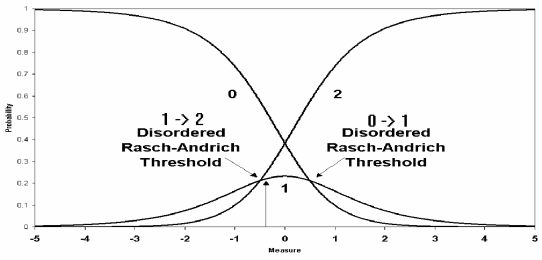
Disordered Rasch-Andrich Thresholds



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 17 No. 4 Spring 2004 ISSN 1051-0796

Disordered Thresholds:

An example from the Functional Independence Measure

A recent paper in Medical Care[1] has raised considerable interest due to its reporting of disordered thresholds in data collected routinely in different countries from patients who have experienced a stroke.

In our work, an adjacent-category-equal-probability Rasch-Andrich threshold defines the boundary between categories, in our case of the polytomous Functional Independence Measure (FIM®)[2]. Where thresholds are ordered, a person location between category boundaries ensures that the probability of a response in that category is larger than of any other single category[3]. However, if thresholds are disordered, a person location between category boundaries will not give that category the greatest probability of being observed. In our work, for example, being observed in a higher category is taken to imply higher independence. But with disordered thresholds, as in the Figure above, an observation of "2" is more likely than an observation of "1" even for a person with a measure of -0.4 (vertical arrow) which is low even for "1". Thus, from this perspective, disordering of thresholds is a violation of the measurement construct in that there is discordance between the category probabilities and the underlying trait.

Table 1. Threshold estimates for FIM motor items.

Item Thresholds	Loc.	1	2	3	4	5	6
*Eating	-1.11	0.51	-1.59	-1.45	-0.16	1.18	1.51
*Grooming	-0.41	-0.36	-0.73	-0.59	-0.12	0.55	1.25
Bathing	0.39	-1.30	-1.10	-0.60	0.12	0.98	1.91
Dressing Upper Body	0.03	-1.09	-0.6	-0.28	0.04	0.54	1.39
Dressing Lower Body	0.47	-1.14	-0.46	-0.20	-0.02	0.41	1.41
*Toileting	0.11	0.09	-0.26	-0.34	-0.21	0.12	0.60
*Bladder Management	-0.64	0.95	-0.27	-0.34	0.04	0.19	-0.58
*Bowel Management	-0.88	0.70	-0.16	-0.35	-0.19	0.02	-0.02
Transfer Bed	-0.15	-1.28	-0.71	-0.29	0.14	0.68	1.47
Transfer Toilet	-0.04	-1.03	-0.50	-0.30	-0.10	0.40	1.52
*Transfer Tub	0.80	0.39	-0.84	-0.97	-0.39	0.50	1.31
*Walk / Wheelchair	0.24	0.16	-0.15	-0.72	-0.96	-0.27	1.95
*Stairs	1.19	2.15	-0.66	-1.73	-1.44	-0.14	1.82

What does disordering look like in practice, and when does it occur? Table 1 gives the estimates for the thresholds taken from the data of 895 stroke patients which formed the basis of the Medical Care paper. The analysis used the unrestricted (partial credit) model. A likelihood ratio test ($p<.001$) showed that the rating scale model was less suitable. The asterisked items have disordered thresholds, with the 'stairs' item displaying a particularly bizarre pattern. At this stage most items misfit the model with an overall standardized mean-square item fit with mean of -0.360 and SD of 4.462, where a mean of 0 and SD of 1 is expected.

Figure 1. Category probability curves for Bathing item.

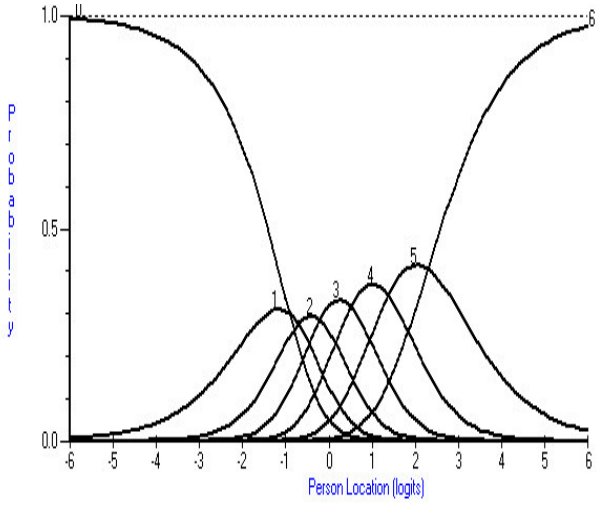


Table of Contents

The items fall into three types with respect to their thresholds; those that are ordered; those that have one or two thresholds disordered and those where many of the thresholds are disordered. Figure 1 shows how categories should work and, in a monotonically increasing fashion, as the trait for independence increases, so does the probability of affirming a higher category. This expected relationship breaks down slightly for the eating item (Figure 2), where at no time would categories one and two be the most probable response. For the bladder management item, this relationship is largely absent, and the item would appear to be working as a dichotomy (Figure 3).

Figure 2. Category probability curves for eating item

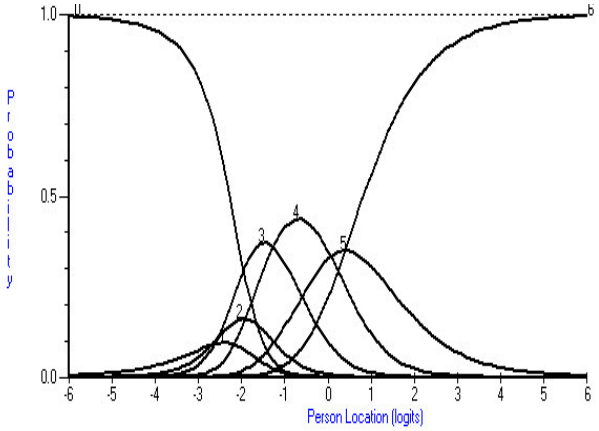
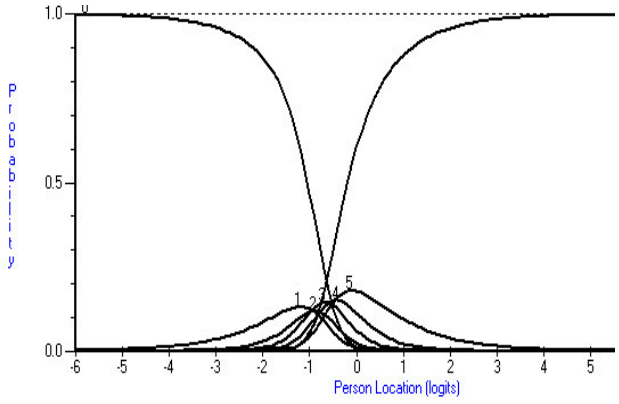


Figure 3. Category probability curves for the Bladder Management item.



How can this deviation from the expected pattern of response come about? An obvious place to start is the distribution of responses across the categories. In the Medical Care paper the analysis was based upon admission data. Might it be that many of the categories implying more independence had null or low frequencies? Table 2 shows that this was not the case, where disordered items are flagged.

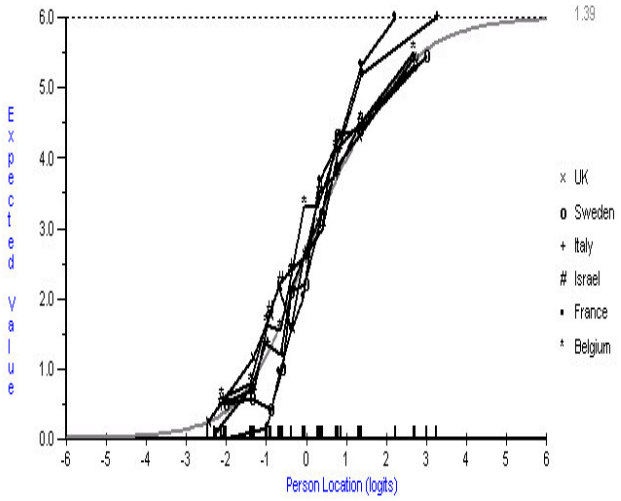
Although there is a clear variation in the distribution of responses across items, all categories had sufficient numbers for estimation[4]. Note that the 'grooming' item which is disordered, has a similar distribution (but in the opposite way) to the 'bathing' item, which is ordered. Furthermore, the conditional pairwise estimation procedure employed in RUMM2020 estimates threshold parameters from all the data, not just from adjacent categories, enhancing the stability of estimates [3] .

Table 2. Category frequencies of FIM motor items.

Item / Category	1	2	3	4	5	6	7
*Eating	26	24	31	74	295	142	230
*Grooming	126	64	63	114	131	149	175
Bathing	173	123	118	160	86	100	62
Dressing Upper Body	159	133	86	115	104	107	118
Dressing Lower Body	233	170	78	102	64	86	89
*Toileting	286	79	60	64	62	99	172
*Bladder Management	171	39	44	49	68	70	379
*Bowel Management	114	29	33	62	57	129	397
Transfer Bed	123	110	111	138	82	129	129
Transfer Toilet	165	117	79	124	69	144	124
*Transfer Tub	406	53	60	84	55	88	60
*Walk / Wheelchair	292	69	41	50	83	181	85
*Stairs	581	14	11	26	58	92	32

Another reason for the disordering may be that different rehabilitation facilities around Europe assign values to the FIM in different ways. Certainly there are different traditions across Europe in the way in which, for example, patients are bathed within rehabilitation facilities[5]. Also the extent of training varies. Two regions, Sweden and Italy have extensive training programs, yet the data from these countries was just as disordered as elsewhere. Furthermore, ordered thresholds were not necessarily associated with the absence of Differential Item Functioning (DIF) across countries. Figure 4 shows the ICC by country for the 'bathing item' which was ordered. However, there was significant DIF for this item ($F=10.22$; $p<0.001$), suggesting that the expected category at any given level could vary by country across the trait.

Figure 4. Plot of ICC by country for 'bathing' item.



The rating scale model has been used previously for analysis of the FIM[6]. Has the use of the unrestricted (partial credit model) contributed to this dilemma? Although the Log Likelihood test shows a significant worse fit for the rating scale model, if used, the extent of disordered thresholds is greater still. Indeed, every item is disordered under the rating scale model. Thus it would seem, in this data set at least, that this is not a reason as to why disordered thresholds are more common than in previous reports.

Prior to seeking a solution to these problems, how does the total raw score reflect the change in category response across the items? At first sight, in Table 3, it would appear that there is an appropriate increase in raw score as each category increases, perhaps with just the exception of the walk/wheelchair item (this is taken from the SPSS file and includes extremes). Thus higher performing patients are rated in higher categories. However, exploratory post-hoc tests suggest that raw scores cannot discriminate across some categories in six of the eight disordered items, but can do so in all the ordered items.

Table 3. FIM average motor raw score for each category of each item.

Item / Category	1	2	3	4	5	6	7
*Eating	16.8	23.2	25.6	36.5	44.7	56.1	70.7
*Grooming	21.3	26.1	35.2	44.0	52.0	64.9	77.1
Bathing	23.6	34.8	42.0	56.3	68.1	78.1	86.5
Dressing Upper Body	22.1	32.9	44.1	52.9	59.4	71.8	82.2
Dressing Lower Body	25.4	38.7	48.5	62.6	67.6	76.3	85.9
*Toileting	26.1	38.4	45.8	55.0	57.7	70.3	80.1
*Bladder Management	23.5	32.6	35.6	37.8	44.0	51.3	69.2
*Bowel Management	20.0	28.6	34.4	35.5	38.7	52.6	66.8
Transfer Bed	20.4	30.0	38.8	48.3	56.4	72.5	83.1
Transfer Toilet	21.2	33.1	41.3	50.1	56.9	71.9	83.4
*Transfer Tub	34.6	38.4	48.2	61.5	70.6	79.5	87.2
*Walk / Wheelchair	28.7	28.6	37.3	42.3	54.2	58.9	84.0
*Stairs	38.5	56.0	59.2	63.0	73.1	81.1	87.0

What can be done about the apparent disordering of thresholds? In the Medical Care paper we rescored items on an individual basis to try and improve fit to the model. As thresholds are estimated with respect to all categories, not just adjacent categories, the final solution was not at all obvious from the category probability curves such as those presented above. For example, the 'bladder' item worked with three categories (Figure 5), while the eating item had to be dichotomized.

In the paper it was shown that the 'eating', 'bowel management' and 'toileting' items had to be dichotomized; 'bladder management' and 'grooming' tritomized; 'walk/wheelchair', 'transfer tub' and stairs were collapsed into four categories, with the remainder working as seven category items. The paper went on to split items for DIF by country, and came up with a working solution, effectively using the FIM motor items at the county level as an item bank, linked by five common items. The final category frequencies for the rescored items are given in Table 4 (This excludes the solution after splitting for country DIF, which makes matters much more complicated; and a couple of additional patients became extreme).

Figure 5. 'Bladder' item after rescoring.

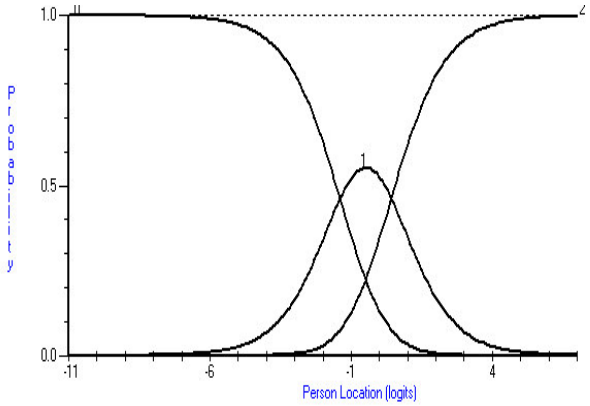
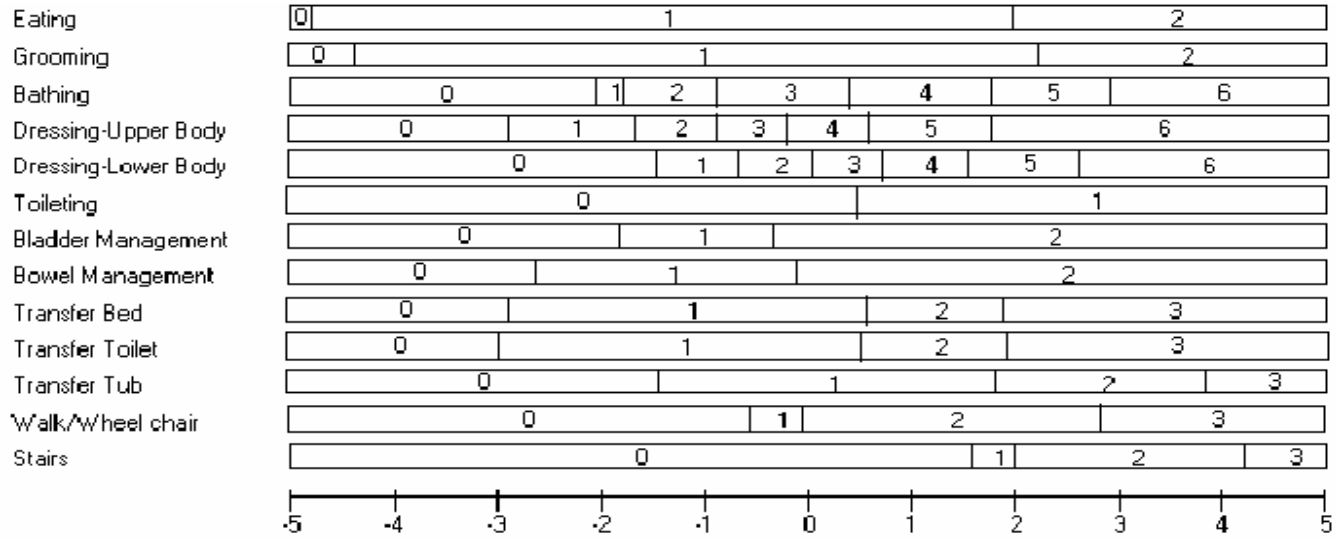


Table 4. Category frequencies after rescoring.

Item/Category	1	2	3	4	5	6	7
*Eating	26	794					
*Grooming	126	521	173				
Bathing	173	123	118	160	86	100	60
Dressing Upper Body	159	133	86	115	104	107	116
Dressing Lower Body	233	170	78	102	64	86	87
*Toileting	286	534					
*Bladder Management	171	270	377				
*Bowel Management	114	705					
Transfer Bed	123	110	111	138	82	129	127
Transfer Toilet	165	117	79	124	69	144	122
*Transfer Tub	406	197	143	58			
*Walk / Wheelchair	292	243	181	83			
*Stairs	581	109	92	30			

The rescoring solution we found is 'messy' in that some items retain all their categories, and then there is a variable reduction in the number of categories for other items. Also, as we have seen, although there is an increase in raw score across all categories for most items, there is a suggestion from post hoc tests that the raw score cannot discriminate across some categories, and these occur where thresholds are disordered. Technically, given fit to the Rasch model, items should not be collapsed further, but prior to splitting for DIF, the case can be made that these data still do not fit the model. Furthermore, there is the issue of differential fit between countries. Single country analysis had shown different fit and different rescoring solutions. For example the UK items 'transfer bed' and 'transfer toilet', which were ordered in the pooled data, were collapsed into four categories for the UK analysis (Figure 6).

Figure 6. Category structure for UK FIM motor scale after rescoring.



What about the use of different software? The results produced by a parallel run with Winsteps are substantively equivalent to those shown here, being limited to minor numerical differences.

It is our contention that scales should work adequately at admission to rehabilitation services, else they should not be used for assessment purposes, or as the basis for outcome measurement. The requirement is for invariance across time. Furthermore, the scale must be invariant across any relevant clinical subtypes if data are to be pooled for the diagnostic group; across diagnostic groups if they are to be pooled at the level of the rehabilitation unit, and across countries if international comparisons are to be made.

The fact that scales work in different ways across different diagnoses and countries should not be surprising given the recent insights provided by modern psychometric methods. The Medical Care paper demonstrated that despite cultural variations, a solution could be found that facilitated the pooling of data. Should we then be so worried about the lack of invariance for some items given we now have the technology to accommodate such variations?

The issue of the disordered thresholds may warrant further effort on the part of FIM users. This involves two aspects; the fundamental aspect of whether or not disordered thresholds are to be taken seriously as a violation of measurement; and the practical aspect of achieving a solution which is common across countries (and the same applies to diagnoses, or within country centers). This will require some clear thinking as to what such disordering means for clinical practice, for outcome measurement, and the pooling of data of the kind undertaken at Buffalo. Given the extent of the FIM database held in Buffalo, at least this is one outcome scale where the users have the capacity to investigate these matters thoroughly, from a well established database.

Alan Tennant BA, PhD. Professor of Rehabilitation Studies, The University of Leeds, UK.

References:

1. Tennant A, Penta M, Tesio L, Grimby G, Thonnard J-L, Slade A, Lawton G, Simone A, Carter J, Lundgren-Nilsson A, Tripolski M, Ring H, Biering-Sørensen F, Marincek C, Burger H, Phillips S. Assessing and adjusting for cross cultural validity of impairment and activity limitation scales through Differential Item Functioning within the framework of the Rasch model : the Pro-ESOR project. *Medical Care* 2004; 42: (Supple 1) 37-48

2. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: A new tool for rehabilitation. In: Eisenberg MG, Grzesiak RC (Eds): *Advances in Clinical Rehabilitation*. New York, Springer Publishing Co; Vol. 1. p. 6-18; 1987.

3. Andrich D, Luo G. Conditional pairwise estimation in the Rasch model for ordered responses using principal components. J Applied Measurement 2003; 4:205-221.

4. Linacre JM. Investigating rating scale category utility. *J Outcome Measurement* 1999; 3:103-122.

5. Küçükdeveci AA, Yavuzer G, Ehan AH, Sonel B, Tennant A. Adaptation of the Functional Independence Measure for use in Turkey. *Clinical Rehabil* 2001; **15**:311-319.

6. Grimby G , Andraon E , Holmgren E , Wright B , Linacre JM , Sundh V. Structure of a combination of Functional Independence Measure and Instrumental Activity Measure items in community-living persons: a study of individuals with cerebral palsy and spina bifida. *Arch Phys Med Rehabil*, 1996; 77(11): 1109-14.

Book Review: *Making Measures*

by Ben Wright and Mark Stone

Published by Phaneron Press (2004), available from Amazon.com - ISBN: 1930847394

This is a relatively short (127 pages) book that attempts to add to the recent literature of readable descriptions of the Rasch model. Using examples refined from years of explanation experience, Ben and Mark produce excellent discussions in short form for objective measurement, conjunction, and item calibration.

Unfortunately, the book has a bit of the Jekyll-and-Hyde syndrome. Between pages 15 and 51, it jumps with minimal explanation into Newton's Laws, principal components analysis, and Winsteps output. The book also has a bit of the Biblical Q syndrome as the reader changes style from Wright to Stone to Stenner. The lack of connectivity (pun intended) is apparent.

Regardless, there are many times that we hear, "So what is the Rasch model? Can you give me a real example?" This inexpensive paperback is a wonderful introduction for those occasions. You could drop off a copy and walk away confident the reader would know much more than they did when they asked the questions.

We could also see this as a supplement to a traditional beginning measurement course for consumers (educators, psychologists, etc.) who want to know a little about the Rasch process, but don't have an extensive background. Fortunately, the examples are mainstream and the references are classic. The historical citations from Peirce, Guttman, Piaget, and Bernoulli are intriguing and worth reading. Each chapter is short and there are plenty of figures, so the total reading time is minimal. Unfortunately, there are also chapters without references (6 and 7) which leave the reader unsatisfied if they cannot understand a technical illustration or concept.

To us, the best thing about *Making Measures* is the refined way that complex concepts and definitions are explained from the perspective of experience and understanding. There are wonderful insights over and over in the book: "The straighter the line [Figure 6], the fewer the distortions and the closer the data points to the line, the more uniform the conjoint relation between items and person" (p. 38). Another: "The ruler does not exist until we imagine it and carve it [from the tree]." (p. 83). The analysis of the thresholds of the *Fear Survey Schedule* is a perfect illustration of the strength of Rasch for rating scale development. Stenner's history of the development of Lexiles should be required reading for all educators who want to measure learned constructs.

The weakest thing is the lack of direction for those who want to learn more. A few pages in the last chapter that would guide one to methods like Facets or authors like Bond, Fox, Linacre, Smith and Andrich are badly needed. Also, there is no mention of IOMW or the Rasch Measurement SIG (AERA). Again, that would be useful for anyone who thought it of interest to read the book.

Steve Lang and Judy Wilkerson
University of South Florida

MOMS

Midwest Objective Measurement Seminar

The next meeting of MOMS will be held on **Friday, May 14, 2004, 9:00 a.m. - 4:00 p.m., at the University of Illinois at Chicago, Chicago,** Illinois USA. There is no registration fee.

Presentations of work by practitioners and students will be of Rasch-related. *Contact me immediately if you would like to present completed or on-going work.*

Lunch: 12:00 noon - 1:00 p.m. Participants are encouraged to bring their own lunches or to stop by one of several Greek, Italian, and American-style restaurants that are within two blocks (five to ten minute walk).

Place: 1040 West Harrison Street in the ECSW building, Room 3427 (third floor) at UIC. Parking available in the parking structure between ECSW and Racine Street on the north side of Harrison. Cost is $8.25 for the full day.

See You There!

Everett Smith, Ph.D.
Associate Professor, Educational Psychology
www.uic.edu/depts/educ/mesalab/

11th Annual Conference of ISOQOL

The International Society for Quality of Life Research

October 16 - 19, 2004 ~ Hong Kong

At the Hong Kong Academy of Medicine's Jockey Club Building, an attractive, modern building with world-class, state-of-the-art conference facilities. Come experience the unique blend of old and new, East and West in this vibrant, cosmopolitan city full of exciting sights and sounds.

The deadline for submission of abstracts is **May 4, 2004.**

www.isoqol.org/2004conf.htm

Madeleine King, PhD, Kwok-fai Leung, PDipOT, and Margaret Tay. Chair and Co-Chairs.

AERA 2004 Rasch-related Papers

Monday, April 12

AERA Roundtable 1. 12:00 p.m. - 12:40 p.m. in Hyatt - Elizabeth Ballroom D, Second Level
Richard P Banghart - Michigan State University
An Examination of the Peer Review System in a Large Research Organization

AERA Roundtable 37. 1:00 p.m. - 1:40 p.m. in Hyatt - Elizabeth Ballroom E, Second Level
Claire Dumont (Université Laval), Richard Bertrand (Université Laval), Marie Gervais (Université Laval), Patrick Fougeyrollas (Institut de réadaptation en déficience physique de Québec)
Comparison Between the Rasch and the Classical Model in the Identification of Social Participation Predictors

SIG - Objective Analysis of Qualitative Research Methods Symposium: Objective Review of Methods for Rating Item Difficulties. 2:15 p.m. - 3:45 p.m. in Hyatt - Windsor C, Third Level
Ning Wang (Widener University)
Subject Matter Expert Rating of Real-Time Questions Compared to Non-Video Item Formats

AERA Poster Fair 1. 2:15 p.m. - 3:45 p.m. in Hyatt - Elizabeth Ballroom G, Second Level
Rebecca A. Goldstein - Montclair State University, Brian D. Bontempo - Microsoft
Conducting a Narrative Analysis on a Tentative Reconciliation Between Opposing Camps: Qualitative Versus Quantitative Research

Tuesday, April 13

SIG - Rasch Measurement Paper Session: Practical Applications of Rasch Measurement: Part I . 8:05 a.m. - 10:15 a.m. in Hyatt - Betsy C, Second Level
Damian P Birney (Yale University), Elena L. Grigorenko (Yale University), Robert J. Sternberg (Yale University)
An Application of the Many-Facet Rasch Measurement Approach to the Evaluation of Triarchic Instruction

Feifei Ye (Ohio State University), William Loadman (Ohio State University)
Assessing Unidimensionality of Dichotomous Item Responses from a Licensure Exam

Donna Surges Tatum, Johnna Gueorguieva (American Society for Clinical Pathology)
Portrait of a Profession: Mapping Laboratory Practice

Shannon Sampson (University of Kentucky), Kelly D Bradley (University of Kentucky)
Measuring Factors Impacting Educator Supply and Demand: An Argument for Rasch Analysis

Richard M. Smith - Data Recognition Corporation (Discussant)
Laurie L. Davis - Pearson Educational Measurement (Chair)

AERA Roundtable 38. 10:35 a.m. - 11:15 a.m. in Hyatt - Elizabeth Ballroom D, Second Level
Mark H Moulton (Educational Data Systems)
Weighting and Calibration: Merging Rasch Reading and Math Subscale Measures into a Composite Measure

Division D - Section 1 - Educational Measurement, Psychometrics, and Assessment Paper Session: DIF Applications. 12:25 p.m. - 1:55 p.m. in Marriott - Point Loma, South Tower, First Level
Lora F. Monfils - ETS K12 Assessments, Gregory Camilli - Rutgers University
Studying School Effects with Item Difficulty Variation: A Simulation Study

AERA Poster Fair 8. 12:25 p.m. - 1:55 p.m. in Hyatt - Elizabeth Ballroom G, Second Level
Kevin Pugh (University of Toledo)
Transformative Experiences in Science: Using Rasch to Develop a Quantitative Measure

Division D - Section 1 - Educational Measurement, Psychometrics, and Assessment Paper Session: Model/Data Fit. 12:25 p.m. - 1:55 p.m. in Marriott - Green Room, South Tower, Third Level
Jing Chen (Ohio State University), Ayres G. D'Costa (Ohio State University)
Effects of Test Anxiety, Time Pressure and Gender on Rasch Person-Fit Measures

Division D - Section 1 - Educational Measurement, Psychometrics, and Assessment Paper Session: Item Context and Item Effects. 2:15 p.m. - 3:45 p.m. in Marriott - Leucadia, South Tower, First Level
Adam N. Prowker - Rutgers University, Gregory Camilli - Rutgers University
Beyond the Composite: An Item Level Methodological Study of NAEP Mathematics Results

SIG - Professional Licensure and Certification Paper Session: Validity Issues in Licensure and Certification. 2:15 p.m. - 3:45 p.m. in Marriott - Laguna, South Tower, First Level

Ning Wang (Widener University)
Obtaining Task Importance Weights from Job Analysis Survey Data: An Application of the Multi-Faceted Rasch Model

Gregory J. Cizek (University of North Carolina at Chapel Hill)
Protecting the Integrity of Computer-Adaptive Licensure Tests: Results of a Legal Challenge

Wednesday, April 14

SIG - Rasch Measurement Paper Session: Measurement Issues in Rasch Models. 8:05 a.m. - 10:15 a.m. in Hyatt - Cunningham A, Fourth Level

Lixiong Gu, Benita J. Barnes, Edward W. Wolfe (Michigan State University)
A Rasch Examination of Psychometric Properties of Mathematic Test Performance Attribution Scale

Mohammed Louguit (Center for Applied Linguistics), Dorry M. Kenyon (Center For Applied Linguistics)
Constructing a Computer-Adaptive Oral Interview (the BEST Plus) Using Many-Facet Rasch Analysis

Steve Stemler (Yale University)
Inter-Rater Reliability and the Many-Facets Rasch Model: A Comparative Example

Linjun Shen (National Board of Osteopathic Medical Examiners)
The Assumption of the Rasch Model-Based Item-Mapping Approach in Setting Pass/Fail Standards

William P. Fisher - MetaMetrics, Inc. (Chair)
Trevor Bond - James Cook University (Discussant)

Division D - Section 1 - Educational Measurement: IRT Estimation. 2:15 p.m. - 3:45 p.m. in Hyatt - Molly A, Second Level

Iasonas Lamprianou (Centre for Formative Assessment Studies, University of Manchester, United Kingdom)
All Rasch Software Packages Are Born Equal: (So, Does It Matter Which One I Use?)

SIG - Rasch Measurement Paper Session: Practical Applications of Rasch Measurement: Part II. 2:15 p.m. - 3:45 p.m. in Hyatt - Betsy A, Second Level

Sue Leibowitz (University of Massachusetts Donahue Institute), Larry H. Ludlow (Boston College)
Measuring Change in Literacy Instruction: The BayState Readers Initiative Classroom Observations

Peter D. Macmillan (University of Northern British Columbia)
Primary School Fluency Measures of Early Literacy: A Many-Faceted Rasch Analysis of DIBELS

William P. Fisher (MetaMetrics, Inc.), Batya E. Elbaum (University of Miami)
Measuring Parent Involvement in and Satisfaction with Special Education Services

Constantia Hadjidemetriou (University of Manchester), Julian S Williams (University of Manchester)
Using Rasch Models to Identify Limitations in Teacher Knowledge

Randall E. Schumacker - University of North Texas (Chair)
Steve Stemler - Yale University (Discussant)

SIG - Rasch Measurement Business Meeting: Rasch SIG/ Business Meeting. 7:45 p.m. in Hyatt - Betsy A, Second Level

Randall E. Schumacker - University of North Texas (Chair) – SIG Program Chair
Richard M. Smith (Data Recognition Corporation) – Editor, Journal of Applied Measurement
Edward W. Wolfe (Michigan State University) – outgoing SIG Secretary
Trevor G. Bond (James Cook University) – outgoing SIG Chair
Election of SIG Officers

Thursday, April 15

AERA Poster Fair 5. 8:05 a.m. - 10:15 a.m. in Hyatt - Elizabeth Ballroom G, Second Level

Russell W. Smith - Prometric
The Impact of Braindump Sites on Item Exposure and Item Parameter Drift

Division C - Section 3 - Mathematics. Paper Session: Curriculum, Assessment and School Reform in Mathematics Education. 10:35 a.m. - 12:05 p.m. in Hyatt - Regency Ballroom E, Fourth Level

Julian S Williams (University of Manchester), Julie T Ryan (Liverpool John Moores University), Constantia Hadjidemetriou, Christina Misailidou, Thekla Afantiti Lamprianou, Maria Pampaka - University of Manchester
Credible Tools for Formative Assessment: Measurement and Qualitative Research Needed for Practice

AERA Poster Fair 14. 12:25 p.m. - 1:55 p.m. in Hyatt - Elizabeth Ballroom G, Second Level

Jale Cakiroglu (Middle East Technical University), Yesim Capa (Ohio State University), Hilal Sarikaya (Middle East Technical University)

Development and Validation of Turkish Version of Teachers' Sense of Efficacy Scale

SIG - Learning Environments (Formerly Study of Learning Environments) Paper Session: Examining the Learning Environment Through Students' Perceptions. 2:15 p.m. - 3:45 p.m. in Marriott - Orlando, North Tower, Lobby Level

Robert F Cavanagh, Graham B. Dellar (Curtin University of Technology, Australia)

Conjoint Use of Rasch and Correlational Analyses in Learning Environment Instrument Development

SIG - Rasch Measurement Paper Session: New Developments in Rasch Measurement. 4:05 p.m. - 6:05 p.m. in Hyatt - Edward A, Second Level

Dimiter M. Dimitrov (George Mason University), Richard M. Smith (Data Recognition Corporation)

Adjusted Rasch Person-Fit Statistics

Dimiter M. Dimitrov (George Mason University)

Ability Re-estimation in the Rasch Model

Chien-lin Yang (American Dental Association), Thomas R. O Neill (NCSBN), Gene A. Kramer (American Dental Association), Carol A. Vanek (American Dental Association)

Applying the Rasch Model to Examine Item Stability: A Longitudinal Study

Timothy W. Pelton (University of Victoria), Leslee G. Pelton (University of Victoria)

Exploring the Rasch Model's Potentials and Limitations by Rediscovering Length

Margo G.H. G.H. Jansen (University of Groningen, Fac. PPSW, GION)

Detecting Model Violations of Rasch's Multiplicative Gamma Model for Speed Tests

Gene A. Kramer - American Dental Association (Chair)
Peter D. Macmillan - University of Northern British Columbia (Discussant)

SIG - Spirituality & Education Network Interactive Symposium: Exploring the Efficacy of Mindfulness in Teaching Education: Diverse Pedagogical and Empirical Perspectives. 4:05 p.m. - 6:05 p.m. in Hyatt - Manchester Ballroom F, Second Level

William P. Fisher (MetaMetrics, Inc.)

Rasch Measurement of the Efficacy of Mindfulness in Education

Friday, April 16

SIG - Learning Environments (Formerly Study of Learning Environments) Paper Session: Technology-Rich Learning Environments. 8:05 a.m. - 10:15 a.m. in Marriott - Solana, South Tower, First Level

Robert F Cavanagh, Graham B. Dellar (Curtin University of Technology, Australia)

Measuring Student Perceptions of Classroom Information and Communication Technology Learning Culture

Division D - Section 1 - Educational Measurement, Psychometrics, and Assessment Paper Session: Equating. 12:25 p.m. - 1:55 p.m. in Hyatt - Madeleine B, Third Level

Husein Taherbhai, Daeryong Seo, Thomas Brooks (Harcourt Educational Measurement)

Comparing Concurrent Versus Fixed Parameter Equating with Common Items for the Equivalent and Nonequivalent Group Designs, Using the One-Parameter Rasch and the Partial Credit Model in a Mixed-Item Format Test

Division D - Section 1 - Educational Measurement, Psychometrics, and Assessment Paper Session: Innovative Item Response Models. 2:15 p.m. - 3:45 p.m. in Hyatt - Annie B, Third Level

Thakur B. Karkee (CTB/McGraw-Hill), Karen R. Wright (CTB/McGraw-Hill)

Evaluation of Linking Methods for Placing Three-Parameter Logistic Item Parameter Estimates Onto a Rasch Scale

Division D - Section 1 - Educational Measurement, Psychometrics, and Assessment Paper Session: Technical Issues in DIF. 2:15 p.m. - 3:45 p.m. in Marriott - Columbia 3, North Tower, Lobby Level

Ya-Hui Su - National Chung Cheng University, Wen-Chung Wang - National Chung Cheng University

Efficiency of the Mantel, Generalized Mantel-Haenszel, and Logistic Discriminant Function Methods in the Assessment of Differential Item Functioning for Polytomous Items

---

At the National Council on Measurement in Education 2004 Annual Meeting

Cees A. W. Glas, Jean-Paul Fox, University of Twente, Netherlands

Analysis of variance and regression using multilevel IRT

The Saltus Model

A fundamental concept underlying most Rasch model applications is Thurstone's 1928 precept: "Within the range of objects for which the measure instrument is intended, its function must be independent of the object of measurement." (*American J. of Sociology*, 33, 529-554)

But what if this isn't so? What if there are two or more types of person to be measured for which the measuring instrument behaves differently? For instance, a language test administered to native and non-native speakers. If the persons can be separated by type, then an immediate solution is to construct a measuring system for each type. If the persons can't be separated, then there is a "mixture" situation. One population (or type) of persons is mixed with another.

A particular "mixture" is that addressed by the Saltus (Lat. "leap") model (Wilson, 1989). The persons are at different levels of psychological development – and the leap from one level to the next changes the way in which the measuring instrument operates. We can guess an individual's level, but not with certainty. We assume that we can identify those items which we expect to operate differently for persons at different levels.

In the simplest case, it is hypothesized that, at each level, the persons can be treated as randomly sampled from a normally distributed population, with a specific mean and standard deviation for that level. Also, while in other mixture models, each dichotomous item has a specific difficulty calibration for persons at that level, the Saltus model is simpler and estimates a much smaller number of parameters: the amount (measured in logits) that **each set of items "**shifts" in difficulty when encountered by persons in each level.

Thus what needs to be estimated are:
(a) The mean and standard deviation of each level's sample.
(b) The proportion of the total sample at each level.
(c) The difficulty calibration of each item at each level.

An estimation approach is to apply the MML (marginal maximum likelihood) formulation with an EM (expectation-maximization) algorithm. At each point, some of the parameters are estimated while others are held steady. Generally this process converges to a stable set of estimates.

From the final estimates, the probability that any particular person belongs to any particular level can be computed. These probabilities can be summed, so that, for instance, the proportion of second-grade boys at a particular level can also be estimated. For individual reporting purposes, it is conventional to consider each person to be at the level with the highest probability for that person, although some people may have a profile of probabilities that are fairly equal across levels.

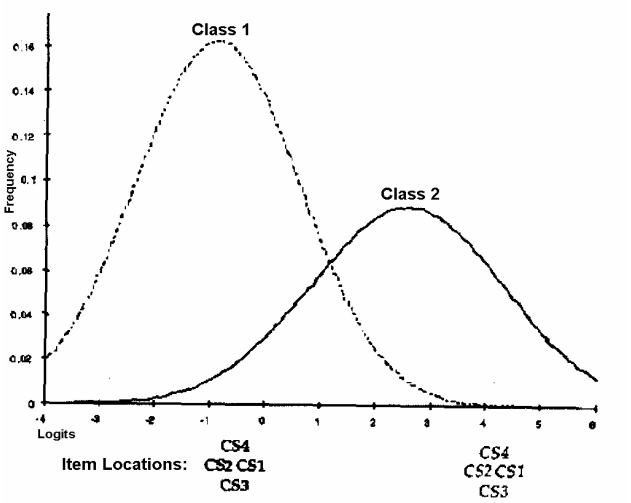
Mark Wilson kindly assisted with this description.

Mislevy, R.J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61*(1), 41-71.

Wilson, M. (1989). Saltus: A psychometric model for discontinuity in cognitive development. Psychological Bulletin, 105, 276-289.

Wilson M. & Draney K. (1997) Partial credit in a developmental context: the case for adopting a mixture model approach. Chap. 18 in M. Wilson, G. Engelhard, Jr., K Draney, Objective Measurement: Theory into Practice, vol. 4. Greenwich CT: Ablex.

A computer program for estimating this model (and a polytomous version) is available from Karen Draney (kdraney *at* clink.berkeley.edu).



Results of a dichotomous Saltus analysis
(Wilson & Draney, 1997)
Items CS1-4 are shown at their difficulty levels for Class 1 (to left) and Class 2 (to right) in the same frame of reference. In this unusual example, the items are more difficult for the higher ability class, because their increased knowledge misleads them into using an incorrect answering strategy.

Rasch Measurement Transactions
P.O. Box 811322, Chicago IL 60681-1322
Tel. & FAX (312) 264-2352
rmt@rasch.org www.rasch.org/rmt/
Editor: John Michael Linacre

SIG Chair: Trevor Bond SIG Secretary: Ed Wolfe

Assessing Statistically and Clinically Meaningful Construct Deficiency/Saturation: Recommended Criteria for Content Coverage and Item Writing

A construct is an underlying latent trait that cannot be directly observed and measured (e.g., a mental property). The goal of measurement, specifically in social science research, is to develop questionnaire or test items to assess those unobservable constructs indirectly. The objective is to have items that cover as much as possible of the construct's continuum to allow for use in collecting information about a wide range of person performance.

In order to correctly estimate a person's location on a construct, it is imperative to define that construct well (Wright & Stone, 1979). When items are developed, they are intended to cover the spectrum of the construct being defined. However, there are instances when this is not the case. The result is insufficient or redundant coverage. The two instances are referred to as: 1) construct deficiency (insufficient coverage) and 2) construct saturation (redundancy). Each has implications for item bank development; which, in turn, impacts development of computer-based and computer-adaptive tests.

An item bank is a comprehensive catalog of items for use in creating psychometrically sound fixed-length, brief form and/or adaptive tests. These items should span the various construct dimensions and function along their respective continua at various difficulty levels. "The idea is that the test user can select test items as required to make up a particular test" (Choppin, 1978). The flexibility provided by an item bank allows the researcher to utilize valid, reliable and well-validated items without being required to re-calibrate those items each time they are used. Items selected for future use can differ, thus allowing optimized use of individual items.

Construct Deficiency:

Under-representation of content area

Construct deficiency, CD, represents 'gaps' on the construct continuum. These 'gaps' represent the points at which the construct is poorly defined by the items (Schulz, 1995). In this situation, the goal is to develop items which fill these 'gaps' at the specified logit value. There are two specific types of CD of interest:

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

Benjamin D. Wright
May, 1976

Long on the wall of Ben Wright's office, Judd 438, at the University of Chicago. Ben became emeritus on January 1, 2004. He continues to live in Hyde Park, Chicago.

1) statistically meaningful construct deficiency (SMCD), and 2) clinically meaningful construct deficiency (CMCD).

SMCD is a flexible index assigned by the principal investigator and item-banking team. A distance of 0.30 to 0.50 logits is a recommendation for SMCD evidence. CMCD is conceptualized on two levels: 1) important content area is not covered, and 2) overall content area is not covered fully. If an item is deemed clinically meaningful, upon consensus, regardless of fit, it is kept in the bank.

Implications for Item Bank

The optimal goal of an item bank is to fully cover the spectrum of a construct, thus producing a reliable measure. When a construct is poorly defined, the implications for future use are: 1) floor and ceiling effects will impact those individuals whose ability levels fall outside of the item difficulty levels, thus providing inadequate information; and 2) individuals whose ability levels are at the location of a 'gap' will be given items that poorly target their ability. Furthermore, there are two specific ramifications for a poorly defined construct: 1) impact on the development of computer-based tests, and 2) on the development of computer-adaptive tests.

Impact on Development of Computer-Based Tests

Construct deficiency can impact the results of a computer-based test because it reduces the amount of information obtained for each individual because the construct is poorly defined. This is problematic on two levels: 1) items are not targeted at the person's ability level, and 2) higher error estimates for the person's ability level, thus lowering precision and interpretability.

Impact on Development of Computer-Adaptive Tests

Construct deficiency impacts computer-adaptive tests in much the same way as it impacts computer-based tests. Maximum-information-based computer-adaptive tests specifically function to target the person at his/her ability level with items at the same level of difficulty. If there is not an item located at that person's ability level, the test is forced to move to an item further away, thus increasing the error of the ability estimate. Items are presented based on responses to the preceding item, therefore, it is necessary to fully define the construct along the continuum before attempting to produce this type of test. A bank of items limited by construct deficiency results in the inability to measure individuals along the entire ability continuum with high precision (Halkitis, 1996).

Setting up a computer adaptive test requires thresholds for item selection (i.e., logit range), and precision (i.e., stopping rules based on individual standard error). When a construct is poorly defined, the individual is forced to

take more items in order to achieve a reliable estimate.

Construct Saturation:
Over-representation of content area

Construct saturation is over-representation by similar items at a specific logit value. This is defined more fully as the point on the construct continuum where several items are measuring the same thing in almost the same way. Overall, the goal is to have all of the items measure the same construct. However, we want them to produce new information at each level of that continuum. "A useful item is 'as similar as possible, but as different as possible' (Linacre, 2000)". An item bank may have many items at the same difficulty level. Over-representation occurs when some of those items are too similar and so are no longer independent. The redundancy incurred by administering two almost identical items slightly distorts the person ability measures, but does not impact the overall measures noticeably.

Implications for Item Banks

The implications of construct saturation in an item bank are more positive than negative. By incorporating items that measure the same thing on a construct, it is possible to extend the choices for item selection by the test developer. But overly similar items should be identified as alternatives when used in the construction of any particular test.

Impact on Development of Computer-Based Tests

The impact of construct saturation on a computer-based test is negative if more than one alternative item is included. Respondents may become frustrated when presented with several items that ask essentially the same thing. Further, statistical information is usually based on regarding the items as independent. It is difficult to make adjustments for non-independent items.

Impact on Development of Computer-Adaptive Tests

Construct saturation on a computer-adaptive test is beneficial for the test developer because it allows different alternative items with similar logit values to be presented to different individuals as they proceed through the test. This overcomes the problem of "tracking", which occurs when all persons of similar ability are administered essentially the same test. Therefore, to avoid over-exposure of individual items and also "tracking", it is actually beneficial to have redundant alternative items.

Construct Coverage Protocol:
Methods for Gap-Filling

In the presence of SMCDs and CMCDs, there are seven steps recommended below as a possible solution:

Step 1: Identification of any clinically or statistically meaningful gaps or redundancies in the continuum. This requires labeling the gaps as statistical, clinical, or both, and identifying sets of alternative items.

Step 2: Determine the number of items needed to fill each gap (e.g., 5-10 items, depending on the gap size).

Step 3: Formulation of new items by a committee comprised of clinical and statistical experts.

Step 4: Review by oversight committee. Reasons for rejection of items recorded in hard copy.

Step 5: Testing of new and revised items with clinical collaborators and selected group of patients.

Step 6: Patient testing utilizing computer-based-testing procedures that incorporate old and new items.

Step 7: Calibration of new items along the anchored continuum of the previous items.

Stacie Hudgens, Kelly Dineen, Kimberly Webster, Jin-Shei Lai, David Cella on behalf of the CORE Item Banking Team

Choppin, B. H. (1978) Item Banking and the Monitoring of Achievement Research in Progress Series, I. NFER.

Halkitis P. N. (1996) CAT with a Limited Item Bank. RMT 9:4 p. 471.

Linacre, J.M. (2000) Redundant Items, Overfit and Measure Bias. RMT 14(3) p.755.

Schulz E. M. (1995) Construct deficiency? RMT 9(3), p. 447.

2nd International Conference on Measurement
in Health, Education, Psychology and Marketing: Developments with Rasch models
Murdoch University, Western Australia, 2004

A personal recollection of the Health Sciences sessions.

It is always a pleasure to visit Perth in the summer time, with the Indian Ocean and white sands to tempt one away from serious business (not to mention the Murdoch Club and its delicious *lattes!).* Nevertheless, the conference program was such that a full program of Rasch applications in Health kept ones full attention. Although it may be mentioned elsewhere, the opening plenary session by Professor David Andrich was based on health data, looking at female maturation. The data was provided by another speaker, Dr Stef van Buuren, from TNO in the Netherlands. David's point was methodological, looking at the difference between statistical approaches whereby the object was to combine items to provide a Maturation Index, and the measurement perspective, which looked at how an item could be used to inform on group differences, rather than be included in the Index. The presentation, using RUMM2020, laid out the inherent tensions which exist in much health outcome work, nicely highlighting the potential differences between measurement and explanation.

The range of applications in health was then illustrated by two presentations in the first parallel session. Carlyne Arnould from the Universitè Catholique de Louvain presented her work on the development of the Abilhand-Kids questionnaire to measure manual ability in children with cerebral palsy. Taking an existing scale for adults as a starting point, the presentation reported on its adaptation, reliability and validity for use with children. There were some differences in the response to items made by the children and their parents (which is not unusual and has been reported in other measures where both perspectives are obtained), but, using Winsteps, the data fitted the Rasch model and good test-retest reliability was achieved (r=0.87). Following this Dr Stef van Buuren from TNO Prevention and Health in the Netherlands presented his work testing Rasch analysis to the extreme by taking data from many different countries, forming an (extremely) ill-conditioned data set, and forging links between countries (sometimes by using a special bridging study) to bring all the various country data together. The approach consisted of two steps, constructing the data matrix, taking items from national disability studies for constructs such as dressing or walking, then creating a conversion key using threshold estimates (from RUMM2020) to link all the items (for a single construct such as dressing) together.

In the next session, Ann Björkdahl from Göteborg University in Sweden presented her work on the European Brain Injury Questionnaire in stroke patients. Using Winsteps, she took data from 59 patients with median age 53 years, and looked at the 63 item scale. With an acceptable fit range of 0.6-1.4, seven items were found to misfit but all items showed stability over time. The Tukey Mean Difference plot was used, which is the same as the Bland & Altman plot commonly found in health studies. Irene Styles from Murdoch University then presented her work on the International Classification of Functioning Disability and Health (ICF) qualifier scale. A checklist of 48 impairments determined what were the patient's problems, and what were the extent of those problems. Data from patients with low back pain, breast cancer and stroke were fitted to the rating scale model. Although data tended to fit the model within diagnosis, extensive DIF was found across conditions.

Elizabeth Betemps from the University of Cincinnati opened the next session with her work on a 13 item Psychiatric Distress Scale, looking at the difference in magnitude of effect by raw scores compared with transformed measures. Based upon 54 subjects on admission and discharge from an outpatient psychiatric treatment program in Ohio, USA, significantly more patients were found to have improved at discharge using the measure as opposed to the raw score. Eric Wong from the Chinese University of Hong Kong then presented his work on a sample of 1956 patients from two stroke databases, comparing the Functional Independence Measure (FIM) with the Barthel Index. With mean age of 72 years; 51% male and 51% with left hemi-paresis, data from the scales were fitted to the partial credit model using Winsteps. Bladder and bowel items were found to misfit on both scales, and the easiest and most difficult items were also consistent. Both measures showed invariance of items across admission and discharge. However, the FIM motor scale showed greater precision (person separation) than the Barthel Index.

The FIM in stroke and head injury formed the next joint presentation between Åsa Lundgren Nilsson from Göteborg University in Sweden and Anita Slade for Sheffield Hallam University in the UK. The thrust of their presentation was similar to the discussion elsewhere in this issue with regard to the disordering of thresholds and DIF by country where pooled international data is required. For the stroke data, there were 2546 patients at admission from six countries. For head injury, 779 patients from 5 countries. The pattern of analysis was consistent across diagnosis; initially with considerable disordering of thresholds, a complex solution involving splitting of items was required to facilitate the pooling of data across countries.

Finally the author presented his work on the development of item banking for quality of life (QoL) across the rheumatic diseases. Using a strong theoretical model of QoL, namely the Needs-Based model of Hunt and McKenna, and the strong mathematical model of Rasch,

an example of building an item bank was given for Psoriatic Arthritis and Systemic Lupus Erythematosus. The basis of this was common item equating, ensuring that links were free of DIF by diagnosis, and that the entire set gave adequate fit to the Rasch model.

Alan Tennant
Professor of Rehabilitation Studies
The University of Leeds. UK.

The three most interesting presentations at 2ICOM:
Although my stay at Perth was very short (from Tuesday morning to Wednesday evening), I enjoyed every minute of my stay in the conference sessions. Among the presentations which I attended (all of which are related to Education, because I am an educator), I found the following three most interesting and useful for my future research. They are:
1) Ted Brown's "The Scalability & Validity of Four Pediatric Visual Perceptual Instruments: A Comparison Using the Rasch Measurement Model", which gave me a hint to reconsider the construct validity of a language test.
2) Juho Loover's "Using Modern Psychometric Theory to Identify Differential Item Functioning
in Polytomously-Scored Constructed-Response Items", which provided me with an illustrative methodology that can be applied to any examination which contains polytomously-scored items.
3) Sergij Garbrscek's "Taking another perspective: Matura examinations in Slovenia", which introduced me to a general ideal of Matura examinations in Slovenia with statistical and educational information.

Yuji Nakamura
Tokyo Keizai (Economics) University

Scale Construction or Analysis?

"The so-called perfect scale, in the scalogram sense, is a one-dimensional structure for data of the kind studied by item analysis. The dimensionality of data is an empirical phenomenon, and not to be determined by fiat. Therefore, I have suggested abandoning the idea and terminology of *scale construction* in favor of *scale analysis*."

Louis Guttman, *Psychometrika, Vol. 36, No. 4, December, 1971.*

Guttman analyzes data hoping to find a latent scale. Rasch constructs the latent scale from data intended to manifest that scale.

"Quality is never an accident. It is always the result of intelligent effort. There must be the will to produce a superior thing."

John Ruskin

From Microscale to Winsteps: 20 years of Rasch Software development

In 1964, Ben Wright, Bruce Choppin and Nargis Panchapakesan began development of Rasch measurement computer programs (RMT 10:2, 494-6). These followed Ben's pioneering factor analysis programs which ran on the University of Chicago's UNIVAC I computer in 1959.

The most successful of Ben's mainframe programs was BICAL (1976, with Ron Mead and later Susan Bell). This was written in FORTRAN IV to run on IBM 370 computers. It constructed measures from complete dichotomous data with a scoring key. It was distributed as source code which the user compiled. There are indications that it is still in use.

In 1983, Ben Wright was consultant to a research company, Mediax, of which Mike Linacre was the Computer Sciences Manager. Mike had written his first computer program in 1965 for the University of Cambridge EDSAC II computer. Mediax was in need of analysis software for educational tests containing dichotomous and rating scale items. Further there were missing data. In a series of meetings, Ben and Mike decided to develop Rasch software, capable of analyzing those data. The software would run on the new business-capable IBM XT personal computer, under the newly stable MS-DOS 2.1 operating system.

Early in 1984 the Rasch computer program Microscale appeared. Ease of data entry and graphical output were important. So Microscale was designed as an add-on to the then widely-used Supercalc3 spreadsheet program. It was popular with test developers, particularly in the language area. A free version, "Student Microscale", was distributed with the free evaluation version of Supercalc3. A drawback of Supercalc3 was its limited dataset size. So "Professional Microscale" was produced as an add-on to the SYSTAT statistics package. Active distribution of Microscale came to an end in 1987.

Since the word "Microscale" was too long for an MS-DOS program name, users entered MSCALE at the DOS prompt to launch it. Ben Wright took this name and applied it to a rewrite of Microscale into Fortran to run on the University of Chicago, Department of Education, UNIX minicomputer. MSCALE appeared around 1987 (Wright, Matt Schulz, Richard Congdon, Mark Rossner, and various authors) and was designed to analyze dichotomies and rating scale data.

MSCALE was distributed as source code which users sometimes had trouble compiling correctly. It had a maximum dataset size which users were starting to exceed. Also personal computers were now becoming the researchers tool of choice. Consequently, Ben and Mike decided to produce a revised and enlarged version of MSCALE, called BIGSCALE, which began to be distributed in 1989 in compiled form for PC computers. Ben had another program, MSTEPS (Wright, Schulz, Congdon, Rossner), for partial credit items. Its functionality was incorporated into BIGSCALE and, with other enhancements, launched as BIGSTEPS in 1991.

In 1998, BIGSTEPS was rewritten for Windows and published as Winsteps. Since a feature of Winsteps is compatibility all the way back to MSCALE, some initially doubted that Winsteps really was a Windows-native program. An advantage of Windows, however, has been the ease with which new capabilities can be introduced into the program. Each time new capabilities have been introduced, users have found innovative uses for those capabilities, and made suggestions for further innovations, and so the range of its Rasch measurement applications is ever widening. Winsteps® is now published by *Winsteps.com.*

For comparison, BICAL had about 1,500 lines of FORTRAN code, MSCALE about 3,000 lines of FORTRAN code. Winsteps has about 65,000 lines of FORTRAN code, 6,000 lines of Visual Basic code, 40 lines of C++ code and also incorporates code modules provided by other software developers.

John Michael Linacre