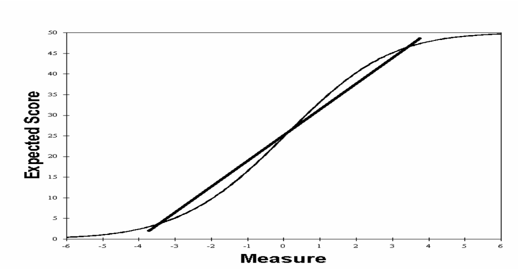
Score-to-Measure Approximation



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

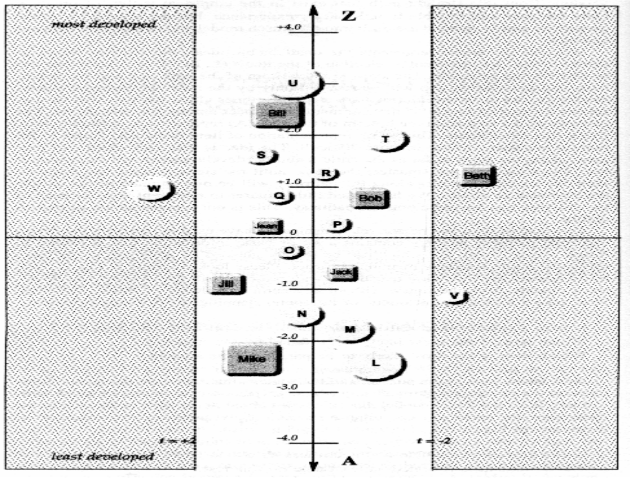
Vol. 16 No. 2 Autumn 2002 ISSN 1051-0796

Review of Reviews of Bond & Fox (2001)

Trevor Bond and Christine Fox have accomplished a remarkable feat: writing an academic best-seller about an obscure area of statistical measurement. Sales figures assert that "**Applying The Rasch Model: Fundamental Measurement in the Human Sciences**" (Mahwah NJ: Lawrence Erlbaum Assoc.) has succeeded in reaching its goal of communicating highly technical material in a non-technical way.

For almost 20 years, Ben Wright toyed with writing a non-technical introductory Rasch text. Ben had a model of what he wanted to do, his "Conversational Statistics in Education & Psychology" (Wright & Mayers, McGraw-Hill, 1984). That book focuses on one dataset and each chapter uses different statistical tools to analyze it. Ben perceived that the "spiral of conversation" develops into the "arrow of knowledge". But Ben's introductory text too quickly became clogged with mathematical and philosophical minutia, and so no progress was made.

Bond & Fox are relatively new to the Rasch field. They remember what they needed to know, and what they didn't need to know; what they needed to understand, and what they didn't need to understand. They follow the "spiral of conversation", reiterating ideas and examples, but from different perspectives, so enabling the reader to accumulate knowledge and experience. The book is a pleasant read. It is nicely laid out and typeset. It has a comprehensive glossary and a suitably non-technical reference list. It can be used for personal study or as a classroom text.



Bond & Fox (2001) Figure 3.1, p. 22

Of course, the book runs into some trouble with expert reviewers, such as Wim van der Linden (2001), Mark Wilson (2002), Ed Wolfe (2002) and myself. I say "of course" because that is always the fate of introductory texts, such as school books. "Not one of the [middle-school science] books we reviewed reached a level that we could call scientifically accurate as far as the physical science contained therein." (John Hubisz, 2001). Sweeping generalizations, substantive short-cuts, and imprecise use of technical terms are features of introductory texts. One of my own pedanticisms is to distinguish between "precision" and "accuracy". Sure enough, the very first time "precision" appears (on p. xvii), "accuracy" is meant. Our reviewers point out more examples.

What is to be done? Hopefully, as with the first 100 years of the "King James" Bible, the most blatant inaccuracies and omissions will be corrected in the next edition. Instructors would be wise to read ahead and provide students with a sheet of annotations for each chapter, focusing on those features relevant to their students. As van der Linden writes, "a statement that is wrong can never be understood." Though van der Linden and I would undoubtedly disagree as to exactly what is right! However,

Table of Contents

we, along with Wilson, definitely agree that European research is conspicuous for its absence. The next edition must include Fischer & Molenaar's (1995) "Rasch Models" in its list of "Classic Reference Texts".

Wolfe correctly says that "there is insufficient information … to allow readers to apply the Rasch model to their own data sets." This flaw I find in many statistics texts, cookery books and guidebooks. So long as you follow closely along with the author, all is well, but branch out on your own, and you are quickly lost. Again, it requires an instructor or coach to help the neophyte, particularly in the operation of ever-changing software.

A fundamental question about "…: Fundamental Measurement …" is what is measurement? We have three answers. The Bond & Fox Glossary answer is "The location of objects along a single dimension on the basis of observations which add together." An amazingly succinct statement from which the Rasch model can be deduced in two ways. If the "observations" are scored qualitative indicators, then the "add together" specifies that the raw score is a sufficient statistic, and the Rasch model follows. On the other hand, if the "observations" are differences between person abilities and item difficulties, then the "add together" is Campbell concatenation of those differences, and the Rasch model follows. Thus Bond & Fox assert, in my words, "if you want to measure, you've got to use a Rasch model!"

Van der Linden answers: "I have difficulty recommending this book as an introductory text to modern measurement. Readers will be much better off with a balanced, elementary text [such] as Hambleton, Swaminathan and Rogers (1991)." But Hambleton's book is actually about IRT, i.e., data description in a quasi-linear framework, which is only measurement in a Stevens sense, as Michell (1999) "Measurement in psychology…", referenced in Bond & Fox, explains. This suggests that the relationship between Rasch and IRT needs more than one page in Bond & Fox, particularly because more advanced books in the field, such as Embretson & Hershberger's (1999) "New Rules" and van der Linden & Hambleton's (1997), "Handbook of Modern Item Response Theory" unabashedly classify Rasch under IRT. Wilson aptly summarizes the problem: "one person's oversimplification is another person's strong measurement philosophy."

Van der Linden's final recommendation is difficult to comprehend, "better still, the introductory chapter and chapters 5 and 6 in the original text by Rasch (1960)." Certainly read them: Chapter 1 ends up with a discussion of binomial trials using mice in a maze. Chapter 5 shows how to draw empirical ICCs. Chapter 6 straightens them out with a logistic transformation. But Wright & Masters (1982) "Rating Scale Analysis" does much, much more. For "non-dichotomous" readers, it is the obvious next step after Bond & Fox.

Wolfe's answer to "what is measurement" restricts its use: "there is no discussion of sample size requirements …. the book could lead a practitioner to erroneously conclude that the Rasch model can be utilized with virtually any data set." Certainly, we need to toss a coin more than once to check that it is fair. But how many times? 3, 5, 20, 1000? After 3 or 4 tosses we have a good idea. By the time we get to 10 we are convinced. Wright & Stone's (1979) "Best Test Design" (mentioned favorably by van der Linden) is based on the analysis of a data set comprising 35 children encountering 18 items. Only one Bond & Fox data set has fewer observations.

But this raises another fundamental question: when is there too little data for Rasch measurement to be informative? Is there a better alternative to Rasch for the analysis of small data sets? In the early days, test developers would make remarks such as "Rasch analysis ruined my perfectly good test!" In fact, Rasch analysis did not change their tests at all. It merely pointed out the flaws that were there all along. Ben Wright advised students to start Rasch analysis of their data just as soon as they started collecting it. Don't wait till you have 1000, 100 or even 10 cases to discover that a typographical error is making the answers to Question 3 unintelligible. Even when precision is lacking due to small sample size, a concern of Wolfe, data-quality-control and construct validity must still be there. If the data can possibly be Rasch analyzed, do it! With computers it takes almost no time, and what you learn may save you weeks of agony later. To take a statement by Fred Lord somewhat out of context, "Small N justifies the Rasch model."

Finally, along with Wolfe, "I applaud the authors …", and, along with Wilson, "I judge that Bond and Fox have largely succeeded …" Purchase this first edition, in anticipation of an even better second edition!

John M. Linacre

Hubisz, J. (2001) Review of Middle School Physical Science Texts. Final Report. The David and Lucile Packard Foundation. Grant #1998-4248.

van der Linden, W.J. (2001) Book Review - Applying the Rasch Model. International Journal Of Testing, 1(3&4), 319–326.

Wilson M. (2002) Book Review - Applying the Rasch Model: ... Applied Psychological Measurement, 26, 2 .

Wolfe E.W. (2002) Book Review - Applying the Rasch Model: ... Journal of Applied Measurement, 3, 4.

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
Copyright © 2002 Rasch Measurement SIG
Submissions welcomed!
Permission to copy is granted.
SIG Chair: Trevor Bond   SIG Secretary: Ed Wolfe

Musical Temperament

Musicians wrestle with equal-interval measurement. The twelve tones in the chromatic scale are shown in column 1 of the Table. "Just intonation" defines the consonance we hear when the two notes sounded together bear a simple numeric ratio to each other. In column 2 of the Table are the well-known ratios from Pythagoras. They are all derived from the ratios of the products of the prime numbers 2, 3, and 5. Unison has a frequency ratio of 1:1 (C:C, i.e., middle C on the piano to middle C) and the octave is 1:2 (C:c, i.e., middle C to the C eight white keys to the right). All tuning systems accept these fundamental ratios. Next comes the perfect fifth (C-G) with the ratio 2:3 and the perfect fourth (C-F) with the ratio 3:4.

Tone	Concordant Frequency Ratio	Exact Ratios
c	8 = 1:2	8.00
B		7.55
A#		7.13
A		6.73
G#		6.61
G	6 = 2:3	5.99
F#		5.66
F	5.33 = 3:4	5.34
E	**5**	**5.04**
D#		4.76
D	4.50 = 8:9	4.49
C#		4.24
C	4 = 1:1	4.00

Column 3 shows the steps of the chromatic scale indicated by equal frequency ratio divisions using the multiplicative constant of $(2)^{1/12}$ = 1.0595. This table shows how the Pythagorean system, derived from studies of the monochord, corresponds to the equal interval system. This also shows why "music" was part of the quantitative "quadrivium", the four liberal arts required in medieval times to advance from B.A. to M.A., the other three being arithmetic, geometry and astronomy. The "trivial" linguistic arts, the "trivium", were grammar, rhetoric and logic.

A natural major third is 5:4 and consonant. In exact ratio tuning, E with C, although a major third, is 5.04 to 4 and very discordant, . An exact cycle of four pure fifths, C-G-d-a-e' produces a major third out of tune when compared to a true major third of e' to c' of 5:4. This discord was known as the "comma of Didymus". Historians of musical theory consider this discord the reason why early medieval music extolled the tone, fourth, fifth, and seventh as "concordant" and treated the major third as "discordant", a *wolf* chord.

Until about the middle of the eighteenth century, ninety-five per cent of all pipe organs were tuned in mean-tone temperament. FinnViderø has made several recordings on just such an organ built in 1616 in the castle at Frederiksborg, Denmark. I have heard the organ in recital there, and it produces a truly beautiful sound.

The commonest mean-tone system uses eight major thirds (C-E, E-G#, Bb-D, D-F#, E-G#, Bb-D, D-F#, F-A, A-C#, Eb-G, G-B). These are exactly in tune and many common chords can be produced. Musicians have long thought these intervals were more pleasant to hear than those in equal temperament where no major third, or any other interval, is in natural tune save the octave. However, mean-tone temperament makes only about a dozen keys available, and the rest don't sound well.

In *Scales: Music and Measurement* (RMT 15:3, p. 838) four interesting comparisons were made between musical tunings and measurement theory:

1. "… mathematical perfections were claimed for 'Pythagorean' tunings, as they are now for some IRT models."

"Just intonation" is both mathematically precise and musically satisfying, but only in certain keys. It does not have objectivity or generality, i.e. it cannot survive a transposition to other keys. The "equal temperament" scheme overcame these drawbacks.

2. "… special practical virtues were perceived in 'just meantone' tunings, as they are now in raw-score-weighting schemes."

Certainly there are unique advantages for specific keys, but no generality is possible. Without generality, our music is restricted.

3. "Pythagorean tuning was simple, and musically effective. Its limitation was that only 11 of the 12 notes of an octave could be in tune simultaneously. Yet it was so easy and familiar, just as raw scores are today, … ."

We usually begin with the system first discovered and advance. Oppenheimer (1955) said, "all sciences, arise as refinements, corrections, and adaptations of common sense.… these are traits that any science must have before it pretends to be one. One is the quest for objectivity. I mean not in a metaphysical sense; but in a very practical sense … ." (p. 128)

4. "… in a remarkable parallel to the current proliferation of psychometric models, 'the history of tuning is saturated with clever and original theories that have no practical application.'"

A host of alternate tunings have been proposed, and many have a long historical lineage, but the major question has always remained, "What generality exists?" Without generality there can be no application.

Mark H. Stone

Oppenheimer, R. (1955). *Analogy in science.* Presented at the 63rd Annual Meeting of the American Psychological Association, San Francisco, CA, September 4, 1955.

The Measurement of Vision Disability

Robert Massof's (2002) article in *Optometry and Vision Science* is a landmark in the history of Rasch measurement publishing, a virtual textbook on what measurement has been and could be. It comprehensively integrates Rasch-calibrated vision disability scales not only into the history of vision measurement, but into the historical role of measurement in both commerce and science. Massof provides excellent accounts of measurement from the perspectives offered by Likert, Thurstone, Classical Test Theory, IRT, and Rasch. His detailed examination of Likert's argument and method is priceless.

Massof's application of five criteria of fundamental measurement theory (additivity, double cancellation, solvability, the "all gaps are finite" Archimedean axiom, and independence) as a basis for model choice is an apparently independent development of the same argument recently presented by George Karabatsos (Bond & Fox, 2001, p. 195), and develops in greater detail the same arguments as those presented by Wright (1985, etc.). Like Karabatsos, Massof shows that the mathematical structure of the 2P IRT model violates each of the requirements for fundamental measurement.

Plots comparing Rasch and 2P IRT analyses of the same data show the results to be much more similar than is the case in my own recent explorations in this area, due in part to Massof's "fortuitous choice of a data set that minimized the differences between models (e.g., there was relatively little variation between items in the discrimination parameter of the IRT model, effectively making it a 'noisy' Rasch model)" (Massof, p. 538).

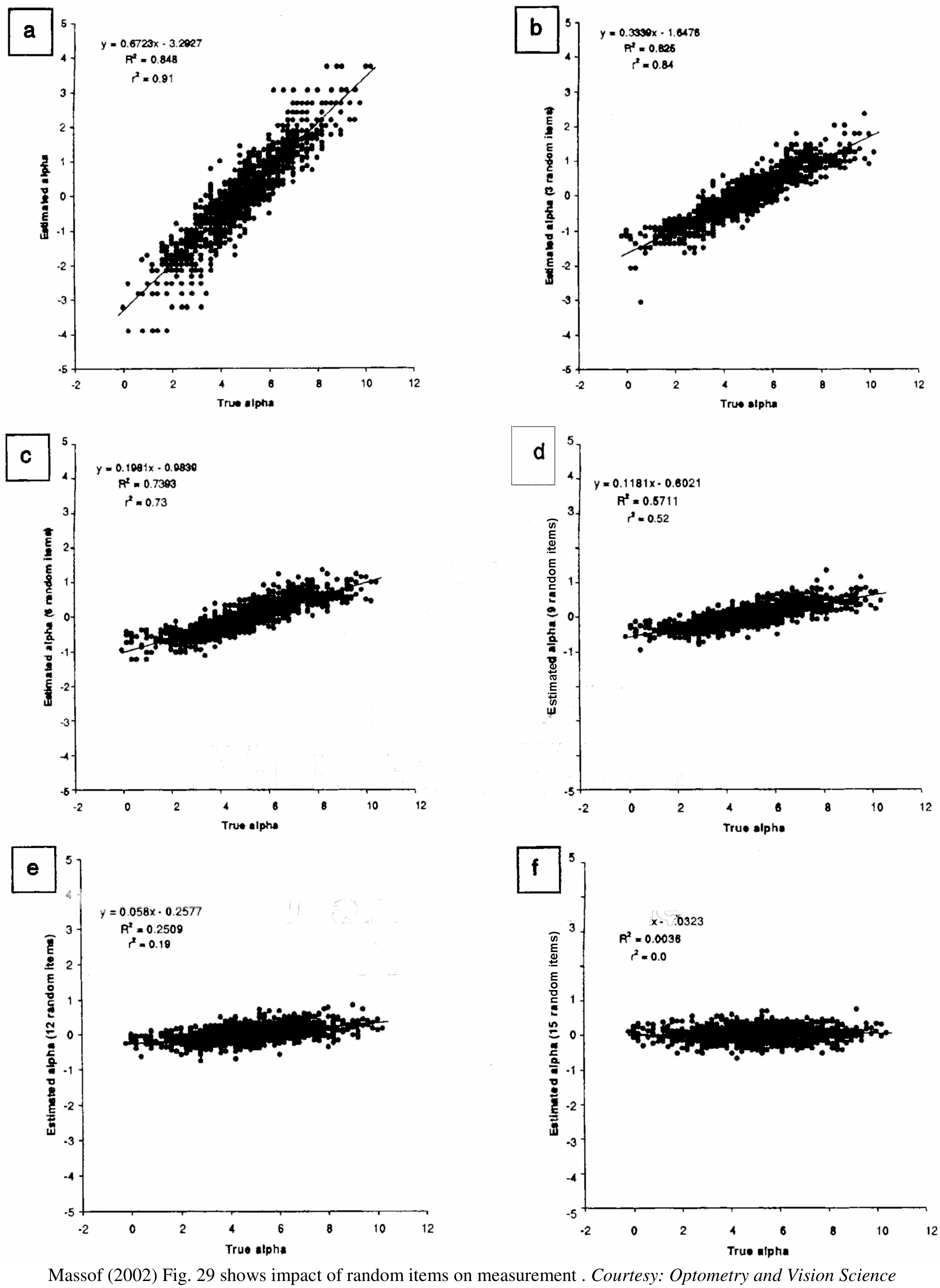
The article does not shy away from mathematical treatments and expositions of principle. It includes 33 equations, unusual for articles presenting measurement theory outside of technical psychometrics journals. Ten of the equations are associated with the IRT presentation, and 15 of them with Rasch models, and associated error, fit, and reliability statistics. Full credit is given where due, with extensive bibliographic citations (107 total) of Andersen, Andrich, Masters, Michell, Schulz, Smith, Wright, and others. Unfortunately, it appears that the article was in press when the Bond & Fox (2001) book came out, and so this resource is left unmentioned.

Empirical evaluations of statistics and models are the order of the day, with 37 numbered graphics in the article, the majority of which are scatterplots. The article includes a section focusing on Monte Carlo simulations that has the aim of demonstrating to the skeptic "that the Rasch model generates verifiable estimates of the latent variable." A data set of simulated observations from 1,000 respondents was designed from known values for 15 items, and was modified five times so as to include random responses for 3, 6, 9, 12, and all 15 items. The resulting calibrations and measures are plotted against their true values and against their fit statistics. Figure 29, reproduced from the article, shows the six plots of the measures versus their true values for each of the variations in the number of random items.

As expected, the scatterplots show a progressive movement away from 1) the identity line to a horizontal line centered at 0.0 logits for the comparisons of the calibrations and measures with themselves; and 2) a largely vertical spread to a horizontal line centered at 0.0 logits for the comparisons of the calibrations and measures with their fit statistics. The latter are interesting for their independent support for work by Richard Smith showing that misfitting anomalous responses are easiest to detect when the proportion of problematic items and/or examinee/respondents is low.

The standardized infit statistics for simulations with fewer random items easily isolate these "noisy" items at the high, positive end, but when there are more random items than not, the fit distribution settles right into the -2.0 to 2.0 range where one might think all is well (apart from the fact that the items all calibrate to 0.0). The results emphasize the value of **strong theory** and **close study of construct validity**, since random data are not likely to be produced from carefully designed questions asked of persons sampled from a relevant population.

The article briefly takes up some neglected history of Rasch



Massof (2002) Fig. 29 shows impact of random items on measurement . *Courtesy: Optometry and Vision Science* applications to vision disability measurements, recounting the work of Wright, Lambert, and Schulz at Hines VA Hospital in the Chicago suburbs in the 1980s. Massof (p. 545) says:

“Like many other milestones in psychometrics, the use of Rasch analysis to measure vision disability can trace its origins to the University of Chicago. Georg Rasch was the father of Rasch analysis, but Benjamin Wright must be considered its legal guardian. Wright and his students and colleagues at the University of Chicago further developed and advanced Rasch’s models, developed and validated analytic tools, and promoted and facilitated applications of Rasch models to a wide variety of fields.”

Massof (p. 548) also makes brief notes of the convergence of different approaches to measuring visual abilities on a common construct, with the realization that the “different measurements can easily be transformed into a common unit.”

The article concludes (p. 550) with strong statements on the value of Rasch measurement, statements that are supported by the thorough and extensive arguments and demonstrations presented:

“Many scientists have long been suspicious of the cavalier assertions by developers and users of visual function questionnaires that the average of patient ratings across questionnaire items is a valid measurement scale. With Rasch analysis, the validity of an instrument does not depend on inferential arguments and correlations with external variables. Rather, it exists on objective statistical tests of the model as an explanation of the data.”

Massof's presentation of this work in the context of a field that has a long history of creating and maintaining reference standard metrics for its primary variables of interest bodes well for the extension of metrological networks away from their historical origins in the domains of physical variables into new homes in the domains of psychosocial variables. Those who act on the opportunity for the advancement of scientific and human values presented by the work of Rasch and others stand to make fundamental contributions. Massof's article will no doubt prove to be a powerful motivation to many who read it.

William P. Fisher, Jr.

Bond, T., & Fox, C. (2001). Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Massof, R. W. (2002). The measurement of vision disability. Optometry and Vision Science, 79(8), 516-52.

Wright, B. D. (1985). Additivity in psychological measurement. In E. Roskam (Ed.), Measurement and personality assessment. North Holland: Elsevier Science.

Report from Italy

Professor David Andrich from Murdoch University, Perth, Western Australia, held a Lecture at the Maugeri Foundation in Pavia, Italy, on June 3rd 2002. The Lecture was entitled "Fundamental measurement: principles and practice of Rasch analysis," and it was nested within a half-day conference entitled "Rasch analysis: the same meter from Human Sciences to Medicine: a meeting with David Andrich."

The Salvatore Maugeri Foundation is the largest Institute for Rehabilitation Medicine in Italy (www.fsm.it). Luigi Tesio, M.D., from the Italian Chapter of the IOM, is heading there a 50-bed inpatient rehabilitation unit.

Some 150 persons attended, coming from all parts of Italy. About half of them came from a mathematical-statistical background, while the other half was made up by physicians, physical therapists, and psychologists: apparently, quite a heterogeneous audience. It was a hard challenge to tell something interesting, yet technically elevated, to either of the halves.

The conference was opened by Prof. Giorgio Vittadini, Director of the Center for Research on Services to the Persons, which is co-sponsored by five Italian Universities. The Center studies the optimization of services such as health care, social assistance, education, and urban life planning. It soon captured the enormous strength of Rasch modeling in the construction of valid scientific measures of person-based variables.

David Andrich introduced the concepts of fundamental measurement. His presentation enabled statisticians to appreciate the originality of the mathematics (basically, the "prescriptive" rather than "descriptive" nature of the model), yet simultaneously health care and education professionals perceived the versatility of the model in constructing person variables. The clue to this success was the emphasis Professor Andrich placed on the philosophical thinking lying behind the mathematics themselves, i.e., what counting and measuring are, why we need to challenge the data with a model , not adapting the model to data, etc.). This allowed the whole audience to perceive the possibility of a measurement paradigm with equally validity for the hard sciences and psychology, education and medicine.

Luigi Tesio, the third and last speaker, focused on the very practical uses of the Rasch model in Rehabilitation Medicine. He presented examples of variables he had constructed through Rasch modeling, and their applications to the management of rehabilitation units.

Rasch modeling was not unknown in Italy; but, it was generally felt that this Conference helped bridge the gaps between different users and opened the door for new ones.

Luigi Tesio, M.D.

Immediate Raw Score to Logit Conversion

A supposed flaw in the Rasch model can be used to great advantage. Bruce Thompson informs us that Fan (1998) and MacDonald and Paunonen (2002) support his perception that the correlation between Rasch measures and raw scores is always .97. ±.02, i.e., is effectively linear. Malec et al. (2000) report a correlation of .98 for their clinical data. If this also holds true for your data then you can immediately convert raw scores to logits!

What conditions must hold for this hold true?
(a) The raw scores are all be on the same set of items.
(b) The proportion of very high and very low scores is low.

Then we have these convenient relationships. For each person *n* and item *i* of a test of length L, there is an observation X_{ni}. Its Rasch model expectation is E_{ni}, and the modeled variance of the observation around its expectation is Q_{ni} (see Wright and Masters, 1982, p. 100). Thus, person *n*'s raw score, R_n , and raw score "error" variance, V_n , are given by:

$$R_n = \sum_{i=1}^{L} E_{ni} \qquad V_n = \sum_{i=1}^{L} Q_{ni}$$

An approximate conversion factor between raw scores and logits for person *n* of ability B_n , at the center of the test characteristic curve is the slope of the curve: $1/V_n$.

Suppose we know the observed standard deviation, *S*, of the raw scores of a sample on a test and the reliability estimate (KR-20, Cronbach Alpha) of the test for the same sample, *R*. Then, from the definition of Reliability as "True Variance" / "Observed Variance", raw score error variance = $S^2(1\text{-}R)$. So that the raw-score-to-Rasch-measure conversion factor is $1/(S^2(1\text{-}R))$.

It is conventional to set the origin of the logit scale in the center of the test, i.e., where the raw score is about 50%. This gives the convenient raw score-to-measure conversion:
$B_n = (R_n – (\text{Maximum score} + \text{Minimum score})/2\ ) / S^2(1\text{-}R)$
And the standard error of B_n is $1/\sqrt{V_n} = 1/(S.\sqrt{(1\text{-}R)})$ logits.

Applying this to the Wright & Masters (1982) "Liking for Science" data: Raw score S.D. = 8.6, Reliability = .87, minimum score = 0, maximum score = 50. Measure for raw score of 20 = -0.52, for 30 = 0.52, with S.E. ±.32. *Winsteps* says –0.55, 0.61 with S.E. ± .34. So that the results are statistically equivalent.

John M. Linacre

Journal of Applied Measurement
Volume 3, Number 3. Autumn 2002

Test Scores, Measurement, and the Use of Analysis of Variance: An Historical Overview. *Joseph Romanoski and Graham Douglas*

Using Rasch Measurement to Investigate the Cross-form Equivalence and Clinical Utility of Spanish and English Versions of a Diabetes Questionnaire: A Pilot Study. *Ben Gerber, Everett V. Smith, Jr., Mariela Girotti, Lourdes Pelaez, Kimberly Lawless, Louanne Smolin, Irwin Brodsky, and Arnold Eiser*

Moving the Cut Score on Rasch Scored Tests. *G. Edward Miller and S. Natasha Beretvas*

Examining Item Difficulty and Response Time on Perceptual Ability Test Items. *Chien-Lin Yang, Thomas R. O'Neill, and Gene A. Kramer*

When Raters Disagree, Then What: Examining a Third-rating Discrepancy Resolution Procedure and Its Utility for Identifying Unusual Patterns of Rating. *Carol M. Myford and Edward V. Wolfe*

Understanding Rasch Measurement: Understanding Resistance to the Data-model Relationship in Rasch's Paradigm: A Reflection for the Next Generation. *David Andrich*

For subscriptions, submissions, back-issues and instructors' sample copies, contact:
Richard M. Smith, Editor
Journal of Applied Measurement
P.O. Box 15171, Sacramento, CA 95851
916-286-8804, rsmith.arm@att.net
http://home.att.net/~rsmith.arm

Congratulations to **George Karabatsos**, author of "The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory." *Journal of Applied Measurement, 2001, 2, 4, 389-423,* for winning the 2002 *New Investigator Award* of the *Society for Mathematical Psychology*.

Fan, X. (1998) Item Response Theory and classical test theory: An empirical comparison of their item/person statistics. Educational and Psychological Measurement, 58, 357-381.

MacDonald, P., & Paunonen, S.V. (2002) A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. Educational and Psychological Measurement, 62.

Malec J. F., Moessner, A. M., Kragness, M., and Lezak, M.D. (2000) Refining a measure of brain injury sequelae to predict postacute rehabilitation outcome: rating scale analysis of the Mayo-Portland Adaptability Inventory (MPAI). Journal of Head Trauma Rehabilitation, 15 (1), 670-682.

What do Infit and Outfit Mean-Square and Standardized mean?

These are all “fit” statistics. In a Rasch context they indicate how accurately or predictably data fit the model.

Infit means inlier-sensitive or information-weighted fit. This is more sensitive to the pattern of responses to items targeted on the person, and *vice-versa*. For example, infit reports overfit for Guttman patterns, underfit for alternative curricula or idiosyncratic clinical groups. These patterns can be hard to diagnose and remedy.

Outfit means outlier-sensitive fit. This is more sensitive to responses to items with difficulty far from a person, and *vice-versa*. For example, outfit reports overfit for imputed responses, underfit for lucky guesses and careless mistakes. These are usually easy to diagnose and remedy.

Mean-square fit statistics show the size of the randomness, i.e., the amount of distortion of the measurement system. 1.0 is their expected values. Values less than 1.0 indicate observations are too predictable (redundancy, data overfit the model). Values greater than 1.0 indicate unpredictability (unmodeled noise, data underfit the model). Statistically, mean-squares are chi-square statistics divided by their degrees of freedom. Mean-squares are always positive.

In general, mean-squares near 1.0 indicate little distortion of the measurement system, regardless of the standardized value. Evaluate mean-squares high above 1.0 before mean-squares much below 1.0, because the average mean-square is usually forced to be near 1.0.

Outfit problems are less of a threat to measurement than Infit ones, but are easier to manage. To evaluate the impact of any misfit, replace suspect responses with missing values and examine the resultant changes to the measures.

Standardized fit statistics (*Zstd* in some computer output) are t-tests of the hypothesis “Do the data fit the model (perfectly)?” These are reported as z-scores, i.e., unit normal deviates. They show the improbability of the data, i.e., its significance, if the data actually did fit the model. 0.0 are their expected values. Less than 0.0 indicates too predictable. More than 0.0 indicates lack of predictability. Standardized values are positive and negative.

Standardized fit statistics are usually obtained by converting the mean-square statistics to the normally-distributed z-standardized ones using the Wilson-Hilferty cube root transformation.

Anchored runs:
Anchor values may not exactly accord with the current data. To the extent that they don't, fit statistics can be misleading. Anchor values that are too central for the current data tend to make the data appear to fit too well. Anchor values that are too dispersed for the current data tend to make the data appear noisy.

John M. Linacre

Mean-square Value	Implication for Measurement
> 2.0	Distorts or degrades the measurement system. May be caused by only one or two observations.
1.5 - 2.0	Unproductive for construction of measurement, but not degrading.
0.5 - 1.5	Productive for measurement.
< 0.5	Less productive for measurement, but not degrading. May produce misleadingly high reliability and separation coefficients.

Standardized Value	Implication for Measurement
≥ 3	Data very unexpected if they fit the model (perfectly), so they probably do not. But, with large sample size, substantive misfit may be small.
2	Data noticeably unpredictable.
-1.9 – 1.9	Data have reasonable predictability.
≤ -2	Data are too predictable. Other “dimensions” may be constraining the response patterns.

April 2003, Chicago

April 19-20, Saturday-Sunday
An Introduction To Rasch Measurement: Theory And Applications.
At the University of Illinois at Chicago. The workshop will be conducted by Dr. Everett V. Smith Jr. and Richard M. Smith. 312/996-5630
evsmith@uic.edu

April 21-25, Monday-Friday
AERA Annual Meeting. *www.aera.net*

April 26-27, Saturday-Sunday
Ben Wright Testimonial, Rehabilitation Institute of Chicago.

April 28-29, Monday-Tuesday
***Facets* Workshop,** CORE, Evanston
www.winsteps.com/seminar.htm

April 30-May 1, Monday-Tuesday
***Winsteps* Workshop**, CORE, Evanston
www.winsteps.com/seminar.htm

The Standardization of Mean-Squares

The reason for standardizing the infit and outfit mean square statistics is to allow their statistical significance, or p-values, to be more conveniently represented. A familiar scale to use for this purpose is the Z-scale, or standard normal scale. Most of us are familiar enough with this scale that we don't even need to look up the p-value of 1.96. And we know that a Z-score over 2.0 is "statistically significant." In contrast, one does not immediately know the statistical significance of variables from other commonly-used reference distributions, such as the chi-square distribution. The distribution changes with its degrees of freedom!

A general formula for converting a variable, X, to the standard normal variate, Z, is:

$$Z(X) = \frac{X - \mu_x}{\sigma_x} \qquad (1)$$

Now one may be certain that *Z*(*X*) has a mean of 0 and a variance of 1, but unless X is normally distributed to begin with, the p-values of Z(X) in a standard normal distribution do not necessarily agree with the p-values of X in its own distribution. For instance, a "normally distributed" variable has no skew, but chi-square distributions are skewed.

Wilson & Hilferty (1931) found a way to transform a chi-square variable to the Z-scale so that their p-values closely approximated. Since chi-square distributions are skewed, the transformation has an extra layer of complexity. The first step in the transformation is to transform the chi-square statistic to a more normally-distributed variable. They showed that the pth root of a chi-square variable divided by its degrees of freedom, *n*, is approximately normally distributed and that

if $$X = \left(\frac{\chi_n^2}{n}\right)^{1/p} \qquad (2)$$

then

$$\sigma_x^2 \approx \left(\frac{1}{p^2}\right)\left(\frac{2}{n}\right), \qquad (3)$$

and

$$\mu_x \approx \left[1 - \sigma^2\right]. \qquad (4)$$

Wilson & Hilferty chose *p*=3 (the cube root) for their transformation. The second step in the transformation is to substitute the results of Equations (2) through (4) into Equation (1). The complete transformation in terms of a chi-square variable, *Y*, with degrees of freedom, *n*, is:

$$W(Y) = \frac{\left(\frac{Y}{n}\right)^{1/3} - \left(1 - \left(\frac{1}{9}\right)\left(\frac{2}{n}\right)\right)}{\sqrt{\left(\frac{1}{9}\right)\left(\frac{2}{n}\right)}} \qquad (5)$$

Notice that Equation (5) has the basic form of a normalizing transformation, but is actually a normalizing transformation of a transformation! The p-values of W(*Y*) are very close to those of a standard normal variable, as desired. That is, if *Z* is a standard normal variable, P(*Z* < W(y)) ≈ P(*Y* < y). So W(Y) approximates a *t* statistic.

The expectation of a chi-square variable, Y, is its degrees of freedom *n*. So the expectation of Y/*n* is 1. Let's call this v_i. The model variance of Y is 2*n*. So the variance of Y/*n is 2/n,* let's call this q_i^2. Substituting in (5) and simplifying, we can see that (5) parallels the formula for the standardized weighted mean square at the bottom of Table 5.4a in *Rating Scale Analysis* (Wright & Masters, 1982, p. 100):

$$t_i = (v_i^{1/3} - 1)(3/q_i) + q_i/3. \qquad (6)$$

In RSA, the residuals comprising the v_i have been weighted, embodying an unstated assumption that the distributional characteristics of weighted and unweighted mean-squares are the same. The unweighted form, which matches (5) exactly, substitutes u_i for v_i and the unweighted mean-square variance for the weighted one. Since the actual degrees of freedom for residual chi-squares are difficult to compute, RSA estimates them from the model distributions of the observations.

Matthew Schulz

Wilson, E. B., & Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences of the United States of America*, *17*, 684-688.

Midwestern Objective Measurement Seminar

The next meeting of the Midwestern Objective Measurement Seminar (MOMS) will be held on **Friday, December 13, 2002, 9:00 a.m.- 4:00 p.m.**, at the University of Illinois at Chicago, **1040 West Harrison Street, Chicago, in the ECSW building, Room 3427** (third floor). It will be jointly sponsored by the University of Illinois at Chicago (UIC) and the Institute for Objective Measurement, Inc. (IOM). Dr. Everett Smith will serve as host. There is no registration fee.

Lunch: 12:00 noon – 1:00 p.m. Participants are encouraged to bring their own lunches or eat at neighboring Greek, Italian, and American-style restaurants

Parking structure is between ECSW and Racine Street on the north side of Harrison. Cost is $7.50 for the day.

Facets, Factors, Elements and Levels

Early test analysis was based on a simple rectangular conception: people encounter items. This could be termed a "two-facet" situation, loosely borrowing a term from Guttman's (1959) "Facet Theory". From a Rasch perspective, the person's ability, competence, motivation, etc., interacts with the item's difficulty, easiness, challenge, etc., to produce the observed outcome. In order to generalize, the individual persons and items are here termed "elements" of the "person" and "item" facets.

Paired comparisons, such as a Chess Tournament or a Football League, are one-facet situations. The ability of one player interacts directly with the ability of another to produce the outcome. The one facet is "players", and each of its elements is a player. This can be extended easily to a non-rectangular two-facet design in order to estimate the advantage of playing first, e.g., playing the white pieces in Chess. The Rasch model then becomes:

$$\log\left(\frac{P_{nm}}{P_{mn}}\right) = B_n - B_m + A_w$$

where player *n* of ability B_n plays the white pieces against player *m* of ability B_m, and A_w is the advantage of playing white.

A three-facet situation occurs when a person encountering an item is rated by a judge. The person's ability interacting with the item's difficulty is rated by a judge with a degree of leniency or severity. A rating in a high category of a rating scale could equally well result from high ability, low difficulty, or high leniency.

Four-facet situations occur when a person performing a task is rated on items of performance by a judge. For instance, in Occupational Therapy, the person is a patient. The rater is a therapist. The task is "make a sandwich". The item is "find materials". A typical Rasch model for a four-facet situation is:

$$\log\left(\frac{P_{nmijk}}{P_{nmij(k-1)}}\right) = B_n - A_m - D_i - C_j - F_{ik}$$

where D_i is the difficulty of item *i*, and F_{ik} specifies that each item *i* has its own rating scale structure, i.e., the "partial credit" model.

And so on, for more facets. In these models, no one facet is treated any differently from the others. This is the conceptualization for "Many-facet Rasch Measurement" (Linacre, 1989) and the *Facets* computer program.

Of course, if all judges are equally severe, then all judge measures will be the same, and they can be omitted from the measurement model without changing the estimates for the other facets. But the inclusion of "dummy" facets, such as equal-severity judges, or gender, age, item type, etc., is often advantageous because their element-level fit statistics are informative.

Multi-facet data can be conceptualized in other ways. In Generalizability theory, one facet is called the "object of measurement". All other facets are called "facets", and are regarded as sources of unwanted variance. Thus, in G-theory, a rectangular data set is a "one-facet design".

In Gerhard Fischer's Linear Logistic Test Model (LLTM), all non-person facets are conceptualized as contributing to item difficulty. So, the dichotomous LLTM model for a four-facet situation (Fischer, 1995) is:

$$\log\left(\frac{P_{nmij}}{1-P_{nmij}}\right) = B_n - \left(\sum_{l=1}^{p} w_{il}\alpha_l + \{c\}\right)$$

where p is the total count of all item, task and judge elements, and w_{il} identifies which item, task and judge elements interact with person *n* to produce the current observation. The normalizing constraints are indicated by {c}. In this model, the components of difficulty are termed "factors" instead of "elements", so the model is said to estimate p factors rather than 4 facets. This is because the factors were originally conceptualized as internal components of item design, rather than external elements of item administration.

David Andrich's Rasch Unidimensional Measurement Models (RUMM) takes a fourth approach. Here the rater etc. facets are termed "factors" when they are modeled within the person or item facets, and the elements within the factors are termed "levels". Our four-facet model is expressed as a two-facet person-item model, with the item facet defined to encompass three factors. The "rating scale" version is:

$$\log\left(\frac{P_{nmijk}}{P_{nmij(k-1)}}\right) = B_n - \delta_{mij} - F_k$$

where D_i is an average of all δ_{mij} for item *i,* A_m is an average of all δ_{mij} for task *m,* etc.

This approach is particularly convenient because it can be applied to the output of any two-facet estimation program, by hand or with a spreadsheet program. Missing δ_{mij} may need to be imputed. With a fully-crossed design, a robust averaging method is standard-error weighting (RMT 8:3 p. 376). With some extra effort, element-level quality-control fit statistics can also be computed.

John M. Linacre

Fischer, G.H., & Molenaar, I.W. (Eds.) (1995) Rasch Models: Foundations, Recent Developments and Applications. New York: Springer.

Guttman, L. (1959) A structural theory for intergroup beliefs and action. American Sociological Review, 24, 318-328.

Random Effects Rasch Model

Georg Rasch was concerned to construct measurement models with "parameters that are specific to the individuals" (1960, p. 9). An early extension of the Rasch model specified that the "individuals" were to be treated as random variables drawn from some distribution. This is now done routinely using the PROX algorithm and Marginal Maximum Likelihood Estimation.

Lalitha Sanathanan in "Some Properties of the Logistic Model for Dichotomous Response" (JASA, 69, 347, 744-749, **1974!**) attempts this, but immediately encounters a major hurdle, the multiplicative form of the Rasch model then in common use. After Herculaean effort, she derives a simple approximation, but is forced to present it in additive form in order to make it tractable. Rewritten, the approximation is:

$$\log\left(\frac{S_i}{N - S_i}\right) = \alpha - \beta D_i$$

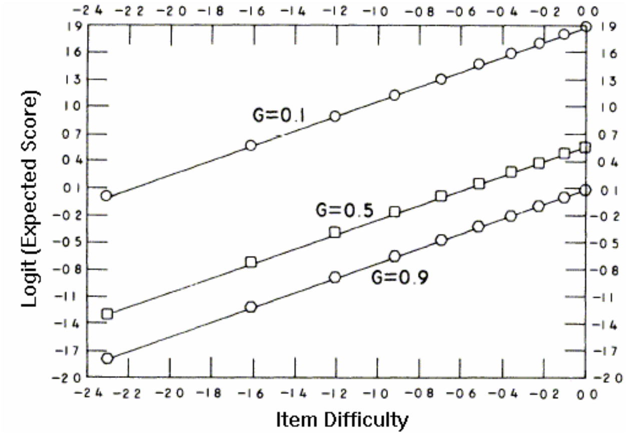
where N is the sample size, S_i is the number of correct answers to item *i,* and D_i is the difficulty of item *i.* α and β are sample dependent, but how they relate to the mean and standard deviation of the distribution is omitted from the paper. In fact, this formulation is equivalent to the PROX equation for a sample distributed N(μ,σ), when

$$\beta = \left(1 + \sigma^2/2.9\right)^{-1/2}$$

and

$$\alpha = \mu\beta$$

Sanathanan realizes that she has "shown how the parameters in the model can be calculated in a rough ready manner" (p. 749), but the utility of her insight was lost due to her abstruse math.



Sanathanan's plot of expected score against item difficulty for different sample distributions.

Australia, 2004

January 5-16. **Rasch Measurement Introductory and Intermediate Courses**, Perth, Western Australia. *chillino@murdoch.edu.au*

January 19. RUMM2010 one-day workshop. Perth, Western Australia. *chillino@murdoch.edu.au*

January 20-22. The 2nd International Conference on Measurement in Health, Education, Psychology and Marketing: Developments with Rasch models. Perth, Western Australia. *chillino@murdoch.edu.au*

June 28 – July 3. Mon-.Sat. IOMW at Quest Marlin Cove Resort, Trinity Beach, Cairns, Queensland. *www.soe.jcu.edu.au/iomw/*

June 28-29, Mon.-Tues. Pre-conference workshops (software etc.)

June 30-July 2. Wed.-Friday. IOMW itself, incl. one 'Teachers' Day' of presentations that would appeal classroom practitioners. And other specialist themes.

July 3. Sat. Post-conference group visits, e.g. to Great Barrier Reef, to rainforest, whale-watching.

Journal of Applied Measurement
Volume 3, Number 4. Winter 2002

For subscriptions, etc., see Box on p. 877

The Illusion of Measurement: Rasch versus 2-PL

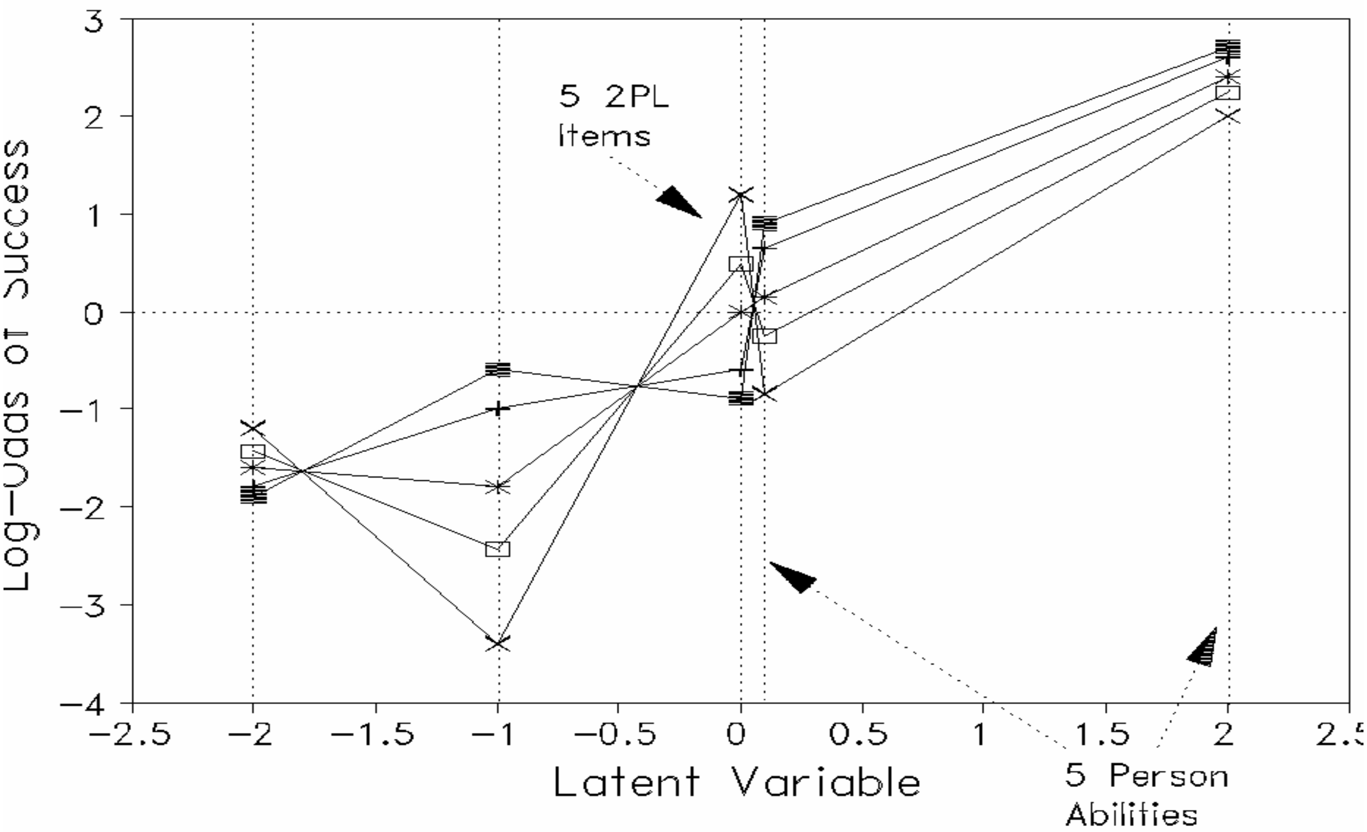
Many researchers who are attempting to measure latent constructs appreciate the special properties of the Rasch model and view it as an ideal model, but, at the same time, they tend to complain about the "inflexibility" of the model when it comes to "explaining" data. The two-parameter logistic model is then seen as a possible resort. However, with discrimination varying from item to item, the very meaning of the construct changes from point to point on the dimension. In other words, measurement in its true sense has not been achieved.

In the dichotomous Rasch model, for each item response only two parameters are relevant, the item location parameter δ and the person location parameter ξ (using Rasch's multiplicative notation). The probability of a correct response then is $\xi/(\xi+\delta)$ (Rasch 1960/1980, p.107). As Georg Rasch points out, if we knew the exact item parameter and the probability of a correct response, the person location could be computed directly. *Vice versa*, if we knew the person parameter and the probability, we could compute the item location. Under the 2PL model, this is not possible without further information because there are infinitely many combination of item difficulty and discrimination which yield the same probability for a given person location.

The plot shows the adjusted log-odds of success on five 2-PL items for persons at five ability levels. The five ability levels are –2, -1, 0, 0.1, and 2 logits. The items have difficulty and (discrimination) of –2 (0.8), -1.0 (1.8), 0.0 (0.4), 0.1 (1.5), and 2 (1.2). For each person-item encounter, the 2-PL probability of success is computed. This is converted into log-odds and adjusted for person ability. The plot thus shows the local Rasch difficulty of each item for each person. If the items were in accord with the Rasch model, this plot would collapse to an identity line. Since the 2-PL item characteristic curves intersect, there is a different "Rasch item difficulty" for each item for each level of person ability. In other words, the meaning of the construct defined by the item difficulty differs for each person location. Thus the apparent advantage of better describing the data set when using the 2PL, rather than a Rasch model, comes at the expense of a highly fuzzy definition of the latent continuum . "Measurement" becomes an illusion, because there is no precise definition of what is being measured.

Thomas Salzberger
Vienna University of Economics
and Business Administration
Austria

Rasch, Georg (1960/1980). Probabilistic Models for Some Intelligence and Attainment Tests. Danish Institute for Educational Research & *www.rasch.org/books.htm*



Plot: Local item difficulty of 5 2PL items for 5 abilities.