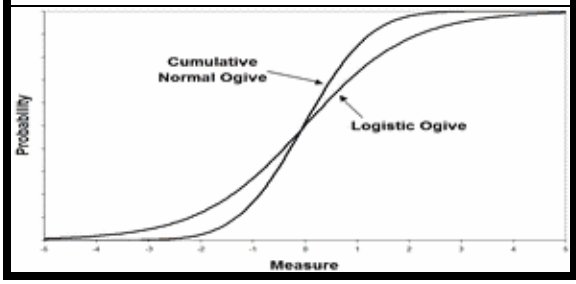
Logistic & Cumulative Normal Ogives



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

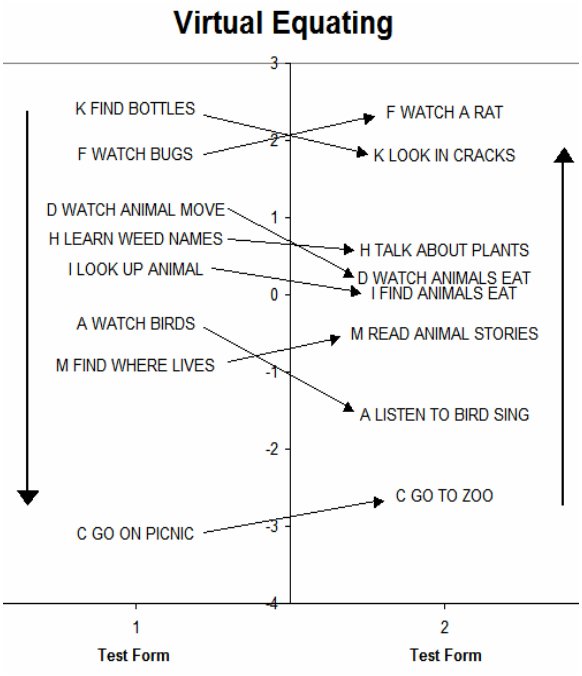
Vol. 19 No. 3 Winter 2005 ISSN 1051-0796

Virtual Equating

When two test forms have no respondents in common and no items in common, then the data for each test form comprises a separate analysis. If the test forms are designed to be parallel, then it may be reasonable to assert that the mean item difficulties of the two test forms are the same. If the distributions of the respondents administered each test forms are considered to be randomly equivalent, then it may be reasonable to assert that the mean respondent abilities are the same. But what if these assertions are unreasonable or need to be verified? Then conduct a qualitatively-based **virtual equating** (Luppescu, 1996).

Step 1. Identify pairs of items of similar content and difficulty in the two tests. Be generous about interpreting "similar" at this stage.

Steps 2-4 by print-out: The two item difficulty hierarchies are printed with the difficulties spaced according to their Rasch measures. Equivalent items are identified. The sheets of paper are moved relative to each other until the overall joint hierarchy makes the most sense. The value on Test A corresponding to the zero on Test B is the equating constant to use for Test B. If the item spacing on one test appears expanded or compressed relative to the other test, then rescale the measures on one test form to compensate.



Or: Step 2 by graphing: From the separate analyses, crossplot the difficulties of the pairs of items, with Test B on the y-axis and Test A on the x-axis. The slope of the best-fit line i.e., the line though the point at the means of the common items and through the (mean ± 1 S.D.) point should have slope near 1.0. If it does, then the intercept of the line with the x-axis is the equating constant.

To place Test B in the Test A frame of reference: add the x-axis intercept to all Test B measures.

Step 3. Examine the scatterplot. Points far away from the best fit line indicate items that are not good pairs. You may wish to consider these to be no longer paired. Drop the items from the plot and redraw the best fit line.

Step 4. The slope of the best fit is: slope = (S.D. of Test B common items) / (S.D. of Test A common items) . So multiply Test B measures by the value of 1/slope, and add the value of the x-intercept. Then reanalyze Test B. Test B is now in the Test A frame of reference, so a cross plot of the equating items should approximate the identity lined. If so, the person and item measures from Test A and Test B can be reported together.

Luppescu, Stuart (1996). Virtual equating: An approach to reading test equating by concept matching of items. Doctoral dissertation, University of Chicago.

Table of Contents

Hong Kong

HKSoQOL Conference on Quality of Life Research in Asia
19-21 May, 2006 (Fri.-Sun.)

More than 15 local and overseas experts present plenary and symposium sessions and also in-depth workshops.
www.hksoqol.org/conference

PROMS
Pacific Rim Objective Measurement Symposium
June 27-29, 2006 (Tues.-Thur.)
Rasch measurement as a tool for scientific progress
www.promshk.org

Rasch Workshops

January 5-6, 2006, Thursday - Friday, Chicago IL
Introduction to *Facets*
conducted by Carol Myford and Lidia Dobria
www.winsteps.com/facwork.htm

January 16-20, 2006, Monday - Friday, Canberra, Australia
Measurement in the Psychosocial Sciences: from Raw Scores to Rasch Measures
conducted by Andrew Stephanou
www.acspri.org.au

February 25, 2006, Saturday, Taiwan
Rasch Analysis Workshop - Introduction Course on Health care
conducted by Wen-Chung Wang
www.healthup.org.tw/rasch/950225.htm

March 20-21, 2006, Monday - Tuesday, Chicago
Introduction to *Winsteps*
conducted by Ken Conrad and Nick Bezruczko
www.winsteps.com/workshop.htm

April 6-7, 2006, Thursday-Friday, San Francisco CA (pre-AERA)
Introduction to Rasch Measurement
conducted by Richard Smith and Everett Smith
www.jampress.org

June 25, 2006, Sunday, Hong Kong (pre-PROMS)
Introduction to *Winsteps*
June 26, 2006, Monday, Hong Kong (pre-PROMS)
Introduction to *Facets*
conducted by Mike Linacre
www.promshk.org

Rasch Online Course "Practical Rasch Measurement"

Feb. 3 - March 3, 2006

www.statistics.com/content/courses/rasch

Aim of the Course: This course covers the practical aspects of data setup, analysis, output interpretation, fit analysis, differential item functioning, dimensionality and reporting. Simple test linking and equating designs are addressed. Supporting theory is presented conceptually. Participants are encouraged to analyze their own datasets in parallel to the course datasets.

Instructor: John "Mike" Linacre

SESSION 1: Basic concepts and operations
- *Ministep* software installation and operation
- Data entry methods
- Scoring data to accord with the latent variable
- Linear measures vs. raw scores
- Simple dichotomous and polytomous analyses
- Item and person maps

SESSION 2: Fit analysis and item structures
- Observations, expectations and residuals
- Quality-control fit statistics for items and persons
- Reliability indexes
- Distracter analysis
- Modeling multi-category items

SESSION 3: Estimation, DIF and dimensionality
- Estimation methods
- How iterative estimation works
- Differential item functioning (DIF)
- Investigating dimensionality

SESSION 4: Test equating, reporting findings.
- Equating, linking and anchoring
- Equating dichotomous and polytomous items
- Analyzing your own data
- Learning from published Rasch-based research
- Reporting Rasch-based findings

Organization of the Course: The course takes place over the Internet at statistics.com. During each course week, you participate at times of your own choosing - there are no set times when you must be online. Course participants will be given access to a private bulletin board that serves as a forum for discussion of ideas, problem solving, and interaction with the instructor. The course is scheduled to take place over four weeks, and should require about 10 hours per week. At the beginning of each week, participants receive the relevant material, in addition to answers to exercises from the previous session. During the week, participants are expected to go over the course materials and work through exercises. Discussion among participants is encouraged. The instructor will provide answers and comments.

Course cost: $399 per participant ($299 academic).

Book Review: Constructing Measures: An Item Response Modeling Approach

The expert Rasch modelers have truly begun the necessary effort to explain and teach the basics to front-line professionals. Mark Wilson (2005, Lawrence Erlbaum Assoc.) has joined Bond & Fox (2002) to help fill the gap of useful introductory material to teach with, their predecessor, *Best Test Design,* was published in the 1970's! *Constructing Measures* is aimed at a first course in measurement, but makes every effort to be consistent with the Rasch model and still remain true *(pun intended)* to the history of measurement in the 20th century.

By proposing a reasonable and understandable scheme of construct modeling using "building blocks", the book is an excellent primer for a basic measurement course. The introduction of construct maps in chapter 2 is subtle and leads easily to recognizable ruler output from Rasch software. The item-design chapter mentions levels of specificity, but stops a little short on the discussion of mixed item types as part of a single instrument.

Outcome space (chapter 4) moves from Likert items to phenomenography to the SOLO taxonomy to Guttman. To me, this would be ordered better *(sorry, I can't help it)* by moving Guttman adjacent to Likert. Chapters 5 and 6 introduce the Rasch model with excellent explanations, appropriate historical references, and plenty of graphs. In chapter 5, I kept expecting a mention of conjoint probabilities but that never appeared.

The concept of logits as units also tiptoes in without much discussion. Between all the other italicized concepts and definitions, this seems a little understated. The basic equations and examples are solid. Almost all the math is in the middle chapters, and most classes will have to slow down here even though the CD that accompanies the text is an excellent teaching tool. *(continues next column)*

Chapter 6 compares the Rasch model to the 2PL-IRT model and provides a very good rational for model choice. The discussion of fit is adequate and introduction of Kidmaps is a plus. Explaining Keyforms to novices has often been the downfall of otherwise excellent presentations for me. Chapter 6 misses an opportunity to mention MFRM as a model, and the book again emphasizes classical reliability coefficients in chapter 7 without discussing Facets for rater effect estimates. Given the previous coverage of discrimination functions and fit, Facets would be no more difficult to comprehend.

The text's return to Construct Maps in chapter 8 on validity is appropriate. The traditional reliability treatment contrasts with the lack of mention of "criterion-related validity" which students are likely to see in other places and so wonder where it is. The introduction of DIF is good. Chapter 9 ends the book with some philosophy (situative), some psychology (cognitive), and some statistics (hierarchical). This chapter has some applied examples that would interest students if the "adding complexity" section doesn't ambush them first.

Packaged with a useful CD, this book would be good for students who had a prior research or basic statistics class, and needed a step-by-step approach to creating a test. Overall, Mark Wilson has obviously worked hard to create understandable examples and a practical process for students to build a test from start to finish. The book accomplishes exactly that with a logical process and basic explanations. As the preface suggests, the text is aimed at active learning by doing, and it would be a shame to use it otherwise.

Steve Lang
University of South Florida, St. Petersburg

Everything Relates to Everything Else

"Our current scientific understanding has moved considerably from the view that the universe, both on the cosmic scale and in the inner workings of matter, is understandable in terms of a sticks-and-balls mechanism, the behavior of which can be elucidated and predicted with greater and greater precision. Instead, we see a large, interactive process with a great deal of unpredictability built into the very nature of things. What is even more fascinating is that *the observer, you and me and the scientist behind the measuring instruments, -- become a part of the process.* The observer, in a curious way, becomes part of what the outcome of the observation is. Mind, in effect, can be seen as an additional reality of the universe, inseparable from its time-space dimensions. Rather than consisting of a lot of separate objects, the universe is comprehensible as a whole of complex events in which everything relates to everything else."

J. John Keggi, "Stillness and the Storm", Augusta, Maine, 22 June 1997

Numerical Nonsense

"In developing procedures, mathematical statisticians have assumed that techniques involving numerical scores [etc.] ... are to be applied where these numbers ... are appropriate and meaningful within the experimenter's problem. If the statistical method involves the procedures of arithmetic used on numerical scores, then the numerical answer is formally correct. Even if the numbers are the purest nonsense, having no relation to real magnitudes or the properties of real things, the answers are still right as numbers. *The difficulty comes with the interpretation of those numbers back into statements about the real world.* If nonsense is put into the mathematical system, nonsense is sure to come out."

Hays, W. L. (1973). Statistics for the social sciences (2d ed.). New York: Holt, Rinehart and Winston. p. 88, as quoted on p. 18 of Shavelson, R. J. (1996). Statistical reasoning for the behavioral sciences (3d Ed.). Boston: Allyn and Bacon. *Courtesy of William P. Fisher, Jr.*

Correlation Coefficients: Describing Relationships

Correlation coefficients summarize the association between two variables. They include:

(a) Both variables are expressed as perfectly precise, normally distributed, real numbers (PPNDRN): the Pearson **product-moment** correlation.

(b) Both variables are PPNDRN, but one is grouped into two classes (high and low): the **biserial** correlation.

(c) Both variables are PPNDRN, and both are grouped into two classes (high and low): the **tetrachoric** correlation, or into multiple ordered classes: the **polychoric** correlation.

(d) One variables is PPNDRN, the other is a discrete variable with only two values (such as gender): the **point-biserial** correlation, or more than two equally spaced values, the **point-polyserial** correlation.

(e) Both variables are discrete with only two values: **phi** correlation.

If the numbers to be correlated are not perfectly precise, then it may be possible to disattenuate the correlation coefficient for measurement error (RMT 10:1, 479).

Pearson's product-moment correlation is the most commonly reported, even for those data for which it is superficially not a good match. Of course, the same is true of other familiar statistics, such as the mean and standard deviation.

So which correlation coefficient is most indicative in any particular instance? Here statistical theory encounters harsh reality. No empirical variable exactly matches the assumptions of a correlation coefficient. Even with natural discrete dichotomies, such as gender, there is always some fuzziness. Mendel's genetic experiments have come under attack for the manner in which he may have manipulated the fuzziness in his data.

So there are two criteria: (i) ease of communication and (ii) protection against misleading inferences.

For ease of communication, the more familiar the coefficient is, the better, provided it does not produce a misleadingly incorrect value.

For those coefficients which produce reasonable values, the temptation is almost always to report the highest (or most significant) relationship possible. This temptation is evident in factor analysis: the choice of communalities, rotation and obliqueness tends to be guided by the desire for a conspicuous finding. Thus, after the correlation reporting the highest or most significant value has been discovered, it is tempting to rationalize why that particular correlation coefficient is the "correct" one for those data.

Guilford (1965, p. 325) points out that if the data accord with the **biserial** correlation, then there is an exact mathematical relationship between the biserial and **point-biserial**. So, if both are computed their ratio must approximate specific values. So when this ratio is observed for empirical data, the biserial may be the correlation of choice. Under essentially all other conditions, Guilford recommends the more conservative point-biserial correlation.

	0	*1*
1	**167**	**374**
0	**203**	**186**

Phi correlation = Pearson = Point-biserial = 0.21
Biserial correlation = 0.27 or 0.31
Tetrachoric correlation = 0.34

An early objection to the **tetrachoric** correlation was that its value could only be approximated. With modern computer power, the approximation can be so precise as to be considered exact. But other objections remain.

Nunnally (1967, 123-4) remarks "There are very strong reasons for *not* [his emphasis] using the biserial and tetrachoric correlations in most of the ways they have been used in the past. …. Unless subsequent steps are made to turn the dichotomous variables into continuous variables, such estimates only serve to fool one into thinking that his variables have explanatory power beyond that which they actually have. It is tempting to employ biserial and tetrachoric correlations rather than phi and point-biserial correlations because the former are usually larger." He adds "When the assumption of normality is not met, the estimates can be off by more than 20 points of correlation."

	1	*2*	*3*	*4*	*5*
1	**0**	**0**	**12**	**32**	**40**
2	**0**	**4**	**23**	**66**	**23**
3	**1**	**10**	**67**	**77**	**15**
4	**1**	**22**	**133**	**40**	**3**
5	**8**	**71**	**125**	**21**	**2**

Pearson correlation = 0.61
Polychoric correlation = 0.67
Computation by Uebersax (2000)

Coote (1998, p. 404) has a provocative paragraph: "Product-moment correlation matrices are often used … although they are only appropriate for continuous variables (Joreskog and Sorbom, 1996). Information collected using five and seven-point Likert scales have ordinal properties (Bollen, 1989). Ordinal variables do not have origins or units of measurement and should not be treated as though they are continuous (Joreskog, 1994). Treating ordinal data as continuous increases the likelihood of correlated error variances, particularly where the initial factor loadings are large. Another disadvantage of using a product-moment correlation matrix with categorical data is that the standard errors and chi-square

test statistics are incorrect (Anderson and Gerbing, 1988). Where Likert scales are used **polychoric correlations** should be computed and analyzed."

But reality is rarely this clear-cut. The conceptualization of the ordinal scale required for the polychoric correlation accords with Samejima's "graded response" model. The categorization is regarded as ordered but the categorization itself is considered arbitrary.

A Rasch-consistent conceptualization would be a variant of the **point-polyserial**, with category intervals consistent with the corresponding Rasch-model ICC. In this regard, integer intervals are exact for two and three category scales, but generally too central for the extreme categories of longer rating scales. Consequently an integer-spaced point-polyserial would tend to misestimate the actual correlation for long rating scales, but generally only by a small amount. The effort required to improve on integer spacing does not appear to have a corresponding benefit.

So, for correlations of Rasch-analyzed data for which the categorization is considered qualitatively substantive, the analyst would need to make a strong case to depart from correlation coefficients with the algebraic form of the Pearson product-moment correlation. These coefficients are the **product-moment** correlation itself, and the **point-biserial, point-polyserial** and **phi** coefficients.

John Michael Linacre

Anderson, J. C. & Gerbing, D. W. (1988) Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach. Psychological Bulletin, 103:3, 411-423.

Bollen, K. A. (1989) Structural Equations with Latent Variables. New York: John Wiley and Sons.

Coote, L. (1998) A Review and Recommended Approach for Estimating Conditional Structural Equation Models. Australia and New Zealand Marketing Academy Conference, University of Otago, Dunedin.

Guilford J. P. (1965) Fundamental Statistics in Psychology and Education. New York: McGraw-Hill.

Joreskog, K. G. (1994) On the Estimation of Polychoric Correlations and their Asymptotic Covariance Matrix. Psychometrika, 59:3, 381-389.

Joreskog, K. G. and Sorbom, D. (1996) PRELIS 2: User's Reference Guide. Chicago: Scientific Software International.

Nunnally, J. (1967) Psychometric Theory. New York: McGraw-Hill.

Olsson U. Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika, 1979, 44(4), 443-460.

Uebersax J.S. (2000) POLYCORR Polychoric Correlation EZ Version software.

Fit to Models: Rasch Model vs. Correlation Model

Viewed as a statistical device, the Rasch model is one of thousands in current use. One of those thousands most frequently employed is the Pearson Correlation Model.

The Correlation Model

The size of the Pearson product-moment correlation between two variables is frequently reported, sometimes accompanied by whether it is significantly different from 0.00. But rarely reported are:

1) whether the observed correlation departs insignificantly from 1.00, which is perfect correlation. But high correlations, regardless of their statistical significance, could be indicative of collinearity. Near-perfect correlation should be regarded with suspicion.

2) whether the observations violate the assumptions underlying the Correlation Model. Violations are rarely tested explicitly because the correlation model is **too useful not to use.** Pearson correlations are often reported for data which are known not to meet its assumptions.

The Rasch Model

The Rasch model is similarly **too useful not to use**. Further, near perfect fit to the Rasch Model should be regarded with suspicion. Empirical processes are uneven. The validity of scientific work has come into question when statistical findings appear to be too perfect.

Taking the same position with regards to the Rasch Model as we do for the Correlation Model, the crucial question is not "Is the correlation statistically 1.0", expressed as "Do the data fit the Rasch model statistically perfectly?" This question has been the focal point of most global fit analysis with the Rasch model. Instead the crucial question becomes "Is the correlation statistically different from 0.00", expressed as "Is there a Rasch dimension which is significantly larger than a point?"

The Rasch dimension reduces to the size of a point when the data are perfectly random. Jacob Cohen (1992) suggests that, for the ratio of explained variance to unexplained variance, 2% is a small effect size, 15% is a medium effect size, and 35% is a large effect size. Recast this as the percentage of total variance explained and 2% is a small effect size, 13% is a medium effect size, and 26% is a large effect size. For comparison, the variance explained by the Rasch measures for the *Liking for Science* data is 81% and for the *Knox Cube Test* data is 99%. Even the variance explained for a relatively central, poorly fitting, clinical data set (a sample of opportunity) is 33%. Rasch papers can routinely report effect statistics, which, if they were the findings of correlation studies, would produce great joy among social scientists.

John M. Linacre

Cohen J. (1992) A Power Primer, Psychological Bulletin, 112, 155-159.

Standard Errors: Means, Measures, Origins and Anchor Values

Statistics text books explain the "standard error of the mean", but are generally silent about the "standard error of a measure". How do they relate?

The standard error is the modeled standard deviation of the observed estimate around the unobservable "true" value. In practice, the observed estimate substitutes for the "true" value and we think of the standard error being centered on observed estimate.

Both the observed estimate and its standard error are computed from the data. Each data point gives us an estimate of the mean or the measure, and the accumulation of the estimates provides the final best estimate along with its precision, its standard error. Thus:

Accumulation of estimates (one per observation)
=> mean parameter estimate ± S.E. of estimate

For a typical "text book" normal distribution, the parameter of interest is the mean, which is the sum of all perfectly-precise observations divided by their count. And its standard error is the sample standard deviation of the observations divided by the square-root of the count.

A Rasch measure has parallels to a sample mean. Conceptually, each qualitative observation ("Right", "Wrong", etc.) provides an estimate of the relevant measure, so

Accumulation of estimates (one per observation)
=> measure estimate ± S.E. of estimate

Implementing this directly is awkward, It is more convenient to rearrange the computation:

Estimate of (accumulation of observations)
=> measure estimate ± S.E. of estimate

Here, the standard error is computed by summing the statistical (Fisher) information across the observations, and then the standard error is the square-root of the inverse of the summed information.

For example, consider 1000 reasonably targeted observations of a dichotomous item. Experience shows that a reasonable p-value for such an item is .8. So the average binomial variance ≅ p-value*(1 - p-value) = .8*.2 = .16 ≅ average Fisher information. So the information in 1000 observations = 1000 * .16 = 160. Standard error of the logit estimate = 1 / square root (Fisher information) = 1 / square-root (160) = .08 logits. The ease of this type of computation is one reason the Rasch model is formulated in logits, rather than in $\log_{10}$, probits, etc.

Local Origins and Standard Errors

The standard error of the mean is usually computed in an absolute frame of reference in which the zero point is defined external to the data. Rasch measures are defined relative to a local zero point. How does this impact standard error computations?

In the same way as the zero point on a temperature scale is an arbitrary point, chosen according to some definition, e.g., "the freezing point of water", the zero point (local origin) of a Rasch measurement scale is an arbitrary point on the latent variable, defined in some manner. Typical choices are "the average difficulty measure of all items", "the difficulty of a specific item" or "the average ability measure of all respondents".

In general, the Rasch local origin is considered to be the absolute location on the latent variable with which the empirically-derived location happens to coincide. Thus the measures and standard errors are considered to be in an absolute frame of reference.

However, when comparing measures across parallel analyses, shifts in the locations of local origins might be crucial. Accordingly the standard error of the empirical zero could be included. This suggests that the most stable possible choice of local origin be made to minimize the need for this computation. In general, if the mean of the item difficulties is chosen, and the same set of items is administered a second time, then the standard error of the mean-item "origin" is the average standard error (root-mean-square-error, RMSE) of the items. Typically, this would be much smaller than the standard error of a person measure. So the joint standard error of the difference between two measures across test forms comprising the same items would approximate:

$$SE(measure_1 - measure_2) = \sqrt{(SE(measure_1)^2 + SE(origin_1)^2 + SE(measure_2)^2 + SE(origin_2)^2)}$$

Anchor Values and Standard Errors

An anchored (fixed) measure is treated as though it is an estimate of the "true" value of the parameter, so it is reported along with the standard error around the "true" value. If the corresponding local empirical value is also computed, this can be compared with the anchor value along with its standard error in order to test the hypothesis that the data were generated by the true (anchor) value.

John Michael Linacre

[My thanks to a reader who pointed out a mistake in the original published version - corrected here.]

Rasch Measurement Transactions
P.O. Box 811322, Chicago IL 60681-1322
www.rasch.org/rmt/
Editor: John Michael Linacre

SIG Chair: Randy Schumacker, Secretary: Steve Stemler
Program Chair: Trevor Bond
SIG website: **www.raschsig.org**

Rasch SIG Elections are underway.
Cast your vote by *December 31, 2005*
www.raschsig.org/news.html

Communicating Conclusions

"We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions."

Ronald A. Fisher, 1955, Statistical methods and scientific induction. Journal of the Royal Statistical Society, B, 17, 69-78.

The Truth about Statistical Models

"All models are wrong. Some are useful."

George E. P. Box, University of Wisconsin

"The hallmark of good science is that it uses models and 'theories' but never believes them"

Martin Bradbury Wilk, in John Tukey, The future of data analysis, Ann. Math. Statist. 33, 1-67, p.7.

"Models must be used but must never be believed."

Attributed to Martin Bradbury Wilk

The Truth about Significance Tests

"Significance tests are things to do
while one is trying to think of something sensible to do."

Attributed to Martin Bradbury Wilk

The Truth about Item Response Theory

"Building statistical models is just like this. You take a real situation with real data, messy as this is, and build a model that works to explain the behavior of real data."

Martha Stocking's summary of Item Response Theory, the statistical methodology of Frederic M. Lord (1912-2000), New York Times, 2-10-2000

The Truth about Factor Analysis

"Factor analysis is useful, especially in those domains where basic and essential concepts are essentially lacking and where crucial experiments are difficult to conceive ... In a domain where fundamental and fruitful concepts are already well formulated and tested, it would be absurd to use the factorial methods except for didactic purposes to illustrate factorial logic."

L.L. Thurstone (1947) Multiple Factor Analysis, Chicago: University of Chicago Press. p. 56

.

The Truth is in the Eye of the Beholder

A reporter showed a photograph of the Earth taken from Space to the late Samuel Shenton, then President of the International Flat Earth (Research) Society. Shenton studied it for a moment and said, **"It's easy to see how such a picture could fool the untrained eye."**

Attributed

Rasch-related Conferences

IOMW 2006

The 13th International Objective Measurement Workshop
April 5–7, 2006 (Wed.-Fri)
University of California, Berkeley, CA, USA
http://bearcenter.berkeley.edu/IOMW2006/

AERA

American Educational Research Association
April 8-12, 2006 (Sat.-Wed.)
San Francisco, CA, USA
www.aera.net

Midwest Objective Measurement Seminar - MOMS

Friday, December 9, 2005
Rehabilitation Institute of Chicago and the Institute for Objective Measurement
at the Northwestern University Kellogg School of Management

Is it Possible To Solve the Incoherence of Rasch Measurement? *George Karabatsos, Ph.D. University of Illinois at Chicago*

Bookmark Standard Setting. *John Stahl, Ph.D., Promissor, Inc.*

Facets Analysis for OSCEs. *Doug Lawson, DC, MSc, PhD(c), University of Calgary, Canada*

Measuring the Impact of Rater Severity Drift on Student Ability Measures. *Lidia Dobria, Everett Smith, Ph.D., Carol Myford, Ph.D., University of Illinois at Chicago*

Score adjustment for rater bias in performance assessment. *Cherdsak Iramaneerat, M.D., M.H.P.E. UIC and Measurement Research Associates, Inc.*

Equating functional status measures across post-acute care rehabilitation settings. *Trudy Mallinson, Ph.D., OTR/L, NZROT, Rehabilitation Institute of Chicago*

A New ADLM for Spinal Cord Injury. *Anne M. Bryden, OTR/L The Cleveland FES Center, MetroHealth Medical Center Rehabilitation Engineering Center and Nikolaus Bezruczko, Ph.D., Measurement and Evaluation Consulting*

Differential Item Functioning with Winsteps and Facets: Implications For Triage in Substance Abuse and Dependence. *Ken Conrad, Ph.D., UIC and Michael Dennis, Ph.D., Chestnut Health Systems*

Study Skills Self-Efficacy of Secondary School Students in Hong Kong. *Qiong (Joan) Fu, UIC; Mantak Yuen, University of Hong Kong; Lidia Dobria, UIC, Everett V. Smith, Ph.D, UIC.*

Aspect	**Rasch Dichotomous Model**	**Item Response Theory: One-Parameter Logistic Model**
Abbreviation	Rasch	1-PL IRT
Motivation	Distribution-free person ability estimates and distribution-free item difficulty estimates on a linear latent variable	Computationally simpler approximation to the Normal Ogive Model of L.L. Thurstone, D.N. Lawley, F.M. Lord
Persons, objects, subjects, cases, etc.	Person n of ability B_n, or Person ν (Greek *nu*) of ability β_ν in logits	Normally-distributed person sample of ability distribution θ, conceptualized as N(0,1), in probits
Items, agents, prompts, probes, multiple-choice questions, etc.	Item i of difficulty D_i, or Item ι (Greek *iota*) of difficulty δ_ι in logits	Item i of difficulty b_i (the "one parameter") in probits
Nature of binary data	1 = "success" - presence of property 0 = "failure" - absence of property	1 = "success" - presence of property 0 = "failure" - absence of property
Probability of binary data	P_{ni} = probability that person n is observed to have the requisite property, "succeeds", when encountering item i	$P_i(\theta)$ = overall probability of "success" by person distribution θ on item i
Formulation: exponential form $e = 2.71828$	$P_{ni} = \frac{e^{B_n - D_i}}{1 + e^{B_n - D_i}}$	$P_i(\theta) = \frac{e^{1.7(\theta - b_i)}}{1 + e^{1.7(\theta - b_i)}}$
Formulation: logit-linear form $\log_e$ = *natural logarithm*	$\log_e \left(\frac{P_{ni}}{1 - P_{ni}} \right) = B_n - D_i$	$\log_e \left(\frac{P_i(\theta)}{1 - P_i(\theta)} \right) = 1.7 (\theta - b_i)$
Local origin of scale: zero of parameter estimates	Average item difficulty, or difficulty of specified item. (Criterion-referenced)	Average person ability. (Norm-referenced)
Item discrimination	Item characteristic curves (ICCs) modeled to be parallel with a slope of 1 (the natural logistic ogive)	ICCs modeled to be parallel with a slope of 1.7 (approximating the slope of the cumulative normal ogive)
Missing data allowed	Yes, depending on estimation method	Yes, depending on estimation method
Fixed (anchored) parameter values for persons and items	Yes, depending on software	Items: depending on software. Persons: only for distributional form.
Fit evaluation	Local, one parameter at a time	Global, accept or reject the model
Data-model mismatch	Defective data do not support parameter separability in a linear framework. Consider editing the data.	Defective model does not adequately describe the data. Consider adding discrimination (2-PL), lower asymptote (guessability, 3-PL) parameters.
Differential item functioning (DIF) detection	Yes, in secondary analysis	Yes, in secondary analysis
First conspicuous appearance	Rasch, Georg. (1960) Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.	Birnbaum, Allan. (1968). Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley.
First conspicuous advocate	Benjamin D. Wright, University of Chicago	Frederic M. Lord, Educational Testing Service
Widely-authoritative currently-active proponent	David Andrich, Murdoch Univ., Perth, Australia	Ronald Hambleton, University of Massachusetts
Introductory textbook	Applying The Rasch Model. *T.G. Bond and C.M. Fox*	Fundamentals of Item Response Theory. *R.K. Hambleton, H. Swaminathan, and H.J. Rogers.*
Widely used software	Winsteps, RUMM, ConQuest	Logist, BILOG