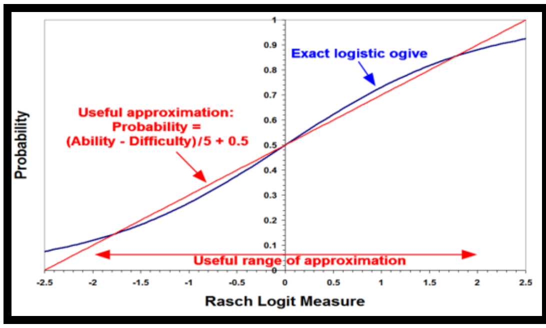


RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 24 No. 4 | Spring 2011 | ISSN 1051-0796

Re-Parameterization of the Partial Credit Model for Estimating Items with Large Values of Maximum Marks

The conventional formulation of Partial Credit Model (PCM) is as follows:

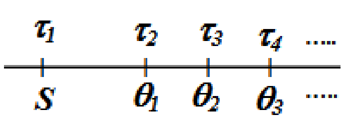
$$P_k(\lambda) = \frac{\exp\{k\lambda - \sum_{j=1}^{k} \tau_j\}}{1 + \sum_{l=1}^{m} \exp\{l\lambda - \sum_{j=1}^{l} \tau_j\}}$$

where $P_k(\lambda)$ is the probability of a student with ability λ obtaining the score k on an item with minimum mark equal to 0 and maximum mark equal to m, and $\{\tau_j\}$ are the non-centralized thresholds (= centralized threshold + item difficulty).

It can be observed that when the maximum mark of an item equals to m, the number of parameters that need to be estimated is also exactly equal to m. When m is very large (e.g., 20 or 30) which may not be uncommon for a non-multiple-choice item, the number of parameters subject to estimation may cause great problems or even a breakdown of commonly available software. In this brief note, a re-parameterization of PCM is proposed in order to cater for items with large values of maximum marks.

Re-parameterization of PCM

We re-formulate PCM using 4 parameters (S, θ, d, c), at most, for items with any values of maximum marks. The idea of the re-formulation is shown in the following diagram:



S: The start point
θ: The first interval (i.e., $\theta_1 = \theta$ and $\theta_{i+1} = \theta_i + d_i$ where $i \geq 1$)
d: The change applied to the first interval to obtain the second one (i.e., $d_1 = d$)
c: The variation of the change as compared with only the previous change (i.e. $d_{i+1} = d_i + c$ where $i \geq 1$)

Note that an approximation (i.e., assuming that the same c is applied for different d_i) is adopted here. By using the approximation, 4 parameters are enough to generate all the thresholds. The derivations of some thresholds are shown below:

$\tau_1 = S$
$\tau_2 = S + \theta_1 = S + \theta$
$\tau_3 = S + \theta_1 + \theta_2 = (S + \theta) + (\theta_1 + d_1) = S + 2\theta + d$
$\tau_4 = S + \theta_1 + \theta_2 + \theta_3 = (S + 2\theta + d) + (\theta_2 + d_2) = (S + 2\theta + d) + (\theta_1 + d_1 + d_1 + c)$
$= (S + 2\theta + d) + (\theta + 2d + c) = S + 3\theta + 3d + c$

It can be shown (by the use of mathematical induction) that the general form is:

$$\tau_j = S + (j-1)\theta + (j-1)(j-2)\, d/2 + (j-1)(j-2)(j-3)c/6$$

Based on the general form, we can immediately derive the following:

$$\sum_{j=1}^{k} \tau_j = kS + k(k-1)\theta/2 + k(k-1)(k-2)d/6 + k(k-1)(k-2)(k-3)c/24$$

Therefore PCM can now be re-formulated using only the 4 parameters: S, θ, d, c.

Parameter Estimation using WinBugs

The estimation of the re-formulated parameters can be achieved directly using the freeware WinBugs. As an illustration, the relevant code to set up the probability model for an item with maximum mark equal to 18 in WinBugs is shown in Figure 2.

In addition, non-informative prior distributions are set up for the parameters concerned. The corresponding WinBugs code is shown in Figure 3.

Table of Contents

Question	S	θ	d	c	Infit Mean-square	Outfit Mean-square
Short Q	-1.499	0.0347	0.0160	0.0001	0.96	1.00
Long Q	-1.434	0.0988	0.0028	0.0003	0.76	0.80

Figure 1. WinBugs parameter estimates and fit statistics.

```
for (i in 1:N) { # N Students
num.p1[i, 1]<-1
for (j in 1:18){ # An item with max mark = 18
fac11[i,j] <-j*base[1]                        # base[1]:S parameter
fac12[i,j]<-j*(j-1)*int[1]/2                  # int[1]:θ parameter
fac13[i,j]<-j*(j-1)*(j-2)*dev[1]/6           # dev[1]:d parameter
fac14[i,j]<-j*(j-1)*(j-2)*(j-3)*chg.dev[1]/24 # chg.dev[1]:c parameter
num.p1[i, j+1] <- exp(j*lambda[i] - (fac11[i,j] +fac12[i,j] +
                  fac13[i,j] + fac14[i,j] ))}
den.p1[i] <- sum(num.p1[i,  1:19] )
for (j in 1:19){p1[i,j] <- num.p1[i,j]/ den.p1[i] } } # normalization
for (i in 1:N)
{r[i, 1] ~ dcat(p1[i, 1:19 ])}# define a categorical distribution
                       # r[i,1] is the response of student i to the item
```

Figure 2. WinBugs code for reparameterized PCM with categories 0-18.

```
for (i in 1:N){lambda[i] ~ dnorm(0,
tau.lambda)}
for (j in 1:T){
  base[t] ~ dnorm (mu.Base, tau.Base)
  int[t] ~dnorm(mu.Int, tau.Int )I(0,)
  dev[t]~dnorm(mu.Dev,tau.Dev)
  chg.dev[t]~dnorm(mu.Chg, tau.Chg)
}
   tau.lambda ~ dgamma(0.001, 0.001)
   mu.Base ~ dnorm(0, 1.E-6)
   tau.Base ~ dgamma(0.001, 0.001)
   mu.Int ~ dnorm(0, 1.E-6)
   tau.Int ~ dgamma(0.001, 0.001)
   mu.Dev ~ dnorm(0, 1.E-6)
   tau.Dev ~ dgamma(0.001, 0.001)
   mu.Chg ~ dnorm(0, 1.E-6)
   tau.Chg ~ dgamma(0.001, 0.001)
```

Figure 3. WinBugs code for non-informative priors.

Then the parameter estimation can be conducted using the built-in Markov Chain Monte Carlo (MCMC) method in WinBugs. The student ability and item parameters (from which all τ_j are derived) can be obtained.

We have applied this novel formulation of PCM to model responses of two items (one is Short Q and the other is Long Q) in the trial run of a test attempted by some 200 students. Short Q has maximum mark equal to 18 and Long Q has maximum mark equal to 20. The outcomes shown in Figure 1 are satisfactory.

All the values of the standard PCM thresholds τ_j derived from these re-formulated parameters (S, θ, d, c) are provided in Figure 4; together with frequency counts for different response categories.

Response Category	Frequency Count for Short Q	Frequency Count for Long Q	Derived PCM Threshold	Short Q	Long Q
0	0	0	-	-	-
1	1	3	**τ1**	-1.50	-1.43
2	1	1	**τ2**	-1.46	-1.34
3	1	3	**τ3**	-1.41	-1.23
4	0	4	**τ4**	-1.35	-1.13
5	5	3	**τ5**	-1.26	-1.02
6	5	5	**τ6**	-1.16	-0.91
7	5	9	**τ7**	-1.05	-0.79
8	9	15	**τ8**	-0.92	-0.67
9	15	11	**τ9**	-0.77	-0.55
10	17	22	**τ10**	-0.60	-0.42
11	24	19	**τ11**	-0.42	-0.29
12	40	29	**τ12**	-0.22	-0.15
13	26	19	**τ13**	0.00	0.00
14	27	18	**τ14**	0.23	0.15
15	20	17	**τ15**	0.48	0.31
16	25	17	**τ16**	0.75	0.47
17	5	12	**τ17**	1.03	0.65
18	3	16	**τ18**	1.34	0.82
19	-	3	**τ19**	-	1.01
20	-	3	**τ20**	-	1.20
Total	229	229	**Average**	-0.46	-0.27

Figure 4. PCM thresholds for 19 categories (Short Q) and 21 categories (Long Q).
Note that category 0 is not observed, and nor is category 4 for Short Q. The reparameterized estimation is robust against unobserved categories.

Summary

In this brief note, we have proposed a novel re-parameterization for PCM in order to handle items with large values of maximum marks. The parameter estimation could be conducted using the freeware, WinBugs. We have applied the re-parameterization to model item responses in some real-life data. The outcomes of the estimations are satisfactory.

Dr. Fung Tze-ho
Manager-Assessment Technology & Research,
Hong Kong Examinations and Assessment Authority

Guttman Parameterization of Rating Scales - Revisited

"A reparameterized form of thresholds into their principal components is the method of estimation operationalized in RUMM2030. This notion of principal components is used in the sense of Guttman (1950), who rearranged ordered categories into successive principal components, beginning with the usual linear one. They are analogous to the use of orthogonal polynomials in regression where the independent variable is ordered. The term does NOT refer to the common "principal components analysis" in which a matrix of correlation coefficients is decomposed by analogy to factor analysis."

from www.rummlab.com.au, January, 2011.

As previously described in Guttman Parameterization of Rating Scales, RMT 17:3,2003, p. 944, Pender Pedler (1987, amended) constructs the Guttman decomposition of the $j = 1$ *to* m Rasch-Andrich thresholds of a rating scale with categories $0,m$. He defines a series of k = 1,K orthogonal polynomials in j,

$T_1(j) = 1$
$T_2(j) = 2 (j - (m+1)/2)$
$T_3(j) = 3 (j - (m+1)/2)^2 - (m^2 - 1)/4$
$T_4(j) = 4 (j - (m+1)/2)^3 - (j - (m+1)/2)(3m^2 - 7)/5$

In general, for polynomial $k+1$ of threshold j,

$T_{k+1}(j) = [(k+1)/k] (j - (m+1)/2)T_k(j) - ([(m^2 - (k-1)^2)(k^2 - 1)]/[4(2k - 1)(2k-3)])T_{k-1}(j)$

So, when $\{F_j\}$ are the Rasch-Andrich thresholds, and $\{c_k\}$ are the coefficients of the polynomials, estimated from the data by, say, Newton-Raphson iteration,

$F_j = \text{sum } (c_k T_k(j))$ for k = 1 to K

Note that there is no requirement that all the categories are observed in the data.

Andrich and Luo (2003) use cumulative thresholds, *kappa(x),* up to threshold x, so that
kappa(x) = - sum(F(j)) for j = 1 to x,
= sum ([sum ($c_k T_k(j)$) for k = 1 to K]) for j=1 to x
= sum ((c_k / A_k) $U_k(x)$) for k = 1 to K
where
$U_k(x) = A_k$.sum($T_k(j)$) for j = 1 to x, and A_k is a constant chosen for convenience. c_1/A_1 is termed the central location, $c_2/A_2 = \theta$ is the dispersion, $c_3/A_3 = \eta$ is the skewness, $c_4/A_4 = \zeta$ is the kurtosis.

Specifically,
$U_1(x) = -x$, with $A_1 = -1$.
$U_2(x) = x(m-x)$, with $A_2 = -1$
$U_3(x) = x(m-x)(2x-m)$ with $A_3 = -2$
$U_4(x) = x(m-x)(5x^2-5xm+m^2+1)$ with $A_4 = -5$

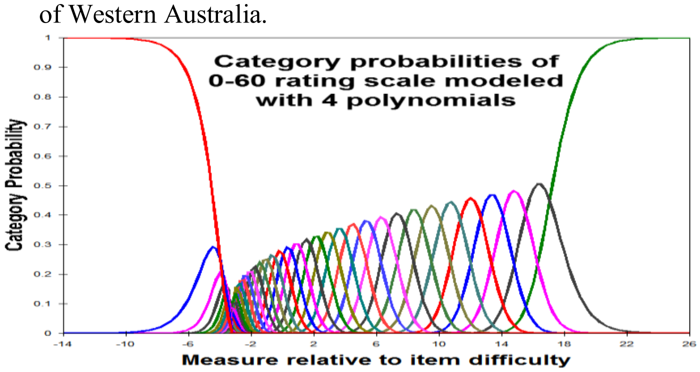
However, the utility of the orthogonal polynomials is that each higher polynomial adds to the lower ones. Accordingly, we can stop when we have estimated enough of the polynomials to give a useful definition of the threshold values. This is especially helpful when estimating long rating scales based on small datasets. The example in the Figures models the thresholds with four polynomials. It is based on ratings of Olympic Ice-Skating and is estimated by Winsteps.

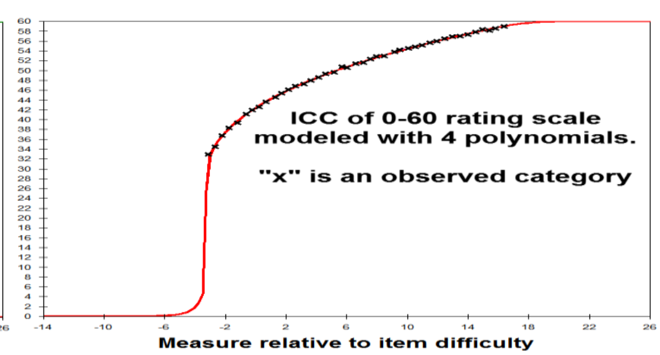
John M. Linacre

Andrich, D. & Luo, G. (2003). Conditional Pairwise Estimation in the Rasch Model for Ordered Response Categories using Principal Components. Journal of Applied Measurement, 4(3), 205-221.

Guttman, L. (1950). The principal components of scale analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star and J.A. Clausen (Eds.), Measurement and Prediction, pp. 312-361. New York: Wiley

Pedler, P.J. (1987) Accounting for psychometric dependence with a class of latent trait models. Ph.D. dissertation. University of Western Australia.





Rasch-related Papers at AERA 2011, New Orleans

Friday, April 8

Friday, April 8 - 12:00 p.m. - 1:30 p.m. Sheraton / Grand Ballroom A
Roundtable: Learning Progressions and Learning Trajectories
Division C - Learning and Instruction. Section 3: Mathematics

Evaluating Learning Progressions in Early Numeration and Computation Development During Elementary School. Joseph Betts (Renaissance Learning)

Friday, April 8 - 2:15 p.m. - 3:45 p.m. Doubletree / Shadows
Issues in Rasch Modeling
Division D - Measurement and Research Methodology. Section 1: Educational Measurement, Psychometrics, and Assessment

A Multilevel Rasch Mixture Testlet Model. Hong Jiao (University of Maryland), Matthias Von Davier (ETS), Akihito Kamata (University of Oregon), Ying-Fang Chen (University of Maryland - College Park)

A Rasch Model for Item Calibration Using Clustered Samples of Examinees. Yeow Meng Thum (Northwest Evaluation Association), Shudong Wang (Northwest Evaluation Association)

Confirmatory Mixture Rasch Models. John T. Willse (University of North Carolina - Greensboro)

Investigation of Precision in Rasch Difficulty Estimation. Mike McGill, Edward W. Wolfe (Pearson)

Random Item Rasch Models in Small-Scale Educational and Psychological Experiments. Feifei Ye (University of Pittsburgh), Qun Guan (University of Pittsburgh)

Friday, April 8 - 2:15 p.m. - 3:45 p.m. Sheraton / Grand Ballroom C
Poster: Strategic Recruitment in Teacher Education
Division K - Teaching and Teacher Education. Section 9: Teacher Education Program Design and Innovations

Admissions to Initial Teacher Education: The Role of Teacher Educators. Amanda K. Ferguson (OISE/University of Toronto)

Friday, April 8 - 4:05 p.m. - 5:35 p.m. Astor Crowne Plaza / Grand Ballroom A
Research on Linking the Moral, Social, and Political in Human Development
Division E - Counseling and Human Development. Section 2: Human Development

Rasch-Based Proficiency Levels as Mixture of Both Civic and Moral Knowledge and Thinking. Fritz K. Oser (University of Fribourg), Horst Biedermann (University of Freiburg)

Friday, April 8 - 4:05 p.m. - 5:35 p.m. Sheraton / Grand Ballroom C
Poster: Applied Research in Secondary Public Schools
Division H - Research, Evaluation and Assessment in Schools. Section 1: Applied Research in the Schools

Getting Kids to Understand Evolution: First-Year Implementation Results. Camelia V. Rosca (Boston College), Laura M. O'Dwyer (Boston College)

Friday, April 8 - 4:05 p.m. - 5:35 p.m. Sheraton / Grand Ballroom D
Roundtable: Examining Language Learning and Proficiency Evaluation Instruments
SIG-Second Language Research

Examination of the Psychometric Properties of a Self-Efficacy Scale. Chuang Wang (University of North Carolina - Charlotte), Do-Hong Kim (University of North Carolina - Charlotte)

Friday, April 8 - 4:05 p.m. - 5:35 p.m. Doubletree / Rosedown A
Assessment of Language and Reading
Division H - Research, Evaluation and Assessment in Schools. Section 3: Assessment in the Schools

Using the Rasch Model to Develop a Screening Measure for At-Risk and Advanced Beginning Readers to Enhance Response-to-Intervention Frameworks. Amy Weisenburgh - Snyder (University of Texas - Austin), Lynn Chen (University of Texas - Austin), Barbara G. Dodd (University of Texas - Austin)

Friday, April 8 - 4:05 p.m. - 5:35 p.m. Doubletree / International Ballroom
Roundtable: Building a Better Model for Testlet-Based Data
Division D - Measurement and Research Methodology. Section 1: Educational Measurement, Psychometrics, and Assessment

A General Framework for Dual Clustering in Item Response Theory Modeling. Hong Jiao (University of Maryland), Robert J. Mislevy (ETS)

An Item Response Model for Testlet-Based Rating Scale Items. Wen-Chung Wang (The Hong Kong Institute of Education), Xuelan Qiu (The Hong Kong Institute of Education)

Friday, April 8 - 4:05 p.m. - 5:35 p.m. Doubletree / International Ballroom
Roundtable: Improving Equating Results Under Less Than Optimal Conditions
Division D - Measurement and Research Methodology. Section 1: Educational Measurement, Psychometrics, and Assessment

Comparing Methods for Detecting Unstable Anchor Items With Net Differential Item Functioning and Global Differential Item Functioning Conceptions. Che-Ming Allen Lau (Pearson), Alvaro J. Arce (Pearson)

Does Removing Anchor Items Based on Statistical Criteria Impact Scale Stability and Student Performance? A Rasch Model Perspective. Thakur B. Karkee (Measurement Inc.), Winnie K. Reid (Measurement Incorporated), Daniel F. Bowen (Measurement Inc.)

Investigating the Effect of Differential Item Functioning (DIF) in Common-Item Nonequivalent Group Equating Design. Tian Song (Pearson)

Several Issues in Reducing Errors of Linking and Equating at All Ability Levels for State Large-Scale High-Stakes K-12 Assessments. Haiyan Lin (University of Illinois - Urbana-Champaign), Hua-Hua Chang (University of Illinois - Urbana-Champaign)

Friday, April 8 - 4:05 p.m. - 5:35 p.m. Sheraton / Grand Ballroom C
Poster Session: Effects of Instructional Format on Learning
Division C - Learning and Instruction. Section 6: Cognitive, Social, and Motivational Processes

Examining the Psychometric Properties of RAT-Chinese Version With Rasch Model. Su Pin Hung (National Taiwan Normal University), Po Seng HAUNG (National Taiwan Normal University), Hsueh-Chi chen (National Taiwan Normal University)

Friday, April 8 - 6:15 p.m. - 7:45 p.m. Doubletree / Rosedown B
Rasch Measurement SIG Business Meeting
SIG-Rasch Measurement

Formulating Latent Growth Using an Explanatory Item Response Model Approach. Mark Wilson, Xiaohui Zheng & Leah Walker McGuire

Saturday, April 9

Saturday, April 9 - 8:15 a.m. - 9:45 a.m. Doubletree / Rosedown B
Dimensionality and Model Fit With Item Response Theory
Division D - Measurement and Research Methodology. Section 1: Educational Measurement, Psychometrics, and Assessment

Dimensionality in Extended Constructed Response Items With Local Dependency. Yongsang Lee (University of California - Berkeley), Jinnie Choi (University of California - Berkeley), Karen L. Draney (University of California - Berkeley), Hyo Jeong Shin (University of California - Berkeley)

Saturday, April 9 - 10:35 a.m. - 12:05 p.m. Sheraton / Grand Ballroom C
Poster: Diverse Topics in Psychometrics and Educational Measurement
Division D - Measurement and Research Methodology. Section 1: Educational Measurement, Psychometrics, and Assessment

Examining Item-Position Effects Within Reading Items: The Linear Logistic Test Model (LLTM) Approach. Okan Bulut (University of Minnesota - Twin Cities)

Measuring Teacher Beliefs About Mathematics Discourse: An Item Response Theory Approach. Heeju Jang (University of California)

Saturday, April 9 - 10:35 a.m. - 12:05 p.m. New Orleans Marriott / Preservation Hall Studio 2
Diversity and Bias
SIG-Science Teaching and Learning

Is the Force Concept Inventory Biased? Investigating Differential Item Functioning on a Test of Conceptual Learning in Physics. Sharon E. Osborn Popp (Arizona State University), David Meltzer (Arizona State University), M. Colleen Megowan-Romanowicz (Arizona State University)

Saturday, April 9 - 2:15 p.m. - 3:45 p.m. Doubletree / Shadows
Issues of Rasch Dimensionality, Scaling, and Fit
SIG-Rasch Measurement
Chair: Shu-Ren Chang (American Dental Association)
Discussant: Matthias Von Davier (ETS)

A Comparison of Item Selection Procedures With Exposure Control Procedures Under Matched and Mismatched Conditions of Item Pool and Ability Distribution: Computerized Adaptive Testing With the Partial Credit Model. Hwa Young Lee (University of Texas - Austin), Barbara G. Dodd (University of Texas - Austin), Tsung-Han Ho (University of Texas - Austin)

A Comparison of Panel Designs in the Multistage Test Based on the Partial Credit Model. Jiseon Kim (University of Washington), Hyewon Chung (John Jay College of Criminal Justice - CUNY), Ryoungsun Park (University of Texas - Austin), Barbara G. Dodd (University of Texas - Austin)

Poor Targeting and CUTLO in Parameter Estimation. Qiong Fu (University of Illinois - Chicago), Everett V. Smith (University of Illinois - Chicago)

Rasch Analysis for the Evaluation of Rank of Student Response Times in Multiple Choice Examinations. James J. Thompson (Louisiana State University - Health Sciences Center), Tong Yang (Louisiana State University - Health Sciences Center), Sheila W. Chauvin (Louisiana State University - Health Sciences Center)

Saturday, April 9 - 4:05 p.m. - 5:35 p.m. Sheraton / Grand Ballroom C
Poster: Rasch SIG
SIG-Rasch Measurement

A Rasch Analysis of Self-Efficacy and Context Beliefs Among Urban Elementary Teachers. Jessica Gale (Emory University)

A Rasch Analysis of the Statistical Anxiety Rating Scale. Eric D. Teman (University of Northern Colorado)

Saturday, April 9 - 4:05 p.m. - 6:05 p.m. Doubletree / Madewood A
Assessment in International Contexts
Division D - Measurement and Research Methodology. Section 1: Educational Measurement, Psychometrics, and Assessment

Quantifying the Difficulty Difference Between Numerical Operations and Word Problem Items Using the Rasch Model. Markus Broer (American Institutes for Research)

Saturday, April 9 - 4:05 p.m. - 6:05 p.m. Hotel Monteleone / Iberville
Assessment to Support Instruction: Advances in Assessing Individual Differences in Reading Performance
Division C - Learning and Instruction. Section 1: Reading, Writing, and Language Arts

Item Response Theory Meets Cognitive Psychology: Analyzing Competencies for Text-Picture Integration From Multiple Perspectives. Wolfgang Schnotz (University of Koblenz-Landau), Holger Horz (University of Koblenz-Landau), Mark Daniel Ullrich (University of Koblenz-Landau), Nele McElvany (Technical University of Dortmund), Sascha Schroeder (Max Planck Institute for Human Development), Juergen Baumert (Max Planck Institute for Human Development)

Sunday, April 12

Sunday, April 10 - 8:00 a.m. - 12:00 p.m. Hotel Monteleone / Riverview
A Hands-on Introduction to Latent Class Models, Rasch Models, and Their Extensions
Professional Development and Training Committee

Director: Matthias Von Davier (ETS)

Sunday, April 10 - 8:15 a.m. - 9:45 a.m. Sheraton / Grand Ballroom C
Poster: College Student Learning and Development
Division J - Postsecondary Education. Section 1: College Student Learning and Development

Using Rasch Measurement to Measure Factors Affecting the Frequency of Academic Misconduct. Kenneth Royal (American Board of Family Medicine), Jennifer Ann Eli (The University of Arizona)

Sunday, April 10 - 10:35 a.m. - 12:05 p.m. Doubletree / Rosedown B
Studies in Rasch Conditions and Applications
SIG-Rasch Measurement
Chair: Kathy E. Green (University of Denver)
Discussant: Shudong Wang (Northwest Evaluation Association)

A Comparison of Two Estimation Methods for the Many-Facet Rasch Model Using Real Data From a Large-Scale Language Assessment. Guangming Ling (ETS)

Cross-Country Comparisons of Inattentive, Hyperactive, and Impulsive Behavior in School-Based Samples of Young Children. Christine Merrell (Durham University), Irene Styles (University of Western Australia), Peter B. Tymms (Durham University), Helen R. Wildy (University of Western Austria), Paul Jones (Durham University)

Exploring the Accuracy of Writing Self-Efficacy Judgments of Eighth Graders Using Rasch Measurement Theory and Qualitative Methods. George Engelhard (Emory University), Nadia Behizadeh (Emory University)

Sunday, April 10 - 10:35 a.m. - 12:05 p.m. Sheraton / Grand Ballroom D
Roundtable: Large Scale Assessment SIG
SIG-Large Scale Assessment

Concurrent Versus Separate Scaling of English Language Proficiency Test Items. Seon-Hi Shin (Korea Institute for Curriculum and Evaluation), Insuk Kim (Korea Institute for Curriculum and Evaluation)

Constructing a Common Scale in a Testing Program to Model Growth: Joint Consideration of Vertical Scaling and Horizontal Equating. Hong Jiao (University of Maryland), Robert W. Lissitz (University of Maryland)

Sunday, April 10 - 10:35 a.m. - 12:05 p.m. Sheraton / Grand Ballroom E
Roundtable: Barriers to and Trends in Professional Certification Throughout the Career
SIG-Professional Licensure and Certification

Extended Time Accommodations and Their Impact on High-Stakes Licensure Examinations Differential Item Functioning. Ada Woo (National Council of State Boards of Nursing), Casimer M. Marks (National Council of State Boards of Nursing), Weiwei Liu, Philip Dickison (National Council of State Boards of Nursing), Sarah L. Hagge (National Council of State Boards of Nursing)

Sunday, April 10 - 2:15 p.m. - 3:45 p.m. Doubletree / Nottaway
Scale Construction and Measurement Invariance in Survey Research
SIG-Survey Research in Education

Survey Research Scales and Score Interpretation: A Rasch Rating Scale Analysis. Randall E. Schumacker (The University of Alabama), Elena C. Papanastasiou (University of Nicosia)

Sunday, April 10 - 2:15 p.m. - 3:45 p.m. Doubletree / International Ballroom
Roundtable: Statistical Methods to Improve and Monitor Rater Behavior
Division D - Measurement and Research Methodology. Section 1: Educational Measurement, Psychometrics, and Assessment

Can We Identify Raters Who Assign Inconsistent Scores? Detecting Rater Inaccuracy Using Simulation Methods. Jessica Yue (Virginia Polytechnic Institute and State University), Edward W. Wolfe (Pearson)

Can We Identify Raters Who Don’t Stand Out? Detecting Rater Centrality Using Simulation Methods. Jessica Yue (Virginia Polytechnic Institute and State University), Edward W. Wolfe (Pearson)

Effects on Scoring Under Rater Drift via Latent Class Signal Detection Theory and Item Response Theory. Yoon Soo Park (Teachers College, Columbia University), Lawrence T. DeCarlo (Teachers College, Columbia University)

Monday, April 11

Monday, April 11 - 10:35 a.m. - 12:05 p.m. Doubletree / Crescent Ballroom
Methodological Issues in Survey Research
SIG-Survey Research in Education

Middle Category or Survey Pitfall: Using Rasch Modeling to Illustrate the Middle Category Measurement Flaw. Kelly D. Bradley (University of Kentucky), Kathryn Shirley Akers (University of Kentucky), Nichole M. Knutson (University of Kentucky), Jessica D. Cunningham (Western Carolina University)

Monday, April 11 - 12:25 p.m. - 1:55 p.m. Astor Crowne Plaza / Bienville
Fun With Test Items: Subgroup Construct Stability, Common and Repeated Items, and Item Relevance Factors
SIG-Professional Licensure and Certification

Construct Stability Across Subgroups: An Evaluation Using Differential Item Functioning. Mikaela Marie Raddatz (University of Kentucky), Thomas R. O'Neill (American Board of Family Medicine)

Evaluating the Performance of Common Items Using Item Parameter Drift, Model-Data Misfit, and Response Time. Brian J. Hess (American Board of Internal Medicine), Renbang Zhu (American Board of Internal Medicine), Louis J. Grosso (American Board of Internal Medicine), Gregory S. Fortna (American Board of Internal Medicine), Rebecca S. Lipner (American Board of Internal Medicine)

The Effect of Different Question Presentation Modes on Relevance Ratings. Louis J. Grosso (American Board of Internal Medicine), Hao Song (American Board of Internal Medicine), Rebecca A. Baranowski (American Board of Internal Medicine), Rebecca S. Lipner (American Board of Internal Medicine), Paul A. Poniatowski (American Board of Internal Medicine)

The Impact of Repeated Exposure to Items. Thomas R. O'Neill (American Board of Family Medicine), Kenneth Royal (American Board of Family Medicine)

Tuesday, April 12

Tuesday, April 12 - 8:15 a.m. - 9:45 a.m. Doubletree / Shadows
Various Differential Item Functioning Angles
Division D - Measurement and Research Methodology. Section 1: Educational Measurement, Psychometrics, and Assessment

Comparison of Rasch-Based and Mantel-Haenszel (MH) Procedures in Detecting Differential Item Functioning. Huiqin Ann Hu (Data Recognition Corporation), Kyoungwon Lee Bishop (Data Recognition Corporation)

Tuesday, April 12 - 10:35 a.m. - 12:05 p.m. Sheraton / Grand Ballroom A
Roundtable: School-Level Social and Emotional Learning Programming and Practice: Development and Implementation
SIG-Social and Emotional Learning

Assessing the Implementation Quality of Social and Emotional Learning Programming Over Time: A Rasch Analysis. Peter Ji (University of Illinois - Chicago)

Tuesday, April 12 - 2:15 p.m. - 3:45 p.m. Doubletree / Madewood A
Rater Cognition and Its Importance for Score Validity: Global Perspectives and Findings
Division D - Measurement and Research Methodology. Section 1: Educational Measurement, Psychometrics

An Application of the Mixed Rasch Model to the Analysis of Rater Characteristics and Rater Effects. Edward W. Wolfe (Pearson)

All Statistical Models are Wrong!

Georg Rasch comments on *"The notion of redundancy and its use as a quantitative measure of the deviation between a statistical hypothesis and a set of observational data,"* a paper presented by Per Martin-Löf, at the Conference on Foundational Questions in Statistical Inference, Aarhus, Denmark, May 7-12, 1973.

http://www.rasch.org/memo19732.pdf

Courtesy of Svend Kreiner.

"I wish to make it quite explicit, that the reason for using both significance and redundancy lies in the contention that **every model is basically wrong** [emphasis author's], i.e., it is bound to fail, given enough data.

When you are in the possession of a set of data you may then either be in the position that your significance test tells you that the model fails, or you may not have got enough observations for that purpose. In the latter case **you cannot yet reject** the model on statistical grounds - which of course should not be construed as meaning that you really **accept** it. In the former case you have to realize that the model fails - and I have no sympathy for relaxing the significance requirement for the reason that the data are substantial enough to show it - but that does not mean that the model is too bad to be applied in the actual case. To take a parallel from elementary physics: A "mathematical pendulum" is defined as "a heavy point, swinging frictionless in a weightless string in vacuum". A contraption like that was never seen; thus as a model for the motion of a real pendulum it is "unrealistic". Notwithstanding, it works quite well for a short time interval, but it begins soon to show a systematic decrease of the oscillation angle. To the model - a second order differential equation - thus requiring an amendment, a Friction term is added, and now it works perfectly well for a long time, even during a few days, until another systematic deviation shows. If needed, a further correction, for air resistance, say, should be attempted - but as a matter of fact, **this is not needed,** because it has worked well enough for the purpose of the geo-physicist, which was to measure the gravity constant ("g") with 7 decimal places !

It is exactly at this point Martin-Löf's redundancy sets in: the model fails - that being demonstrated by some significance test - but does it matter for its purposes ? Taking his cue from Information Theory, Martin-Löf uses the redundancy, as there defined, for measuring the deviation of the model from the data, in the sense of determining the relative decrease of the amount of information in the data which is caused by the departure from the null hypothesis.

Taken literally, the redundancy as a tool may be a rather gross evaluation of the loss suffered by replacing the data by the model. Even if it seems small **the parts lost** may affect some of the use of the model quite appreciably. Therefore it may be necessary to undertake a careful analysis in order to localize the losses and consider what to do about them.

In this connection I may touch upon Weldon's dice throwing experiment with a redundancy of 0.000024. But what if we on several repetitions found the same result and it turned out, that the deviations of the observed distributions from the model distributions persisted in the same parts of them ?

I do not know of any repetition of the experiment, neither of any detailed report on fractions of it as they were produced during some years, but I do happen to know (see Steffensen (1923)) that in a similar case the deviations were taken sufficiently seriously by statisticians to attempt fitting them with a number of alternative distributions, any particular justification of which I do not recall having seen.

Let me end up with the scale of redundancies presented by the speaker. It did leave me with the notion of new horrors of conventional limits ! In **this** connection we may, however, have a chance of doing it more rationally by analyzing just which sort of damage and how much of it is invoked by using the model for specified purposes.

I do look forward to the contribution of the redundancy concept to articulating my vague thesis, that we should never succumb to the illusion that any of our models are correct, but we should certainly aim at **making them adequate for our purposes** - the redundancy possibly being a useful measuring instrument in that connection."

References: Johan Frederik Steffensen, Factorial moments and discontinuous frequency functions, *Skandinavisk Aktuarietidsskrift, vol. 6* (1923), pp. 73-89

Walter Frank Raphael Weldon, Letter to Francis Galton, Feb.2, 1894, reporting 26,306 rolls of 12 dice.

Svend Kreiner adds:

Let me also point out that David Cox in his book on "Applied Statistics" with E.J. Snell (1981, Chapman and Hall, p. 42) does not talk about model fitting or model testing. He talks about how to examine the *adequacy* (my italics) of models. That, I think, is the way we should understand what we are doing, when we test the fit of the Rasch model. If we want to use the Rasch model, despite the fact that items do not fit the model, we are obliged to provide some evidence that it is adequate for the purpose we have with the items. It is not enough to say that we want to disregard the statistical test results because models are always rejected if we have enough data.

Rasch Methodology and the Law

S. E. Phillips (Editor). Defending A High School Graduation Test: GI Forum V. Texas Education Agency. A Special Issue of Applied Measurement in Education [Paperback], 2000, Lawrence-Erlbaum.

AERA-APA-NCME *Standards for Educational and Psychological Testing*

teststandards.org tells us: The *Standards for Educational and Psychological Testing* was developed to "promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices" (AERA, APA, & NCME, 1999, p. 1). The *Standards* provides criteria for the "evaluation of tests, testing practices, and the effects of test use" (AERA, APA, & NCME, 1999, p. 2).
Comments on the current revision of the *Standards* are requested by **Wednesday, April 20, 2011.**

Let me share my thoughts with you ....

There's a fundamentally qualitative difference between measurement standards as they are defined by the AERA-APA-NCME and as they are defined in sciences that prioritize measurement as quantification involving the equal ratio divisibility of magnitude differences. The *Standards,* to date, try to control the process with operational recommendations. Standards in the natural sciences, in contrast, focus on the metrics.

The psychosocial sciences focus on process because of the near-universal assumption that quantitative standards like those of the natural sciences are not feasible, notwithstanding 80+ years of theory, data, and calibrated instruments to the contrary. In contrast, the natural sciences do not need to bother with processual recommendations as standards because traceability to reference-standard metrics mediates the way natural laws and theories structure the relation of observations to expectations. Of course, a tape measure or an ammeter must be used correctly, but no one worries about whether the tape measure or the ammeter is itself "fair".

What we need to do is to establish the viability of an alternative form of standards, ones dominated by rigorously articulated and predictive construct theories, instruments calibrated to universally uniform metrics, and data fit to models specifying the requirements for objective inference. Georg Rasch was well aware that his models are, in fact, statements of psychosocial laws (Rasch, 1960, p. 10-11). When data fit a Rasch model, the investigator has effectively discovered a new law, or failed to falsify an existing one.

In summary:

AERA-APA-NCME: Mandated Fairness →
Consequence: (Hoped for) Good Measures

Rasch: Constructed Good Measures →
Consequence: (Verifiable) Fairness

With the widespread awareness of the dominant paradigm's role in maintaining the *status quo* in various disciplines, it has become commonplace to observe that the adherents of a paradigm are almost never persuaded to abandon it in favor of another. New paradigms replace old ones as the proponents and adherents of the old one retire and are replaced by people who grew up familiar with the new one (Kuhn, 1962).

My personal body of work is aimed at building up documentation, language, theory, evidence, and instruments that would constitute the beginnings of a history and tradition of measurement research and practice capable of offering an alternative to the *status quo* of the dominant psychosocial *Standards* paradigm.

The August 2011 International Measurement Confederation (IMEKO) meeting in Jena, Germany is an opportunity for Rasch measurement theoreticians and practitioners to interact with natural scientists and engineers who are especially attuned to, informed about, and interested in the possibilities for unifying the language, concepts, and practice of measurement across the sciences. The call for papers is available at www.tu-ilmenau.de/fakmb/Call-for-Paper.cfps.0.html.
The submission deadline is **Thursday, March 31st, 2011.** I hope to see many of you there.

William P. Fisher, Jr.

T. Kuhn (1962) The Structure of Scientific Revolutions. Chicago: Univ. of Chicago Press.

Rasch Measures and Unidimensionality

Rasch measures never lose their unidimensionality (nor their linearity for the additive form of the Rasch model). Those properties are forced by the Rasch model. But we can lose the connection between the Rasch measures and the intended unidimensional latent variable.

Example: we want to measure "arithmetic ability". 1,000 children take our arithmetic test. 500 children respond to the items carefully. 500 children guess at random.

If we Rasch-analyze the 500 careful children, then we will obtain ability measures and item difficulties on the intended unidimensional, linear "arithmetic" latent variable with good fit of the data to the Rasch measures.

If we Rasch-analyze the 500 guessing children, then we will obtain person measures and item difficulties on a unidimensional, linear "random guessing" latent variable with good fit of the data to the Rasch measures..

If we Rasch-analyze 500 careful children + 500 guessing children, then we will obtain "ability" measures and item difficulties on a unidimensional, linear "arithmetic + random guessing" latent variable with poor fit of the data to the Rasch measures.

We might say, but "arithmetic + random guessing" is not substantively unidimensional! We know that, but the Rasch model does not. It analyzes the data as though they are unidimensional, and then the fit statistics report how well the data match the mathematically unidimensional framework that the Rasch analysis has constructed.

John M. Linacre

Beware of complexifiers and complicators. Truly "smart people" simplify things. *Tom Peters , Business Guru*

Standard Setting, Cut-Scores, and Incorrect Decisions

Anthony James asks:
Is there a term in the testing literature to refer to the stability, accuracy or consistency of pass/fail decisions in high-stakes tests when we compare candidates scores with a cut-score?

I have come up with some terms such as 'false positives', 'false negatives' and 'decision validity'. Is there a more precise term?

Gregory Stone answers:
There are several concepts we must consider when setting standards.

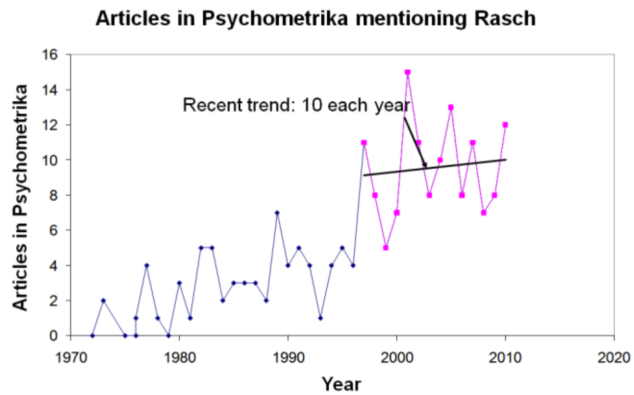
First, standard setting is an evaluative decision. Measurement assists us (extremely well if justifiable, valid models are used) but ultimately it is an evaluative decision. We cannot be slaves to calculations. Instead, assuming you have a construct, and can therefore describe what a person who passes has mastered, changes to the derived cut score should be considered in terms of content, and realistically, political reality. "If I reduce the score to X, I am giving up mastery of Y sort of content," for example. If such a loss is OK, then proceed. If not, consider more than just your standard - consider your expectations, development of the content, task analysis, etc. We cannot put the weight of these qualitative decisions on the back of the quantification.

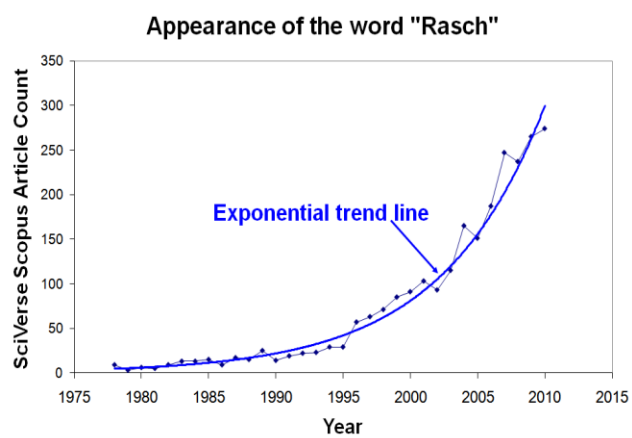
Second, "stability" and "consistency," and to a lesser extent accuracy are really parameters of validity (or validation). Reasoned standard setting models provide error terms. Reasoned standard setting models demonstrate the description of a meaningful, content-based standard. Reasoned standard setting excludes iterative processes that simply introduce external norming, and, like IRT (2-3PL) introduce sample/item specific information that negating the possibility of generalization, equating, etc. All such conversations revolve around "Construct Validity" but construct validity in Messick's holistic expression, not simply a collection of pieces. Whether epistemological (Messick) or ontological (Borsboom) the idea of construct validity is the same. Therefore, assuming a reasonable model is used, there is no "false," because the standard is defined as a particular set of content. It is what it is. If we disagree, it doesn't mean the process has produced a false result.

Third, you ask about fairness. That's an excellent point. Reasonable models include an accounting of error as said. However, more importantly why are we giving or denying a person a job on the basis of one test score, whatever the cut score? Why do we hold back children, or prevent them from graduating on the basis of one score? The premise is that a single test score (a measure of mastery) is equivalent to "competency." It is not. Competency involves much more than a single score, regardless of how fair the cut score and well-developed the test may be. We too often consider mastery and competency as interchangeable. This is a problem. So if you deny a person a job or reject an applicant from college, it does not mean the standard on the exam is problematic; rather, it reflects a process of hiring/admission that produces results that fail the tests of validity and validation. Would we, for example, involuntarily hospitalize an individual on the basis of one psychological assessment tool? Of course not. We would review their overall case file. We would talk with them at length during a session. Why then do we believe one exam should wield so much power in achievement or employment or certification?

Construct Validity (and Validation) are the only terms we really need I would suggest. This isn't a statistical problem (with false x's) but an evaluative one.

(Excerpted from a conversation on the Rasch Listserv)





Alan Tennant reports on his investigation into the publication of Rasch papers in Journals according to the SciVerse Scopus bibliographic database. As of December 2010, the *Journal of Applied Measurement* (and its predecessor, the *Journal of Outcome Measurement)* had published the most Rasch papers (200 in the Scopus database). Second was Psychometrika (106). Overall, the publication of Rasch-related articles is growing exponentially each year, reaching 274 in 2010. The author with the most published Rasch papers (58) is Alan himself. Reviewers who claim that Rasch methodology is esoteric or insignificant are out of touch with the Rasch revolution that is well underway.

The Measurement Papers of Louis Leon Thurstone

Online at the Mead Project, Brock University: www.brocku.ca/MeadProject/inventory5.html

1. * "A Law of Comparative Judgment." Psychology Review, 34 (1927): 273-286.
2. "A Mental Unit of Measurement." Psychological Review 34 (1927): 415-423.
3. * "A Method of Scaling Psychological and Educational Tests." Journal of Educational Psychology 16 (1925): 433- 451
4. "A Scale for Measuring Attitude toward the Movies." Journal of Educational Research 22 (1930): 89-94.
5. "An Experimental Study of Nationality Preferences." Journal of General Psychology 1 (1928): 405-423.
6. "Aspects of Public Opinion." (Report from the Round Table on Politics and Psychology at the Third National Conference on the Science of Politics). American Political Science Review 20 (1926): 126-127
7. * "Attitudes Can Be Measured." American Journal of Sociology 33, (1928): 529-554.
8. "Commentary." In Stuart A. Rice (ed.). Statistics in Social Studies. Philadelphia: University of Pennsylvania Press (1930): 192-196.
9. "Equally Often Noticed Differences." Journal of Educational Psychology 18 (1927): 289-293.
10. "Fechner's Law and the Method of Equal-Appearing Intervals." Journal of Experimental Psychology, 12 (1929): 215-223.
11. "Influence of Freudism on Theoretical Psychology." Psychological Review 31 (1924): 175-183.
12. "Influence of Motion Pictures on Children's Attitudes." Journal of Social Psychology 2 (1931): 291- 304.
13. "Intelligence and Its Measurement." Journal of Educational Psychology 12 (1921): 201-207.
14. "Introduction. Part II: Subjective Measurement." *The Measurement of Value.* Chicago: University of Chicago (1959)
15. "Introduction. Part III: Attitude Measurement." *The Measurement of Value.* Chicago: University of Chicago (1959)
16. "L.L. Thurstone." In Gardner Lindzey (ed.) A History of Psychology in Autobiography Vol. 4.. Englewood Cliffs, NJ: Prentice Hall (1952): 294 - 321.
17. "Motion Pictures and the Social Attitudes of Children: A Payne Fund Study." New York: Macmillan & Company (1933) [co-authored with Ruth C. Peterson]
18. "Psychophysical Analysis." American Journal of Psychology 38 (1927): 368-89.
19. "Rank Order as a Psychophysical Method." Journal of Experimental Psychology 14 (1931): 187 - 201.
20. "Stimulus Dispersion in the Method of Constant Stimuli." Journal of Experimental Psychology 15 (1932): 284-297.
21. "The Anticipatory Aspect of Consciousness." Journal of Philosophy, Psychology and Scientific Methods 16 (1919): 561-568.
22. "The Effect of a Motion Picture Film on Children's Attitudes Toward Germans." Journal of Educational Psychology 23 (1932): 241-246.
23. "The Effect of Motion Pictures on the Social Attitudes of High School Children." Ann Arbor: Edwards Brothers (1932)
24. "The Intelligence of Policemen." Journal of Personnel Research 1 (1922): 64-74.
25. * "The Measurement of Attitude." with E.J. Chave (1929). Chicago: University of Chicago.
26. "The Measurement of Change in Social Attitude." Journal of Social Psychology 2 (1931): 230-235.
27. * "The Measurement of Opinion." Journal of Abnormal and Social Psychology 22 (1928): 415-430.
28. "The Measurement of Psychological Value." In Thomas V. Smith and William K. Wright (eds.), Essays in Philosophy by Seventeen Doctors of Philosophy of the University of Chicago. Chicago: Open Court (1929): 157-174.
29. * "The Measurement of Social Attitudes." Journal of Abnormal and Social Psychology 26 (1931): 249-269.
30. "The Measurement of Value." Psychological Review 61 (1954): 47 - 58.
31. "The Mental Age Concept." Psychological Review 33 (1926): 268-278.
32. * "The Method of Paired Comparisons for Social Values." Journal of Abnormal and Social Psychology, 21, (1927): 384-400.
33. "The Nature of General Intelligence and Ability." British Journal of Psychology 14 (1924): 243-247.
34. "The Nature of Intelligence." London: Kegan Paul, Trench Trubner & Co., (1924).
35. "The Phi Gamma Hypothesis." Journal of Experimental Psychology, 11 (1928): 293-305.
36. * "The Scoring of Individual Performance." Journal of Educational Psychology 17 (1926): 446-457.
37. "The Significance of Psychology For the Study of Government and Certain Specific Problems Involving Psychology and Politics." (Report from the Round Table on Politics and Psychology at the Second National Conference on the Science of Politics, Chicago, Sept
38. "The Stimulus-Response Fallacy in Psychology." Psychological Review 30 (1923): 354-369.
39. "The Unit of Measurement in Educational Scales." Journal of Educational Psychology 18 (1927): 505-524.
40. * "Theory of Attitude Measurement." Psychological Review 36 (1929): 222-241.
41. "Three Psychophysical Laws." Psychological Review, 34 (1927): 424-432.

Courtesy of Serkan Dolma

** indicates Thurstone papers cited by Benjamin D. Wright*

Thurstone and Controversy

L.L. Thurstone (in "L.L. Thurstone" 1952, pp.310-312):

"I wrote a paper entitled 'Attitudes can be measured.' Instead of gaining some approval for this effort, I found myself in a storm of criticism and controversy. The critics assumed that the essence of social attitudes was by definition something unmeasurable."

"There was heavy correspondence with people who were interested in attitude measurement, but they were concerned mostly with the selection of attitude scales on particular issues to be used on particular groups of people."

"There seemed to be very little interest in developing the theory of the subject. The construction of more and more attitude scales seemed to be unproductive, and I decided to stop any further work of this kind. Incomplete material for a dozen more attitude scales was thrown in the wastebasket and I discouraged any further work of that kind in my laboratory. I wanted to clear the place for work in developing multiple factor analysis."

"The excuse is often made that social phenomena are so complex that the relatively simple methods of the older sciences do not apply. This argument is probably false. The analytical study of social phenomena is probably not so difficult as is commonly believed. The principal difficulty is that the experts in social studies are frequently hostile to science. They try to describe the totality of a situation and their orientation is often to the market place or the election next week. They do not understand the thrill of discovering an invariance of some kind which never covers the totality of any situation. Social studies will not become science until students of social phenomena learn to appreciate this essential aspect of science."

Later, L.L. Thurstone *(1959, " Introduction to Part III: Attitude measurement" p. 321)* said that he had "tried to avoid controversy when it would have been better to ignore it."

I wonder why he did not seem to consider the value of engaging with controversy?

William P. Fisher, Jr.

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

Good Measures from Bad Data

In many assessments, there are examinees who misbehave, and items that are poorly constructed. Nevertheless, everyone must be measured, and every item must be included except those that are obviously, blatantly faulty.

Blatantly faulty items are those that we can show to a content expert (who knows nothing about statistics) and say: "Do you see this ... (typographical error, ambiguity, scoring problem, irrelevant content, ...). This item is obviously wrong or off-topic!"

Items with conspicuous DIF are more awkward to handle, and depend on the policy of the testing agency. It is easiest to treat them as blatantly faulty and omit them, but they can be split into separate items for separate DIF groups.

But what about random guessing, doubtful items and other problematic data? A three-stage estimation process provides a solution:

i) Analyze all the data. Identify problems.

ii) Reanalyze all the data, but with items and persons with misfit problems deleted and obviously errant or off-target responses omitted. This is the "good" dataset. Save the estimates of the item difficulties and Rasch-Andrich thresholds (for polytomies).

iii) Analyze all the data. Delete only obviously, blatantly faulty items. Anchor (fix) the "good" items at their "good" difficulties, and the polytomies at their "good" thresholds. Output the final set of person measures and item difficulties.

The measure for each person is now estimated in the frame-of-reference of the "good" data with the minimum of distortion of that measure by irrelevant (to that person) "bad" data.

Timed Tests

If we have a timed test, and score all incorrect answers and all item-not-reached answers as "0", then the final items have few correct answers, "1", even if the very last item is the conceptually easiest item on the test.

To get around this problem we do the three-stage analysis. In the second stage, we use only data from examinees who have definitely reached an item (right or wrong). All unreached responses are coded "not administered" (e.g., M for missing) and excluded from the analysis. This analysis gives us the best estimates of the difficulties of the items. We save these "good" item difficulties.

In the third stage, we score all the data 0-1, but use the "good" item difficulties, so that the measures of students who responded to most of the items are not distorted by the performances of students who responded to fewer items.

John M. Linacre

Additive Conjoint Measurement and Rasch

New insights into Additive Conjoint Measurement (Luce & Tukey, 1964) are provided by Newby et al. (2010):

1. "an ordered conjoint structure has an additive representation if and only if it has a Rasch representation" (p. 10)

2. "not all data to which the Rasch model *could* be applied is data to which the Rasch model *should* be applied" (p. 5, italics authors')

Luce RD & Tukey JW. 1964. Simultaneous conjoint measurement. Journal of Mathematical Psychology, 1, 1-27.

Newby, V., Conner, G., Grant, C., & Bunderson, V. (2010). Rasch model and additive conjoint measurement (pp.1-11). In M. Garner, G. Engelhard, M. Wilson, & W. Fisher (Eds.), Advances in Rasch measurement (Vol. 1) Maple Grove, MN: JAM Press.

Rasch-related Papers in Full

An online resource is: "Free Full Text" –
http://www.knowmade.fr/results.html?q=rasch

Journal of Applied Measurement
Vol. 11, No. 4 Winter 2010

Richard M. Smith, Editor
JAM web site: http://www.jampress.org

Rasch-related Coming Events

March 15 - June 30, 2011, Tues. - Thur.
KDD-Cup 2011 Yahoo! Music Competition, Rasch Team on FaceBook

March 23-25, 2011, Mon.-Wed. In-person workshop: Introductory Rasch (A. Tennant, RUMM), UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Apr. 8-12, 2011, Fri.-Tues. AERA Annual Meeting, New Orleans, LA, www.aera.net

April 29 - May 27, 2011, Fri.-Fri. Online course: Rasch (Winsteps, introductory) online course (M. Linacre, Winsteps), www.statistics.com

May 4-6, 2011, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), UK,
May 9-11, 2011, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

June 20-21, 2011, Mon.-Tues. Summer Institute on Measuring Rehabilitation Outcomes (Gershon, Northrock), Chicago, USA, Rehabilitation Inst. of Chicago, CROR

June 23-25, 2011, Thurs.-Sat. 33rd Language Testing Research Colloquium LTRC, Ann Arbor, MI, USA, www.lsa.umich.edu/eli/LTRC2011

July 4-5, 2011, Mon.-Tues. International Workshop on Patient Reported Outcomes and Quality of Life, France, www.lsta.upmc.fr/mesbah/PROQOL/

July 8 - Aug. 5, 2011, Fri.-Fri. Online course: Rasch - Further Topics (Winsteps, Advanced) online course (M. Linacre, Winsteps), www.statistics.com

July 11-15, 2011, Mon.-Fri. PROMS-2011 Pacific Rim Objective Measurement Symposium, Singapore proms2011.nie.edu.sg

Aug. 31 - Sept. 2, 2011, Wed.-Fri. IMEKO Conference, Jena, Germany, www.tu-ilmenau.de

Sept. 14-16, 2011, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), UK,
Sept. 19-21, 2011, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), UK,
Sept. 22-23, 2011, Wed.-Fri. In-person workshop: Advanced Rasch (A. Tennant, RUMM), UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Jan. 9-15, 2012, Mon.-Wed. In-person workshop: Introductory Rasch course (Andrich, RUMM2030),
Jan. 16-20, 2012, Mon.-Wed. In-person workshop: Advanced Rasch course (Andrich, RUMM2030), Perth, Australia, www.education.uwa.edu.au

Jan. 23-25, 2012, Mon.-Wed. Fifth International Conference on Probabilistic Models for Measurement in Education, Psychology, Social Science and Health, Perth, Australia, www.education.uwa.edu.au