. Galison suggests an open-ended model that allows partial autonomy to each area, with revolutionary transformations occurring with different periodizations, and with each in relative parity with the other two.

Galison's account of science in general seems to be in accord with Engelhard's observations of the situation with reading research, theory, and measurement.

Neglecting Scales

Now, consider a second factor, namely that "in quoting quantitative empirical laws, scientists frequently neglect to specify the various scales entering in the equations" (Falmagne & Narens, 1983, p. 287). This unstated invariant proportionality of scientific laws underlies the value of standards, which, "to do their job...must operate as a set of shared assumptions, the unexamined background against which we strike agreements and make distinctions" (Alder, 2002, p. 2). In leaving measurement scales and standards unstated and unexamined, in the background, as shared assumptions, we find ourselves in a situation in which

"...the absence of metrological information in scientific papers is simply a part of the culture of science that effaces the work needed to make its universality self-evident. This culture is reinforced by the division of labor within the lab; metrological activity is largely invisible to the scientists who write papers simply because it is performed by their technicians, and at time different from when experiments are performed" (O'Connell, 1993, p. 159).

The technical work done by instrumentalists is not only done at times different from when experiments are performed, but is likely done in a different place by persons unknown to the experimentalist. In the history of science, the technical means by which experimental results were produced were sometimes literally cut out of the picture (Shapin, 1989) as were the roles of everyone but the propertied gentleman who sponsored and directed the research (Shapin, 1991).

And so, in general,

"Metrology has not often been granted much historical significance. ... Intellectualist condescension distracts our attention from these everyday practices, from their technical staff, and from the work which makes results count outside laboratory walls" (Schaffer, 1992, pp. 23-4).

In the natural sciences, there are commercially available precision tools calibrated to universally uniform reference standards built up out of scale-free laws. The transparency of the substantive qualitative meanings shared in an elaborated metrological network renders the effects of metrology's lack of historical significance relatively harmless.

But what might the consequences of this intellectualist condescension be for the work that makes Rasch scaling results count outside laboratory walls? Given that metrological activity is largely invisible to scientists writing papers, what happens to fields in which this unexamined background activity is assumed, as it has always been assumed in all scientific fields, but is now being taken for granted in fields in which it does not exist, in which it has never existed?

In this scenario, should not we expect just what we have? Each psychosocial field's batteries of incommensurable instruments measuring in locally-dependent, nonadditive, statistically insufficient, and variable metrics are akin to so many Towers of Babel. Because metrological work is discounted, ignored, and cut out of the picture, no one noticed its importance to the success of science, and no one noticed its absence when it was not being done.

We have here nothing less than an answer to the question raised by Joel Michell (1990, 2000) concerning how psychology's methodological thought disorder became such a pathological episode in the history of science. That is, it became possible for psychology to establish itself as a putative quantitative science without systematic tests of the hypothesis that its variables are quantitative precisely because that work has historically never been done by theoreticians or experimental laboratorians. It has always been performed by metrological engineers and instrumentalists, working within their own subculture according to its standards and traditions. These subcultures, as Galison points out, are so separate that, in psychology's case, the absence of the metrologists was never noticed!

But what do we have without them? We can only begin to estimate what we are missing by comparing science's loftiest achievements to what would have been possible in a world without metrology. Is it, after all, a mere coincidence that the birth in the early 19th century of the second scientific revolution and of the industrial revolution coincides with the birth of the concept of objectivity as we understand it today (Daston and Galison, 1992) and also with the birth of metrology as a professional discipline? Metrology is a necessary factor in all monumental architectural accomplishments, from the Great Wall to the Great Pyramid, and in all major industrial and engineering accomplishments, from the auto industry to interstate highway systems. The rise of western Europe as a world power in the years from 1250 to 1600 is held to be due to the unity of mathematics and measurement in a quantitative model of the world; that model made it possible for Europeans "to organize large collections of people and capital and to exploit physical reality for useful knowledge and for power more efficiently than any other people of the time" (Crosby, 1997, p. x). Also consider that we spend two to three times as much on creating and maintaining measurement standards as we do on scientific research as a whole (Latour, 1987, p. 251). From all of this, we can surmise that the world would be vastly different without metrology and metrologists. The entire cumulative history of science would disappear in one fell swoop.

The Vanishing Trick

In his history of the Ohm, Schaffer (1992, p. 42) observes that

"Immense labor had been performed to achieve the vanishing trick through which the local practices needed to make standards had simply disappeared. ...the absolute system depended on no particular instrument, or technique, or institution. This helps account for metrology's power. Metrology involves work which sets up values and then makes their origin invisible."

At a deeper philosophical level than that plumbed by Michell, then, we can see a way toward accounting for what Husserl (1970) termed Galileo's "fateful omission" of the means by which mathematical understandings of nature were formulated (Fisher, 2003). It seems as though the greatest strength of transparent measurement-its capacity to bring encapsulated theoretical and inferential power to end users ignorant of theory and technicalities-is also its greatest weakness.

In not requiring an understanding of optics of telescope or microscope users, in making thermometers useful to those unschooled in thermodynamics, in bringing high fidelity music into the homes of millions with no clue as to how lasers can translate pits in plastic-coated aluminum foil into arias and drumbeats, technoscience simultaneously erases the conditions of its possibilities as it writes out the terms of new realities.

Rasch's probabilistic models tap into and exploit deeply rooted, widespread, and usually unarticulated and unexamined assumptions about what makes words, numbers, and measures meaningful and useful. As these assumptions are progressively and increasingly made more explicit, conceptually and practically, in theory and experiment, in a wide array of fields and applications, the value of the models will accordingly also become more apparent, and their range of application will deepen and broaden.

But we have more and higher hurdles to cross in the psychosocial sciences than in the natural sciences. In the psychosocial sciences, but not in the natural sciences, the invisibility of metrology is debilitating because of the way measurement becomes assumed even when it is absent. In addition, in the psychosocial sciences, there are many putative variables, and associated collections of observations hardly of sufficient value to call data. These would fail to scale in the natural sciences, and would be much less likely to form the basis of entire communities of research in the way they have in the psychosocial sciences.

How will these hurdles be surmounted? How might intellectualist condescension toward metrology and the discipline's own vanishing tricks be turned from weaknesses into strengths? Probably through the creation

of value, value not obtainable anywhere else, or by any other means. What that value is and how it is produced is another story for another time.

William P. Fisher, Jr., Avatar International Inc.
WFisher -x- avatar-intl.com

Alder, K. (2002). The measure of all things: The seven-year odyssey and hidden error that transformed the world. New York: The Free Press.

Burdick, D. S., Stone, M. H., & Stenner, A. J. (2006). The Combined Gas Law and a Rasch Reading Law. Rasch Measurement Transactions, 20(2), 1059-60 http://www.rasch.org/rmt/rmt202.pdf

Crosby, A. W. (1997). The measure of reality: Quantification and Western society, 1250-1600. Cambridge: Cambridge University Press.

Daston, L., & Galison, P. (1992, Fall). The image of objectivity. Representations, 40, 81-128.

Engelhard, G., Jr. (2001). Historical view of the influences of measurement and reading theories on the assessment of reading. Journal of Applied Measurement, 2(1), 1-26.

Falmagne, J.-C., & Narens, L. (1983). Scales and meaningfulness of quantitative laws. Synthese, 55, 287-325.

Fisher, W. P., Jr. (2003, December). Mathematics, measurement, metaphor, metaphysics: Part II. Accounting for Galileo's "fateful omission." Theory & Psychology, 13(6), 791-828.

Galison, P. (1999). Trading zone: Coordinating action and belief. In M. Biagioli (Ed.), The science studies reader (pp. 137-160). New York, New York: Routledge.

Husserl, E. (1954, 1970). The crisis of European sciences and transcendental phenomenology: An introduction to phenomenological philosophy (D. Carr, Trans.). Evanston, Illinois: Northwestern University Press.

Jaeger, R. M. (1973). The national test equating study in reading (The Anchor Test Study). Measurement in Education, 4, 1-8.

Kuhn, T. S. (1961). The function of measurement in modern physical science. Isis, 52(168), 161-193. (Rpt. in T. S. Kuhn, (Ed.). (1977). The essential tension: Selected studies in scientific tradition and change (pp. 178-224). Chicago: University of Chicago Press.

Loevinger, J. (1965). Person and population as psychometric concepts. Psychological Review, 72(2), 143-155.

Michell, J. (1990). An introduction to the logic of psychological measurement. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Michell, J. (2000, October). Normal science, pathological science and psychometrics. Theory & Psychology, 10(5), 639-667.

O'Connell, J. (1993). Metrology: The creation of universality by the circulation of particulars. Social Studies of Science, 23, 129-173.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Schaffer, S. (1992). Late Victorian metrology and its instrumentation: A manufactory of Ohms. In R. Bud & S. E. Cozzens (Eds.), Invisible connections: Instruments, institutions, and science (pp. 23-56). Bellingham, WA: SPIE Optical Engineering Press.

Shapin, S. (1989, November-December). The invisible technician. American Scientist, 77, 554-563.

Shapin, S. (1991). 'A Scholar and a Gentleman': The problematic identity of the scientific practitioner in early modern England. History of Science, 29, 279-327.

Stenner, A. J., & Smith III, M. (1982). Testing construct theories. Perceptual and Motor Skills, 55, 415-426.

Stenner, A. J., Smith, M., III, & Burdick, D. S. (1983, Winter). Toward a theory of construct definition. Journal of Educational Measurement, 20(4), 305-316.

Rasch Analysis is Important to Understand and Use for Measurement

- In measurement, our intent is to use numbers (which are really raw scores/ratings) to indicate "more" or "less" of the trait that is presumed to be homogeneous; actually an important part of investigation is to verify that the data reflect that homogeneity.

- Rasch Analysis (RA) is a unique approach of mathematical modeling based upon a latent trait and accomplishes stochastic (probabilistic) conjoint additivity (conjoint means measurement of persons and items on the same scale and additivity is the equal-interval property of the scale).

- The purposes of RA are to maximize the homogeneity of the trait and to allow greater reduction of redundancy at no sacrifice of measurement information by decreasing items and/or scoring levels to yield a more valid and simple measure. At times this requires extracting from messy data measures that conform to a homogeneous latent variable and/or identifying for removal features of the data (e.g., bad items, mis-categorization) which contradict measure homogeneity.

- RA permits rating of a limited set of attributes that are representative of the underlying trait, limited means that a small set may be sufficient.

- Whether observed or self-reported, the summed rating of the attributes represents how much of the trait has been mastered, since the raw score is the "sufficient statistic" for the Rasch measure.

- The model assumes that the probability of a given person/item interaction (in terms of rating high or low) is only governed by the difficulty of the item and the ability of the person, that are determined by the item locations on the presumed latent variable along with the rating scale structure.

- Raw scores have unknown spacing between them. Rasch builds estimates of true intervals of item difficulty and person ability by creating linear measures.

- In this process, item values are calibrated and person abilities are measured on a shared continuum that accounts for the latent trait. Should an item rating be missing, the model estimates the person's probable rating without imputing the missing data.

- Concurrently, the improbability of a person's passing or failing a particular item is estimated item by item in terms of fit statistics. This is a comparison between what actually happened and what the model predicts should

have happened based on the estimated measures.

- INFIT and OUTFIT statistics are the most widely used diagnostic Rasch fit statistics. Comparison is with an estimated value that is near to or far from the expected value. INFIT is more diagnostic when item measures are close to the person measures. OUTFIT is more diagnostic when item measures are far from the person measures. But, for long rating scales, like the FIM™ instrument, this difference tends to disappear.

- The fit statistics indicate where the operator should decide whether to either delete, rescore, or reword an item. Deciding to how to select the number and cut-points of the rating categories is more complex, requiring a combination of fit, reliability and substantive meaning.
See http://www.rasch.org/rmt/rmt101k.htm .

- The Rasch linear measures are originally expressed in log-odd units but may be rescaled to suit conventional scaling, as from 0 to 100 while still retaining conjoint additivity. The model also estimates the scoring error at each level as standard errors of the measure.

- Error is always greater at the upper and lower ends of a scale because the Rasch model is not limited at the extremes, but measures from the middle of the range of values and anticipates infinity in both directions. Measurement is better when the middle values of subjects lie close to the middle values of the measure. In other words, the true score is more uncertain as the limits of the scale are approached.
See http://www.rasch.org/rmt/rmt204f.htm .

- RA transforms ordinal scales into interval measures that may be used in parametric statistical analyses and the measures are characterized with standard errors for even more sophisticated analyses. Patient measures and calibration of individual item values are measured on the same metric and are locally independent, provided that Rasch criteria are met.

- Measures constructed using RA are unidimensional and have predictable hierarchies of item calibrations that span the range of difficulty within a domain of assessment.

- Final measures are built by the operator based upon the best judgments of:
  - spread of item values (evenness of steps)
  - reduced error of measurement (precision)
  - probability and improbability (fit) of item and person values to that expected from the model
  - overall reliability (noise)
  - simplicity, and
  - conformity to the nature of the clinical values that are being measured

- Building measures using RA requires that the data fit the model, not that the model fit the data.

- Rasch modeling facilitates analysis of responsiveness of individual items with respect to their calibrated positions within a measure.

- In summary, Rasch analysis provides an internally valid measure that, when developed from an appropriate sample, is independent of the particular sample to which it is applied, meaning that the findings for the sample extrapolate to its population.

Dr. Carl V. Granger, SUNY at Buffalo

Mathematics Education Research Journal Volume 18, Number 2, 2006

Research in Mathematics Education and Rasch Measurement. *Rosemary Callingham and Trevor Bond*

A Case of the Inapplicability of the Rasch Model: Mapping Conceptual Learning. *Kaye Stacey and Vicki Steinle*

A Longitudinal Study of Student Understanding of Chance and Data. *Jane Watson, Ben Kelly and John Izard*

Applying the Rasch Rating Scale Model to Gain Insights into Students? Conceptualization of Quality Mathematics Instruction. *Kelly Bradley, Shannon Sampson and Kenneth Royal*

Easier Analysis and Better Reporting: Modeling Ordinal Data in Mathematics Education Research. *Brian Doig and Susie Groves*

Modeling Mathematics Problem Solving Item Responses Using a Multidimensional IRT Model. *Margaret Wu and Raymond Adams*

Surveying Primary Teachers about Compulsory Numeracy Testing: Combining Factor Analysis with Rasch Analysis. *Peter Grimbeek and Steven Nisbet*

Free download from:
http://www.merga.net.au/publications/merj_display.php?volume=18&number=2

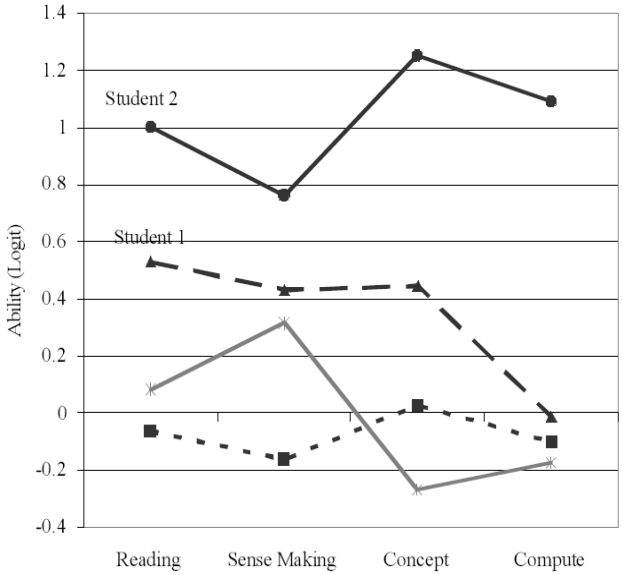


Figure from Wu and Adams (2006)

Differential Rater Functioning

Monitoring the quality of ratings obtained within the context of rater-mediated assessments is of major importance (Engelhard, 2002). One of the areas of concern is differential rater functioning (DRF). DRF focuses on whether or not raters show evidence of exercising differential severity/leniency when rating students within different subgroups. For example, a rater my rate male students' essays (or female students' essays) more severely or leniently than expected. Ideally, each rater's level of severity/leniency should be invariant across gender subgroups. Residual analyses of raters flagged with DRF can be used to provide a detailed exploration of potential rater biases, and they can also form the basis for conducting mixed-methods study (Creswell & Plano-Clark, 2007).

In order to illustrate the use of residual analyses to examine DRF, data from (Engelhard & Myford, 2003) are used. The purpose of the original study was to examine the rating behavior of raters who scored essays written for the Advanced Placement® English Literature and Composition (AP ELC) exam. Data from the 1999 AP ELC exam were analyzed using the FACETS model. One of the sections of this report focused on DRF among raters scoring the AP ELC exam.

A rater x student gender bias analysis was conducted to determine whether or not raters were rating essays composed by male and female students in a similar fashion. Were there raters who were more prone to gender bias than other raters? The FACETS analyses identified 18 raters that, based on statistical criteria, may have exhibited DRF related to student gender.

Table 1. Summary of Differential Rater Functioning Statistics (Student-Gender Interactions) for Rater 108

Student Subgroup	Count	Mean Observed	Mean Expected	Mean Residual	Bias Logit	Bias SE	Bias *z*-statistic
Male	9	5.33	4.56	.77	-.56	.28	-2.00*
Female	23	4.83	5.13	-.30	.23	.18	1.25
Bias against Females relative to Males				1.07	.79	.33	2.39*

* | Z | > = 2.00

Based on the overall fit statistics (INFIT MNSQ = 1.1, OUTFIT MNSQ = 1.1), Rater 108 did not appear to be rating in an unusual fashion. However, when the interaction between rater and student gender is specifically examined, Table 1, a different story emerges. Rater 108 tended to rate the male students' essays higher on average (5.33) than expected (4.56). For females, the observed average (4.83) is less than the expected average (5.13). In summary, there is a statistically significant gender-difference in the rater's severity (z = 2.39).

Figure 1 shows that Rater 108 assigned higher-than-expected ratings to 8 of the 9 male students' essays, but lower than expected ratings to 13 of the 23 female students' essays. This highlights the importance of

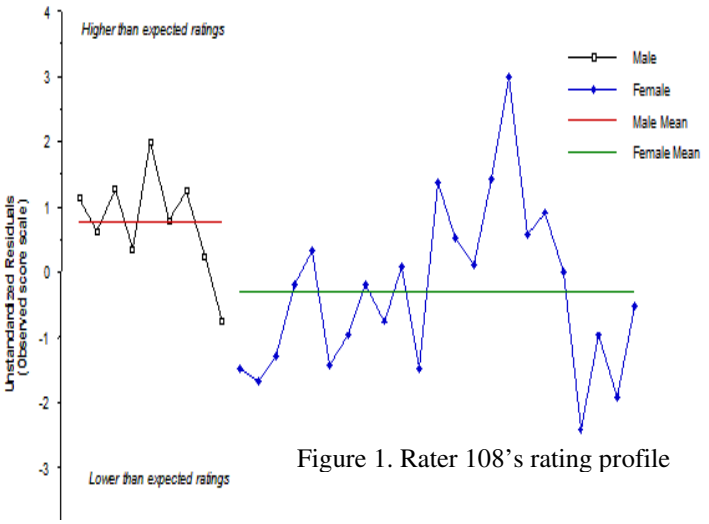


Figure 1. Rater 108's rating profile

exploring not only mean differences between observed and expected ratings within each subgroup category but also the variability and spread of residuals within subgroups. Ultimately, DRF involves looking at discrepancies between observed and expected ratings at the individual level. As pointed out many years ago by Wright (1984, p. 285),

"bias found for groups is never uniformly present among members of the groups or uniformly absent among those not in the group. For the analysis of item bias to do individuals any good, say, by removing the bias from their measures, it will have to be done on the individual level."

In rater-mediated assessments, it is very important to conduct group-level analyses of DRF, but use caution if routine statistical adjustments are made for rater severity. The full interpretation of these effects require a detailed examination of residuals for each rater. Using a mixed-methods framework, suspect raters that can then be investigated in more detail using case studies and other qualitative analyses.

George Engelhard, Jr., Emory University

Creswell J.W. & Plano-Clark V.L. (2007). Designing and conducting mixed methods research. Sage.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal and T. Haladyna (Eds.), Large-scale Assessment Programs for ALL Students: Development, Implementation, and Analysis, (pp. 261-287). Mahwah, NJ: Erlbaum.

Engelhard, G, & Myford, C.M. (2003). Monitoring rater performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model. NY: College Entrance Examination Board. http://professionals.collegeboard.com/research/pdf/cbresearchreport20031_22204.pdf

Wright, B.D. (1984). Despair and hope for educational measurement. Contemporary Education Review, 3(1), 281-285. http:// www.rasch.org/memo41.htm

How to Simulate Rasch Data

Dichotomous data:
1. Decide about the items. They are usually uniformly distributed. How many items? How wide the interval? The item mean is usually set at 0 logits. Simulate the item difficulties.

2. Decide about the person sample. This is usually normally distributed. How big a sample? What is the mean? What is the standard deviation? Simulate the person abilities.

4. For each response by a person to an item:

4A. Generate a random number U = uniform [0,1]

4B. Probability of failure = 1/(1 + exp(ability - difficulty))

4C. If U > Probability of failure, then X=1 else X=0.

4D. X is the simulated observation.

5. Check this by simulating data for a very high ability person (logit = 10): the data should all be "1".
Simulate data for a very low ability person (logit = -10): the data should all be "0"

Polytomous (rating scale or partial credit) data:
1. Decide about the items. They are usually uniformly distributed. How many items? How wide the interval? The item mean is usually set at 0 logits. Simulate the item difficulties.

2. Decide about the person sample. This is usually normally distributed. How big a sample? What is the mean? What is the standard deviation? Simulate the person abilities.

3. Decide about the number of categories, m. The higher categories, 2 to m, have Rasch-Andrich threshold values that are usually ascending and sum to zero across all the categories. Simulate the threshold values.

4. For each response by a person to an item:

4A. Generate a random number U = uniform [0,1]

4B. Compute the cumulative exponential of observing each category:

```
measure = 0
cumexp(1) = 1
Compute for category j = 2 to m
  measure = measure + ability - difficulty - threshold(j)
 cumexp(j) = cumexp(j-1) + exponential(measure)
```

4C. Identify the simulated observation:

```
U = U * cumexp(m)
For category j = 1 to m
    if U <= cumexp(j) then X = j: exit
```

4D. X is the simulated observation.

5. Check this by simulating data for a very high ability person (logit = 10): the data should all be "m" (the top category).
Simulate data for a very low ability person (logit = -10): the data should all be "1" (the bottom category).

John M. Linacre

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 Chairs: Sharon Solloway & Ed Wolfe
SIG website: http://www.raschsig.org/

Rasch Measurement SIG Business

Issue 1: I am writing to announce that we received a single nomination for each AERA Rasch SIG leadership position. These were accepted. As a result, according to the SIG By-laws there will be no election held for officers. At the 2008 Annual Business Meeting, the following people will begin a two-year term as SIG officers.

Chair: Edward W. Wolfe
Secretary/Treasurer: Timothy Muckle

Issue 2: Dues for the Rasch SIG are currently $15 for one year or $25 for two years - a rate that is slightly higher than typical for AERA SIGs ($5 to $10 per year, although some are higher than ours). Currently, our bank balance is about $7,700. Our annual expenditures include (approximate values): Website ($200), Annual Meeting rentals ($350), AERA fees ($225) - totaling about $800 per year.

The only potential additional expense that we can foresee is a stipend for an award and the cost of a plaque for that award. This is a matter to be discussed at the Annual Business Meeting at the SIG Business Meeting in March.

Our December 2007 membership is about 180 members, which is up only slightly from April of 2007.

Members of the SIG in good standing: Please vote YES or NO on the following proposal by sending an email to edwolfe -x- vt.edu. Voting will be closed March 1st, 2008, and results will be announced at the Annual Business Meeting.

Proposal: ***In April of 2008, the Rasch SIG will reduce its annual membership rate to $10 per year.***

The relevant SIG By-law for this vote is:
"Section 4 -- Dues. The amount of Rasch SIG dues may be modified by a majority of the Rasch SIG members voting by e-mail or at the annual Rasch SIG business meeting held during the AERA annual meeting."

Edward W. Wolfe
Secretary/Treasurer
Rasch Measurement SIG, AERA

Investigating Displacement in Rasch Item Calibrations

Item drift analyses that use displacement have reported displacement distributions symmetrically distributed around zero (Jones and Smith, 2006). Approximately equal numbers of items appear to drift in both a positive (harder) and a negative (easier) direction, despite hypotheses suggesting systematic drift in one direction. A portion of this displacement distribution can be shown to be statistical artifact resulting from the way the statistic is calculated.

Displacement is a useful statistic generated from the Winsteps analysis program. The displacement statistic "approximates the displacement of the estimate away from the statistically better value which would result from the best fit of your data to the model." (Linacre, Winsteps User Manual). In any analysis featuring anchored items, Winsteps simultaneously performs a free (unanchored) parameter estimation for all of the items. The displacement statistic results from a direct comparison of the anchored difficulty value with the value from the free estimation arising from the current data. Due to the re-centering procedures in Winsteps, the free parameter estimates are constrained to be centered around a mean of 0. Accordingly, all displacement values also sum to 0. As a result, in a dataset featuring systematic drift in one direction (i.e. easier), it is possible to observe, in stable items, drift in the opposite direction (i.e. harder) resulting from a statistical artifact.

Since the displacement statistic contains an artifact that does not represent actual item difficulty drift, the interpretation of the statistic becomes problematic and its usefulness is diminished. Simulation data was used to replicate certain conditions of item difficulty drift and to assess the impact of these drift conditions on the displacement statistic and its interpretation.

The candidate sample was chosen to have a standard normal distribution N(0,1). Three candidate samples were selected, one having 200 individuals, one having 500 individuals and one having 1,000 individuals. Three item samples were also selected, each having a standard normal item difficulty distribution N(0,1). The item sample sizes were 30 items, 100 items and 200 items. Each candidate sample size was then matched against each item sample size, resulting in nine combinations. The Promissor simulator (Becker, 2006) was used to generate an initial response string data sample for each of these combinations using Rasch probability as the basis for assigning a right/wrong response for each candidate/item interaction. The nine response strings generated by the above procedure were analyzed using the Winsteps program, and item difficulty calibrations were obtained for each item as it was used in each of the nine combinations. These item calibrations were then used to create item anchor files to be used in subsequent drift analyses.

In order to minimize the impact of outside variation, the same response string data sample was used to simulate item drift. Answers were systematically changed in the response strings to simulate a drift in an easier and/or harder direction by changing answers from wrong to right or vice versa.

The number of items that were simulated to drift was systematically varied. In the first condition, 10% of the items were simulated to drift. In the second condition 20% of the items were simulated to drift. In the final, most extreme condition, 50% of the items were simulated to drift. The first two conditions are probably more reflective of normal drift conditions. The final condition would be more reflective of a serious security breach.

Within each of the above conditions the direction of drift was also varied. In one situation, the drift was all in a single direction with the items becoming easier. In the second situation, the drift was symmetrically balanced with half the items drifting easier and half the items drifting harder. In the final situation, the drift was asymmetrical with 70% of the items drifting easier and 30% of the items drifting harder. The final condition is a combination of the first two with more emphasis on the condition of compromised items.

Combining the 9 candidate/item combination with the 3 percentages of drift and the 3 directions of drift resulted in 81 unique conditions that were simulated.

To simulate drift, the desired number of items and candidates was randomly selected without replacement from the total candidate and item samples. Each response string was examined and, for each interaction of a selected candidate and a selected item, the answer was examined and changed appropriately to simulate the desired drift. If the answer for that particular candidate/item interaction was already in the direction of the desired drift no action was taken.

The modified data sets were then reanalyzed using Winsteps. The item difficulties of all of the items were anchored to the item calibrations obtained from the initial analysis. The impact of the drift on the displacement distribution was then assessed.

The results of the simulations and analyses are summarized in Tables 1 and 2. Tables 1 and 2 contain average displacement values observed on items whose response strings were not modified to mimic drift (i.e. items hypothesized to remain stable over time). The mean displacement values are replicated for each condition of the simulations. Table 1 represents simulation cases where hypothetically drifting items were all displacing exclusively in a negative (easier) direction.

Table 1 also demonstrates that when systematic drift in one direction was present in a data set, then hypothetically stable items in every test condition exhibited artificial positive displacement. The artifact was more pronounced

Table 1. Mean displacement values for hypothetically stable items (unidirectional drift)

Condition	Examinee Sample	30 item test	100 item test	200 item test
10% modified, all easier	N=200	0.01	0.03	0.03
	N=500	0.02	0.02	0.02
	N=1000	0.03	0.02	0.03
20% modified, all easier	N=200	0.11	0.12	0.13
	N=500	0.09	0.11	0.10
	N=1000	0.12	0.09	0.10
50% modified, all easier	N=200	0.66	0.70	0.77
	N=500	0.55	0.72	0.74
	N=1000	0.61	0.78	0.68

in conditions with increased test length and increased proportion of drifting items. The amount of artificial positive drift appeared to be unrelated to examinee sample size. To provide some indication of significance, we compared the Table 1 displacements to the value corresponding to two times the average standard error (SE) of item calibrations for each examinee sample condition (for N=200, 2*SE = 0.34; N=500, 0.22; N=1000, 0.14). Highlighted cells in Table 1 reflect displacement values that might be interpreted as significant because they are more then more than two SE from the original calibration. Obviously, these potential false positives occur exclusively in the extreme simulation conditions where 50% of the test items were modified to show easier drift.

Table 2 summarizes simulation cases where hypothetically drifting items were displacing asymmetrically (70% easier, 30% harder). The patterns of average displacement for hypothetically stable items (i.e. items whose response strings underwent no modification) are similar in Table 2, yet not as pronounced. The artificial positive displacement is detected more in data sets with a large amount of systematic drift. Again, artifact that could be interpreted as significant is highlighted. The problem posed by artifact is somewhat ameliorated in the simulations where all of the drift was not in one direction.

Not surprisingly, in simulations featuring balanced drift – or equal amounts of drift in both directions – the problem of artifact was completely ameliorated. In these simulations, the average displacement value for items hypothesized to remain stable was consistently 0, so their values are not tabulated. This statistical artifact is a result of the way the displacement statistic is calculated and the fact that the results are mean centered. In situations in which the degree of drift is symmetrically distributed in both an easier and a harder direction, the impact is relatively minor. As the drift becomes more asymmetrically distributed, the impact of the artifact becomes more noticeable so that non-drifted items may be flagged as significantly and substantially drifted.

Table 2. Mean displacement values for hypothetically stable items (asymmetrical drift)

Condition	Examinee Sample	30 item test	100 item test	200 item test
10% modified, all easier	N=200	0.00	0.02	0.01
	N=500	0.00	0.01	0.01
	N=1000	0.01	0.01	0.01
20% modified, all easier	N=200	0.04	0.05	0.05
	N=500	0.03	0.06	0.04
	N=1000	0.02	0.03	0.03
50% modified, all easier	N=200	0.12	0.29	0.39
	N=500	0.00	0.30	0.25
	N=1000	0.12	0.21	0.29

However, the impact can be easily detected by plotting the displacement against the sequence number. Items with a very consistent drift are items that are being affected by the artifact. It is recommended that such a plot be used as a part of any drift analysis. It may also be possible to detect the artifact during the displacement calculation by determining the variability of the drift within subsets. A subset that exhibits little variability may reflect artifact.

John A. Stahl, Timothy Muckle
Pearson VUE, Evanston, Illinois

Becker, Kirk (2006) *Promissor CAT Simulator*

Jones, P.E. & Russell W. Smith (2006) *Item Parameter Drift in Certification Exams and Its Impact on Pass-Fail Decision Making,* Paper presented NCME, San Francisco.

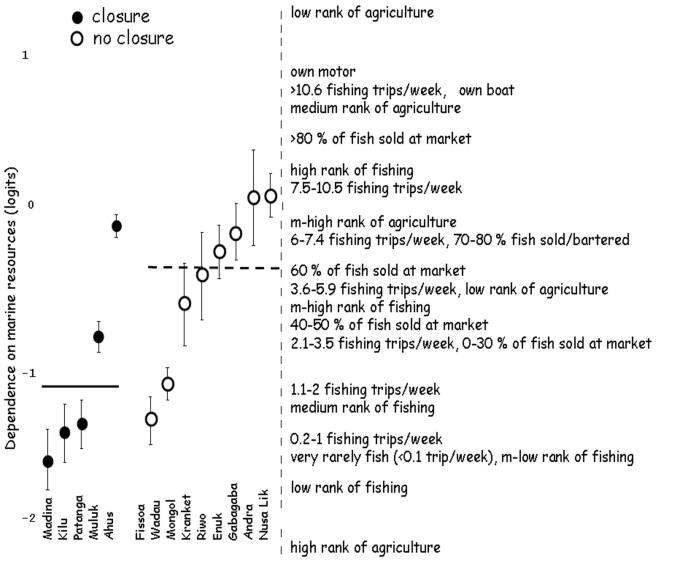


Figure. Comparison of dependence in marine communities. Cinner, J., S. Sutton, T. Bond (2007) Socioeconomic thresholds that affect use of customary fisheries management tools. Conservation Biology: 21(6): 1603-1611

Book: Assessing and Modeling Cognitive Development in School: Intellectual Growth and Standard Setting

JAM Press, http://www.jampress.org/ (click on: JAM Press Books), is pleased to announce this new book which presents a series of papers that examine the area of cognitive modeling in assessment with a particular emphasis on standard setting. These papers present the most up to date information on modeling student learning using multivariate IRT models, progress variable mapping, value-based approaches, content trajectories, on line tutoring records, and vertically articulated performance standards.

The No Child Left Behind (NCLB) legislation has encouraged a keen interest in standard setting. At the same time, there has been a steady increase in the use of cognitive models to understand student performance. These models are being used to characterize the patterns of problem solving that a student utilizes to solve the test items with which he or she is faced in an assessment. This book combines these two interests in a way that gives the reader an overview of the current literature as well as the issues that remain unresolved. This book helps one to understand the standard setting problem as one of characterizing the expert student's problem solving strategies and differentiating these strategies from those used by the inexpert student. The result is a view of standard setting and student progress that takes on a very different appearance from that traditionally used in psychometrics.

This book is based on the very well received conference of the same name held on the University of Maryland Campus on October 19 and 20, 2006.

The titles and authors of the eleven chapters are as follows:

1. A Prospective, Progressive, and Predictive Approach to Standard Setting
   Isaac I. Bejar, Henry I. Braun, and Richard J. Tannenbaum, Educational Testing Service
2. Vertically Articulated Performance Standards: An Exploratory Study of Inferences about Achievement and Growth
   Steve Ferrara. Gary W. Phillips, Paul L. Williams, Steven Leinwand, Shannon Mahoney, and Stephan Ahad, American Institutes for Research
3. Using On-line Tutoring Records to Predict End-of-Year Exam Scores: Experience with the ASSISTments Project and MCAS 8th Grade Mathematics
   Brian W. Junker, Carnegie Mellon University
4. Non-Linear Unidimensional Scale Trajectories through Multidimensional Content Spaces: A Critical Examination of the Common Psychometric Claims of Unidimensionality, Linearity, and Interval-Level Measurement
   Joseph A. Martineau, Michigan Department of Education; Dipendra Raj Subedi, Michigan State University; Kyle H. Ward, Michigan Department of Education; Tianli Li, Yang Lu, Qi Diao, Feng-Hsien Pang, Samuel Drake, Tian Song, Shu-Chuan Kao, Yan Zheng, and Xin Li, Michigan State University
5. Item Response Theory and Longitudinal Modeling: The Real World is Less Complicated than We Fear
   Marty McCall, and Carl Hauser, Northwest Evaluation Association
6. A Culture of Remembering: Contexts of Mathematical Development and their Implications for Assessment and Standard-Setting
   Christopher A. Correa, and Kevin F. Miller, University of Michigan
7. Estimating Gain in Achievement when Content Specifications Change: A Multidimensional Item Response Theory Approach
   Mark D. Reckase, Michigan State University, and Tianli Li, ACT, Inc.
8. Implementing Cognition-Based Learning Goals in Classrooms: The State Role
   Mark Moody, Hillcrest and Main, Inc., William D. Schafer, University of Maryland, and Lani Seikaly, Hillcrest and Main, Inc.
9. A Value-Based Approach for Quantifying Student's Scientific Problem Solving Efficiency and Effectiveness Within and Across Educational Systems
   Ron Stevens, IMMEX Project, UCLA
10. Once You Know What They've Learned, What Do You Do Next? Designing Curriculum and Assessment for Growth
    Dylan Wiliam, Institute of Education, University of London
11. Using Progress Variables to Map Intellectual Development
    Cathleen A. Kennedy, and Mark Wilson, University of California, Berkeley