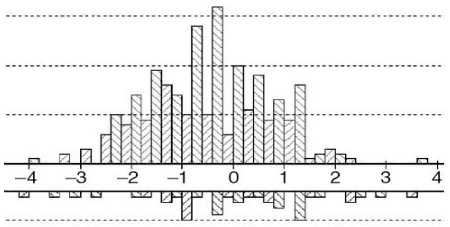
Measure Distributions



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 21 No. 4 Spring 2008 ISSN 1051-0796

Something about Bridge-Building [Test Equating] Techniques - a sensational new creation by Dr. Rasch

From the 23 May, 1957 issue of "Folkeskolen", the Danish elementary school journal.
Translated by Cecilie Kreiner, courtesy of Svend Kreiner

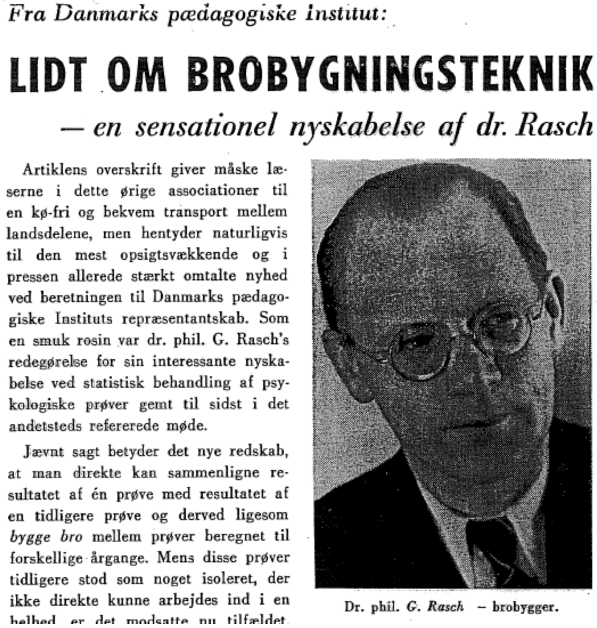
The headline of this article may give readers in this island kingdom associations to comfortable and queue-free transportation between the islands. However, it refers, of course, to the most sensational news at the report to the council at the Danish Institute for Educational Research, news which has already been much spoken of in the press. As a last treat, G. Rasch's account of his interesting new creation for statistical processing of psychological tests concluded the meeting covered elsewhere.

Plainly put, this new tool means that one can compare the result of one test directly with the result of a previous test, thereby *building a bridge* between tests meant for different grades. Whereas these tests were previously something isolated which could not be incorporated into a whole, the opposite is now the case, and one can thus produce a development curve.

An attempt at an explanation

After the council meeting, we had an interview with statistician Dr. Rasch, who is not only connected with the Danish Institute for Educational Research but also a lecturer at University of Copenhagen where, amongst other things, he teaches statistics to psychology students. The work on shaping the new tool has been going on for several years, and a full account of the work will obviously contain a terminology which will make it unintelligible outside the circle of colleagues. Dr. Rasch's explanation below is, however, largely comprehensible to the interviewer, and that should guarantee the reader's understanding.

- The work with the bridge building method commenced at a study of children who were slow readers carried out for the Ministry of social Affairs, Dr. Rasch says. During the work on this study, the problem of transferring the result from one test to another in order to create a basis for comparison arose. However, no measuring instrument existed, because a standardizing is not an actual measurement, as two children in different grades receive different scores for the same performance.



Fra Danmarks pædagogiske Institut:

LIDT OM BROBYGNINGSTEKNIK

— en sensationel nyskabelse af dr. Rasch

Artiklens overskrift giver måske læserne i dette ørige associationer til en kø-fri og bekvem transport mellem landsdelene, men hentyder naturligvis til den mest opsigtsvækkende og i pressen allerede stærkt omtalte nyhed ved beretningen til Danmarks pædagogiske Instituts repræsentantskab. Som en smuk rosin var dr. phil. G. Rasch's redegørelse for sin interessante nyskabelse ved statistisk behandling af psykologiske prøver gemt til sidst i det andetsteds refererede møde.

Jævnt sagt betyder det nye redskab, at man direkte kan sammenligne resultatet af én prøve med resultatet af en tidligere prøve og derved ligesom *bygge bro* mellem prøver beregnet til forskellige årgange. Mens disse prøver tidligere stod som noget isoleret, der ikke direkte kunne arbejdes ind i en helhed, er det modsatte nu tilfældet

Dr. phil. *G. Rasch* – brobygger.

Picture caption: Dr. Phil. G. Rasch - bridge-builder.

- In collaboration with head of department, Master of Psychology *Carl Åge Larsen,* I carried out reading tests on a large number of children in the 2nd to 7th grades, in which children in the same grade were given two or more tests. Thus the tests "T 5" and "S" were compared directly in the 4th, 5th and 6th grades. When points for the illustration of the amount of errors were subsequently placed in a diagram with two figure axes so that the horizontal distances illustrated the number of errors in the "S"-test, and the vertical distances illustrated the number of errors in the "T 5"-test, the obvious result emerged at

Table of Contents

first: with the same test, the lower grades had a high percentage of errors, the higher grades had a lower percentage of errors, and the highest grades had the lowest percentage of errors. The interesting thing was, however, that the three assemblies of points which illustrated the amount of errors in the different grades immediately succeeded each other and pointed towards the 0-point. A line through them showed that approximately 12 errors in the "T 5"-test corresponded to 10 errors in the "S"-test, and that approximately 24 errors in the "T 5"-test corresponded to 20 errors in the "S"-test.

- In short: 1 error in the one item corresponds to 1.2 errors in the other item. It has turned out that the ratio remains the same when comparing two items by means of a third. An expression for the degree of difficulty of one test in relation to another has thus been found here. - Similar ratios have been found for further tests which, like these reading tests, are not systematically constructed *[Rasch's Poisson model]*.

- Other tests, such as, for example, the "F"-test, consist of a number of individual items (texts) of continuously increasing degrees of difficulty, and here the calculations become considerable more complicated, since the individual items have to be considered first *[Rasch's dichotomous model]*.

Two questions

- One may then ask: Is it possible to speak of the degree of difficulty of two items relative to each other? Can a person be given a number for their proficiency in solving this item? For example, can A be twice as good as B? Not just at playing football and drinking coffee, but at solving a number of items of the same kind but of different degrees of difficulty?

- Then another question immediately follows: Can item I be twice as difficult as item II?

Let us say that:

A's degree of proficiency is D,
B's degree of proficiency is d,
D must be 2 x d.
As for the items, the ratio is:
I's degree of difficulty is S,
II's degree of difficulty is s,
3
S must be 2 x s.

In order to answer these questions, I propose: A must solve I just as easily as B solves II.

But what is "just as easily"? This vicious question will not be avoided by applying an ordinary view to what people do. Precise physical laws may be laid down for the movements of the planets. But people? They can think of doing anything, both when solving items and in other situations. By making foolish mistakes, the proficient may accidentally solve an easy item in a wrong way. Conversely, the slow learner may chance on a correct solution to a difficult item. But in both cases there is a chance – a likelihood – great or small that the item is solved correctly. This idea, the chance that a person solves a given item correctly, can be used for giving meaning to the notorious expression "A solves I just as easily as B solves II", since quite simply it is to be understood as their chances being equally good.

Now the chances that A solves I should be determined by his proficiency D and the difficulty of the item S, and it should be the same for B with I, i.e. when both proficiency and degree of difficulty is bisected, or for that matter multiplied by another factor. The chances of solving an item thus come to depend only on the ratio between proficiency and degree of difficulty, and this turns out to be the crux of the matter.

- A detailed investigation of the individual items in the "F"-test has shown that these ideas are applicable to the "F"-test. Thereby, it follows that the result of an "F"-test is evaluated as a whole, i.e. as a measure of his proficiency in these kinds of tests. Had he been given a different "F"-test, the result of the measurement would most likely have been the same.

- A bridge has thus been built between the different "F"-tests. The result of one test can be translated into the result of another.

With this illustration, Dr. Rasch has at any rate removed the entirely abstract from the concept of bridge building techniques. This ground-breaking work will be studied and employed, not only in Denmark, but in the whole world. The significance it can have practically within school work will be accounted for in a following article.

Finn Jolander

1st International Conference on High School and College Educational Evaluation
17, 18, 19, September 2008

Veracruz City, Mexico. (in Spanish).

The *Colegio de Bachilleres de Veracruz, COBAEV* (Main Mid-Higher Level Institution of the State of Veracruz) and the *IEIA (Institute of Evaluation and Advanced Engineering)* encourage you to participate, as we have extended the call for papers up to **May 30** for the abstracts and up to **June 3**0 for the complete paper.

Topics:

1. Evaluation of competencies at High School level and for selection at College undergraduate level.
2. Teacher evaluation at High School level
3. Standards for quality of tests
4. New trends on evaluation at High School level
5. Qualitative evaluation through objective instruments
6. New technologies, software, projects and materials.

Workshops:

1. Design of Evaluation Centers for Mid-Higher level Education
2. Introduction to Rasch analysis for objective tests
3. Analysis of the initial profile of the undergraduate student in relation to Mid-Higher level education.
4. Organization of evaluation instruments (tests, portfolios, etc.) for the classroom.
5. Design of portfolios for the classroom evaluation.
6. Test design for competencies with free-response items.
7. Test scoring and Objective Standard-Setting for Judge-Mediated Examinations with free-response items

Featured Presenters:

· Debbie Reese, PhD / Project Manager of the Selene project on evaluation of competencies for the sciences, sponsored by the NASA. Center for Educational Technologies Wheeling Jesuit University, West Virginia (EUA).

· Margarita Peña-Borrero, PhD / Director of the Colombian Institute for the Improvement of Higher Education, ICFES (Colombia).

· Gregory E. Stone, Ph.D. / Ass. Professor of Research and Measurement. J.H. College of Education. University of Toledo, Ohio (EUA)

· Héctor Valdés-Veloz, PhD / Director of the LatinAmerican Laboratory for the Evaluation of the Quality of education, LLECE (Chile).

· Lic. Ana Ma. Aceves-Estrada / Director of Evaluation and Educational Politics of the Secretary of Education, (México).

· Lic. Ma. Antonieta Díaz-Gutiérrez / National Project Manager from México, PISA project (OECD), Director of International and Special Projects, National Institute for the Educational Evaluation. INEE (México).

Introduction to Rasch Measurement of Modern Test Theory

Online Course, July-November, 2008

Coordinators:
Professor David Andrich and Dr Ida Marais.

More information about the course can be obtained from Ida Marais: (ida.marais -at- uwa.edu.au)
http://www.education.uwa.edu.au/httpwww.education.uwa.edu.aunews/on_line_course

The Course Of Study - Background

In the Australian Semester 2, 2008 (July to November), a graduate level course of study introducing Rasch measurement is available in the external study mode. This mode of study means that it can be studied from anywhere in the world. A discussion group will operate for online interaction as part of the course of study.

Those enrolled obtain (i) a set of lecture materials, which includes hard copy of all of the lectures, (ii) details of the assignments you will be required to submit, (iii) the necessary reading materials, and (iv) the Study Guide setting out the steps you will need to follow to successfully complete the course.

The course is available to graduate students but also appropriate for professionals (who are not enrolled students at UWA) who have become interested in applying Rasch models and would like to improve their understanding of first principles and the relationship between the Rasch models and traditional test theory.

This course has been presented in the same period every year from 2000. In each year, people from many parts of the world have taken the opportunity to enroll. Because of the success of the previous presentations, the course is being offered again this year.

The use of the program RUMM2020 is available throughout the course.

The RUMM program is a very easy to use interactive program that permits learning many features of the Rasch measurement model by working around the program's menus – for example the effects of rescoring any item, deleting items, studying alternatives in distractors, assessing differential item functioning, automatic linking of different sets of items, effects of deleting samples or individuals, taking account of missing data, and so on. To enhance understanding all of the information is available both graphically and statistically, including item characteristic curves, person item maps, etc. The program comes with three manuals and analyzed data sets.

Standard Errors for Performance Standards based on Bookmark Judgments

A variety of methods can be used to estimate the standard errors of performance standards or cut scores. Historically, these methods have ranged from classical methods based on the standard errors of mean panelist judgments (Jaeger, 1991) to more elaborate approaches based on generalizability theory (Yin and Sconing, 2008). Engelhard (2007) and his colleagues (Sullivan, Caines, Tucker, & Engelhard, 2008) recently described the use of Rasch measurement theory as a conceptual framework for evaluating the quality of panelist judgments within the context of bookmark and other item mapping based methods. The multifaceted Rasch measurement (MFR) model provides another approach for estimating the standard errors of performance standards. The MFR model can be used to model judgments collected from modified-Angoff procedures, as well as procedures based on item maps, such as bookmark and mapmark procedures (Schulz and Mitzel, in press).

Modified-Angoff and item-map based procedures are the two most popular methods for collecting judgments from standard-setting panelists (Cizek and Bunch, 2007). The bookmark procedure (Mitzel, Lewis, Patz, and Green, 2001) is becoming the standard-setting method of choice in many statewide assessment programs. For example one possible MRM model for bookmark judgments is:

$$Ln\ [P_{nijk} / P_{nijk-1}] = \theta_n - \delta_i - \omega_j - \tau_k \qquad [1]$$

where

P_{nijk} = probability of panelist n giving a bookmark rating of k on item i for round j,

P_{nijk-1} = probability of panelist n giving a bookmark rating of k-1 on item i for round j,

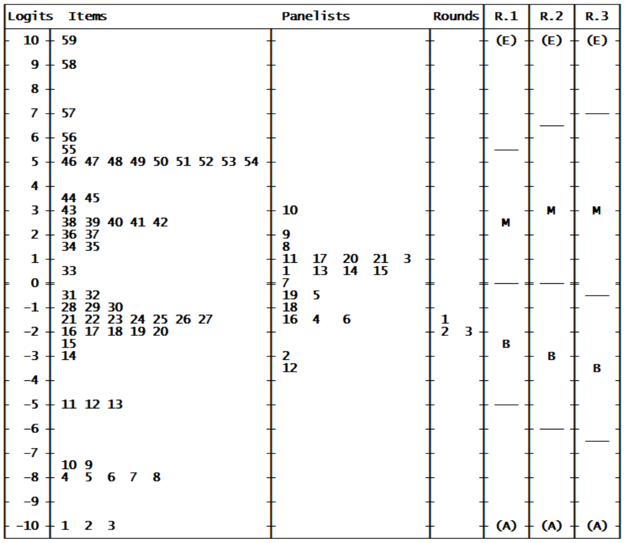
θ_n = judged performance level for panelist n,

δ_i = judged difficulty for item i,

ω_j = judged performance level for round j, and

τ_k = judged performance standard for bookmark rating category k relative to category k-1.

Figure 1. *Wright Map (Grade 3 mathematics)*



The rating category coefficients, τ_k, define the performance standards or cut scores.

Table 1. *Calibration of performance standards (Rasch cut scores)*

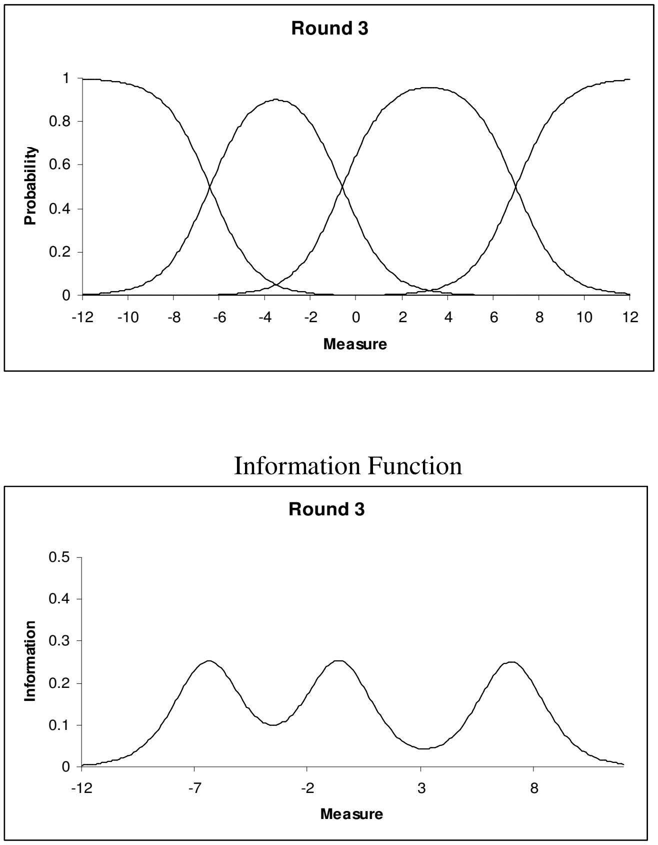
Category	Count	Percent	Mean Measure	OUTFIT MnSq	Rasch Cut Score	S.E.
Round 1						
Apprentice	265	23%	-7.55	0.90		
Basic	416	36%	-2.52	3.00	-5.20	0.13
Met	392	34%	2.18	0.90	-0.18	0.11
Exceeded	82	7%	5.52	1.00	5.38	0.15
Round 2						
Apprentice	230	20%	-8.42	0.60		
Basic	470	41%	-2.84	1.60	-6.13	0.14
Met	413	36%	2.35	0.60	-0.22	0.11
Exceeded	42	4%	6.15	0.70	6.35	0.20
Round 3						
Apprentice	231	20%	-8.68	0.60		
Basic	462	40%	-3.14	1.60	-6.40	0.14
Met	438	38%	2.18	0.60	-0.61	0.11
Exceeded	24	2%	5.55	1.20	7.01	0.25

The Rasch cut scores are the recommended performance standards.

In order to illustrate the use of the MFR model for estimating standard errors, data from the Michigan Educational Assessment Program are used (website: http://www.michigan.gov/mde/). There were 21 panelists on the standard-setting panel. The instrument examined in this study is the Grade 3 mathematics test used in the Michigan Educational Assessment Program (MEAP). The judgments were obtained based on a modified bookmark approach called Item Mapping. The standard-setting judgments were obtained in three separate rounds.

The Wright map with the calibrations of the items, panelists, rounds, and performance standards is presented in Figure 1. The judged locations of the items represent the shared understandings of the standard-setting panelists for students within the four performance levels. Panelist locations represent their severities, while round locations represent average difficulties of judgments for each round. Finally, the category coefficients represent the performance standards by round (R.1, R.2, and R3 for these panelists on this assessment (A=Apprentice, B=Basic, M=Met, and E=Exceeded).

Figure 2. *Category functions*


Category Response Function

Information Function

Table 1 presents the category statistics with the category coefficients defined as the performance standards or cut scores. The performance standards change over rounds, and the most disagreement is found in Round 1 for the Apprentice category (OUTFIT=3.00). The final column in Table 1 gives the standard errors. The standard errors for the performance standards do not vary much over rounds for the apprentice/basic cut score or the basic/met cut score. However, uncertainty regarding the met/exceeded category increases significantly over rounds. The error variance at Round 3 is three times larger than the error variance at Round 1 (.0625/.0225 = 2.7777). The top panel in Figure 2 presents the category response function for the performance standards for Round 3. The bottom panel presents the information function with a very distinctive shape with a peak at each of the performance standards. The information function shows graphically the spread in the information function at each performance standard.

Additional work is needed to compare different approaches for estimating standard errors for performance standards. Given the high-stakes decisions made on the basis of assessments in education, health, and the professions, it is essential to develop procedures for conveying the uncertainty inherent in the estimated performance standards. The standard errors are readily obtained using the MFR model, and the MFR model offers additional information about the quality of standard-setting judgments that is not available with approaches based on classical or generalizability theory.

George Engelhard, Jr., Ph.D.
Emory University

Cizek, G.J., & Bunch, M.B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage Publications, Inc.

Engelhard, G. (2007). Evaluating bookmark judgments. *Rasch measurement: Transactions of the Rasch measurement SIG*, American Educational Research Association, 21(2), 1097-1098.

Jaeger, R.M. (1991). Selection of judges for standard-setting. *Educational Measurement*, Spring, 3-6, 10, 14.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed), *Setting performance standards: Concepts, methods and perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Schulz, E.M., & Mitzel, H.C. (in press). A mapmark method of standard setting as implemented for the National Assessment Governing Board. In E. V. Smith, Jr., and G. E. Stone (Eds.), *Applications of Rasch measurement in criterion-referenced testing*, JAM Press.

Sullivan, R., Caines, J., Tucker, C., Engelhard, G. (March 2008). *Examining the bookmark ratings of standard-setting panelists: An approach based on the multifaceted Rasch measurement model.* Paper presented at the 2008 meeting of the International Objective Measurement Workshop. New York: New York University.

Yin, P., & Sconing, J. (2007). Evaluating standard errors of cut scores for Item Rating and Mapmark procedures: A Generalizability Theory approach. *Educational and Psychological Measurement, 68*(1), 25-41.

National Certification Board for Alzheimer Care
NCBAC http://www.ncbac.net/

This is such a Rasch/University of Chicago story. At the University of Chicago Graham School in the 2004 Spring semester, Janis Nowak took my *Persuasive Communication* class. We had a conversation the first night. She is president of 8 assisted living homes in Wisconsin, and was talking about certifying her caregivers. After establishing that it was a "certificate of training", we said, "Let's put on a show!"

We want everyone to know that these certifications exist. Please help us spread the word.

Donna Surges Tatum

Using a Partial-Credit Rasch Model to Detect Social Desirability Bias

In Public Health, since many causes of disease, disability and death are preventable by behavior changes there is a strong reliance on designing and evaluating prevention-focused interventions. In designing these evaluations, researchers rely heavily on scales to assess complex variables including knowledge, attitudes and even behavior. Oftentimes, sensitive topics or behaviors are measured using these self-reported scales. Under Classical Test Theory (CTT) and using traditional statistical tests (e.g. t-tests on raw scores), evaluators assume equal interval properties of Likert-scales without ever systematically assessing whether the scales truly fit that assumption. The danger of this assumption is that if respondents are not using the rating scale categories in the hypothesized manner, results obtained from these analyses may be misleading or incorrect.

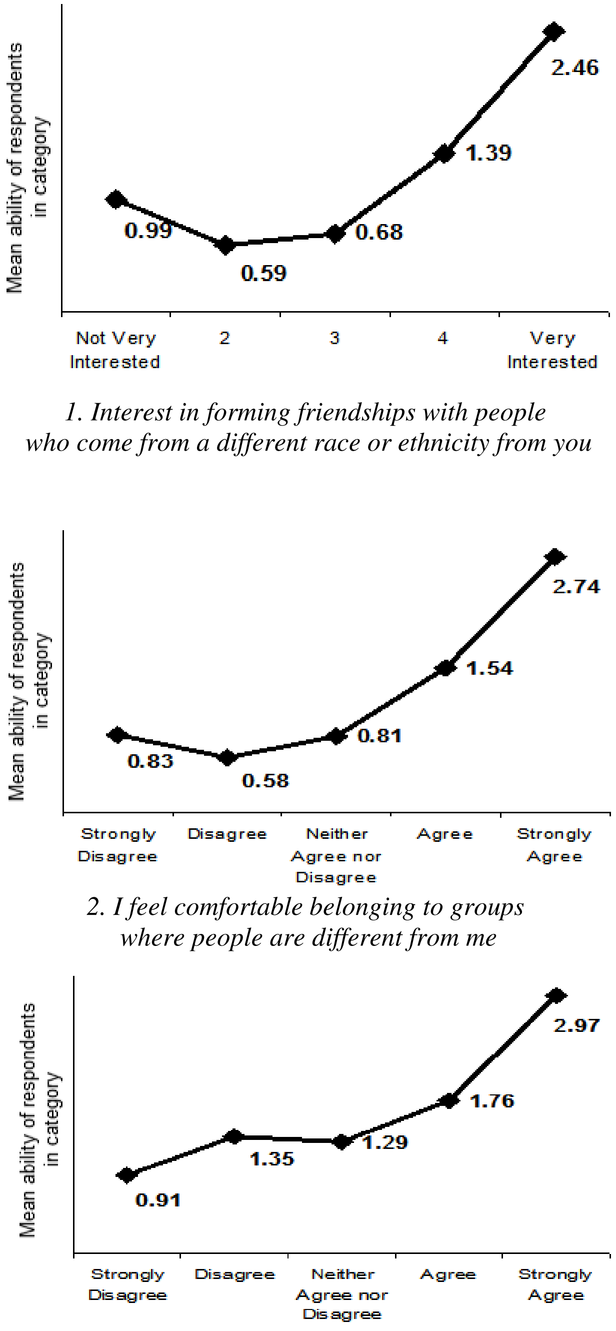
To illustrate this challenge, data from the longitudinal evaluation of the Americorps Program were used (CNS, 2004). Among the immediate hypothesized outcomes of the Americorps program are Awareness of Others/Diversity, impacted by both specific educational activities offered by the program and by the diversity of the Corps itself. The program hopes to improve participants' understanding of diverse cultures and backgrounds, and appreciation of the value of diverse people and opinions. The evaluation of the program included an 11-item Appreciation of Ethnic and Cultural Diversity Scale to assess the change in this latent variable.

While many interventions have a similar desired impact on appreciation of diversity and other potentially sensitive topics, measurement of this latent variable is fraught with challenges. For one, social norms impact respondent behavior on self-reported surveys, even those that employ validated scales. The extent to which these norms influence respondent behavior in self-administered surveys comprises Social Desirability Bias (e.g. Nederhof, 1985). Those interested in assessing variables highly susceptible to social norms should be particularly interested in detecting whether or not Social Desirability Bias is at work in their sample.

A partial credit FACETS model was run using the data from this evaluation (n=4,016). The model included three facets (respondents, items and time period/group) to account for the design of the original evaluation (which included a pre-test and post-test for both Americorps members and a comparison group). Overall fit for the model varied for each of the three facets (Table 1). While the mean OUTFIT mean-square for respondents was 1.08, close to its expected value of 1.0, the S.D. is somewhat larger than is the typically encountered for well-behaved data. 8% of the people had alarmingly large mean-square values over 2.0, forcing another 12% to have mean-square values less than 0.5. This misfit prompts an investigation

Table 1: Fit Statistics		Facet		
		Respondents "ability"	Time-point	Items
Measure	Mean	0.42	-2.10	0
	SD	1.54	0.01	0.24
	N	4,016	4	11
OUTFIT Mean-square	Mean	1.08	1.08	1.08
	SD	0.84	0.03	0.22
INFIT Mean-square	Mean	1.07	1.01	1.00
	SD	0.76	0.01	0.11

Figure 1: Rating Scale Categories



1. Interest in forming friendships with people who come from a different race or ethnicity from you

2. I feel comfortable belonging to groups where people are different from me

3. I am comfortable interacting with people from a different racial or ethnic background into whether Social Desirability Bias may have influence these people's responses.

The items were asked using a five-point Likert-type scale. A FACETS model was used to estimate the mean "ability" (location on the latent variable) of those who responded in each category of each item's rating scale. If the mean abilities are disordered, this could indicate that our respondents did not treat the rating scale as strictly monotonic, resulting in empirically disordered categories. Consequently we could have reason to believe that Social Desirability Bias may have skewed the use of the rating scales.

Figure 1 shows three examples of the distribution of ability estimates for each of the five rating scale options. For example, for the second item the ability estimate that corresponds with Strongly Disagree is 0.83 logits, the ability estimate for Strongly Agree is 2.74 logits. Six of the eleven items show a similar disordering of the rating scale categories; in other words, for those items there is at least one point in the rating scale where the ability estimate that corresponds to the category goes *down* while the category goes *up*. This indicates that respondents chose a higher rating for those items than their actual ability; a sign that Social Desirability Bias is likely in play.

Journal of Applied Measurement Volume 9, Number 2. Summer 2008

Effects of Varying Magnitude and Patterns of Response Dependence in the Unidimensional Rasch Model. *Ida Marais and David Andrich*

Fisher's Information Function and Rasch Measurement. *Mark H. Stone*

A Rasch Analysis for Classification of Systemic Lupus Erythematosus and Mixed Connective Tissue Disease. *Kyle Perkins, Robert W. Hoffman, and Nikolaus Bezruczko*

Magnitude Estimation and Categorical Rating Scaling in Social Sciences: A Theoretical and Psychometric Controversy. *Svetlana Beltyukova, Gregory E. Stone, and Christine M. Fox*

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

Rasch Measurement in Developing Faculty Ratings of Student s Applying to Graduate School. *Sooyeon Kim and Patrick C. Kyllonen*

Understanding Rasch Measurement: Using Rasch Scaled Stage Scores to Validate Orders of Hierarchical Complexity of Balance Beam Task Sequences. *Michael Lamport Commons, Eric Andrew Goodheart, Alexander Pekker, Theo Linda Dawson, Karen Draney, and Kathryn Marie Adams*

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

Using these methods to detect Social Desirability Bias may also provide opportunities to correct the analysis plan. Replacing rating scale category values (e.g. 1 for Strongly Disagree) with the estimated ability from the Rasch model, for example, will allow the analysis to take into account the disordered ratings of participants. This allows researchers to account for the impact of inaccurate self-assessment without altering the format of the scale itself.

Using CTT and t-tests on raw scores for these items assumes that Strongly Disagree is the lowest rating, but these results show that for some items *Disagree* represented the lowest rating. In the Americorps evaluation, participants from certain programs performed significantly worse on this variable at follow-up as compared to baseline. The evaluators concluded that program-related experiences, "may [have led] to short-term disillusion with the concept of working in diverse groups." This analysis, however, indicates that an alternate analysis approach that incorporates the ability estimates for each item's rating scale may provide a more accurate impact of the program that is less impacted by Social Desirability Bias.

Laura M. Lessard, MPH
Department of Behavioral Sciences and Health Education
Rollins School of Public Health
Emory University
Atlanta, Georgia

Nederhof, AJ. (1985). Methods of coping with social desirability bias: a review. European Journal of Social Psychology, 15: 263-280

Serving Country and Community: A Longitudinal Study of Service in Americorps. 2004, Corporation for National and Community Service (CNS), Office of Research and Policy Development: Washington, DC. Available at: http://www.americorps.org/pdf/06_1223_longstudy_report.pdf

Rasch as a World-View

Rasch measurement is really about a more powerful way of thinking about the world around us. Rasch models imagine the world to be composed of perfect equal-interval latent variables. These express the meaning in everything that is around us. We can reconstruct the perfect variables from imperfect data. The perfect variables give us great insights into why things are the way they are. Their use gives us the security that comes from knowing what will probably happen. They even gives us power to change the future. All this sounds impossible, far out of the reach of mere humans. But "You have to believe the impossible" (Howard Head, inventor of Head skis and Prince tennis rackets). Rasch measurement is about the impossible.

John Michael Linacre

International Conference On Outcomes Measurement

Behavioral Health Measurement: From Theory To Application

September 11th-13th, Thursday-Saturday, 2008

Bethesda, Maryland at the Hyatt Regency Bethesda

Call For Papers:
The primary focus of ICOM is behavioral health, especially substance abuse, but papers with particularly interesting methodological insights are also welcome even if they address other substantive areas. Specifically, we are seeking speakers who can present papers addressing conference objectives on the following key topics:

§ Scoring and clinical interpretation issues in the IRT/Rasch framework.

§ Ensuring the comparability of health outcomes assessment across diverse population sub-groups.

§ Interpretation of changes.

§ Measurement criteria for model choice.

§ Using person and item fit statistics to refine measures and target interventions more appropriately.

§ Integrating measurement into multilevel modeling. Multi-level modeling makes assumptions regarding scale construction.

§ Computer Adaptive Testing (CAT) Applications.

§ Unidimensional and multidimensional models.

§ Other innovative applications of IRT/Rasch or approaches to improving the validity and efficiency of measurement.

The Call for Papers deadline is June 6, 2008. To submit an abstract, please visit the ICOM website at http://www.icom-2008.org/ and fill out a presenter proposal form.

Workshop Opportunities:
Pre-/post-session workshops will be held on September 9th, 10th, and 13th, 2008. These opportunities include *An Introduction to Rasch Measurement: Theory and Applications* taught by Everett Smith and Richard Smith; *An Introduction to RUMM 2020* taught by David Andrich; and a special session free to NIH employees of *An Introduction to IRT/Rasch Measurement Using Winsteps* taught by Kendon Conrad, Barth Riley, and Michael Dennis. Full descriptions of the workshops can be found at the ICOM website

Conference Registration:
Please register on-line for the main conference and workshop sessions. You will also be able to reserve a room at the Hyatt Regency Bethesda with the ICOM rate at the website: http://www.icom-2008.org/

Hosted By National Institute On Drug Abuse (Nida), Office Of Behavioral Social Science Research (Obssr), & University Of Illinois At Chicago

International Symposium on Measurement of Participation in Rehabilitation Research

October 14-15, Tuesday-Wednesday, 15 October 2008

Toronto, Ontario, Canada at the Delta Chelsea Hotel

Pre-Meeting Symposium to the 2008 ACRM-ASNR Joint Educational Conference, October 15-19, 2008.

What is the symposium about?
This symposium will examine the construct of participation and its measurement, and nurture the development of an international consortium on the measurement of this important outcome by bringing together leaders in the field and establishing working groups on the key issues of participation measurement: conceptualization, operationalization, environmental influences, and personal characteristics.

Who should attend?
Researchers and clinicians interested in exploring the development and application of participation measures. Registration is limited to 200.

What will I learn?
The objectives are to define and discuss the state-of-the-art in the measurement of participation, as well as its utility as an outcome measure for individuals with physical and cognitive disabilities who receive rehabilitation services.

Who are the faculty?
Confirmed faculty include Rita Bode, PhD; Jennifer Bogner, PhD; Margaret Brown, PhD; Alarcos Cieza, PhD, MPH; Marcel Dijkers, PhD; Joy Hammel, PhD; OTR/L, Allen Heinemann, PhD; Alan Jette, PhD; Don Lollar, EdD; Susan Magasi, PhD, OTR; Trudy Mallinson, PhD, OTR/L, NZROT; Mary Ann McColl, PhD; Tim Muzzio, PhD; Luc Noreau, PhD; Marcel Post, PhD; Carolyn Schwartz, ScD; David Tulsky, PhD; Gale Whiteneck, PhD.

To register
Please visit http://www.acrm.org/annual_conference/ for conference information. Registration will be available online after **July 7, 2008.**

For more information about this symposium, please contact Allen Heinemann at (312) 238-2802 or a-heinemann ~at~ northwestern.edu.

Funding is provided by the National Institute on Disability and Rehabilitation Research, the Ontario Neurotrauma Foundation, the Veteran's Administration Rehabilitation Research and Development Service, and the PVA Education Foundation.

Presented by The Rehabilitation Research and Training Center on Measuring Outcomes and Effectiveness at the Rehabilitation Institute of Chicago.

Alternative Approaches to Finding Happiness

Daniel Gilbert

Harvard psychologist Daniel Gilbert (2007) explores the way people find themselves more often *Stumbling on Happiness* than successfully planning for and achieving it. Gilbert's main argument as to why we more often stumble on happiness than arrive at it deliberately follows from the firmness with which we all believe our individual uniqueness makes comparison impossible. On page 252, Gilbert cites a series of research studies showing that "the average person doesn't see herself as average." The *Lake Wobegon* effect apparently extends into almost every area of life, with most people thinking they are more intelligent, fair, attractive, skilled, etc. than average.

Gilbert offers three reasons why we think of ourselves as uniquely special: 1) because we know ourselves so much better than we know anyone else; 2) because we value individuality and are uncomfortable with too much conformity; and 3) because we focus more on the interesting features that set individuals apart from others than we do on what everyone has in common.

We are so tuned in to differences, and we blow them so wildly out of proportion relative to what we have in common, that we wind up unable to learn as much from others' experiences as we ought to. The book's key point comes on pp. 255-6, where Gilbert says [my emphasis]:
"Our mythical belief in the variability and uniqueness of individuals is the main reason why **we refuse to use others as surrogates.** After all, surrogation is only useful when we can count on a surrogate to react to an event roughly as we would, and if we believe that people's emotional reactions are more varied than they actually are, then surrogation will seem less useful to us than it actually is. The irony, or course, is that surrogation is a cheap and effective way to predict one's future emotions, but because we don't realize just how similar we all are, we reject this reliable method and rely instead on our imaginations, as flawed and fallible as they may be."

The Afterword of the book touches on the key issues, too, addressing "a formula for predicting utility" (p. 262) after introducing Daniel Bernoulli's ideas on the probabilistic estimation of utilities. Gilbert concludes that we are left dependent on our fallible imaginations for predicting future happiness.

Benjamin Wright

Gilbert could have reached a far different conclusion if his research had been pushed so far as to have found Wright (1997), which traces developments from Daniel Bernoulli's father, Jacob. Wright makes two relevant points. First, any measurement worthy of the name has to produce the same results no matter which particular instrument is used to measure the construct of interest. That is, we have to a) be able to conceive of any given collection of statements concerning a coherent domain of utilities, for instance, as representing the entire universe or population of all possible ways of articulating that domain, and then b) show that the same measures are in fact produced by different collections of those statements.

Demythologizing ...

Gilbert's overall point as to our unwillingness to rely on surrogates for information on the choices likely to make us happy stems from the fact that a) and b) are so rarely undertaken in psychology and social science. Everyone is using different words, phrases, and languages to talk about the same thing, and we focus on widely different ranges of the overall continuum of less and more utility. We naturally assume, as Gilbert says, that the myth of variability and individual uniqueness makes it impossible to apply a variation on the Golden Rule (Fisher, 1994 in RMT 7:4) and so take a surrogate's sense of what's good for them as an analogy for what's good for us.

A properly constituted economic science, however, builds on proven instances in which a) and b) hold, which leads to Wright's second point, namely, that our goal in science is to learn from the data we have in order to make inferences about data we don't have. A measuring instrument is a tool that embodies a formula for predicting utility. **What we would need to do is research into the differences between what we say we want, what we objectively get, and what we subjectively experience.** First, we would establish that the differences exist and in what forms. Second, we would measure those differences, and third, we would study variation in the differences by various demographics. The results would be the information we need to trust surrogates and let go of the myth of incomparable individual variability.

The existence of a formula for predicting utility will not result in simplistic or obvious recommendations for choices any more than it will result in unidimensional reductions of individual uniqueness to homogenized sameness. Anyone who has much experience at all with test, survey, or assessment data has likely been struck by the fact that good model fit in no way entails some kind of rigid conformity with an externally imposed standard. Rather, natural laws of human behavior are defined by and emerge from within the behaviors themselves.

There has never been greater potential for the emergence of a science of psychology capable of bringing useful technologies to bear on the life problems of everyday people. But as long as even the Harvard psychologists studying those problems themselves buy into the myth of the variability and uniqueness of individuals, we will not see much progress in the direction of using others' experiences as surrogates for our own.

William P. Fisher, Jr.

Gilbert, Daniel. (2007). Stumbling on Happiness. New York: Vintage.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16(4), 33-45,52.* www.rasch.org/memo62.htm

Les informo que se está organizando
por el IUDE el **3er WORKSHOP** sobre

Modelos De Rasch En Administración De Empresas

para el **10 de noviembre de 2008**.

Los investigadores que deseen participar pueden enviar sus trabajos, en español o inglés, a la atención de la Comisión Científica del Workshop en iude ~at~ ull.es, antes del **15 de septiembre de 2008**, indicando a qué sesión se dirigen.

Las sesiones previstas son: Metodología; Dirección y Estrategias Empresariales; Comercialización e Investigación de Mercados; Sistemas y Tecnologías de la Información; Organización de Empresas, Cultura Estratégica y Recursos Humanos; Sectores y Nuevos Desarrollos.

Los trabajos admitidos está previsto sean publicados en una monografía por la Fundación Canaria para la Formación y el Desarrollo Empresarial (FYDE CajaCanarias) en su colección de E-Books.

La asistencia al Workshop es libre y sin gastos, previa inscripción en iude ~at~ ull.es . La información relativa al III Workshop estará disponible en la página web del IUDE http://www.iude.ull.es/

Research Collaboration

Drs. Carl Granger and Paulette Niewczyk are interested in working with instrument developers who may wish to **collaborate in tool construction, refinement and dissemination for practical uses.**

We and the staff of the Uniform Data System for Medical Rehabilitation (UDSmr) and the Center for Functional Assessment Research (CFAR) specialize in transforming health-related functional status data into Rasch-derived measures. These are used in practice environments for clinical research and to guide practice, utilization management, quality of care assessment and improvement, patient education and examination credentialing, and treatment planning.

We employ Rasch modeling techniques to build functional assessment instruments and to compare new tools with existing, standardized tools. We have extensive experience with inpatient and outpatient measures covering physical functioning, pain experience, quality of daily living, mood, social interaction, and spirituality for adults and children with neurological and musculoskeletal disorders and other disablement conditions.

Please contact: Paulette Niewczyk, MPH, PHD at pniewczy ~at~ udsmr.org or Carl Granger, MD at cgranger ~at~ udsmr.org for more information.

El Instituto Universitario de la Empresa (IUDE)
de la Universidad de La Laguna (Tenerife-España)
www.iude.ull.es
comunica que ha publicado **el libro:**

Modelos De Rasch En Administración De Empresas. Aplicaciones Avanzadas

Coordinado por el Profesor Dr. Jaime Febles, IUDE-Universidad de La Laguna, España

La edición ha sido financiada por la fundación canaria FYDE-CajaCanarias que atenderá peticiones de ejemplares hasta agotar las existencias. **Los interesados pueden solicitar un ejemplar, libre de cargos, a:** administracion ~at~ fyde-cajacanarias.es

La Laguna, 3 de abril de 2008
Profesora Dra. Isabel Montero-Muradas
Directora del IUDE - ULL

Contenido:
Prologo.
Introducción: Los Modelos de Rasch en Administración de Empresas.
El Modelo de Rasch como herramienta para obtener una única prioridad entre varias.
Configuración estratégica a partir del rendimiento ambiental: Un análisis aplicado a la empresa hotelera canaria.
Visión de los responsables de las tecnologías de la información sobre sus categorías y usos en las Pymes.
Los valores subyacentes de la cultura estratégica de las Pymes canarias y su adecuación a los objetivos estratégicos, el estilo de dirección y las dimensiones organizativas.
La capacidad de relacionarse con el cliente en las empresas alojativas del turismo rural de Tenerife, según el tipo de alojamiento.
La comunicación en la cadena de suministros agroalimentaria en Canarias.
La presencia en Internet de las empresas canarias: Una aproximación mediante análisis estadístico y el Modelo de Rasch.
Marketing interno: Calidad de vida laboral y remuneración. Un análisis efectuado bajo la aplicación del Modelo de Rasch.
Media y análisis de la fidelidad del turista a un destino mediante el Modelo de Rasch.
Análisis de la cooperación entre artistas y galerías en el mercado del arte.
Benchmarking estratégico entre los destinos turísticos de Tenerife: Análisis de los factores que determinan su capacidad de atracción.
La coordinación proveedor-industria agroalimentaria en la implantación de un sistema de trazabilidad hacia atrás.
Los análisis *Rack* y *Stack* del dinamismo del entorno.

Ratios and Meaningfulness in Measurement

Back in 1993, a writer on the STAT-L discussion list posted comments that included the following paragraph:

"Measurement level must be considered to avoid making meaningless statements. A typical example of a meaningless statement is the claim by the weatherman on the local TV station that it was twice as warm today as yesterday because it was 40 degrees Fahrenheit today but only 20 degrees yesterday. Fahrenheit is not a ratio scale, and there is no meaningful sense in which 40 degrees is twice as warm as 20 degrees. It would be just as meaningless to compute the geometric mean or coefficient of variation of a set of temperatures in degrees Fahrenheit, since these statistics are not invariant or equivariant under change of origin. There are many other statistics that can be meaningfully applied only to data at a sufficiently strong level of measurement."

A similar statement was made by Roberts (1985, p. 312):

"To illustrate the definition [of measurement as meaningful when invariant across changes in scale], we note that it is meaningful to assert that I weigh more than the elephant in the zoo. This is meaningful because it is false under all acceptable scales of weight. Meaningfulness is not the same as truth. On the other hand, it is meaningless to assert that the temperature of this room is twice the temperature outside. For this might be true under one scale of temperature, e.g., Fahrenheit, while false under another scale, e.g., Centigrade. Yet, it is meaningful to assert that I weigh twice as much as the elephant. For if this statement is true in pounds, it is also true in grams, kilograms, etc."

Both of these statements have treated the arbitrary origins of temperature scales as though they are absolute origins. Roberts explicitly compares the absolute origin of weight measures (all scales start from no weight) with the arbitrary origins of different temperature scales (only degrees Kelvin starts from the theoretical absolute origin of no temperature).

Let's take the STAT-L writer's statement that "there is no meaningful sense in which 40 degrees is twice as warm as 20 degrees." It is obviously true that just because 40 degrees Fahrenheit is twice 20 degrees Fahrenheit, we cannot expect the associated Celsius measures (4.44 and -6.67) to be in the same relation. Where we divide 40 by 20 and get 2 in Fahrenheit, we divide 4.44 by -6.67 and get -.67 in Celsius.

But this procedure is misconceived. Measures are not just numbers but representations of amount. The statement that today is twice as warm as yesterday because today it is 40 and yesterday it was 20 is meaningful because the amount of temperature represented by the difference between 0 and 20 degrees temperature is indeed half of what is represented by the difference between 0 and 40, and that ratio difference will indeed remain constant across any scales that actually measure temperature.

For instance, the differences between the relevant associated Celsius measures are as follows:

0 F is -17.78 C.
20 F is -6.67 C.
40 F is 4.44 C.

Then,

-6.67 - (-17.78) = 11.11.
4.44 - (-17.78) = 22.22.

And

22.22 / 11.11 = 2.

Just as

40 / 20 = 2.

Temperature is invariant or equivariant under changes of origin, and so is measured on an equal-interval ratio scale. Ratio relationships such as 'twice-as' have to apply as much to the attribute being measured as they do to the numbers. A temperature of 40 degrees is twice that of 20 degrees no matter which scale it is measured on.

It would seem that familiarity with a difference model of measurement, such as Rasch's, can lead to psychometric lessons on measurement for thermometrics. For more on the lessons on measurement that thermometry can offer psychometrics, see Choppin (1985).

If we choose a suitable "origin" for our current purposes, such as "sea level" for measuring the height of mountains, then "twice as high" is definitely meaningful. If 72°F (22°C) is "comfortable", then 82°F (27°C) is "hot", and 92°F (32°C) is "twice as hot" relative to our "origin" of "comfortable". This is exactly how to choose origins for Rasch measures, so that ratio statements become meaningful, e.g., "Mary is twice as able as Joe relative to the difficulty of the test".

William P. Fisher, Jr., Ph.D.
Avatar International Inc.

*William Fisher has been selected as an outstanding reviewer for **Quality of Life Research Journal**'s 2007 editorial year.*

Choppin, B. H. L. (1985). Lessons for psychometrics from thermometry. Evaluation in Education, 9(1), 9-12.

Roberts, F. S. (1985). Applications of the theory of meaningfulness to psychology. Journal of Mathematical Psychology, 29, 311-32.

Rasch Measurement Transactions
P.O. Box 811322, Chicago IL 60681-1322
www.rasch.org/rmt
Editor: John Michael Linacre
Copyright © 2008 Rasch Measurement SIG
Permission to copy is granted.
SIG Chair: Ed Wolfe, Secretary: Timothy Muckle
Program Chairs: Dimiter Dimitrov & Diana Bernbaum
SIG website: http://www.raschsig.org/

Assessing Psychiatric Patient Self-Awareness Behavior with Many-Facet Rasch Analysis

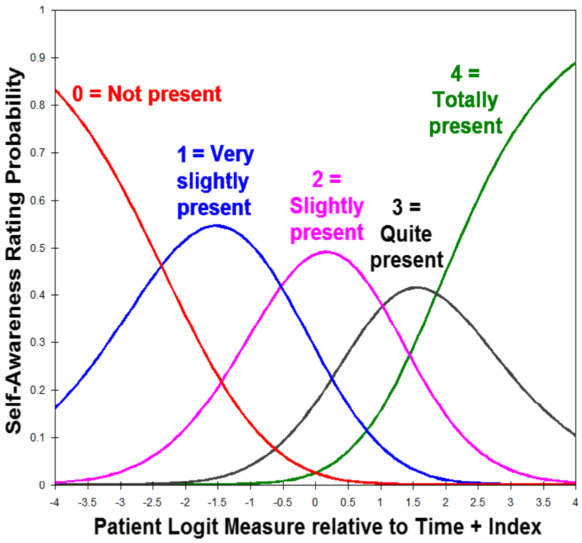


Figure 1. Category probability curves.

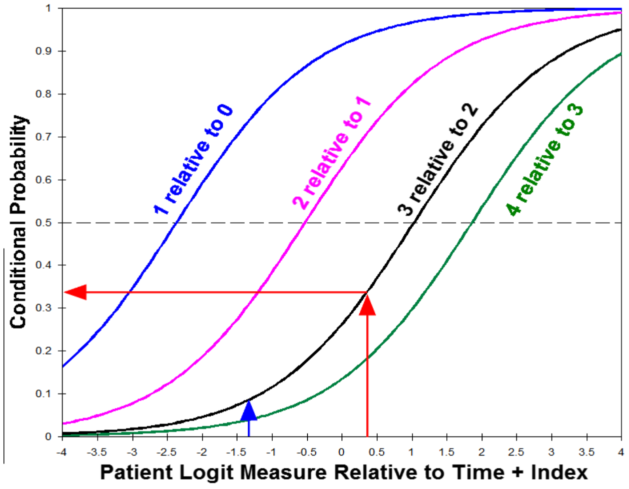


Figure 2. Adjacent-category probability curves.

The Many-Facet Rasch Model (MFRM) has the great advantage for clinical practice that it allows the practitioner not only to examine and assess the patient behavior patterns, but also to analyze the patient behavior on different occasions. In this study, building on previous research (e. g. Rangell, 1981; Markova & Berrios, 1995) and on psychiatric practice, the Psychiatric Patient Self-Awareness (PPSA) behavior is identified through five self-awareness indexes, namely: 1. *Request* (the patient decides autonomously to ask for help); 2. *Autonomy* (the patient is aware of his health status); 3. *Content* (the reasons why a request for help is advanced); 4. *Relations* (the patient is able to communicate with the others); 5. *Context* (the patient is aware of the context where he is acting).

The patients are 48 Italian adult females, mean age 50.29; their school levels are low (65%), medium (30%) and high (5%) . All patients are evaluated by a team of experts (psychiatrists and psychologists) at two successive occasions (time 1: the medical team visits the patient at his/her arrival at the medical center; time 2: the patient is revisited after a period of time which can vary from 20 days to 88 days). At each time-point, the experts rate each patient on the five indexes using a self-awareness rating scale.

There are three facets in the model: 1. patient (48), 2. time-point (2) and 3. index (5). The analysis produced patient measures, index calibration measures on a hypothesized PPSA behavior variable, and time-point measures. The infit statistics and the outfit statistics are satisfactory for time-points, also for the index measures, except for index 1 (request) for which the mean-square fit statistics are slightly above the upper criterion of 1.30 (RMT 8:3, 370 - http://www.rasch.org/rmt/rmt83b.htm). The majority of the patient measures fit statistics are also satisfactory.

Figure 1 shows the category probability curves for the self-awareness rating scale (0 = not present, 1 = very slightly present, 2 = slightly present, 3 = quite present, 4 = totally present) according to the Andrich rating-scale model. These curves indicate that the experts were able to discriminate the category hierarchy. Figure 2 shows the category relationships, but depicted as the probabilities of the higher ratings in each pair of adjacent categories of the rating scale. These have the form of the familiar Rasch dichotomous logistic ogives. The pairwise ogives have probability 0.5 at the Rasch-Andrich thresholds, where the adjacent categories are equally probable.

The Table shows the five index measures, the mean raw ratings received by the patients on the Likert scale, the corresponding patient measures and the time measures. Each cell contains the probability of presenting an index which is rated 3 by the experts, relative to a rating of 2, given the index's calibration on the PPSA variable, the overall measure for the patient and the relevant time-point measure.

Table 1, based on Figure 2, is a useful tool for psychiatric assessment of PPSA behavior. Suppose that at time 1 (0.24 logits), a patient with an overall self-awareness mean rating of 1.7 (patient measure = -.33), is rated on index 1 (-0.88 logits) in category 3 ("quite present") . Then the combined measure is -.33 - (0.24 + -0.88) = 0.31. This corresponds to a pairwise probability of 0.33 (arrows in plot, and bold cells in Table). This rating of the index has to be considered quite usual because its relative probability of occurrence on the PPSA variable is rather high (p = .33). But the same cannot be said when a patient rated 3 on index 4 (i.e. the patient is able to communicate with the others) because this has a low probability (p = .08) at a mean score of 1.7 (-1.4 logits in Figure 2).

Table 1. Probability of Self-awareness Indexes calculated for a rating of 3 (Quite Present) relative to a rating of 2 (Slightly Present) on the self-awareness rating scale: 0 1 2 3 4

Patient's average rating (on 5 indexes at 2 time-points) Range: 0-4	**Patient** measure logits	**Time 1** = .24 logits					**Time 2** = -.24 logits				
		+ Index calibration measure in logits					+ Index calibration measure in logits				
		Request 1	Autonomy 2	Content 3	Context 5	Relations 4	Request 1	Autonomy 2	Content 3	Context 5	Relations 4
		-.88	-.61	.32	.34	.83	-.88	-.61	.32	.34	.83
3.3	2.16	.85	.82	.64	.63	.52	.90	.88	.74	.74	.63
3.2	1.97	.83	.79	.59	.59	.47	.89	.86	.70	.70	.59
3.1	1.79	.80	.76	.55	.54	.43	.87	.83	.66	.66	.54
3.0	1.63	.77	.72	.51	.49	.38	.85	.81	.63	.62	.50
2.9	1.47	.75	.69	.47	.46	.35	.83	.78	.59	.58	.46
2.8	1.31	.71	.66	.43	.43	.31	.80	.78	.55	.54	.42
2.7	1.16	.68	.63	.39	.39	.28	.78	.73	.51	.51	.39
2.6	1.01	.65	.59	.36	.35	.25	.75	.70	.48	.47	.35
2.5	.87	.63	.55	.33	.32	.23	.72	.67	.44	.44	.32
2.4	.72	.58	.52	.30	.29	.20	.69	.63	.40	.40	.29
2.3	.58	.55	.48	.27	.26	.18	.66	.60	.37	.37	.26
2.2	.43	.51	.44	.24	.23	.16	.63	.56	.34	.33	.23
2.0	.13	.44	.37	.19	.18	.12	.55	.49	.27	.27	.18
1.9	-.02	.40	.34	.17	.16	.11	.52	.45	.24	.24	.16
1.8	-.17	.36	.30	.15	.14	.09	.48	.41	.22	.21	.14
1.7	**-.33**	**.33**	.27	.13	.13	.08	.44	.38	.19	.19	.12
1.6	-.49	.29	.24	.11	.11	.07	.40	.34	.17	.16	.11
1.5	-.65	.26	.21	.10	.09	.06	.36	.30	.15	.14	.09
1.4	-.82	.23	.19	.08	.08	.05	.33	.27	.13	.12	.08
1.3	-1.00	.20	.16	.07	.07	.04	.29	.23	.11	.11	.07
1.2	-1.18	.17	.14	.06	.06	.04	.25	.20	.09	.09	.06
1.1	-1.37	.15	.12	.05	.04	.03	.22	.18	.08	.08	.05
0.9	-1.77	.10	.08	.03	.03	.02	.16	.12	.05	.05	.03
0.7	-2.23	.09	.05	.02	.02	.01	.11	.08	.03	.03	.02

At time-point 2, the probability of a higher relative rating is always higher than at time 1 in accordance with the difference between the time-point measures, 0.48 logits.

The combination of Rasch analysis and expert clinical knowledge allows us to predict clinical diagnosis of PPSA behavior. Further the inclusion of a time-point facet enables us to investigate and diagnose patient behavior longitudinally, which is helpful in patient treatment and predicting the usage of clinical resources.

Stefania Mannarini
University of Padova - Dept of General Psychology

Renato Lalli
Casa di Cura Parco dei Tigli –Teolo (Pd)

Markova, I.S. & Berrios G.E. (1995). Insight in clinical psychiatry. A new model. *The Journal of Nervous and Mental Disease*, 183, 12, 743-751.

Rangell, L. (1981). From Insight to Change. *Journal of American Psychoanalytic Association*, 29, 119-141.

Nursing Treatment Matches the Rasch Model

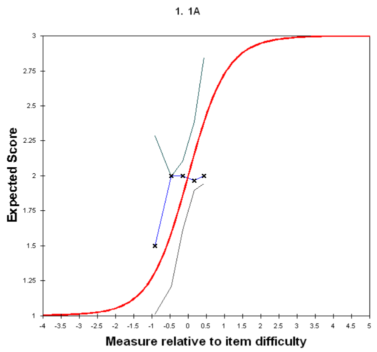
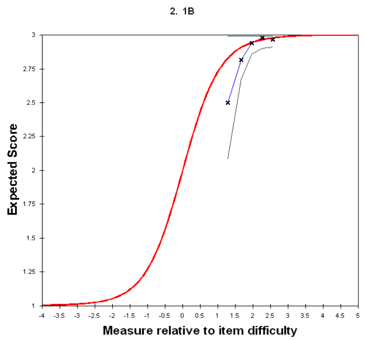
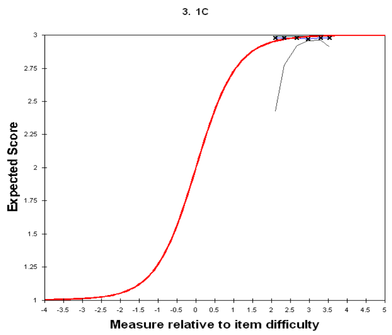
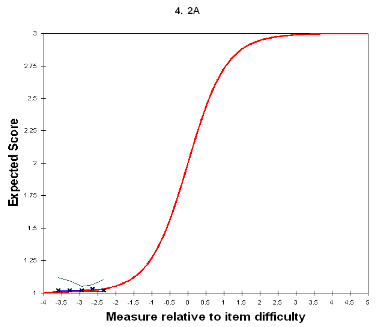
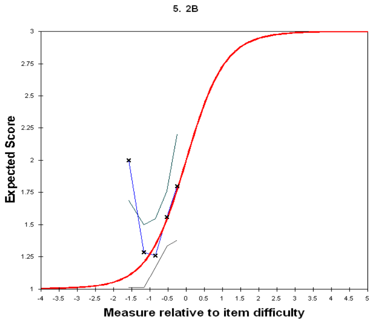
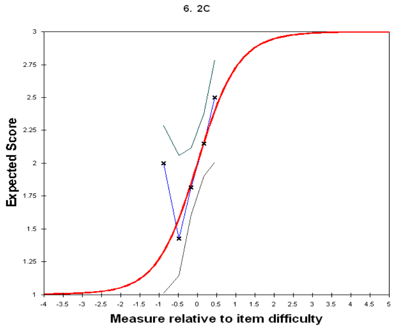
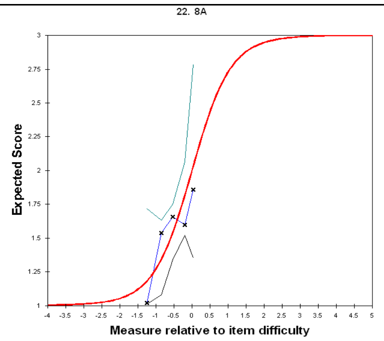
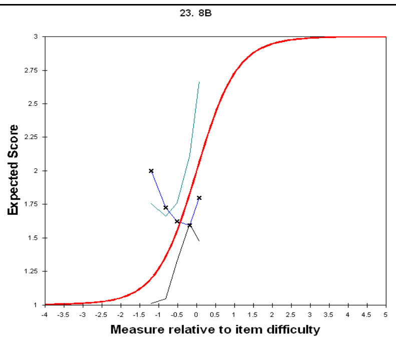
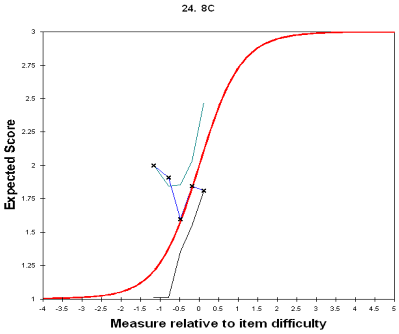
A research project to define the best nursing practice (while taking care of cardiovascular post-surgery patients) is focused on the measurement of their progress 24, 48, and 72-92 hours after leaving the intensive care unit.

These measures are expected to discriminate between traits resulting from the natural improvement of the health status of the patient and other traits that are not improving and that require special nursing care or medical treatment. A 44-item questionnaire including two or three categories has been developed in Colombia and has been administered to approximately 250 patients over a one-year term. This questionnaire includes 23 items of clinical events in the main body systems (neurological, cardiovascular, respiratory, and skin, among others) and 21 diagnosis items (mainly regarding urine and blood laboratory results and X-Rays). The nurse administers the questionnaire at three different time-points in order to study the progress of a patient. On the two- or three-category rating scale (from "low to high" or "poor to good" health condition) lower categories indicate poor condition of the patient, while higher categories indicate normal health conditions.

Analysis proceeds in tow stages: (a) improvement of the health status of the patient from 0 to 96 hours (the measured trait should show higher values at the end of the period), and (b) identifying critical variables where this progress does not occur or when an irregular condition is found, requiring the intervention of a nurse to help the patient reach a better health level.

Plots showing the model and empirical ICCs were used. The results for three items at the three different time-points are shown. The red continuous line is the Rasch-model prediction. The blue line with x's are the empirical patient statuses. The thinner grey lines are confidence intervals around the model predicted line.

Item 1 – "Conscious level of the patient": The patients show positive progress from A to B to C (x's ascending left-to-right) following the Rasch model prediction very closely. The lowest conscious level takes place during time-point A, higher at time-point B, and all patients are in the best condition at time-point C. For treatment related

Item	A=24 hours	B=48 hours	C=72-96 hours
1			
2			
8			

to this item, the participation of a nurse is only required to check the condition of the patient.

Item 2 – "Sleep and rest": At time-point A, about 98% of the patients show an irregular condition of sleep and rest; at time-point B (x's very low), 77% of the patients continue experiencing trouble. Misfitting x's can be seen in the patients with lower measures. These indicate an unexpectedly healthy condition on this item, so worth investigating further. At time-point C, 49% of the patients show no noticeable improvement with regard to time-point B (although no irregularity is found yet). Treatment related to this item requires the nurse to help the patient, if necessary, after 48 hours.

Item 8 – "Blood pressure": This is critical to the well-being of the patients, but the progress of the patients does not reach the regular level during the whole period. Irregularity in time-points B and C, as well as no significant changes in the measures, show that blood pressure should be closely supervised by the nurse and even by the physician from the start. Misfit (x's below expectation) identifies patients who have a deficit in blood pressure. Clinical intervention is needed.

Agustin Tristan (Instituto de Evaluación e Ingeniería Avanzada, S.C., Mexico), Claudia Ariza, Doctorate candidate, and Maria Mercedes Duran, Ph.D., Universidad Nacional de Colombia.

Rasch Measurement SIG, AERA Annual Report 2007-2008

At the Rasch Measurement SIG Business Meeting at the AERA Annual Meeting, New York, 2008, Secretary/Treasurer Ed Wolfe summarized the current status of the SIG.

Membership: There are currently 197 members, an increase of 59 in the last year with 87 lapsed memberships.

Finances: The January 2008 balance of the SIG account is $8,649.10, an increase of $2,281.95 since January of 2007.

Officers: Ed Wolfe was nominated (unopposed) and appointed to the office of Chair of the SIG, and Timothy Muckle was nominated (unopposed) and appointed to the office of Secretary/Treasurer. Dimiter Dimitrov and Diana Bernbaum were appointed as the SIG's 2009 Program Co-Chairs. Mike Linacre will continue as the Editor of *Rasch Measurement Transactions.* William Fisher was appointed to an *ad hoc* committee to explore the development of SIG-sponsored awards.

Invitation: William Fisher introduced the idea of developing **SIG-sponsored awards.** Anyone interested in being a member of this ad hoc committee should contact Ed Wolfe (edwolfe ~at ~ vt.edu).

IRT and The Rasch Model in Disguise

"The three-parameter logistic (3PL) model (...) was used to analyze item responses on the multiple choice items. For analysis of the constructed response items, the two-parameter partial credit model (2PPC) (...) was used." (pp. 8-9)

"In IRT, all the item characteristic curves for the items on a test can be added together to yield a function - the test characteristic curve (TCC) - that shows the expected raw score for each given scale score. By inverting the TCC, an expected scale score can be computed for each raw score. This new function - the inverse of the TCC - can be summarized in an RS-SS table. An advantage of RS-SS tables is that they make scoring relatively straightforward: With number-correct scoring, it is sufficient to know how many raw score points a student obtained on the test to determine a student's scale score." (p. 19)

New York State Testing Program. Mathematics Grade 8. Technical Report 2003.
http://www.emsc.nysed.gov/osa/inform/2003nys-gr8math-satr.pdf

Comment: "sufficient to know how many raw score points" is a criterion for the Rasch model: RMT 3:2 p. 62.
http://www.rasch.org/rmt/rmt32e.htm

By setting up a one-to-one correspondence between raw scores and scale scores, New York State approximated the Rasch model and, as they say, this is "an advantage"!

The Truth about IRT Scaling

"The [2-PL, 3-PL] theta-scale, or any linear transformation of it, however, does not possess the properties of a ratio or interval scale, although it is popular and reasonable to assume that the theta-scale has equal-interval properties"
Hambleton, Swaminathan and Rogers, "Fundamentals of Item Response Theory", 1991, p.87.

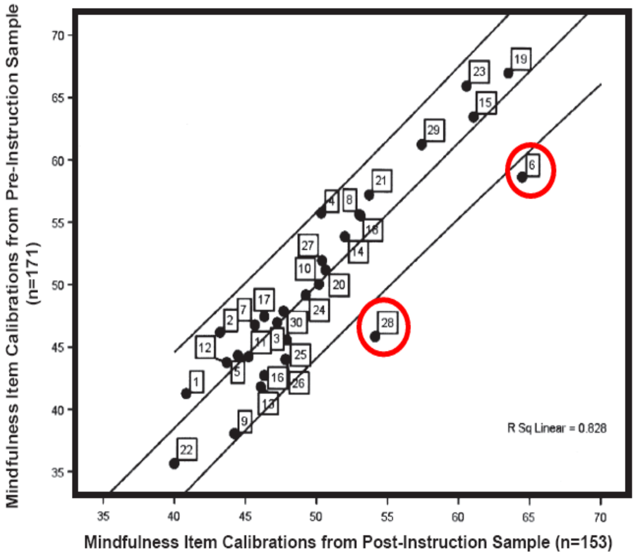
"Liking for Science" - the inside scoop!

The book "Rating Scale Analysis" (Wright & Masters, 1982) features a data set in which 75 children give their reaction to 25 items relating to science activities. Julian Mingus (Ph.D., U. Toledo, 1975) was in the Dept. of Education at Cleveland State University, Cleveland, Ohio, when he conducted the data collection. He remembers that both boys and girls were in the study. They had mixed ethnicity. Their families were of lower to middle socio-econimic status and their grade levels were probably 1st to 6th (6 yrs - 11 yrs) .

Mingus co-authored "Comparative perceptions of Elementary, Junior High, and Senior High school principals on selected work related variables." *Brent Poppenhagen, Julian Mingus, Joseph Rogus. Journal Of Educational Administration, 1980, 18, 1, 69 - 87*

One of his interests in photography:
http://www.jmingus.com/

Courtesy of Christine Fox, U. Toledo



“Omitting the items [6, 28] from the scale makes virtually no difference in the resulting measures, suggesting that there is no substantive significance to the items’ statistically significant changes in position.”

Solloway, S.G., & Fisher, W.P. (2007) Mindfulness in Measurement. International Journal of Transpersonal Studies, 26, 58-81

Case-Control Studies

“Statistically, the Rasch model we outline is equivalent to the ‘conditional logit model’ or the ‘conditional logistic regression for matched case-control groups model’, as it is referred to by statisticians and epidemiologists. It can be estimated using a conditional maximum likelihood method.”

Gautschi, T. (2001, March). Trust over time: The effects of dyadic social capital., Department of Sociology, Utrecht University, The Netherlands.
www.soz.unibe.ch/personal/gautschi/downloads/Trust over Time.pdf

“This article presents new results on the standard techniques used in analysis of data from Rasch models, and the special case of Rasch models typically used with data from matched case-control studies.”

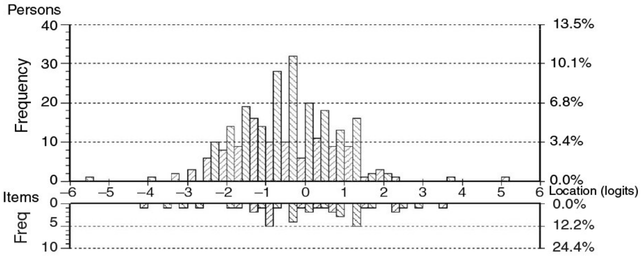
Rice, K. M. (2004). Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications. *Journal of the American Statistical Association, 99(466), 510-22.*

Courtesy of William P. Fisher, Jr

Misfitting Observations

“When you have something simple that agrees with all the rest of physics [i.e., all the rest of the latent trait] and really seems to explain what is going on, a few experimental data against it are no objection whatsoever.”

Nobel Laureate Murray Gell-Mann (1969, Physics) quoted in “The Evidential Power of Beauty: Science and Theology” (Thomas Dubay), p. 115.



Person item distribution graph for Total HADS Psychological Distress Scale.

Pallant F. & Tennant A. (2007) An introduction to the Rasch measurement model : An example using the Hospital Anxiety and Depression Scale (HADS). British Journal of Clinical Psychology, 46, 1–18

Rasch-related Coming Events

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

July 4 - Aug. 9, 2008, Fri.-Fri. Practical Rasch Measurement online course, (M. Linacre, Winsteps), http://www.statistics.com/courses/rasch

July 7-8, 2008, Mon.-Tues. ASEASA Rasch Workshop (B. Sheridan, RUMM), South Africa

July 9-11, 2008, Wed.-Fri. ASEASA Evaluation & Assessment Conference, South Africa

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

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

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

Sept. 17-19, 2008, Wed.-Fri. 1st International Conference on High School and College Educational Evaluation, Veracruz, Mexico http://www.ieia.mx.com

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

Oct. 14-15, 2008, Tues.-Wed. International Symposium on Measurement of Participation in Rehabilitation Research, Toronto, Canada
http://www.acrm.org/annual_conference/Precourses.cfm

Nov. 10, 2008, Monday III Workshop “Modelos de Rasch en Administración de Empresas”, Tenerife, Spain. http://www.iude.ull.es/

April 13-17, 2009, Mon.-Fri. AERA Annual Meeting, San Diego, CA, USA, http://www.aera.net/

August, 2010 Probabilistic models for Measurement - 50 years, Conference, Copenhagen, Denmark