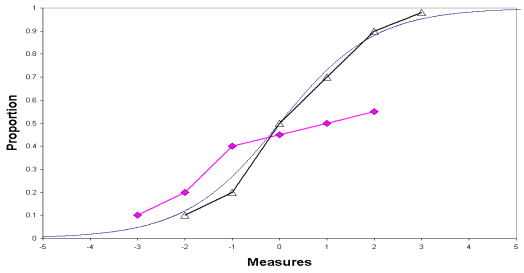
Differential Item Functioning



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 18 No. 1 Summer 2004 ISSN 1051-0796

Item Discrimination, Guessing and Carelessness:
Estimating IRT Parameters with Rasch

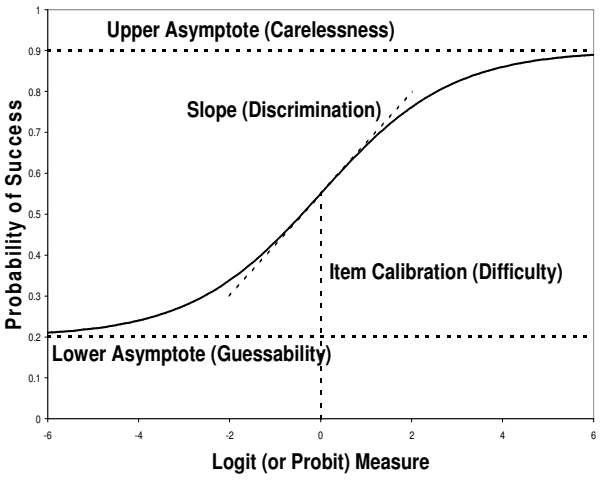
Fred Lord's three-parameter-logistic Item Response Theory (3-PL IRT) model (Birnbaum, 1968) incorporates an item discrimination parameter, modeling the slope of the item characteristic curve, and a lower asymptote parameter modeling "guessing" or, better, "item guessability". Here is a 3-PL model, written in log-odds format, with c_i as the lower asymptote, a_i as the item discrimination, θ_n as the person ability and b_i as the item difficulty:

$$\log\left(\frac{P_{ni}-c_i}{1-P_{ni}}\right)=a_i(\theta_n-b_i)$$

Lord's 4-PL model (Barton & Lord, 1981) incorporates an upper asymptote parameter for item-specific "carelessness". Here is a "carelessness" model, written in log-odds format, with d_i as the upper asymptote:

$$\log\left(\frac{P_{ni}}{d_i-P_{ni}}\right)=a_i(\theta_n-b_i)$$

Upper and lower asymptotes are notoriously difficult to estimate, so it appears that Lord abandoned his 4-PL model, and the value of c_i in the 3-PL model is, on occasion, imputed from the number of options in a multiple-choice item, instead of being estimated directly from the data. Even the estimation of item discrimination usually requires constraints, such as "a_i cannot be negative or too big."



4-PL IRT Item Characteristic Curve

The dichotomous Rasch model, however, provides an opportunity to estimate a first approximation to these parameters. These estimates can be useful in diagnosing whether the behavior they reflect could be distorting the Rasch measures. In the dichotomous Rasch model, c_i=0, d_i=1 and a_i=1. We can, however, treat the Rasch values as starting values in a Newton-Raphson iterative processed apparently intended to find the maximum-likelihood values of each of these parameters, in a context in which all other parameter values are known.

Following Wright & Masters (1982, 72-77), and using the standard approach of first and second derivatives of the log-likelihood of the data with respect to the parameter of interest, we obtain the following Newton-Raphson estimation equations for the first approximations:

Item discrimination (ICC slope):

$$\hat{a}_i = 1+\left[\frac{\sum_n (X_{ni}-P_{ni})(\theta_n-b_i)}{\sum_n P_{ni}(1-P_{ni})(\theta_n-b_i)^2}\right]$$

Table of Contents

$$\text{with S.E.} \cong 1/\sqrt{\sum_n P_{ni}(1-P_{ni})(\theta_n - b_i)^2}$$

The Rasch expectation of a_i is 1.

A corollary is that, when data fit the dichotomous Rasch model, there is zero correlation between the observation residuals and their generating measure differences.

There is a similar result for polytomous items. The Generalized Partial Credit can be written:

$$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = a_i(\theta_n - b_i - \tau_{ij}) \qquad \text{j=1,m}$$

The "generalized" item discrimination (ICC slope) is:

$$\hat{a}_i = 1 + \left[\frac{\sum_n \left(M_{niX_{ni}} - \sum_{k=1}^{m} P_{nik} M_{nik}\right)}{\sum_n \left(\sum_{k=1}^{m} M_{nik}^2 P_{nik} - \left(\sum_{k=1}^{m} M_{nik} P_{nik}\right)^2\right)}\right]$$

where $M_{nik} = k(\theta_n - b_i) - \sum_{j=1}^{k} \tau_{ij}$ and $M_{ni0} = 0$

And for the discrimination of polytomous inter-category "generalized" thresholds:

$$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = a_{ij}(\theta_n - b_i - \tau_{ij}) \qquad \text{j=0,m}$$

the "generalized" threshold discrimination is:

$$\hat{a}_{ij} = 1 + \left[\frac{\sum_{X_{ni} \geq j}(\theta_n - b_i - \tau_{ij}) - \sum_n (\theta_n - b_i - \tau_{ik}) \sum_{k=j}^{m} P_{nik}}{\sum_n (\theta_n - b_i - \tau_{ik})^2 \left(\sum_{k=j}^{m} P_{nik} - \left(\sum_{k=j}^{m} P_{nik}\right)^2\right)}\right]$$

Returning to the dichotomous model:

the lower asymptote (guessability) is:

$$\hat{c}_i = \left[\frac{\sum_{n,x=1} e^{-(\theta_n - b_i)} - \sum_{n,x=0} 1}{\sum_{n,x=1} e^{-2(\theta_n - b_i)} + \sum_{n,x=0} 1}\right]$$

where $0 \leq c_i \leq 1$

$$\text{with S.E.} \cong 1/\sqrt{\left(\sum_{n,x=1} e^{-2(\theta_n - b_i)} + \sum_{n,x=0} 1\right)}$$

The Rasch expectation of c_i is 0.

The upper asymptote (carelessness) is:

$$\hat{d}_i = 1 - \left[\frac{\sum_{n,x=0} e^{(\theta_n - b_i)} - \sum_{n,x=1} 1}{\sum_{n,x=0} e^{2(\theta_n - b_i)} + \sum_{n,x=1} 1}\right]$$

where $0 \leq d_i \leq 1$

$$\text{with S.E.} \cong 1/\sqrt{\left(\sum_{n,x=0} e^{2(\theta_n - b_i)} + \sum_{n,x=1} 1\right)}$$

The Rasch expectation of d_i is 1.

John Michael Linacre

Birnbaum A. (1968) Some latent trait models and their uses in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores (pp. 395-479). Reading, MA: Addison-Wesley.

Barton M.A. & Lord F.M. (1981) An upper asymptote for the three-parameter logistic item-response model. Princeton, N.J.: Educational Testing Service.

IOMW XII Activities and Presentations
June 28 – July 2, Cairns, Australia

Monday, June 28

Afternoon: Winsteps and Facets, Mike Linacre

Tuesday, June 29

Morning: RUMM, Barry Sheridan

Afternoon: ConQuest, Margaret Wu & Mark Wilson

Wednesday, June 30

Vice-Chancellor Bernard Moulden of James Cook University, introduced by Trevor Bond
Opening Address

Brent Michael Duckor
Knowing what we know about educational measurement knowledge: a case for developing measures in the classroom and in the field

Judy R. Wilkerson & William Steve Lang
Measuring teacher dispositions with different item structure: an application of the Rasch model to a complex accreditation requirement

Magdalena M. C. Mok, Hazel M. Y. Lam & S. E. Audrey Lim
Validation of scores from Early Literacy Development Checklist for kindergarten students from Hong Kong using Rasch measurement

A. Jackson Stenner
Does the reader comprehend the text because the text is easy or the reader is able?

Mark Wilson
On choosing a model for measuring. Part II: polytomous data

Noor Lide & Abu Kassim
Scaling of grammatical structures: implications for assessment and instruction

William Steve Lang & Judy R. Wilkerson
Measuring College sailing teams ability: an application of the many-facet Rasch model to ordinal data

Gage Kingsbury
A comparison of state proficiency levels

James Sick
Assessing willingness to communicate in a second language

Robert Cavanagh, Joseph Romanoski & Russell Waugh
The design, implementation and evaluation of a research methodology utilizing deterministic and probabilistic analytic techniques

Symposium 1: Rasch Measurement: Issues of model fit
Richard M. Smith: Assessing the fit of data to the family of Rasch measurement models: Interpreting residual-based fit statistics
J. Michael Linacre: Principal components factor analysis of item and person residuals
Barry Sheridan: Using item parameter invariance to formalize and account for DIF

Thursday July 1

Mark Wilson & Nathaniel Brown
Measurement as struggle

Andrew Kyngdon
The Rasch model from the perspective of the representational theory of measurement.

Mark H. Moulton
One use of a Non-Unidimensional Scaling (NOUS) model: transferring information across dimensions and subscales

Wen-Chung Wang & Mark Wilson
The Random-Effects Rasch model

Trevor Bond
Is Rasch modeling just the new factor analysis?

J. Michael Linacre & Benjamin D. Wright
Predicting rating scale category usage and measurement implications

David D. Curtis & Peter Boman
The identification of misfitting response patterns to, and their influences on the calibration of, attitude survey instruments

Jean-Guy Blais, Nathalie Loye & Gilles Raîche
A four years study of a 20-items scale program evaluation questionnaire with the Rating Scale model

Juho Looveer
Using Rasch analyses with satisfaction surveys to assess change

Rassoul Sadeghi & Jim Tognolini
Rasch model and equating: practical issues

Ainol Madziah Zubairi, Noor Lide & Abu Kassim
Equating tests across alternate forms and different cohorts over time: the IIUM experience

Juho Looveer
Theoretical implications of equating methodologies

Symposium 2: Rasch Measurement: Test equating
Peter Congdon & Magda Lees: Impact of student engagement in equating tests at three year levels on cohort performance measures
Renee Chow & Peter Congdon: Writing performance - online versus pen-and-paper format
Cathy Boldiston & Peter Congdon: Equating state-wide test results
Rassoul Sadeghi & Jim Tognolini: Rasch Model and equating: practical issues
Ainol Madziah Zubairi & Noor Lide Abu Kassim; Juho Looveer.

Friday, July 2

William Fisher
Relational networks and trust in the measurement of social capital

Pedro Alvarez
Detecting diet by food-purchasing habit

Pedro Alvarez & Jorge M. S. Honorio.
Measuring the monetary and financial capacity of Portuguese regions.

Pedro Alvarez, Rafael De Reyna Zaballa & Julio García Del Junco
Can entrepreneurship be measured?

Thomas Salzberger
A Rasch analysis of customer satisfaction as a latent construct in consumer research

Pedro Alvarez, Carlos García-Zorita & Elias Sanz-Casado.
Formulating a measure: a case study.

Sun-Geun Baek & Hyesook Kim
The relationship between classroom teachers' judgments and fit statistics of the Partial Credit model

Robert W. Massof
Dimensions of functional ability in low vision

Ching-Lin Hsieh, Wen-Chung Wang, Ching-Fan Sheu & Jau-Hong Lin
A Rasch analysis of a self-perceived change in the quality of life scale in patients with stroke

Curt Hagquist & David Andrich
Optimal categorization of ordered items – a comparison of different response formats

Ted Brown
An evaluation of the construct validity of the Test of Visual Perceptual Skills – Revised using the Rasch measurement model

Aleksandar Baucal & Trevor Bond
Rasch measurement: zone of proximal development of the measurement of the ZPD

Jean-Guy Blais, Linda Drouin & Michel D. Laurier
Using the many-facet Rasch measurement model to develop a large-scale performance-based test of writing skills in Quebec

Kari Tormakangas
Comparing item level achievement of 8th and 9th graders using Finnish IEA Civics data

Hsueh-Chu Chen & Wen-Chung Wang
The Pronoun-drop Test for Chinese learners of English

Nordin Abd Razak
Investigating the appropriateness of a school organizational culture questionnaire for use in a multi-ethnic setting

Robert Cavanagh & Joseph Romanoski
The influences on the effective use of information and communication technology in elementary and secondary school classrooms

Teachers' Day: Friday, July 2

Juho Looveer
How measurement is fundamental to state and national evaluation systems: a nontechnical explanation

Trevor Bond
Good testing should help teachers teach: the role of immediate detailed feedback

Nathaniel Brown
The development and refinement of construct maps: what we are learning about the BEAR assessment system in the field

Judith Murphy & Barbara Dodd
Understanding children's ability to draw inferences from text: how Rasch measurement can help classroom teachers

Marie Bond
Monitoring children's classroom music performances in light of curriculum requirements

Magdalena M. C. Mok, S. E. Audrey Lim & Hazel M. Y. Lam
Rasch measurement of literacy development over 12 months of kindergarten students from Hong Kong

Rick Dills
An Oregon school district's efforts at growth-focused measurement vs. the public accountability sanctioned by No Child Left Behind

Judy R. Wilkerson, William Steve Lang & Jerome Wilkerson
Measuring teacher ability: an application of the Rasch model to teacher certification using performance measures

Martin Caust
Trusting teachers: a proposal for maximizing the value of classroom teachers' own evaluations and linking them to state-wide assessments

Jack Stenner & William Fisher
Measuring reading ability: the Lexile® framework

Jack Stenner & William Fisher
The Lexile® framework: teachers' workshop.

IOMW XII
Coordinator: Trevor Bond
Website: www.soe.jcu.edu.au/iomw2004

Thinking about Validity: The Case of Functional Assessment

The word **"validity"** has its roots in the Latin *ualere,* "to be strong". Other words sharing the same root include available, convalesce, prevail, valiant, valor, and value. A valid measure's value could well be said to reside in the strength with which it makes an intended effect or phenomenon available for examination, experimental comparison, and application. Highly valid measures robustly resist tests of their strength and persistently prevail in stable states across samples, instruments, researchers, time, space, etc. Invalid measures, then, are weak and of less value because they provide less evidence that the thing measured is what is supposed to be measured, and do not hold up when subjected to the stresses of application.

For instance, a 15mm wrench fits with a small degree of error around the head of a 15mm machine screw or bolt. The strength and value of this measure are tested by the extent to which the fit of the wrench on the bolt head (and the structural integrity of the wrench handle) provides leverage for turning the bolt and screwing it in place, or removing it. The validity and practical value of the wrench as a measure of the bolt head and of the screw's leverageable capacity to function as an inclined plane follow from the extent to which it repeatedly facilitates the production of a particular effect (torque) at the point of use. The validity and value of the theory informing the process stem from the extent to which the mathematical relations of force, mass, and acceleration can be predicted for any combination of wrench, bolt, and application, anywhere and any time.

Similarly, a functional assessment adaptively targeted in a medical rehabilitation context at 350 PAR (Physical Activity Rehabits) brings the mobility and ADL skills of a 350 PAR stroke survivor into sharp focus for the informed therapist. The strength and value of this measure are tested, in one way, by the extent to which the targeting of the assessment provides leverage for moving the stroke survivor's mobility and ADL skills higher up the PAR scale. The validity and value of the assessment as a measure of physical activity follow from the extent to which it repeatedly facilitates the production of the desired effect at the point of use. In the absence of a valid qualitative or quantitative conceptual measure of physical activity, it would be possible neither to assess how much functionality the stroke survivor possesses, nor how much, if any, change in functionality occurs over time.

It also follows, then, that the validity and value of the theory informing the process stem from the extent to which the mathematical relations of functional ability, task difficulty, rater harshness, and expected percent independent can be predicted for any combination of rehabilitation candidate, physical activity, therapist, and functional independence. In the absence of quantitative measures, theory remains mathematical to the extent that some degree of transparency in the relevant relations is obtained (Fisher, 2003). Therapists can (and routinely do), for instance, intuit whether any given patient will be able to perform any given task with a given degree of independence. Valid intuitions concerning the correspondence between patients' abilities and various task difficulties provide an initial degree of the mathematical clarity and proportionate rationality that enable a field of practice to take on a coherent identity as a community.

Locally more advanced degrees of clarity of mathematical views of functional independence provide more value to rehabilitation practitioners by providing a stronger, experimentally-based quantitative measure of constant amounts. The validity of qualitative intuitions is compromised by their variability across therapists and by the lack of a systematic frame of reference for communicating their meaning, and the same problems are associated with functional assessments that stop with the method of summated ratings (Merbitz, Morris, & Grip, 1989; Michell, 2003). Calibrated additive representations overcome these limitations by locating patients' abilities, task difficulties, and sometimes rater harshness on a common continuum capable of providing quantitative measures (among many others, see Silverstein, Fisher, Kilgore, et al., 1992; Heinemann, Linacre, Wright, et al., 1993; Fisher, Bryze, Granger, et al., 1994; Velozo, Kielhofner, & Lai, 1999). The local validity of these measures for distinguishing between various groups of rehabilitation clients and predicting the relevant level of care is well established (Harvey, Silverstein, Venzon, et al., 1992; Heinemann, Linacre, Wright, et al., 1994).

But yet more advanced degrees of such clarity and rationality have become available as different instruments intended to measure the same physical functioning construct have been shown to do so in linearly transformed versions of the same metric (Grimby, Andrén, Holmgren, et al., 1996; Fisher, 1997; Fisher, Eubanks, & Marier, 1997; Segal, Heinemann, Schall, et al., 1997; Wolfe, Hawley, Goldenberg, et al., 2000; Wolfe, 2001; Zhu, 2001), which might be termed the Rehabit (Fisher, Harvey, Taylor, et al., 1995). Evidence strongly supports the possibility that several, if not many, of the functional assessment instruments currently in use could be equated to a common reference standard. In this context, it would become possible for all users of functional assessment measures to use the same numeric language for referring to demonstrably constant amounts of more and less functionality.

Measurement validity is inherently a matter of the value of the practical consequences that follow from the application of an instrument. The full potential of integrated instruction or rehabilitation and assessment will be realized only when three steps are taken. First,

experimental assessments of instruments must focus on establishing the existence of a single one thing that adds and divides up consistently and proportionately enough to be represented by numbers (Rasch, 1960). Second, different instruments supposed to measure the same variable ought to be examined for convergence on a common construct (Fisher, 1997) and equated if the evidence supports that course of action.

Third, every class of potential and actual users of functional assessment measures, from the treatment teams and the clients to researchers, disability advocacy groups, educators, payors, accreditors, administrators, and accountants, all need to agree on basic conventions of data quality, the quantitative unit's size and range, valid applications and inferences, and systems for maintaining and improving the metric across instruments. When all three of these steps are taken, we will arrive at a system of functional metrology with the widely distributed strength and generalized value of other metrological systems, such as the one that makes it possible in principle for any metric wrench manufactured by any tool company to fit any metric bolt anywhere in the world on any hour of any day. If and when we can also arrive at a pure mathematical theory of functionometric relationships, then we will have opened the door to a new kind of scientific revolution, one like the second scientific revolution of the nineteenth century in being provoked by "the immense efficacy of quantitative experimentation undertaken within the context of a fully mathematized theory" (Kuhn, 1977, pp. 219-20).

After all, what might we expect to happen if and when everyone researching or practicing physical rehabilitation thinks about the constructs of functional assessment in a common language? What might follow from everyone repeatedly seeing the consistency with which experimentally controlled, and even everyday variations in, treatment, initial status, length of stay, etc. do, or do not, affect functional assessment measures? Research in cognitive psychology (for instance, among many others, Hutchins, 1995; Latour, 1995) suggests that we are highly likely to also see a manifestation of the collective, group-level effect characteristic of distributed thinking. The technologically-embodied cognition effected by a standardized metric gives birth to a propagation of one and the same construct through different media.

This process, and not metaphysically vapid claims about unobservable mental events, provides the only documentable evidence of representation that anyone has made available to date. Navigational charts, for instance, do not make anything observable in and of themselves. No, a tool like a chart functions only insofar as a navigator, a pilot, and the chart maker are able make features on the landscape correspond with the features on the chart en route to achieving some change in position relative to those features. In other words, a map mediates relationships between people with different perspectives, and so validly provides practical value and supports strong inferences, only insofar as it helps them get where they want to go. Insofar as maps of functional assessment variables are valid, should we expect any less strength and value from them?

William P. Fisher

Fisher, A. G., Bryze, K. A., Granger, C. V., Haley, S. M., Hamilton, B. B., Heinemann, A. W., Puderbaugh, J. K., Linacre, J. M., Ludlow, L. H., McCabe, M. A., Wright, B. D. (1994) Applications of conjoint measurement to the development of functional assessments. International Journal of Educational Research, 21(6), 579-593.

Fisher, W. P., Jr. (1997) Physical disability construct convergence across instruments: Towards a universal metric. Journal of Outcome Measurement, 1(2), 87-113.

Fisher, W. P., Jr. (2003) Mathematics, measurement, metaphor, metaphysics: Part I. Implications for method in postmodern science. Theory & Psychology, 13(6), 753-90.

Fisher, W. P., Jr., Eubanks, R. L., Marier, R. L. (1997) Equating the MOS SF36 and the LSU HSI physical functioning scales. Journal of Outcome Measurement, 1(4), 329-362.

Fisher, W. P., Jr., Harvey, R. F., Taylor, P., Kilgore, K. M., Kelly, C. K. (1995) Rehabits: A common language of functional assessment. Archives of Physical Medicine and Rehabilitation, 76(2), 113-122.

Grimby, G., Andrén, E., Holmgren, E., Wright, B., Linacre, J. M., Sundh, V. (1996) Structure of a combination of Functional Independence Measure and Instrumental Activity Measure items in community-living persons: A study of individuals with spina bifida. Archives of Physical Medicine and Rehabilitation, 77(11), 1109-1114.

Harvey, R. F., Silverstein, B., Venzon, M. A., Kilgore, K. M., Fisher, W. P., Jr., Steiner, M., Harley, J. P. (1992, October) Applying psychometric criteria to functional assessment in medical rehabilitation: III. construct validity and predicting level of care. Archives of Physical Medicine and Rehabilitation, 73(10), 887-892.

Heinemann, A. W., Linacre, J. M., Wright, B. D., Hamilton, B. B., Granger, C. V. (1993) Relationships between impairment and physical disability as measured by the Functional Independence Measure. Archives of Physical Medicine and Rehabilitation, 74(6), 566-573.

Heinemann, A. W., Linacre, J. M., Wright, B. D., Hamilton, B. B., Granger, C. V. (1994) Prediction of rehabilitation outcomes with disability measures. Archives of Physical Medicine and Rehabilitation, 75(2), 133-143.

Hutchins, E. (1995) Cognition in the wild. Cambridge, Massachusetts: MIT Press.

Kuhn, T. S. (1977) The function of measurement in modern physical science. In T. S. Kuhn, The essential tension: Selected studies in scientific tradition and change (pp. 178-224) Chicago: University of Chicago Press.

Latour, B. (1995) Cogito ergo sumus! Or psychology swept inside out by the fresh air of the upper deck: Review of Hutchins' Cognition in the Wild, MIT Press, 1995. Mind, Culture, and Activity, 3(1), 54-63.

Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V., Hamilton, B. B. (1994) The structure and stability of the Functional Independence Measure. Archives of Physical Medicine and Rehabilitation, 75(2), 127-132 [http://www.rasch.org/memo50.htm].

Merbitz, C., Morris, J., Grip, J. (1989) Ordinal scales and the foundations of misinference. Archives of Physical Medicine and Rehabilitation, 70, 308-312.

Michell, J. (2003) Measurement: A beginner's guide. Journal of Applied Measurement, 4(4), 298-308.

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

Segal, M., Heinemann, A., Schall, R. R., Wright, B. D. (1997) Extending the range of the Functional Independence Measure with SF-36 items. Physical Medicine & Rehabilitation: State of the Art Reviews, 11(2), 385-396.

Velozo, C. A., Kielhofner, G., Lai, J. S. (1999) The use of Rasch analysis to produce scale-free measurement of functional ability. American Journal of Occupational Therapy, 53(1), 83-90.

Wolfe, F. (2001) Which HAQ is best? A comparison of the HAQ, MHAQ and RA-HAQ, a difficult 8 item HAQ (DHAQ), and a rescored 20 item HAQ (HAQ20): Analyses in 2,491 rheumatoid arthritis patients following leflunomide initiation. Journal of Rheumatology, 28(5), 982-9.

Wolfe, F., Hawley, D., Goldenberg, D., Russell, I., Buskila, D., Neumann, L. (2000) The assessment of functional impairment in fibromyalgia (FM): Rasch analyses of 5 functional scales and the development of the FM Health Assessment Questionnaire. Journal of Rheumatology, 27(8), 1989-99.

Zhu, W. (2001, Jan). An empirical investigation of Rasch equating of motor function tasks. Adapted Physical Activity Quarterly, 18(1), 72-89.

Rasch Measurement SIG
From the Secretary

Dear Colleagues:
I hope that those of you who attended this year's annual AERA meeting in San Diego had a successful conference and made it back home safely. There were a number of very interesting papers presented at the sessions sponsored by the Rasch SIG. Thanks to all who attended and presented at these sessions!

ELECTIONS

Elections were held at the SIG Annual Business Meeting held during the AERA Annual Meeting, and the following officers were elected:

Chair:
Randy Schumacker, University of North Texas
Secretary/Treasurer:
Steve Stemler, Yale University
Program Chair:
Trevor Bond, James Cook University

NEWSLETTER

Mike Linacre has graciously agreed to continuing serving as the Newsletter editor. He asked me to remind SIG members that the current and back-issues of Rasch Measurement Transactions (RMT) can be downloaded from www.rasch.org/rmt/

SIG MEMBERSHIP

I would like to encourage all AERA members to join the Rasch SIG (if you haven't yet). Membership in the SIG is critically important as the number of paper presentation slots allocated by AERA is solely determined by the number of paid members of the SIG. So, in order to ensure that we have a quality program for Montreal next year, we encourage everyone to make sure that they are SIG members, and to encourage others who may be interested to join as well.

1. Log onto the members only site of AERA
   https://www.aera.net/member/index.htm
2. Once you are logged in, choose the option that says "Join a special interest group".
3. Check the box next to the Rasch Measurement SIG (83). If you have already paid your dues, the box will be grayed out so can't select it - that is one way to check your status.
4. Charge the $10 dues to your credit card using AERA's secure website, and *voila!,* you are a full-fledged SIG member.

Steve Stemler
Secretary, Rasch Measurement SIG
Assistant Director, PACE Center, Yale University
www.yale.edu/pace

Proper Measurement is "Universally Reproducible"

The late John A. Simpson was a physicist and metrologist associated with the University of Chicago, the Enrico Fermi Institute, and the US National Institute for Standards and Technology.

In the following definition of measurement, taken from the *Metrology* subject heading in the *Encyclopedia of Physics,* note that Simpson makes repeated explicit references to the concept of **quantity** without specifically invoking any tests of additivity or divisibility, though these, along with other similar tests, are indeed implicit in the concept of a **"continuous scale of magnitude."**

Simpson places far more emphasis on **common units, and common methods** of obtaining them and determining ordinal relations, than he does on knowing how and when additive relations have been established. He goes so far as to hold that "a proper measurement" is one that is "universally reproducible" "wherever and whenever the measurement process is repeated."

Psychometricians might then do well to shift some of their resources toward deployment of common units and methods for each major measurable variable, and away from the generation of ever more different units and methods. In what follows, the emphasis is mine.

William Fisher

Simpson, J. A. (1991). Metrology. In R. G. Lerner & G. L. Trigg (Eds.), Encyclopedia of physics, 2d Ed. (pp. 723-5). New York, New York: VCH Publishers, Inc.

p. 723-4: "A measurement is a series of manipulations of physical objects or systems according to defined protocols that result in a number. The objects or systems involved are test objects, measuring devices, and computational operations. The objects and devices exist in and are influenced by some environment. The value obtained is purported to represent uniquely the magnitude, or intensity, of some quantity embodied in the test object. This number is acquired to form the basis of a decision affecting some human goal or satisfying some human need that depends on the properties of the test object.

In order to attain this goal of **useful decision making,** metrology has focused on the task of assuring that the value obtained for a given quantity of a given object is **functionally identical wherever and whenever the measurement process is repeated.** Only then can all parties to the decision work from a concordant data base. Such a universally reproducible measurement is called a **proper measurement.**

An analysis of the logical conditions that must be satisfied to achieve a proper measurement shows that three independent arbitrary axioms must be universally agreed upon:

1. All parties must agree upon and have access to a **common unit** in which the results will be expressed.

2. There must be an agreed-upon physically realizable method of obtaining a **continuous scale** of magnitude based on the unit.

3. There must be an agreed-upon physically realizable method of determining when the quantity of interest, as embodied in a physical object or system, is equal to, less than, or greater than, some **fixed point** on this realized scale.

The principal activity of metrologists consists of generating, propagating, testing, and applying to an object or system of interest sets of these measurement axioms for all quantities and all useful magnitudes of those quantities…..

Fundamental to the success of such a system is the development, at each transfer [points through which the unit is traceable to the reference standard from secondary standards and the point of use], of **realistic estimates of uncertainty.**"

p. 725: "By far **the greatest activity in metrology is that performed in the service of quality control.** Manufacturing establishments of any size maintain standards laboratories and/or metrology laboratories. The laboratories maintain the company master standards, gauges, and measuring instruments, which are periodically calibrated against the national standards. The working measuring equipment on the shop floor is calibrated by the metrology laboratory on a scheduled basis. ... In this manner the measurements made for quality control are considered 'traceable' to national standards."

William P. Fisher, Jr.

"Measurement lies at the heart of genuine quality improvement, the kind that healthcare organizations undertake on behalf of their patients and communities, not simply to ensure accreditation. When delivery systems get ready to transition from talking about continuous quality improvement to really practicing it, learning to measure and manage care processes and outcomes becomes the first priority. If quality is Job One, measurement is Job Zero."

Carl Stevens, M.D. (UCLA Medical Center) in the *Foreword* to *Statistical Process Control for Healthcare,* Marilyn K. Hart & Robert F. Hart, Brooks Cole, 2001

And it is now agreed that measurement is not merely, as S.S. Stevens mistakenly leads people to believe, the arbitrary assignment of numbers to observations.

William P. Fisher, Jr.

EXTERNAL STUDY/ONLINE UNIT, JULY 26 – NOVEMBER 1, 2004

INTRODUCTION TO RASCH MEASUREMENT AND TRADITIONAL TEST THEORY

Unit Coordinators: Professor David Andrich and Associate Professor Guanzhong Luo

THE UNIT OF STUDY - BACKGROUND

In the Australian Semester 2, 2004 (July 26 to November 26), a graduate unit of study introducing Rasch measurement is available in the *external study* mode. This mode of study means that the unit can be studied from anywhere in the world. A discussion group will operate for online interaction as part of the course.

Students 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 unit.

This unit has been presented in the same period every year from 2000. In each of 2002 and 2003, over 50 people from many parts of the world took the opportunity to enroll. Because of the success of the previous presentations, the course is being offered again this year. See below for a list of the enrolment formats available to you.

FEATURES OF THE UNIT

(i) it begins from first principles,
(ii) exercises at the end of each lecture consolidate the ideas,
(iii) it introduces the Guttman structure as a lead into both traditional test theory and Rasch measurement,
(iv) it reviews elementary traditional test theory in a way that it relates to the Rasch models,
(v) it reviews the necessary elementary statistics,
(vi) it studies the dichotomous model and the model for ordered response categories,
(vii) it studies model fit, including differential item functioning,
(viii) it involves discussion group which permits you to interact with other students in the class
(ix) it provides a full version of the interactive, Windows-based program RUMM for analyzing data. (The use of the program is available throughout the unit)

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 distracters, 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.

TOPICS COVERED

Topic 1	Review of measurement and statistics in education and social science
Topic 2	Reliability and validity
Topic 3	Formalization of traditional reliability
Topic 4	Calculation of reliability
Topic 5	The Rasch model for dichotomous responses: The simplest latent trait model
Topic 6	Separation of person and item parameters
Topic 7	The significance of total scores
Topic 8	Estimating person ability and item difficulty
Topic 9	Fit of the data to the model: general fit.
Topic 10	The Rasch model for ordered response categories: Analysis of partial credit or rated items
Topic 11	Fit of the data to the model: Differential Item Functioning (DIF)
Topic 12	(a) A relationship between the reliability of traditional test theory and Rasch latent trait theory (b) Linking using the Rasch model

Information about the course can be obtained from *www.education.murdoch.edu.au/educ_RaschCourse2004.html*

EXAMPLES OF POSITIVE RESPONSES TO THE UNIT IN 2002.

"This is by far one of the best courses on measurement theory I have ever enrolled in!"

"Despite it being a distance course, I learned a great deal."

"Both unit materials and assignments allowed me to learn the essential aspects of the subject."

"The lecture materials were well organized, logical, and easy to follow."

WHO SHOULD ENROLL

The unit is suitable for people from many social research backgrounds, but four in particular have been seen to gain most benefit from their enrolment.

(i) Professionals engaged in assessment and measurement of performance and attitude, interested in learning the principles of modern test theory and Rasch measurement in particular.

(ii) People in education, psychology, health care, health sciences who are concerned with outcome measurement.

(iii) People who have become familiar with Rasch measurement and item response theory through professional exposure, but would like to consolidate their understanding of its first principles.

(iv) Students who are involved in higher degree studies and require knowledge and evidence of studying educational and psychological measurement, in particular introduction to traditional and modern test theory.

THREE METHODS OF ENROLLING

1. As a professional taking the unit as a professional development course. The only difference between this enrolment and the next two kinds of enrolment is that in this one no formal assessment and grading is carried out, although work handed in is marked. Those who have participated effectively are given a certificate of participation.
2. As a student from a university, other than Murdoch University for which the student would like credit towards their degree. Students should check in advance if their university will give them credit for the unit.
3. As a student enrolled at Murdoch University.

The cost of the unit for enrolment as professional development or from a non-Australian university is $US700.00. Students from Australian universities may wish to enroll using the cross-institutional enrolment facility. For further information regarding costs, timetable, enrolment procedures, and so on, please contact Jan Christie (christie@murdoch.edu.au) using the SUBJECT: "RaschOnLine2004S2" who will respond to your query and put you on our mailing list.

David Andrich
andrich@murdoch.edu.au

Midwestern Objective Measurement Seminar

Sponsored by the University of Illinois at Chicago
and the Institute for Objective Measurement
Friday, May 14, 2004
University of Illinois at Chicago

Investigation of Organization and Presentation of Scores to Oral Examinations.
Jim Houston, University of Illinois at Chicago

Measuring Change Across Four Time Points
Lidia Dobria, University of Illinois at Chicago

Developing Content Guidelines using Rasch Analysis
Lidia Martinez and Amy Mericle, Measurement Research Associates, Inc.

Computer Familiarity and Test Performance on Computer-Based Tests.
Surintorn Suanthong, Ph.D. and Tanya Joosten, Measurement Research Associates, Inc.

Determining the True Confidence Interval of IRT statistics through Parametric Bootstrapping
Kirk Becker, Promissor/UIC and George Karabatsos, Ph.D, UIC

Using Paired Comparisons and a One-Faceted Rasch Model to Create the Semantic Construct of Frequency
Thomas R. O'Neil, Ph.D., National Council of State Boards of Nursing

Comparison of the Mood and Anxiety Symptom Questionnaire (MASQ) and the Beck Depression Inventory (BDI) Using Rasch Analysis
Z. Wang, K.J. Conrad, B.L. Hankin, Z. Huang, UIC

Can You Measure Change Using Three Different Measures Over Time?
Nikolaus Bezruczko, Ph.D and Ken Conrad, Ph.D., UIC

Differential Item Functioning for Women and Men in the Assessment of Depression
Kendon J. Conrad, PhD, Benjamin Hankin, PhD, Zhixiao Wang, PharmD, School of Public Health (MC 923), University of Illinois at Chicago

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

SIG Chair: Randy Schumacker, Secretary: Steve Stemler

Test Validity and Rasch Measurement: Construct, Content, etc.

Early in his career, Lee J. Cronbach made a perceptive statement, "A test is valid to the degree that *we know* what it measures or predicts" (1949, emphasis his). In the ensuing 50 years, test validity has become an evermore complex topic. Here is an interpretation of Sam Messick's (1989 etc.) conceptualization:

		Purpose	
		Interpretation	*Use*
Justification	*Evidence*	**Construct validity** **Content validity** **Face validity**	**Utility** **Predictive validity** **Concurrent validity** **Criterion-oriented validity** **Statistical reliability**
	Consequence	**Value implications**	**Social consequences**

Rasch measurement, as a means of test analysis, parallels physical measurement processes. Both are largely concerned with the construction of accurate, precise, linear measures along specific, unidimensional constructs. Even in those instance when a multi-dimensional Rasch approach is employed, the assumption is that the multi-dimensional space is a composite of unidimensional variables.

Consider the beginning of large-scale precise and accurate physical linear measurement for industrial purposes. This was an accomplishment of F. A. Pratt and Amos Whitney in the 1870s. But were their "comparator" and its resultant "standard inch" valid as a "test of length"? Not according to Messick's summary, because early applications were to the manufacture of military equipment including German Mauser rifles and British naval guns. Thus the "comparator" facilitated the carnage of the First World War. Its social consequences were dire. Surely Pratt and Whitney should have abandoned their project! But then the modern age of precision technology, mass production, speedy transportation and computers might never have occurred. Should development of tests of literacy be abandoned because such tests have been used to disenfranchise the illiterate? Surely it is impossible for a Test Constructor to predict the social consequences of a Test in any other than a short-sighted and limited way.

The value implications of a bathroom weight-scale can also be profound. Low numbers possibly indicate anorexia, high numbers probably indicate obesity. Both of these have negative stereotypical implications, i.e., detrimental value implications.

In Messick's scheme, uses and consequences, even when intended, recommended or foreseen by the constructor, are largely beyond the constructor's control. Only the "construct validity" cell is strictly within the control of the Test constructor.

The motivation for test construction comes from its hoped-for consequences. Those consequences suggest a Test's intended uses. But the history of science indicates that actual uses can be far wider than those original intended uses. Newton's Laws of Motion originated in astronomy. Computers were not conceptualized as a means of entertainment.

Content validity is an initial screening device. It verifies that extraneous material has been omitted, and that the test is representative of all relevant material. The history of the development of the thermometer indicates that the definition of what is relevant content can change as test development progresses. Thermometry now encompasses measuring the temperature of stars, but now excludes the impact of atmospheric pressure. Careful development of an educational achievement test may identify both gaps and irrelevancies in the material being taught.

Rasch measurement produces a hierarchy of persons along the latent variable. Are those persons regarded as high performers at one end of the hierarchy, and those regarded as low performers at the other with a gradation in between? If so, this indicates "Use-Evidence" of validity (predictive, concurrent, criterion-oriented, etc. – depending on the source of the external information about the sample.) But samples have their idiosyncrasies, as do external indicators, so, more important is ….

The hierarchy of items along the latent variable. This is the progression form "easy" to "hard", "common" to "rare", "general" to "specific", etc. Before (or without knowledge of specific details of the) data collection, experts should predict the difficulty ordering of the items (according to the intended construct theory). This is then compared with the items' empirical difficulties. Coincidence confirms construct validity as demonstrated in the books by Wright & Masters' (1982) "Rating Scale Analysis" and also Wright & Stone' (1979) "Best Test Design". Correlations are not important here (but can be computed, if desired). More important is that empirical disordering of one or more items in the overall hierarchy indicates that those particular items may be exhibiting unintended features - or that the construct theory is deficient.

Figure 1 is illustrative of the investigation of construct validity. It is typical of scatterplots of item difficulties for Pre-test and Post-test administrations, or at-Admission and at-Discharge. In the Figure, the item spread is wider for the high group (6 logits) than for the low group (4 logits). So the high group discriminate item difficulty more strongly. This is typical of educational tests, e.g., of

Chinese characters, where, as knowledge increases, the difference between easy and hard items becomes more pronounced. Quality-of-life assessment during rehabilitation shows the opposite characteristic. As patient status returns to normal, all regular tasks become equally easy. The variable defined by the widest spread of item difficulty is usually the most relevant.

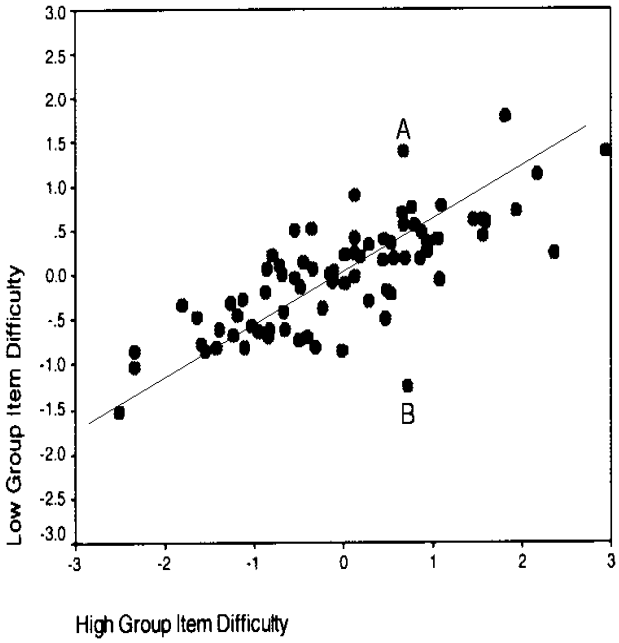


Figure 1. High group vs. low group item difficulty. (Smith & Suh, 2003, Journal of Applied Measurement 4:2, 159)

In Figure 1, however, two somewhat different variables have been defined. For the high group, items A and B are equally difficult. For the low group, those same items A and B are almost at the extremes of the variable as defined by these items. Which is the intended variable? If the order of items had been predicted *a-priori* according to some construct theory, then the hierarchy more closely matching the intended variable could be identified immediately. The best result, from a construct validity perspective, would be that the intended variable follows the "best fit" diagonal line on the plot. Since items A and B are so markedly misplaced, it is likely that they contain flaws or features which make them essentially different items for the two performance groups. Construct validity must be carefully constructed, it is unlikely to emerge fortuitously from a collection of test items.

John Michael Linacre

Messick, S. (1989). Validity. In R.L. Linn (ed.) *Educational measurement*. Third edition. New York: Macmillan, 13-103.

Smith R.M. & Suh K.K. (2003) Rasch fit statistics as a test of the invariance of item parameter estimates. Journal of Applied Measurement, 4:2, 159.

Winsteps and Facets Workshops

August 5, 2004 - Thursday:
Introductory Winsteps workshop, Chicago
conducted by Ken Conrad and Nick Bezruczko
www.winsteps.com/workshop.htm

October 11-12 & 13-14, 2004 – Monday-Thursday
Winsteps and Facets workshops, Durham NC
conducted by Mike Linacre
www.winsteps.com/seminar.htm

Interchangeable Parts: "Accuracy" and "Dependability"

In 1879, Pratt & Whitney financed the efforts of Harvard Professor William A. Rogers and George M. Bond from Stevens Institute of Technology to develop a comparator for [physical linear] measurements accurate within one-50 thousandths of an inch. In addition, the P&W Company established the standard inch. By 1885 the P&W standard measuring machine was beginning to be known all over the world as the basis of the construction of recognized standards of length – by then accurate to one hundred thousandth of an inch!

The new idea of interchangeability of parts had been thought of, and talked about to some extent, by Eli Whitney and Samuel Colt, but it remained for Amos Whitney and F. A. Pratt to make the idea practical on a large scale. As a result the Pratt & Whitney Company became pioneers and leaders in developing and applying the new system of interchangeable manufacture. Much of the success of this system depended upon the development and use of accurate gages and trustworthy standards of length.

Aircraft Engine Historical Society, www.enginehistory.org

Report from the Outgoing Secretary of the Rasch Measurement SIG

Financial Report:
On about April 1st, 2003 — the opening account balance for the last year — the balance was $454.62.
On January 29th, 2004 — the closing account balance for my tenure as SIG secretary — the balance was $908.28.
Currently, the only pending charge to the SIG is the July 2004 SIG-payable dues to AERA, which will equal $100. This is $50 less than in previous years. In 2004, AERA will increase SIG individual membership dues by $5 per year.

Membership Report:
Total: 198. AERA 143 (72%), SIG only 55 (28%)

Edward W. Wolfe - April 2004

Predicting Measures from Rating Scale or Partial Credit Categories for Samples and Individuals

We have collected data and analyzed it. We believe that the our findings are a reasonable basis for predicting the future. Now we want to make that prediction.

Here is what our analysis has told us about a particular partial-credit item of our instrument:

Observed Partial Credit Category	Observed frequency	Observed sample average measure	Expected sample average measure	*Mean* Expected rating measure	*Median* Rasch-Thurstone threshold	*Modal* (Rasch-Andrich?) threshold
1	14%	-.51	-.42	(-2.22)	---	---
---	---	---	---	-1.50	-1.18	-.79
2	26%	.39	.04	-.61	---	---
---	---	---	---	.28	-.04	-.43
3	60%	.73	.86	(1.00)	---	---

Our example uses a partial-credit item, but this discussion is equally applicable to predicting from "rating scale" data, and much of it applies to dichotomous data.

I. Sample-level

A. The conventional descriptive-statistical approach of, for instance, Generalizability Theory, is to assume that the next sample will **exactly resemble** the current one. In which case, the first three columns will suffice. For persons rated in category 1, we would predict a measure of -.51, which was the average measure of those observed in category 1 of this item in the earlier sample.

B. The earlier sample performed largely as the Rasch model predicts, but not exactly. We assume that the next sample will have the **same measure distribution** and exhibit the same Rasch-coherent behavior as the earlier sample, but the next sample's idiosyncratic non-Rasch behavior is unpredictable. In which case, the fourth column, the "expected sample average measure" is our prediction. It reflects only the Rasch-coherent aspect of the current sample. We expect that the next sample will exhibit small, but different, idiosyncratic departures from these measures, but, since we don't know what these idiosyncrasies will be, for persons rated in category 1, we would predict a measure of -.42, which would have been the average measure of those observed in category 1 of this item in the earlier sample, if that sample had followed exact Rasch-model predictions.

II. Individual-level

We expect the next person to behave in the same Rasch-conforming way as the previous sample, but we can make no distributional assumptions relevant to the next person. This is a "non-informative Bayesian prior" and parallels our use of a tape measure or bathroom scale.

C. For the next individual who receives a rating of 2, we predict the measure corresponding to the point on the latent variable where a rating of 2 is **most probable** to be observed (or where the average of the ratings expected to be observed has the value of the category). This is -.61, its "expected rating measure". From this perspective, the measures corresponding to extreme categories, 1 and 3, are infinite, so the "expected rating measure" reported in the Table for category 1 corresponds to an expected rating of 1.25 (at measure -2.22), conceptually half-way between 1 and 1.5 (at measure -1.50), a boundary between category 1 and category 2. Similarly for category 3, the reported expected rating is for 2.75 (at measure 1.00).

D. Our prediction for a person in a particular category is the **range** of measures for which there is a 50% or greater chance that the person would be observed in this category or above, and also a 50% or greater chance that the person would be observed in this category or below. For this, the range boundaries are the **Rasch-Thurstone thresholds**. In our Table, the range of measures for category 1 would be from -∞ to -1.18. These values appear on cumulative probability plots as the points where a .5 probability line intercepts the cumulative curves.

E. Our prediction is the range of measures for which the observed category is the most likely category to be observed. These are the **modal thresholds,** and are the Rasch-Andrich thresholds (when those are ordered). When the Rasch-Andrich thresholds are disordered, some categories will never be the ones most likely to be observed. In our Table, the range of measures for category 1 would be from -∞ to -.79. These values appear on category probability plots as the abscissae of the points where the probability curves for modal categories meet.

F. Our prediction is the range of measures for which the **average rating is in the neighborhood** of this category. The Rasch model predicts the probability of any category being observed anywhere along the latent variable. From these probabilities, the average value of the ratings at any point along the variable can be predicted. For any intermediate category, its neighborhood can be defined as the interval from "category value – 0.5" to "category value + 0.5". For extreme categories, the outer ends of the intervals are infinite. These "neighborhoods" are shown as transition values in the *Mean* column. . Here, the range of measures for category 1 would be from -∞ to -1.50. These values can be seen on the "expected score ogive" (the "model" item characteristic curve).

John Michael Linacre