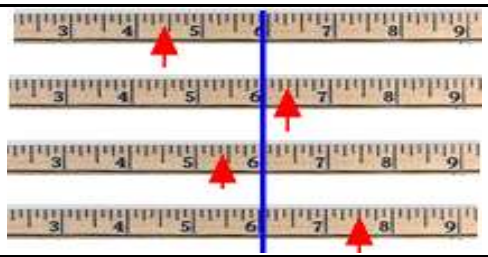
Diagnostic Profiles



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 25 No. 4 Spring 2012 ISSN 1051-0796

Convergence, Collapsed Categories and Construct Validity

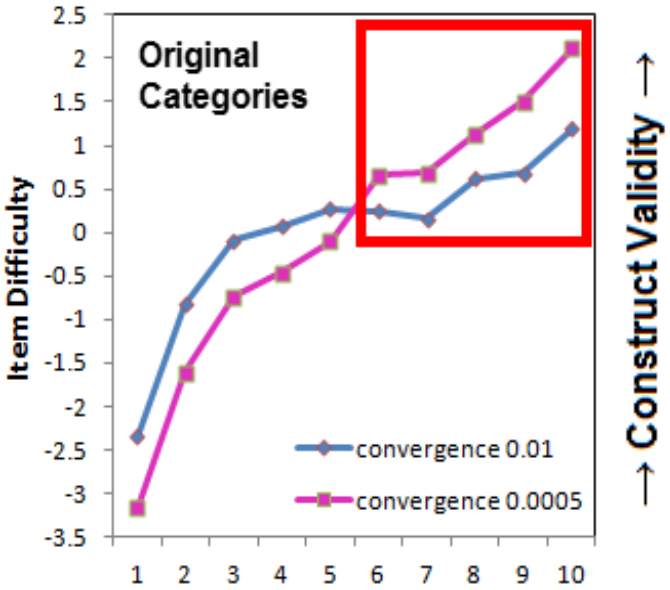


Figure 1. Item locations with original categories, including unobserved categories.

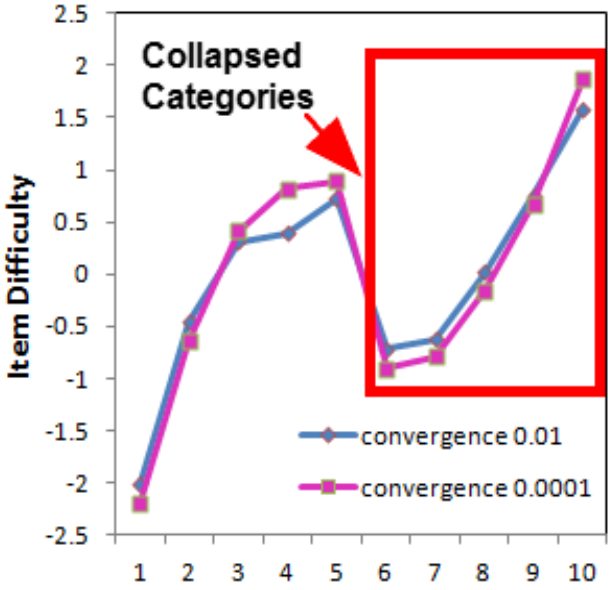


Figure 2. Item locations with unobserved extreme categories collapsed with neighboring categories

While analyzing a dataset of 10 polytomous partial-credit items, I found that the estimates of item difficulties and their ordering varied depending on the convergence limits set for estimation. The ordering is important because it is used as evidence for the construct validity of the instrument. In my investigation, the item locations were calibrated twice. Once with the item convergence limits set at 0.01 and again at 0.0005. The sample size was 6,520 and the person ability distribution was roughly normal.

Figure 1 is a graphical presentation of the differences between item locations at the two convergence limits. As expected, the tighter (smaller) convergence limit resulted in more dispersed item difficulty estimates. The location differences between convergence at 0.01 and convergence at 0.0005 are surprisingly large. The absolute values of the differences vary from 0.38 to 0.92 logits. What could be the reason?

When category frequencies were examined, it was found that items 6-10 had no observations in their extreme highest categories (see Table 1). These had been automatically accommodated by my RUMM2030 analysis. To examine the impact of the unobserved categories on the item locations, the unobserved extreme category 5s were combined with their adjacent category 4s. After collapsing those extreme categories, the item locations were again estimated twice with convergence limits at 0.01 and more tightly at 0.0001.

Figure 2 shows the resulting item estimates. Compared with the estimates in Figure 1, the location differences for each item at the two convergence limits are much smaller. This time, the absolute differences varied from 0.10 to 0.42 logits. Although, there were no changes made to items 1-5, the location differences for most of these items

Table of Contents

Item	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5
1	640	3928	1952		
2	768	5539	213		
3	46	1305	5050	118	1
4	31	5581	907	1	
5	450	3886	1997	185	2
6	500	1940	3982	98	0#
7	83	4719	1685	33	0#
8	918	3313	2222	67	0#
9	400	5289	824	7	0#
10	1654	4425	437	4	0#

Table 1. Original category frequencies of the data

Note: # this unobserved category collapsed with adjacent category in the second analysis.

also reduced between the two convergence limits. The difference for item 4 remained somewhat large, perhaps because item 4 has only 1 observation in category 4, its top category, making estimation of its difficulty location less stable.

As expected, the items with collapsed categories, items 6-10, have become relatively easier than in the first, uncollapsed analysis. This is because the definition of item difficulty is "the location on the latent variable at which the top and bottom categories are equally probable." Collapsing the two highest categories for each item has moved the combined top category toward the middle of the original rating scale, and so moved the item location down the latent variable.

In conclusion, these analyses indicate that convergence limits should be set tightly enough to be substantively stable. These analyses also show that collapsing categories can make conspicuous changes to the item difficulty hierarchy. If categories are collapsed and the item hierarchy must be maintained for measure interpretation and construct validity, then pivot-anchoring (RMT 11:3 p. 576-7) may be required.

Edward Li

Report on the Kaggle-Grockit Competition

The Grokit competition featured in RMT 25:3, p. 1329. Contestants were asked to predict students' responses based on those students' previous responses. The competition ended on Feb. 29, 2012. Many contestants used Rasch models. The most successful contestant was Steffen Rendle, Social Network Analysis, University of Konstanz. He used a "Factorization Machine" which estimates higher order interactions in very sparse datasets in order to predict new data. It is not known whether the increased complexity of the winning algorithm is matched by an increased utility in its predictions beyond those of a simple Rasch model.

Rasch-related Coming Events

March 20, 2012, Tues. 6th UK Rasch User Group Meeting, Leeds, UK, www.rasch.org.uk

March 21-23, 2012, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Apr. 11-12, 2012, Wed.-Thurs. IOMW International Objective Measurement Workshop, Vancouver BC, Canada, www.iomw2012.com

Apr. 13-17, 2012, Fri.-Tues. AERA Annual Meeting, Vancouver BC, Canada, www.aera.net

May 18, 2012, Fri. Ohio River Valley Objective Measurement Seminar (ORVOMS), Lexington, KY, Announcement

May 23-25, 2012, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

May 28-30, 2012, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

July-Nov., 2012 On-line course: Introduction to Rasch Measurement of Modern Test Theory (D. Andrich, RUMM2030), Perth, Australia, www.education.uwa.edu.au/ppl/courses/introduction

Aug. 6-9, 2012, Mon.-Thur. PROMS2012, Jiaxing University, Zhejiang Province, P.R.China, http://cfs.zjxu.edu.cn/proms/

Aug. 12-14. 2012, Sun.-Tues. IACAT 2012, International Association for Computer Adaptive Testing, Sydney, Australia, www.iacat.org

Sept. 5-7, 2012, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Sept. 10-12, 2012, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Sept. 13-14, 2012, Thurs.-Fri. In-person workshop: Advanced Rasch (A. Tennant, RUMM), Leeds, UK,

Dec. 5-7, 2012, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK,

Dec. 10-12, 2012, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Apr. 27 – May 1, 2013, Sat.-Wed. AERA Annual Meeting, San Francisco, CA, www.aera.net

A Suggestion for Taking Rasch-based Survey Results Even Further

Researchers across disciplines regularly publish articles that investigate the psychometric properties of a survey instrument, commonly referred to as "validation studies". Although researchers seem well-versed in making arguments for the various aspects of construct validity and addressing the technical specifics of their findings, one glaring omission seems predominant in most articles: Researchers fail to address how others can use the results for direct and meaningful comparisons.

The concept of anchoring is certainly nothing new to the measurement community. Likewise, research has long touted that Rasch models produce sample-free calibrations (meaning as long as the predominant latent trait is sufficiently detectable the construct should be defined in both an accurate and stable manner across samples, thus negating the need for representative samples). Despite the Rasch community being well aware of both of these important concepts, rarely are these important concepts extended to their utmost utility.

I argue that instead of simply making the case that one's instrument appears psychometrically sound and encouraging others to adopt it for studies of their own, researchers should consider going a step farther. When researchers are confident that they have defined the construct based on sufficiently unidimensional measures, others may benefit by not only using the same instrument, but also by linking their results onto the same scale for direct comparisons. In order to do this, researchers need to report the rating scale categories with threshold calibrations and item calibrations so that these estimates can serve as anchors for other researchers who wish to bring their measures onto the same scale. This will allow for direct comparisons across administrations of the instrument. Of course, the reverse is true as well. Researchers looking to replicate findings can easily create rating scale and item anchors and bring their sample of respondents onto the same scale as presented in the initial study for direct comparison. In all instances, the concept of exchangeability is taking place and researchers are able to essentially use the same "currency" to investigate findings. Furthermore, when a common currency is available, substantive and theoretical differences and similarities can be better detected, thus potentially advancing the knowledge base within a field at a much quicker rate.

An example might include an instrument that measures mental toughness among collegiate athletes. With appropriate anchoring, members of two sporting teams who have completed the instrument could be compared. These athletes performance in competition could then be coupled with the mental toughness findings to determine the extent to which mental toughness seems to matter in competitive sports. Do people who are identified as having the greatest amount of mental toughness seem to shine in competition, as theory might suggest? Of course, this is just a hypothetical example, but the possibilities are rather endless when one considers the wide array of academic disciplines in which Rasch models are now used.

Of course, there are caveats to this approach. Persons conducting studies of their own need to ensure the instrument is functioning as desired given the particular sample of respondents. Typical quality control checks should be executed upon initial unanchored analyses of the data, as well as after the rating scale threshold and item calibrations have been anchored. Should data fit the model adequately and other indicators suggest the scores are sufficiently reproducible and valid in both scenarios, the suggestion to take findings a step farther could have a number of meaningful consequences for knowledge production and information discernment.

With regard to future directions, we know that the concept of exchangeability is not just something that Rasch advocates value. People from all walks of life also value the simplicity and utility of having common frames of reference. I believe this topic is one that the Rasch community has yet to fully realize in practice, and one that could potentially help others who are uninformed about Rasch models better appreciate their beauty and utility as well.

Kenneth D. Royal

United Kingdom Rasch Day 2012

Leeds, Tuesday, March 20th, 2012

at Weetwood Hall www.weetwood.co.uk

hosted by *Professor Alan Tennant,* Psychometric Laboratory for Health Sciences in Leeds

Highlights from the program

Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment. *Svend Kreiner. Dept. of Biostatistics, University of Copenhagen.*

Cross-country Comparisons of Inattentive, Hyperactive and Impulsive Behaviour in School-Based Samples of Young Children. *Christine Merrell , Peter Tymms and Irene Styles, Universities of Durham, and of Western Australia.*

Is Aberrant Response Behaviour an Inherent Characteristic of Students Taking Classroom Maths Tests? *Dr Panayiotis Panayides, Lyceum of Polemidia.*

Rasch Analysis of the Intermittent and Constant Osteoarthritis Pain Questionnaire. *Bryan, J. Moreton, University of Nottingham.*

Rasch Theory in Product Design Applications. *Fabio Camargo, Brian Henson. University of Leeds.*

UK Rasch User Group, www.rasch.org.uk

Fifth International Conference on Probabilistic Models for Measurement

January 23-25, 2012, Perth, Australia

Authors and Papers

Henrik Albeck - Large scale adaptive testing of students grade (1-9)
David Andrich - Rasch's measurement theory and R. A. Fisher's experimental epistemology
David Andrich - Equating of high stakes university selection tests by applying the Rasch model using RUMM2030 software
Carolina Ballert - Developing clinical measures of functioning for SCI based on the international classification of functioning, disability and health
Skye Barbic - Challenges in exploring emotional vitality items over time using Rasch analysis
Daniel Bergh - Measuring psychosocial work environment - an analysis of the psychometric properties of a scale using Swedish data
Bipin Bhakta - A test of invariance for different versions of the same test in an undergraduate mechanical engineering exam
Fabio R Camargo - A rationale for comparing affective responses to stimulus objects using the faceted Rasch model
Stefan J. Cano - Combining clinical hypotheses with Rasch measurement theory: how we built the BREAST-Q©
Yuk Fai Cheong - Hierarchical Rasch models for rater-mediated assessments
Jonathan David Comins - Construct validity in patient-related outcome scores for anterior cruciate ligament deficiency - a matter of content!
Dawson Cooke - Investigating parental reflective functioning
Serkan Dolma - A pragmatic approach to the problem of choosing a polytomous model in organizational behavioral studies
George Engelhard, Jr. - Rater-invariant assessments in the human sciences
William P. Fisher, Jr. - Geometrical and social aspects of measurement: on the potential for metrological traceability in the social sciences
Masoud Geramipour - Comparison of Confirmatory Factor Analysis and IRT based Likelihood Ratio in detecting Differential Item Functioning
Curt Hagquist - Using the polytomous Rasch model to distinguish between real and artificial DIF: An illustrative example based on Swedish adolescent data
Clayon Hamilton - Rasch Analysis of the Upper Extremity Functional Index (UEFI)
Joanne Hardy - Putting Rasch into context: construct validity of the Context Assessment Index in a Western Australian tertiary hospital
Jeremy C Hobart - Quantifying clinical change: The responsiveness paradox
Jeremy Hobart - And exactly what does that mean for me doctor? Using Rasch analysis to extract clinical meaning from the numbers generated by treatment trials
Mike Horton - investigating the patient scar assessment questionnaire using Rasch analysis
Mike Horton - Development of the Stroke-QoL: A needs based quality of life measure specific to stroke
Steve Humphry - Is it possible to develop a system of units in the social sciences?
Ingvar Johansson - Problems in the SI System and their relevance to social-scientific measurement
Paula Kersten - The cognitive behavioral responses to symptoms questionnaire (CBSQ), a first validation study
Paula Kersten - A cross-cultural validation study of the MSIS-29
Paula Kersten - The MS Fatigue Self-Efficacy Scale, a valid measure for MS populations
Tetsuo Kimura - Moodle UCAT: a computer-adaptive test module for Moodle based on the Rasch model
Svend Kreiner - On a shaky foundation? A critical look at PISA's scaling model
Andrew Kyngdon - Event splitting effects, violation of stochastic dominance and the role of the unit in utility measurement
Becky Lau - A Rasch measure of young children's temperament in Hong Kong
Joseph P. Lavallee - Restricted range bias in the Angoff method: Causes and consequences
Goran Lazendic - The implementation of the Rasch model in construction of NAPLAN assessment scales
Caroline Long - Insights into the multiplicative conceptual field informed by an application of the Rasch measurement model
Juho Looveer - Development of a Rasch scale for an interview-based pattern and structure assessment (PASA) for early years mathematics
Guanzhong Luo - Roles of the Rasch analysis in the grading of high stake examinations and their relationships with professional judgments
Ida Marais - Quantifying local, response dependence between two polytomous items using the Rasch model
Ida Marais - The effect of students' random guessing on Rasch item and person estimates
Nicholas Marosszeky - Assessing unidimensionality using the Rasch extended logistic model
Joshua McGrane - Stevens' forgotten crossroads: The divergent paths of measurement in the physical and psychological sciences in the latter half of the 20th century
Mounir Mesbah - Choice of a logit link function: adjacent or cumulative?
Joel Michell - On heterogeneous orders
Marianne Mueller - Nursing competence: Psychometric evaluation using Rasch modeling

Åsa Lundgren Nilsson - The internal (factorial) construct validity of the multidimensional fatigue inventory (MFI-20) in a sample of patients treated for myocardial infarction.
Maja Olsbjerg - Coefficient alpha and latent correlation in multidimensional Rasch models and unidimensional multi frame of reference Rasch models
Robyn Owen - Welcome Address by Pro-Vice Chancellor (Research), University of Western Australia
Imogene Rothnie - The utility of the multi-faceted Rasch model for evaluating construct validity, rater effects and DIF on the multiple mini interview for medical school selection
Rassoul Sadeghi - Investigating the cross-lingual measurement invariance using Rasch's simple logistic model
Thomas Salzberger - Comparing Rasch item measures and measures from best-worst-scaling (BWS) using a scale assessing perceived corporate social responsibility
Charles Sèbiyo Batcho - ACTIVLIM-Stroke: A cross-cultural Rasch-built scale of activity limitations in stroke patients
Anita Slade - Less is more: Do fewer scoring categories in the BERG deliver better measurement in multiple sclerosis?
Anita Slade - The ALPS CIPA Tool: Development of an inter-professional self-assessment measure of student's perceived competence to practice.
A. Jackson Stenner - Causal Rasch models
A. Jackson Stenner - Using the quantile framework in curriculum embedded mathematics assessment
Khairiah Syahabuddin - Differential item functioning of English reading comprehension in a second language context when fit to the Rasch model is not ideal
Alan Tennant - Internal construct validity of the Wimbledon self report scale in a subarachnoid hemorrhage population
Jean-Louis Thonnard - Manual ability unbiased by diagnosis: dream or reality?
Emese Verdes - Health state estimation in WHO's multi-country survey studies
Russell Waugh - Prestest/posttest, control/experimental group Rasch measures of attitude and behavior to physics at year 9 level
W. Denny Way - A multi-stage adaptive testing model for PISA
Edward W. Wolfe - Comparison of confirmatory factor analyses via ConQuest and Mplus
Edward W. Wolfe - Application of the Rasch model to measuring the performance of cognitive radios
Edward W. Wolfe - Rater effect comparability in local independence and rater bundle models

AERA 2012 Rasch-related Papers

Vancouver, British Columbia, Canada
Friday, April 13 - Tuesday, April 17, 2012

Friday, April 13

12:00 p.m. - 1:30 p.m. Marriott Pinnacle, Floor Third Level - Pinnacle I
Division D - Measurement and Research Methodology.
Section 1: Educational Measurement, Psychometrics, and Assessment
Rasch Measurement Models and the Advanced Placement Program Examinations

Rating Quality Studies Using Rasch Measurement Theory. *George Engelhard and Stefanie Anne Wind (Emory University)*

Comparative Analyses of Generalizability Theory and the Many-Facet Rasch Model. *Amy B. Hendrickson (The College Board), George Engelhard (Emory University)*

Hierarchical Rasch Models for Rater-Mediated Assessments. *George Engelhard and Yuk F. Cheong (Emory University)*

Using the Many-Facet Rasch Model to Inform Standard-Setting Procedures: Setting performance standards for Advanced Placement examinations. *Pamela K. Kaliski (The College Board), George Engelhard (Emory University), Deanna Lynn Morgan and Rosemary A. Reshetar (The College Board), Barbara S. Plake (University of Nebraska - Lincoln)*

12:00 p.m. - 1:30 p.m. Vancouver Convention Center, Floor Second Level - East Room 12
SIG-Second Language Research: **Measurement in the Second Language Classroom**

Assessing Learning Outcomes in Short-Term Foreign Language Programs: Validation Results of a Triangulated Assessment System. *Megan Masters (University of Maryland), Steven J. Ross (University of Maryland)*

2:15 p.m. - 3:45 p.m. Marriott Pinnacle, Floor Third Level - Dundarave
SIG-Rasch Measurement: Studies in Rasch Conditions and Applications
Chair: *Shungwon Ro (Kenexa)*. Discussant: *Nathaniel J.S. Brown (Indiana University - Bloomington)*

Rasch Analysis of the Outcome Questionnaire with African Americans. *Ruth C.L. Chao (University of Denver), Kathy E. Green (University of Denver)*

Differential Item and Person Functioning in Large-Scale Writing Assessments Within the Context of the SAT Reasoning Test. *George Engelhard (Emory University), Stefanie Anne Wind (Emory University), Jennifer L. Kobrin (The College Board), Michael Chajewski (The College Board)*

A Study of Rasch, Partial Credit, and Rating Scale Model Parameter Recovery in WINSTEPS and jMetrik. *Patrick Meyer and Emily Hailey (University of Virginia)*

Measuring Student Perceptions of Adult Influences on Their Classroom Learning. *Robert Frederick Cavanagh (Curtin University), Graham B. Dellar (Curtin University)*

2:15 p.m. - 3:45 p.m. Sheraton Wall Centre, Floor Grand Ballroom Level - North Grand Ballroom B
SIG-Research in Mathematics Education
Harnessing Psychometric Models to Develop Next-Generation, Research-Based Assessments of Rational Number Knowledge

Testing the Reorganization of the Equipartitioning Learning Trajectory Using Rasch Item Response Theory Modeling. *Kenny Huy Nguyen (North Carolina State University), Andre A. Rupp (University of Maryland), Jere Confrey (North Carolina State University), Alan Maloney (North Carolina State University)*

Apr 13 - 6:15 p.m. - 7:45 p.m. Marriott Pinnacle, Floor Third Level - Dundarave
Rasch Measurement SIG Business Meeting
SIG Chair: *Michael Young;* Secretary: *Kenneth Royal*
Program Chairs: *Daeryong Seo & Stephen Jirka*

Invited speaker: *John H. A. L. de Jong (Pearson)*

IOMW 2012: "Diversity and Inclusion"

Vancouver, BC, Canada

April 11-12, 2012

at Vancouver Public Library, 8:00 a.m.

This *International Objective Measurement Workshop* includes a wide range of presentations in health and education from around the globe. We have a strong graduate student contingent with poster and oral presentations along with a prize for best student submission. We are fortunate to have *Jean Guy Blais* from the *Universite de Montreal* as our closing keynote speaker. Biographical information and the abstract of the keynote presentation can be found on the Program page of the Conference website: **www.iomw2012.com**

An IOMW hosted reception at the end of Day One will provide an opportunity for networking and discussion.

For Conference registration (early-bird rates until March 15) and special accommodation rates at the Georgian Court Hotel , see the Conference website:
www.iomw2012.com

We look forward to seeing you in Vancouver.

Peter MacMillan, Lois Lochhead and Stefanie Sebok
Conference Organizing Committee

Saturday, April 14

10:35 a.m. - 12:05 p.m. Vancouver Convention Center, Floor Second Level - East Room 2&3
SIG-Rasch Measurement
Roundtable Session 26. Chair: *Kwang-Lee Chu (Pearson)*

Effect of Missing Data in Computerized Adaptive Testing on Accuracy of Item Parameter Estimation: A Comparison of NWEA (Northwest Evaluation Association) and WINSTEPS Item Parameter Calibration Procedures. *Shudong Wang (Northwest Evaluation Association), Gregg Harris (Northwest Evaluation Association)*

Using Rasch Measurement Theory to Validate the Student Performance Character and Student Moral Character Scales. *Jade Caines (University of Pennsylvania)*

The Development of the Teaching Economic Literacy: Confidence and Anxiety Scale. *Julia Rollison, Larry H. Ludlow (Boston College)*

The Effects of the Sample selection on Item Parameter Estimation. *Lixiong Gu (ETS), Venessa F. Lall (ETS), Maxwell D. Wise (ETS)*

Cognitive Diagnostic Assessment of TIMSS (Trends in International Mathematics and Science Study) 2007 Mathematics Achievement Items for Eighth Graders in Turkey. *Turker Toker (University of Denver), Kathy E. Green (University of Denver)*

12:25 p.m. - 1:55 p.m. Sheraton Wall Centre, Floor Third Level - South Azure
Division C - Learning and Instruction. Section 4: Science
Assessments Serving Science Learning and Instruction

Measuring Student Perceptions of Constructivism within the Science Classroom: Development and Application of the Elementary School Science Classroom Environment Scale. *Laura M. O'Dwyer (Boston College), Shelagh M. Peoples (Boston College), Yang Wang (Boston College), Katherine Shields (Boston College)*

12:25 p.m. - 1:55 p.m. Vancouver Convention Center, Floor First Level - East Ballroom B
Division H - Research, Evaluation and Assessment in Schools. Section 3: Assessment in the Schools
Assessment in the Schools Poster Session 2

Model Competence: A Valid Learning Progression for Biology Lessons. *Dirk Krueger (Freie Universität Berlin), Annette (Upmeier zu Belzen, Humboldt University – Berlin)*

2:15 p.m. - 3:45 p.m. Marriott Pinnacle, Floor Third Level - Dundarave
SIG-Rasch Measurement: Issues of Rasch Dimensionality, Scaling, and Fit
Chair: *Mary Garner (Kennesaw State University)*. Discussant: *Lihshing Leigh Wang (University of Cincinnati)*

Assessing the Effects of Different Item Parameter Profiles in Mixture Rasch Models. *Youngmi Cho (University of Maryland), Hong Jiao (University of Maryland), George B. Macready (University of Maryland)*

Comparing Panel Designs With Routing Methods in the Multistage Test With the Partial Credit Model. *Jiseon Kim (University of Washington - Seattle), Hyewon Chung (John Jay College of Criminal Justice - CUNY), Ryoungsun Park (The University of Texas - Austin), Barbara G. Dodd (The University of Texas - Austin)*

Comparison of Priors in Bayesian Estimation of 1-PL (One-Parameter Logistic) Item Response Models. *Prathiba Natesan (University of North Texas), Ratna Nandakumar (University of Delaware), Tom Minka (Microsoft Research), Xiaoyu Qian, Jonathan D. Rubright (University of Delaware)*

The Distribution of Between-Dimension Correlation in Mis-specified Multidimensional Rasch Models in Unidimensional Data. *Leigh M. Harrell-Williams (Virginia Polytechnic Institute and State University)*

4:05 p.m. - 5:35 p.m. Marriott Pinnacle, Floor Third Level - Pinnacle I
Division D - Measurement and Research Methodology.
Section 1: Educational Measurement, Psychometrics, and Assessment
Assessment of Special Populations

Scale Comparability for Accommodated Forms in the Rasch Model: A Person-Fit Approach. *Dong Gi Seo (Michigan Department of Education), Shiqi Hao (Michigan Department of Education), Steven Guy Viger (Michigan State University)*

4:05 p.m. - 6:05 p.m. Marriott Pinnacle, Floor Third Level - Dundarave
Division D - Measurement and Research Methodology
Section 1: Educational Measurement, Psychometrics, and Assessment
Validity Investigations

Multidimensional Rasch Model for Analysis of Growth in Career Maturity. *Hyo Jeong Shin (University of California – Berkeley)*

Sunday, April 15

8:15 a.m. - 9:45am. Marriott Pinnacle, Floor Third Level - Pinnacle II
Division D - Measurement and Research Methodology
Section 1: Educational Measurement, Psychometrics, and Assessment
Assessments in International Settings

Validation of Creative Achievement Questionnaire: Through A Rasch Perspective. *Chia-chi Wang (National Sun Yat-Sen University), Hsiao-Chi Ho (National Sun Yat-Sen University), Chih-Ling Cheng (National Sun Yat-Sen University), Ying-Yao Cheng (National Sun Yat-Sen University), Chih-Wen Kuo (Institute of Education National Sun Yat-sen University)*

10:35 a.m. - 12:05 p.m. Marriott Pinnacle, Floor Third Level - Shaughnessy I
SIG-Academic Audit Research in Teacher Education
Assessment and Accreditation: How Do Instruments and Procedures Relate to Policy and Performance Indicators?

Scale Functioning and Licensure Invariance of the Student Teaching Exit Survey: A Rasch Analysis. *Noela A. Haughton (University of Toledo), Peter Paprzycki (University of Toledo)*

12:25 p.m. - 1:55 p.m. Pan Pacific, Floor Lobby Level - Oceanview 1&2
Division C - Learning and Instruction. Section 7: Technology Research
New Measurement Paradigms: Psychometric Methods for Technology-Based Assessments

From Rasch Models to Rule Space and Poset-Based Adaptive Testing. *Douglas H. Clements (University at Buffalo – SUNY), Curtis Tatsuoka (Case Western Reserve University), Kikum Tatsuoka (Teachers College, Columbia University)*

12:25 p.m. - 1:55 p.m. Vancouver Convention Center, Floor First Level - East Ballroom B
Division D - Measurement and Research Methodology
Section 1: Educational Measurement, Psychometrics, and Assessment
Diverse Topics in Psychometrics and Educational Measurement

Sensitivity of Anchor Designs on Scaling and Proficiency Classifications in the Rasch Model. *Thakur B. Karkee, (Measurement Incorporated), Winnie K. Reid (Measurement Incorporated)*

12:25 p.m. - 1:55 p.m. Vancouver Convention Center, Floor Second Level - East Room 2&3
Division G - Social Context of Education. Section 4: Social Context of Educational Policy, Politics, and Praxis
Higher Education, Diversity, and Equity in Critical Perspective

Exploring the Black, White, and Gray Areas of Faculty Perceptions of Inclusiveness. *Kelly D. Bradley (University of Kentucky), Sonja Feist-Price (University of Kentucky), Nancye E. McCrary (University of Kentucky), Jessica D. Cunningham (Western Carolina University)*

2:15 p.m. - 3:45 p.m. Pan Pacific, Floor Restaurant Level - Pacific Rim Suite 2
Division I - Education in the Professions
Measuring the Noncognitive Traits of Students in the Professions

Many-Facet Rasch Analysis of Standardized Patient Ratings of Students' Humanistic Competence on a Medical Licensure Examination. *Xiuyuan Zhang (National Board of Osteopathic Medical Examiners), William L. Roberts (National Board of Osteopathic Medical Examiners)*

2:15 p.m. - 3:45 p.m. Sheraton Wall Centre, Floor Third Level - South Pavilion Ballroom C
Division C - Learning and Instruction. Section 6a: Cognitive, Social, and Motivational Processes
Evaluating the Psychometric Quality of Self-Efficacy Measures With Diverse Item-Analysis Methods

Using the Many Facet Rasch Model to Evaluate the Psychometric Quality of Teacher Sense of Efficacy Scale. *Mei-Lin Chang (Emory University), George Engelhard (Emory University)*

Monday, April 16

10:35 a.m. - 12:05 p.m. Marriott Pinnacle, Floor Third Level - Shaughnessy I
SIG-Survey Research in Education
Measurement Issues in Survey Research

An Empirical Study of Response Category Effects: A Rasch Rating Scale Analysis. *Zongmin Kang (DePaul University)*

Survey Analysis With Mixture Rasch Models. *John T. Willse (University of North Carolina at Greensboro), Andrew Dallas (University of North Carolina – Greensboro)*

With Hiccups and Bumps: An Innovative Measure of Student Understanding of the Nature of Science. *Shelagh M. Peoples (Boston College), Katherine Shields (Boston College), Laura M. O'Dwyer (Boston College), Yang Wang (Boston College)*

Tuesday, April 17

12:25 p.m. - 1:55 p.m. Vancouver Convention Center, Floor First Level - East Ballroom C
Division D - Measurement and Research Methodology
Section 1: Educational Measurement, Psychometrics, and Assessment
Estimation Issues in Item Response Theory

Comparison of Four Maximum-Likelihood Methods in Estimating the Rasch Model. *Tianshu Pan (Pearson)*

12:25 p.m. - 1:55 p.m. Vancouver Convention Center, Floor First Level - East Ballroom C
Division D - Measurement and Research Methodology
Section 1: Educational Measurement, Psychometrics, and Assessment
Validation of Scales

New Evidence on the Validity of the Classroom Assessment Scoring System. *Nicole Makas Colwell (University of Illinois at Chicago)*

12:25 p.m. - 1:55 p.m. Building/Room: Vancouver Convention Center, Floor First Level - East Ballroom C
Division D - Measurement and Research Methodology
Section 1: Educational Measurement, Psychometrics, and Assessment
Discussions in Item Response Theory

A New Tool for Fitting Polytomous Item Response Theory Models. *Zhushan Mandy Li (Boston College)*

Book: Advances in Rasch Measurement, Volume 2

Edited by Nathaniel J. S. Brown, Brent Duckor, Karen Draney, and Mark Wilson, 2011, www.jampress.org

1. Bringing Human, Social, and Natural Capital to Life: Practical Consequences and Opportunities, William P. Fisher
2. From Model to Measurement with Dichotomous Items, Don Burdick, A. Jackson Stenner, and Andrew Kyngdon
3. Measuring Measuring: Toward a Theory of Proficiency with the Constructing Measures Framework, Brent Duckor, Karen Draney, and Mark Wilson
4. Predicting Responses from Rasch Measures, John M. Linacre
5. Random Parameter Structure and the Testlet Model: Extension of the Rasch Testlet Model, Insu Paek, Haniza Yon, Mark Wilson, and Taehoon Kang
6. Estimating Tests Including Subtests, Steffen Brandt
7. The Construction and Implementation of User-Defined Fit tests for Use with Marginal Maximum Likelihood Estimation and Generalised Item Response Models, Raymond J. Adams and Margaret L. Wu
8. The Efficacy of Link Items in the Construction of a Numeracy Achievement Scale from Kindergarten to Year 6, Juho Looveer and Joanne Mulligan
9. Rasch Model's Contribution to the Study of Items and Item Response Scales Formulation in Opinion/Perception Questionnaires, Jean-Guy Blais, Julie Grondin, Nathalie Loye, and Gilles Raîche
10. On the Factor Structure of Standardized Educational Achievement Tests, Tim W. Gaffey, Robert Cudeck, Emilio Ferrer, and Keith F. Widaman
11. Optimizing the Compatibility between Rating Scales and Measures of Productive Second Language Competence, Christopher Weaver
12. Assessment of English Language Development: A Validity Study of a District Initiative, Juan D. Sanchez
13. Using FACETS to Inform Decisions on Staff Development and Remuneration: A Case Study of Student Rating of Teaching Effectiveness Survey, Nuraihan Mat Daud and Noor Lide Abu Kassim
14. Using Guttman's Mapping Sentences and Many Facet Rasch Measurement Theory to Develop an Instrument that Examines the Grading Philosophies of Teachers, Jennifer Randall and George Engelhard, Jr.
15. Measure for Measure: Curriculum Requirements and Children's Achievement in Music Education, Trevor Bond and Marie Bond
16. Development of a Multidimensional Measure of Academic Engagement, Kyra Caspary and Maria Veronica Santelices
17. Rasch Family Models in e-Learning: Analyzing Architectural Sketching with a Digital Pen, Kathleen Scalise, Nancy Yen-wen Cheng and Nargas Oskui
18. Using Item Response Modeling Methods to Test Theory Related to Human Performance, Diane D. Allen
19. Sources of Self-efficacy Belief: Development and Validation of Two Scales, Ou Lydia Liu and Mark Wilson

The Root Mean Square Error of Approximation (RMSEA)

as a supplementary statistic to determine fit to the Rasch model with large sample sizes

Georg Rasch mentioned chi-square statistics as a way of evaluating fit of data to the model (Rasch, 1980, p. 25). Ben Wright's Infit and Outfit mean-square statistics are the chi-square divided by their degrees of freedom. However, large sample sizes have always posed problems for significance tests based on chi-square statistics. The issue is that, the larger the sample, the greater the power, and so ever smaller differences are reported as indicating statistically significant misfit between the data and the model. Thus very large sample sizes can detect miniscule differences, and with such samples there is almost no need to undertake a chi-square test as we know that it will be significant (P. Martin-Löf (1974). Indeed, Georg Rasch himself remarked: "On the whole we should not overlook that since a model is never true, but only more or less adequate, deficiencies are bound to show, given sufficient data" (Rasch, 1980, p. 92).

Smith et al. (1998) show that the critical interval values for a Type I error (rejection of a true hypothesis) associated with these statistics varies with sample size. Experience indicates that, while the value of mean-square tends to increase only slowly with sample size, the critical interval associated with a 5% significance level shrinks considerably as sample size increases. Thus a sample of 50 would have a 5% range for Infit of 0.72-1.28, whereas a sample of 500 would have a 5% range of 0.91-1.09. A sample size of 5000 would have a 5% range of 0.97-1.03 (RMT 17:1 p. 918).

In general, large sample sizes will cause most chi-square-based statistics to almost always report a statistically significant difference between the observed data and model expectations, suggesting misfit, regardless of the true situation.

One potential mechanism for accommodating large sample sizes may be to use the Root Mean Square Error of Approximation (RMSEA, Steiger and Lind, 1980) as a supplementary fit. The RMSEA is widely used in Structural Equation Modeling to provide a mechanism for adjusting for sample size where chi-square statistics are used.

Consequently, we set out to test the potential of the RMSEA to supplement the chi-square fit tests reported for Rasch analyses performed by RUMM2030. This investigation focuses on the "summary fit chi-square" (the item trait interaction statistic). The utility of the RMSEA to supplement the interpretation of the chi square fit in larger samples was assessed, along with determination of the level of RMSEA that is consistent with fit to the Rasch model.

Methods.

A number of simulations were undertaken with the RUMMss simulation package (Marais I, Andrich D, 2007). Two polytomous item sets of 10 and 20 items with five response categories were simulated with different degrees of fit to the Rasch model. In addition, a set of dichotomous (30) items were also simulated. Perfect fit (100% of the items with simulated discriminations of 1.0), minor deviations (90% with 1.0, 10% with 3.0) and more serious deviations from model expectations (80% with 1.0, 20% with 3.0) were simulated. Each set of simulations was repeated for 200, 500, 2000, 5000, and 10,000 cases. All other parameters were held constant.

Table 1. RMSEA Results for Set 1
(10 polytomous items)

Sample Size	No Misfit	10% Misfit	20% Misfit
200	0.000	*0.000*	0.033
500	0.004	0.024	0.035
2000	0.011	0.024	0.030
5000	0.014	0.024	0.031
10000	0.014	0.024	0.031

Table 2. RMSEA Results for Set 2
(20 polytomous items)

Sample Size	No Misfit	10% Misfit	20% Misfit
200	0.000	*0.053*	0.043
500	0.000	0.024	0.040
2000	0.004	0.031	0.038
5000	0.006	0.030	0.038
10000	0.009	0.031	0.038

Table 3. RMSEA Results for Set 3
(20 dichotomous items)

Sample Size	No Misfit	10% Misfit	20% Misfit
200	0.000	*0.061*	*0.073*
500	0.016	0.019	0.035
2000	0.013	0.026	0.040
5000	0.011	0.027	0.040
10000	0.012	0.027	0.041

The RMSEA was calculated for each simulation, based upon the summary chi-square interaction statistic reported by RUMM2030. The RMSEA formulae can be shown to be equal to:

$$RMSEA = \sqrt{} \max([((\chi^2/df) - 1)/(N - 1)] , 0)$$

where χ^2 is the RUMM2030 chi-square value, df is its degrees of freedom and N is the sample size. Notice that the RMSEA has an expected value of zero when the data fit the model. Overfit of the data to the model, $\chi^2/df < 1$, is ignored. For a given χ^2/df, RMSEA decreases as sample size, N, increases.

Results

In Tables 1-3, the average RMSEA for each simulated condition is reported. Within each column of each Table,

the RMSEA is largely invariant as the sample size increases, as we had hoped.

Across each row of each Table, for sample sizes of 500 or more, the RMSEA is sensitive to increasing misfit. Thus it may be appropriate to use this supplementary fit statistic in the presence of sample sizes of 500 or more cases, to inform if sample size is inflating the chi-square statistic, and hence its significance.

Conclusion

The results of this study suggest that investigations of fit to the Rasch model using RUMM2030 and specifically the item-trait interaction chi-square fit statistic, in the presence of large sample sizes, can be supplemented through applying the RMSEA statistic. RMSEA values of < 0.2 with sample sizes of 500+, and certainly 1000+, may indicate that the data do not underfit the model, and that the chi-square was inflated by sample size.

Alan Tennant, Department of Rehabilitation Medicine, Faculty of Medicine and Health, The University of Leeds, UK

Julie F. Pallant, Rural Health Academic Centre, University of Melbourne, Australia.

References

Marais I, Andrich D (2007)\: RUMMss. Rasch Unidimensional Measurement Models Simulation Studies Software. The University of Western Australia, Perth.

Martin-Löf P. (1974). The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and observational data. Scandinavian Journal of Statistics, 1:3.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Smith, R. M, Schumacker RE, Bush MJ. (1998). Using item mean squares to evaluate fit to the Rasch model. Journal of Outcome Measurement, 2: 66-78.

Steiger, J. H. and Lind, J. (1980), "Statistically-based tests for the number of common factors," Paper presented at the Annual Spring Meeting of the Psychometric Society, Iowa City.

2011 IMEKO Conference Proceedings Available Online

Papers from the Joint International IMEKO TC1+TC7+TC13 Symposium held August 31st to September 2nd, 2011, in Jena, Germany are available online at www.db-thueringen.de/servlets/DerivateServlet/Derivate-24575/IMEKO2011_TOC.pdf.

Rasch-related papers:

A Clinical Scale for Measuring Functional Caregiving of Children Assisted with Medical Technologies. *Nikolaus Bezruczko, Shu-Pi C. Chen, Connie Hill, Joyce M. Chesniak*

A Technology Roadmap for Intangible Assets Metrology. *William .P Fisher, Jr., A. Jackson Stenner.*

Body, Mind, and Spirit are Instrumental to Functional Health: A Case Study. *Carl V. Granger, Nikolaus Bezruczko.*

Continuous Quantity and Unit; Their Centrality to Measurement. *Gordon A. Cooper, William P. Fisher, Jr.*

Foundational Imperatives for Measurement with Mathematical Models. *Nikolaus Bezruczko.*

From Breast-Q © to Q-Score ©: Using Rasch Measurement to Better Capture Breast Surgery Outcomes. *Stefan Cano, Anne F. Klassen, Andrea L. Pusic, Andrea*

How to Model and Test for the Mechanisms that Make Measurement Systems Tick. *A. Jackson Stenner, Mark Stone, Donald Burdick.*

Measurement, Metrology and the Coordination of Sociotechnical Networks. *William P. Fisher, Jr.*

The Quantification of Latent Variables in the Social Sciences: Requirements for Scientific Measurement and Shortcomings of Current Procedures. *Thomas Salzberger.*

The Role of Mathematical Models in Measurement: A Perspective from Psychometrics. *Mark Wilson.*

Also of interest are papers on the Fundamentals of Measurement Science and Mathematical Models in Measurement:

Application of Mathematical Models in Optical Coordinate Metrology. *Susanne C.N. Töpfer.*

Application-Oriented Approach to Mathematical Modelling of Measurement Processes. *Roman Z. Morawski.*

From Verbal Models to Mathematical Models – A Didactical Concept not just in Metrology. *Karl H. Ruhm*

Measurement Modelling: Foundations and Probabilistic Approach. *Giovanni Battista Rossi.*

Quantity and Quantity Value. *Alessandro Giordani, Luca Mari.*

The Role of Mathematical Modelling in the Analysis and Design of Measurement Systems. *Sanowar H. Khan, Ludwik Finkelstein.*

Uncertainty in Fuzzy Scales Based Measurements. *Eric Benoit.*

William P. Fisher, Jr.

Rasch Measurement Transactions
www.rasch.org/rmt
Editor: John Michael Linacre

SIG Chair: Michael Young
Secretary: Kenneth Royal
Program Chairs: Daeryong Seo & Stephen Jirka
SIG website: www.raschsig.org

Benjamin D. Wright's Annotated KeyMath Diagnostic Profile

Among Ben's favorite teaching aids is the KeyMath Diagnostic Profile, designed by Richard Woodcock at an AERA Presession on the Rasch Model in Spring 1969. Here we can see its essential features. An arithmetic test of 209 items has been Rasch-calibrated in one analysis. The items are displayed in 14 content strands (rows) with each item at its own difficulty in the overall frame-of-reference (circles above the lines). The raw scores on each strand have been positioned at their ability measures (numbers beneath the lines). Raw scores on all 209 items are shown on the "TOTAL TEST" line, positioned at their ability estimates. Ben has ringed the raw scores of a child on each strand, and some of the child's scored responses. We see a profile of the child's performance across the strands, along with unexpected successes and failures.