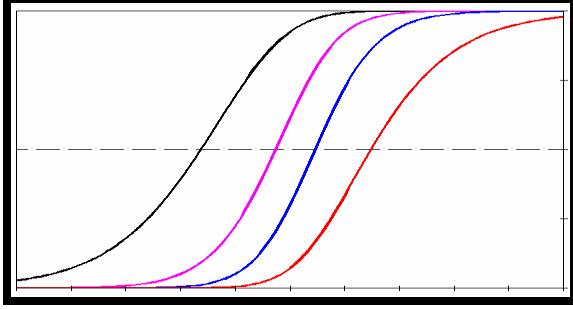
Thurstone-type Thresholds



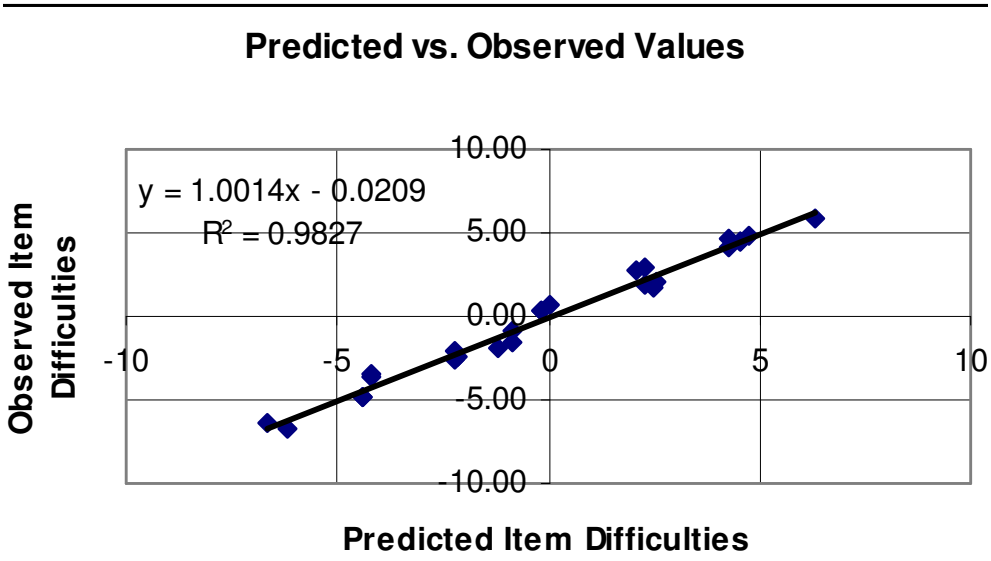
RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 17 No. 3 Winter 2003 ISSN 1051-0796

Item Specification vs. Item Banking

Our thesis is simple and straightforward. It is not necessary to have a bank of items for measuring a construct when we possess an algorithm for writing an item at any desired level of difficulty. The algorithm is the *key to the bank*, so to speak. If one has the key, the bank is open.



Bruce Choppin (1968) was an early Rasch pioneer who promoted item bank development. Items representative of the variable of interest are banked and selected for use as required. Leveled paper-pencil tests can be quickly assembled from the bank of items based on their associated item calibrations and item use histories. Also, computer based adaptive tests can be assembled electronically and targeted to each examinee. As useful as item banking has proven to be it is possible to move beyond the banking of individual items and their associated item statistics.

When enough is known about what causes item difficulty a specification equation can be written that yields a theory based item calibration for any item the computer software designs. An item's calibration is seen to be the consequence of decisions the computer software makes in constructing the item. This process mimics the steps a human item writer takes in constructing an item, albeit, with more control over the causal recipe for item difficulty. A thesis of this paper is that when asserting that a measure possesses construct validity there is no better evidence than demonstrated experimental control over the causes of item difficulty.

A measurement instrument embodies a construct theory; a story about what it means to move up and down a scale (Stenner, Smith & Burdick, 1983). Such a theory should be vigorously tested. In a demonstration of these methods Stone (2002) theorized that the difficulty of short term memory and attention items (Knox Cube Test) was caused by (1) number of taps, (2) number of reverses in the direction of the tapping pattern and (3) total distance in taps for the pattern. This theory was tested by regressing the observed item difficulties on the above mentioned three variables. The Figure plots the correspondence between predicted (theoretical) item difficulties and observed item difficulties. Ninety-eight percent (98%) of the variation in observed item difficulties was explained by number of taps

Table of Contents

(standardized Beta=.80) and distance covered (standardized Beta=.20). Number of reverses in the context of these two predictors made no independent contributions. An earlier study (Stenner and Smith, 1982) using different samples of items and persons found that an equation employing the same two variables explained 93% of the item difficulty variance. Finally, Stone (2002) re-analyzed KCT-like items developed over the last century and found a striking correspondence between the two variable theory and observation. We should note that there is some uncertainty in the observed item difficulties analyzed in these studies, suggesting that the dis-attenuated correlation between theory and observation approaches unity.

When item difficulties and by implication person measures are under control of a construct theory and associated specification equation it becomes possible to engineer items on demand. No need to develop more items than you need, pilot test these items, estimate item calibrations and then bank the best of these items for use on future instruments. Rather, when an instrument is needed an algorithm generates items to a target test specification along with calibrations for each item.

Applications that incorporate the above ideas are under development for the next KCT revision and for an on line reading program that builds reading items real time as the reader progresses through an electronic text.

Some of the practical benefits of what might be called theory referenced measurement are (1) if the process yields reproducible person measures, then evidence for construct validity is strong, (2) test security is facilitated because there are no extant instruments that would be compromised upon release, and (3) a fully computerized procedure keeps the process under tight quality control at a fraction of the cost of traditional item standardization procedures.

Rasch Measurement SIG Chair and Secretary Elections

Election of the new SIG officers will take place at the SIG Business Meeting during the AERA Annual Meeting in San Diego, April 2004.

The SIG *Chair* and *Secretary* commence their 2 year terms at the AERA Meeting. They must be AERA and SIG members. The **Chair** oversees SIG activities, represents the SIG to AERA and chairs the Annual Business Meeting at AERA. The **Secretary** oversees the SIG mailing list and bank account. These are maintained by AERA. **Please email *Chair* and *Secretary* nominations to the address below. Self-nomination is welcome.**

Trevor Bond
Rasch Measurement SIG Chair
Trevor.Bond@jcu.edu.au

Finally, one well-recognized means of supporting an inference about what causes item difficulty is to experimentally manipulate the variables in the specification equation and observe whether the predicted item difficulties materialize when examinees take the items. In building the latest version of the KCT a part of the scale had an insufficient number of items. The specification equation was used to engineer candidate items to fill in the space. Subsequent data collection confirmed that the items behaved in accord with theoretical predictions (Stone, 2002). Although this exercise involved only four items, it suggests that the construct specification equation is a causal representation (rather than merely descriptive) of the construct variance.

Reflecting on this extraordinary agreement between observation and theory suggests two conclusions: (1) the specification equation affords a nearly complete account of what makes items difficult, and (2) the Rasch model used to linearize the ratios of counts correct/counts incorrect must be producing an equal interval scale or a linear equation could not account for such a high proportion of the reliable variation in item difficulties.

Measurement of constructs evolves along a predictable course. Early in a constructs history measurements are subjective, awkward to implement, inaccurate and poorly understood. The king's foot as a measure of length is an illustration. With time, standards are introduced, common metrics are imposed, artifacts are adopted, (e.g. the meter bar) precision is increased and use becomes ubiquitous. Finally, the process of abstraction leaps forward again and the concrete artifact based framework is left behind in favor of a theoretical process for defining and maintaining a unit of length (oscillations of a cesium atom). Human science instrumentation similarly evolves along this pathway of increasing abstraction. In the early stages a construct and unit of measurement are inseparable from a single instrument. In time multiple instruments come to share a common metric, item banking becomes commonplace and finally, the construct is specified. When a specification equation exists for a construct and accounts for a high percentage of the reliable variance in item difficulties (or ensembles) the construct is no longer operationalized by a bank of items but rather by the causal recipe for generating items with pre-specified attributes.

Jack Stenner & Mark Stone

Choppin, B. (1968). Item banking using sample-free calibration. *Nature, 219 (5156)*, 870-872.

Stenner, A. J. & Smith, M. (1982). Testing construct theories. *Perceptual and Motor Skills, 55*, 415-426.

Stenner, A. J., Smith, M. & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement, 20* (4), 305-315.

Stone, M. H. (2002). Quality control in testing. *Popular Measurement, 4* (1), 15-23.

Stone, M. H.(2002). *Knox's cube test – revised*. Wood Dale: Stoelting.

IOMW-XII: An Australian Contribution

June 30th, Wednesday – July 2nd, Friday, 2004
Cairns, Queensland, Australia

The Twelfth International Objective Measurement Workshop will focus on developments in four professional strands as well as the usual more general Rasch-based presentations.

Education Chair:	**Psychology** Chair:	**Business** Chair:	**Health Care** Chair:
Juho Looveer	Karen M. Schmidt	Thomas Salzberger	Robert W. Massof

Conference Registration Fee:
$75.00 US international delegates; $75.00 AUD for Australian & New Zealand delegates
$50 AUD for full-time students; $20.00 AUD for Teachers' Day

Abstracts deadline: 31 January 2004 Acceptances notified: 28 February 2004.
Registration process & proposal submission will be conducted online:
Website: www.soe.jcu.edu.au/iomw2004/ email: iomw@jcu.edu.au

Delegates requiring early acceptance in order to commit funds, seek employer support / approval, etc. should email: iomw@jcu.edu.au

Proposals after the due date may be accepted, subject to program / accommodation availability.

IOMW Pre-Conference Workshops:
Learn the features of the software from the people who wrote it:
Winsteps & Facets - June 28 Monday afternoon - Linacre
RUMM - June 29 Tuesday morning - Sheridan / Andrich
ConQuest - June 29 Tuesday afternoon - Adams / Wu / Wilson

Teachers' Workshops: July 2, Friday

Accommodation: Just 15 minutes drive north of Cairns, Quest Marlin Cove Resort is situated in the lush tropical surrounds of Cairns' favorite beach, Trinity Beach. The Resort comprises of a combination of one, two and three bedroom apartments and offers the convenience of Hotel Service with the flexibility and value of apartment living. Special rates are offered for IOMW delegates. Please book now. Email jindorato@questapartments.com.au for details (mention **IOMW** - request shared accommodation, if you wish). Website: www.questapartments.com.au – Far North Queensland.

Transport: The Resort is 18 km (11 miles) from Cairns International Airport (CNS) served by Qantas (codesharing with American Airlines). Rental cars available (Australia drives on the left). Much more about local transport and amenities at www.cairns.aust.com/about/trinity.htm

Weather: Rain: very little. Sunshine: 7+ hours per day. Temperature: 17ºC - 26ºC daily (63ºF – 78ºF). Humidity: low.

IOMW-XII Chair: **Trevor Bond**

IOMW-XII is supported by the School of Education, James Cook University

SECOND INTERNATIONAL CONFERENCE ON MEASUREMENT

In Health, Education, Psychology and Marketing: Developments with Rasch and Unfolding Models

January 20-22, 2004 Murdoch University, Perth, Australia

***Conference website*: www.education.murdoch.edu.au/educ_RaschJanuary2004.html**

Monday, January 19, 2004

Pre-conference Workshop: Using the Rasch analysis program RUMM2020

Tuesday, January 20, 2004

Opening and Introduction. 8:45-9:00. *Chair: D. Andrich*

Session 1.1 9:00-9:45 Plenary. *Chair: D. Andrich*
Two approaches to studying indices of development. *David Andrich, A. Slade, A. Tennant & S. van Buuren*

Session 1.2A 9:45-10:45. Rasch models in Education, Psychology and Sociology. *Chair: J. Tognolini*
A three-way comparison between rating scale, pairwise comparison, and direct response data in the setting of educational standards. *S. Heldsinger & S.Humphry*
Integrating quantitative and qualitative approaches to the study of intelligence: the relationship between the algorithms of Raven's Progressive Matrices and Piagetian stages. *I. Styles*

Session 1.2B 9:45-10:45. Rasch models in Health Sciences. *Chair: A. Bjorkdah*
ABILHAND-Kids questionnaire: a Rasch measure of manual ability for cerebral palsy. *C. Arnould & J. Thonnard*
Improving comparability with response conversion: a new application of IRT. *S. van Buuren & A. Tennant*

Session 1.3A 11:15-12:15. Rasch models: Theoretical and technical perspectives. *Chair: D. Andrich*
Dealing with differential item functioning and local dependence in measurements of quality of life by analysis using graphical loglinear Rasch models. *S. Kreiner*
Detecting Differential Item Functioning in the Dichotomous Rasch Measurement Model. *R. M. Smith*

Session 1.3B 11:15-12:15. Rasch models in Attitude Measurement. *Chair: G. Luo*
Student surveys of teaching effectiveness: A Rasch measurement approach. *T. Bond*
Determining the Frame of Reference of a Scale Measuring the Attitude towards Advertising. *T. Salzberger*

Session 1.4A 2:00-3:00. Rasch models in Education, Psychology and Sociology. *Chair: P. Titmanis*
Measuring Academic Motivation to Achieve in Malaysia using a Rasch Measurement Model. *R. F. Waugh & J. N. Njiru*
Application of the Rasch model to develop a measure of classroom information and communication technology learning culture. *R. F. Cavanagh & J. Romanoski*

Session 1.4B 2:00-3:00. Rasch models in Health Sciences. *Chair: A. Slade*
The structural properties of European Brain Injury Questionnaire in patients with stroke. *A. Bjorkdah*
Rasch modeling of "international classification of functioning disability and health" qualifiers scale. *A. Cieza & I. Styles*

Session 1.5A 3:00-4:00. Rasch models: History and Philosophy. *Chair: J. Michell*
Some introductory remarks on Probability, Invariance and Measurement. *A. Leplege*
Meaning and Method in the Human Sciences. *William P. Fisher, Jr.*

Session 1.5B 3:00-4:00. Rasch models in Education, Psychology and Sociology. *Chair: C. Hagquist*
The Scalability & Validity of Four Paediatric Visual Perceptual Instruments: A Comparison Using the Rasch Measurement Model. *G. T. Brown*
Psychometric Properties of the Korean Version of Beck Depression Inventory: Rasch Rating Scale Modeling. *S. H. Hong*

Session 1.6A 4:00-4:30. Rasch models in Education, Psychology and Sociology. *Chair: I. Styles*
Philadelphia Geriatric Center Morale Scale in noninstitutionalized and institutionalized elderly Chinese in Hong Kong: a differential item functioning analysis. *E. Wong*

Session 1.6B 4:00-4:30. Theoretical and Technical Perspectives. *Chair: T. Salzberger*
Modeling subjective use of an ordinal response scale in a many-period crossover experiment. *R. Wolfe & D. Firth*

Wednesday, January 21, 2004

Session 2.1 9:00-9:45 Plenary. *Chair: D. Andrich*
The distinction between order And quantity: its history, philosophy and significance for the Rasch model. Joel Michell

Session 2.2A 9:45-10:45. Rasch models in Education, Psychology and Sociology. *Chair: T. Bond*
Multi-Facet Rasch analysis of Three-Dimensional Speaking test data. *Yuji Nakamura*
Measuring coping at a university using a Rasch model. *R. F. Waugh*

Session 2.2B 9:45-10:45. Unfolding models. *Chair: J. Michell*
Examination of the relationship between some IRT models of unidimensional unfolding and Coombs's (1964) deterministic theory. *A. Kyngdon*
Psychological scales of preference and choice which take individual difference into account. *D. Andrich & G. Luo*

Session 2.3A 11:15-12:15. Unfolding models. *Chair: G. Luo*
Unidimensional unfolding theories and quantitative differences between attitudes. *B. Richards*
Using unfolding models for personality scale construction. *O. S. Chernyshenko & S. Stark*

Session 2.3B. 11:15-12:15. Rasch models in Health Sciences. *Chair: A. Tennant*
Why Functional Independence Measure is better than Barthel Index Motor performance in assessing motor performance for stroke patients in Hong Kong: A Rasch perspective. *E. Wong, C. Chan, A. Chan, B. Ng, L. Li & J. Woo*
Mental health services evaluation - Measures or total scores. *E. Betemps*

Session 2.4A 2:00-3:00. Rasch models in Education, Psychology and Sociology. *Chair: J. Thonnard*
Differential item functioning of Triandis' instruments of Individualism and Collectivism. *P. Snider & I. Styles*
Comparison of person ability logit scores of scales measuring the same visual perceptual construct: Common subject test equating. *G. T. Brown*

Session 2.4B 2:00-3:00. Rasch models in Health Sciences. *Chair: A. Tennant*
Cross-cultural validity of Functional Independence Measure (FIM) items in stroke. *Isa Lundgren Nilsson*
Cross-cultural validity of the FIM in traumatic brain injury (TBI). *A. Slade & A. Tennant*

Session 2.5A 3:00-4:00. Rasch models: History and Philosophy. *Chair: A. Leplege*
On latent structures and models. *D. Andrich*
Consequences of standardized technical effects for scientific advancement. *W. Fisher*

Session 2.5B 3:00-4:00. Rasch models: Theoretical and technical perspectives. *Chair: T. Bond*
Notes on artificial results of Andersen's Likelihood Ratio Test and on the Mixed Rasch Model as a model check of the dichotomous Rasch Model. *C. Draxler & K. D. Kubinger*
Correcting for person misfit in aggregated score reporting using the Rasch model. *R. S. Brown*

Session 2.5C 3:00-4:00. Theoretical and technical perspectives. *Chair: W. Fisher*
From Rasch scores to regression. *K. B. Christensen*
Does the Rasch model work for equating? *R. Sadeghi & J. Tognolini*

Session 2.6A 4:00-4:30. Rasch models in Health Sciences *Chair: A. Tennant*
Psychosocial consequences of false positive screening mammography - an adaptation of the Psychological Consequences Questionnaire (PCQ) into Danish. *John Brodersen*

Session 2.6B 4:00-4:30. Rasch models in Education, Psychology and Sociology. *Chair: 1. Styles*
Taking another perspective: Matura examinations in Slovenia. *S. Gabrscek*

Thursday, January 22, 2004

Session 3.1 9:00-9:45 Plenary. *Chair: D. Andrich*
Conceptual and methodological issues in establishing an item bank for quality of life in the rheumatic diseases. *A. Tennant, D. Yeale, S.P. McKenna, L. C. Doward & P. Emery*

Session 3.2A 9:45-10:45. Rasch models: Theoretical and technical perspectives. *Chair: R. Smith*
Modifying or replacing items: A suggestion for a strategy. *T. Nielsen & S. Kreiner*
Weighted likelihood estimation of person locations in the Rasch model. *G. Luo & D. Andrich*

Session 3.2B 9:45-10:45. Rasch models in Education, Psychology, Sociology and Attitude Measurement. *Chair: J. Thonnard*
Internal and external construct validity of the Objective Structured Clinical Examination (OSCE) in undergraduate medical education. *B. Bhakta, M. C. Horton & A.Tennant*
Measuring Primary students' attitude to mathematics using a Rasch measurement model. *R. F. Waugh & E. Chapman*

Session 3.3A 11:15-12:45. Rasch models in Education, Psychology and Sociology. *Chair: T. Bond*
Learning environment instrument calibration and validation: the case of measuring elementary school classroom learning culture. *R. F. Cavanagh, J. Romanoski & R. F. Waugh*
Measuring economic status of students in rural Vietnam. *Nguyen Thi Kim Cuc & Patrick Griffin*
Constructing Cognitive Entry Behaviour Tests (CEBT) for guidance school students. *Masoud Fazilatpour*

Session 3.3B 11:15-12:45. Rasch models in Economics and marketing. *Chair: R. Smith*
An investigation of the psychometric properties of a multiple-choice test of marketing knowledge. *T. Salzberger*
Bringing capital to life via measurement: a contribution to the New Economics. *William P. Fisher, Jr.*

Complex IRT = Simple Rasch

Reise et al. (2001) present what, at first blush, appears to be a complex multi-group IRT analysis, defying objective measurement criteria and beyond the capabilities of standard Rasch software. But closer inspection reveals that their analysis is Wright and Stone (1979) p. 94 *redivivus.*

Reise at al. administer the same instrument, comprising multiple "facet" strands, to two gender groups, male and female. They want to examine differential test functioning. Here is the core of their description of what they did using PARSCALE:

"First, within each facet scale we estimated a multiple-group IRT model in which the item location parameters (λ_i) were freely estimated within groups and all slope parameters (a_i) and category parameters (τ_{iv}) were constrained to equality across gender. In this multiple-group model the mean and standard deviation on the latent variable is fixed at 0 and 1 for men but are estimated parameters for women." (p. 96).

Let us deconstruct this paragraph:
"within each facet scale" – i.e., analyzing one strand at a time.

"item location parameters (λ_i) were freely estimated within groups" – since the groups share no common persons, this is equivalent to doing separate gender-group analyses.

"all slope parameters (a_i) and category parameters (τ_{iv}) were constrained to equality across gender." – these are exactly the specifications for an "Andrich Rating Scale" Rasch analysis of a gender group.

"In this multiple-group model the mean and standard deviation on the latent variable is fixed at 0 and 1 for men" - so the scale origin is chosen so that the mean measure for men is zero, and the measure scaling factor is 1/(men s.d.).

"but are estimated parameters for women." – so the mean of the women's measures is relative to the men's mean, and the men's scaling factor is applied to the women's analysis.

Further, inspection of the Tables in Reise et al. reveals that, for each "facet" strand, the mean of the item location parameters for the men and the women is constrained to be the same.

So, what analysis did Reise et al. actually perform? Two separate "Andrich" Rasch analyses in which the item means are constrained to be the same (the Rasch default), and the scaling factor for both analyses is chosen such that the "men" group's standard deviation is 1.

Moral of the story: if you need to squeeze a "parallel runs" Rasch DIF analysis passed reluctant reviewers, dress it up with a light reworking of another sentence from Reise et al. (p. 96):

"The DIF detection procedure implemented [here is similar to that implemented] by PARSCALE (Muraki & Bock, 1998) [and] is similar to the DIF detection routine implemented by BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) in that ... maximum likelihood estimation routines are used and contrasts of item parameter estimates are developed."

And then paste in your own version of the Reise et al. paragraph quoted above!

John Michael Linacre
Courtesy of Deon De Bruin

Muraki, E. & Bock, R. D. (1998). PARSCALE (version 3.5): Parameter scaling of rating data. Chicago, IL: Scientific Software, Inc.

Reise, S.P., Smith, L., Furr, R.M. (2001) Invariance on the NEO PI-R Neuroticism Scale. *Multivariate Behavioral Research*, *36* (1), 83-110

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items. Chicago, IL: Scientific Software International.

The Bruce H. Choppin Memorial Award

Bruce Choppin studied mathematics at Cambridge University in England before attending the University of Chicago, where he earned his PhD in the area of measurement, evaluation and statistical analysis. He was closely connected with IEA (International Association for the Evaluation of Educational Achievement) from 1965 until his premature death in 1983. His first work with IEA involved data analysis for the English national report of the First IEA Mathematics Study. Along with Dr Alan Purves, he later undertook a small-scale exploratory study designed to measure student understanding and appreciation of literary prose and poetry. He also constructed several attitude and descriptive scales of the IEA Six-Subject Survey and headed its Data Processing Center in New York from 1969 to 1972.

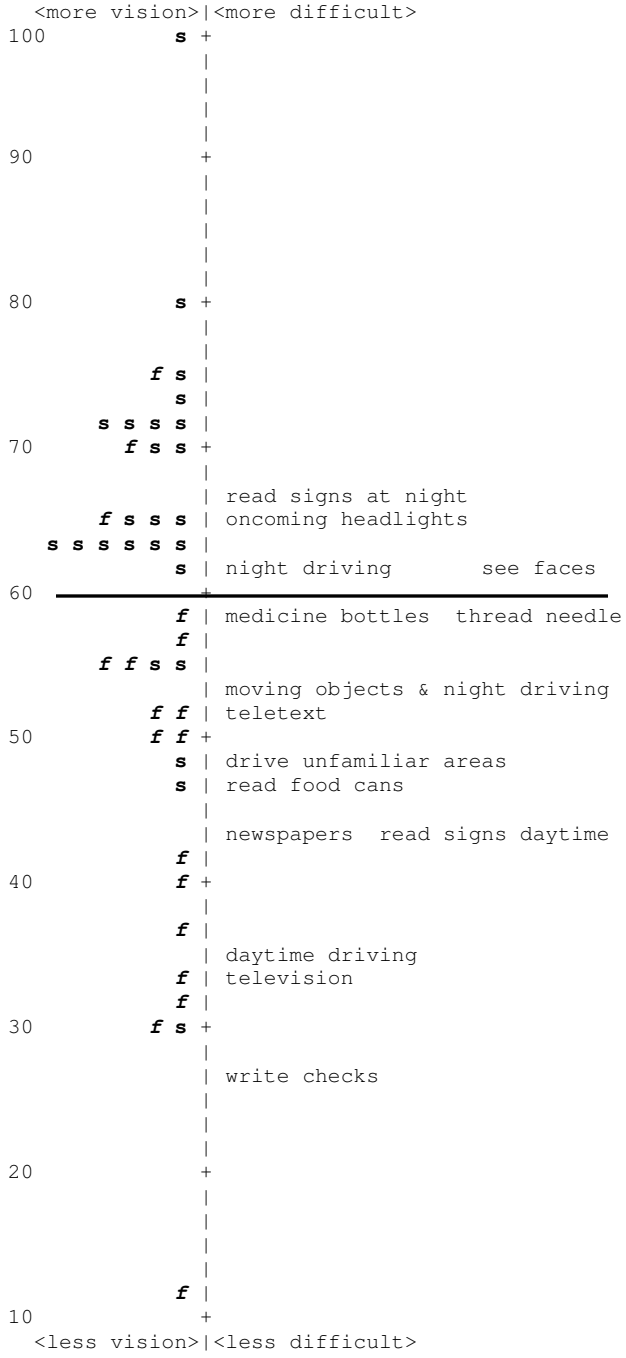
Dr Choppin was a proponent of the Rasch method of scaling aptitude and achievement test scores (having come under the influence of Benjamin Wright). He was at the center of a debate about Rasch scaling at a time (the 1970s) when this method was still looked upon with skepticism by those in the field of testing. He wrote a monograph for IEA titled Correction for Guessing and, with Neville Postlethwaite as co-editor, began the journal Evaluation in Education, which latter became the International Journal of Educational Research. In addition to his work with the New York Data Processing Center, He worked for several years at the National Foundation for Educational Research (NFER) in England, the Science Education Centre in Israel and the University of California and Cornell University in the United States. He died in Chile, having gone there to help the country's National Research Coordinator for the IEA Study on Written Composition. His ashes are buried in London.

The Bruce H Choppin Award is given annually to an author who makes use of data from any of the IEA studies and employs empirical research methods in his or her master's or doctoral thesis, written within the three years preceding the entry date (31 March of that year). For further details see: **www.iea.nl/Home/IEA/**

Bruce H. Choppin Awards - Past Awardees

1985 Ingrid Munck, U. of Stockholm, Sweden
1986 Lauren Sontag, Columbia U.
1987 Nongnuch Wattanowaba, U. of Illinois
1989 Marilda Chandvarkar, Columbia U.
1990 KC Cheung, U. of London, UK
1991 Hans Pelgrum, U. of Twente, Netherlands
1992 Norbert Sellin, U. of Hamburg, Germany
1993 Andreas Schleicher, Deakin U., Australia
1994 Diedra Young, Curtin U., Australia
1997 Petra Leitz, Flinders U., Australia;
Ingeborg Janssen Reinen, U. of Twente, Netherlands
1999 Dana Kelly, Boston College
2002 Laura M O'Dwyer, Boston College

Maps for Diagnosis and Prediction



The "bilateral cataract" patients are waiting for first (**f**) or second (**s**) eye cataract surgery. The line at 60 units roughly divides "**f**" from "**s**" patients. It indicates an expected vision level after first surgery. Vision may be good enough for day-time driving, but not for night-time.

Pesudovs K., Garamendi E., Keeves J.P., & Elliott D.B. (2003) The Activities of Daily Vision Scale for cataract surgery outcomes: Re-evaluating validity with Rasch analysis. *Investigative Ophthalmology & Visual Science,* 44, 7, 2892-2899.

Measurement and Communities of Inquiry

Metrological standards have long been recognized as essential to fair, just economic and legal relations. The French revolution gave rise not only to nascent democratic institutions but to the metric system, and US Presidents Washington and Jefferson were intensely interested in the standardization of currency and weights and measures as necessary for promoting the greater good of society.

On the downside, one of the most important complaints about the emerging global economy is that exclusive focus on manufacturing, product, and financial standards often has profound negative consequences for human well-being, social relations, and the natural environment. And quite apart from our apparent incapacity to export democratically and environmentally sound values, as De Soto (1989, 2000) says, we don't even export capitalism very well, since World Bank and IMF policies are imposed on many countries that have not yet built up the infrastructure of financially and legally negotiable titles and deeds necessary for the successful implementation of those policies.

The problem with globalization may not be capitalism itself, but its incomplete state of development (see De Soto, as well as Hawken, Lovins, & Lovins 1999). The general failure of the various communist and socialist experiments suggests that the only way to counteract the negative consequences of capitalism may be to trace out its root metaphor in the natural reproductive capacities of livestock to its logical consequences. As De Soto points out, following Latour (1987, p. 223), the infrastructure of fungible financial instruments is after all predicated on an abstract model of capital in which value is brought to life only when it is expressed in a stable metric that can be added up across properties owned, or divided into shares and sold, without any physical change to, or manipulation of, the property itself.

Though neither De Soto nor Latour recognize it, this metaphor of living capital is itself rooted in the Socratic art of midwifery, which has its "highest point" in "the power to prove by every test whether the offspring of a young man's thought is a false phantom or instinct with life and truth" (Plato 1961, p. 855). Midwifery's proofs and tests are fundamentally mathematical in the sense of requiring a transparent clarity essential to the simplest and most fully achieved forms understanding, those of arithmetic and geometry. Because of its role in bringing understanding to life, mathematical clarity constitutes the metaphysical foundations of not only academia and science (Heidegger 1977; Fisher 1992, 2003a, b, c, d), but also of the various survey-based measures essential to establishing capitalist property rights.

Rasch's models for measurement make it possible to deploy, explicitly and deliberately, these metaphysical principles in ways that extend capital accounting and management methods (Fisher 2002) to the domains of human (Fitz-enz 2000), social (Putnam 1993), and natural capital (Hawken, et al. 1999). But the implications of Rasch's models for redressing the imbalances created by the currently incomplete implementation of capitalist principles cannot be appreciated until the models are more fully integrated into a larger metrological framework.

This integration can begin in a number of different ways, but to pick one, it is of interest that a significant body of research (for instance, Rogoff, Matusov, & White, 1996) stresses the participatory involvement of learners and teachers in communities of inquiry. The questions they leave largely unasked are:

- What is the medium of this involvement? In other words, what are the behaviors, signs, and symbols through which this involvement is coordinated and mediated?
- How is learning and/or development expressed in the medium?
- How do we know learning and/or development when we see it?
- How do we locate one another relative to this medium?
- How should this medium be structured, distributed, and maintained in order to maximize the cohesiveness of the community of inquiry?

Is the community of inquiry defined by a focus on a common question or set of questions expressed in a common language? If so, is the language defined vaguely, as "in English" or "in Urdu"? Is the set of relevant questions defined concretely as a particular collection of test items dealing with "reading ability," for instance? Is the common language for expressing this construct constituted only by test-dependent ordinal scores that require complete data from one single instrument's set of questions? Or is the language defined precisely, as "in Lexiles" (Burdick & Stenner 1996, Stenner & Burdick 1997; Smith 1998; Wright 1995), an abstract unit of linear measurement that can be read off any properly calibrated reading test, and that can be universally interpreted as predicting 75% reading comprehension for any reader with a Lexile reading ability measure that matches any book's Lexile readability measure?

Even if the common questions and language are relatively precisely defined in terms of content, by what criteria does anyone know whether all participants in the dialogue are talking about the same thing? Via vague criteria, such as "using the same words for roughly the same behaviors" in the context of different tests giving incomparable scores in nonlinear metrics? Or via precise criteria, as in the context of having the same measure for the same readability or reading ability everywhere and any time?

Communities of inquiry probably cannot begin to realize their potential for collective, distributed thinking (Latour 1995) until metrological networks of evidence experimentally test the hypotheses that a single object dominates the conversation, and that there is in fact a stable, additive, and divisible line of inquiry functioning as a reference standard. As Alder (2002, p. 2) puts it,

> "To do their job, standards must operate as a set of shared assumptions, the unexamined background against which we strike agreements and make distinctions. So it is not surprising that we take measurement for granted and consider it banal. Yet the use a society makes of its measures expresses its sense of fair dealing. That is why the balance scale is a widespread symbol of justice. ... Our methods of measurement define who we are and what we value."

The disciplinary, professional identity of communities of inquiry, and their effectiveness in creating new learning, would then seem to rest on the Socratic proofs and tests of hypothesized mathematical clarity, and the shared languages universal uniform metrics make possible (Fisher 2003a, b, c, d).

When both the intra- and inter-laboratory aspects of metrology in the human sciences are achieved, Rasch measurement will have been expanded from its current focus on within-laboratory instrument ruggedness tests (Wernimont 1977, 1978) to also include the between-laboratory (Mandel 1977, 1978) equating studies and the item content prediction theories necessary for universal uniform metrics (for more on this and relevant historical/philosophical considerations, see Fisher 1992, 1993, 1995, etc.; if anyone else is doing work in this area, please let me know so I can cite it). And when:

- Rasch's models for measurement are integrated into the larger metrological framework,
- the variables specific to each different form of capital (human, social, and natural) are expressed in universal uniform metrics (so far as this turns out to be possible),
- systems for maintaining, improving, applying, and learning from these metrics are implemented,
- and the metrics are deployed everywhere they are needed in forms that provide the relevant quantitative and qualitative information at the point of use,

then we will be en route to completing capitalism in a way that promotes the growth of healthy, fulfilled human beings living in balanced, sustainable social and natural ecologies.

That at least is a dream of an epic adventure, a goal worthy of people great enough to pursue it. It has long been argued that the development of stable, coherent individual and group identities follows from the ways in which a self is tested by the challenges it faces (Bettelheim 1967; Zaner 1981) and is shaped the stories it can tell about itself (Ricoeur 1992). The emergence of a coherent and friendly global human identity depends on what challenges we select as testing grounds and how we yield to or aggressively meet those challenges. To make no choice at all is still to choose to fail. It would be quite another thing, however, to choose to follow through on the principles implied and assumed in not only the deep structures of our economic and democratic institutions, but also in the structures of nature's balanced and sustainable ecologies.

There is no doubt need to further validate the relevant principles and methods, but their logic, their history of practical successes to date, and the ubiquity of human suffering, social injustice, and environmental disasters in the world today strongly suggest that the day may be approaching when we will deploy a systematic program for tuning the instruments of the human, social, and ecological sciences, with the aim of harmonizing human, social, and ecological relations on a global scale. Whether any of us alive today will be around to play in the resulting ensemble and/or dance to its music may largely depend on how much energy we put into making it happen.

William P. Fisher, Jr.

Alder, K. (2002). The measure of all things: The seven-year odyssey and hidden error that transformed the world. New York, New York: The Free Press.

Bettelheim, B. (1967). The empty fortress: Infantile autism and the birth of the self. New York, New York: The Free Press.

Burdick, H., & Stenner, A. J. (1996). Theoretical prediction of test items. Rasch Measurement Transactions, 10(1), 475

De Soto, H. (1989). The other path: The economic answer to terrorism. New York, New York: Basic Books.

De Soto, H. (2000). The mystery of capital: Why capitalism triumphs in the West and fails everywhere else. New York, New York: Basic Books.

Fisher, W. P., Jr. (1992). Objectivity in measurement: A philosophical history of Rasch's separability theorem. In M. Wilson (Ed.), Objective measurement: Theory into practice. Vol. I (pp. 29-58). Norwood, New Jersey: Ablex Publishing Corporation.

Fisher, W. P., Jr. (1993). Scale-free measurement revisited. Rasch Measurement Transactions, 7(1), 272-3

Fisher, W. P., Jr. (1995). Opportunism, a first step to inevitability? Rasch Measurement Transactions, 9(2), 426

Fisher, W. P., Jr. (1996, Winter). The Rasch alternative. Rasch Measurement Transactions, 9(4), 466-467

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

Fisher, W. P., Jr. (1997b, June). What scale-free measurement means to health outcomes research. Physical Medicine & Rehabilitation State of the Art

Reviews, 11(2), 357-373.

Fisher, W. P., Jr. (1999). Foundations for health status metrology: The stability of MOS SF-36 PF-10 calibrations across samples. Journal of the Louisiana State Medical Society, 151(11), 566-578.

Fisher, W. P., Jr. (2000a). Objectivity in psychosocial measurement: What, why, how. Journal of Outcome Measurement, 4(2), 527-563.

Fisher, W. P., Jr. (2000b). Rasch measurement as the definition of scientific agency. Rasch Measurement Transactions, 14(3), 761.

Fisher, W. P., Jr. (2001). Review of John Roche's The Mathematics of Measurement: A Critical History. Journal of Applied Measurement, 2(4), 426-440.

Fisher, W. P., Jr. (2002, Spring). "The Mystery of Capital" and the human sciences. Rasch Measurement Transactions, 15(4), 854

Fisher, W. P., Jr. (2003a). The mathematical metaphysics of measurement and metrology: Towards meaningful quantification in the human sciences. In A. Morales (Ed.), Renascent pragmatism: Studies in law and social science (p. in press). Brookfield, VT: Ashgate Publishing Co.

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

Fisher, W. P., Jr. (2003c, December). Mathematics, measurement, metaphor, metaphysics: Part II. Accounting for Galileo's "fateful omission." Theory & Psychology, 13(6), in press.

Fisher, W. P., Jr. (2003d, April 26-7). Provoking professional identity development: The postmodern legacy of Benjamin Drake Wright. In E. Smith (Chair), Presentations 7. A Celebration of the Career and Contributions of Benjamin D. Wright, Rehabilitation Institute of Chicago and the Institute for Objective Measurement, Chicago, Illinois.

Fisher, W. P., Jr. (2004). Meaning and method in the human sciences. Human Studies: A Journal for Philosophy and the Social Sciences, 27, in press.

Fitz-enz, J. (2000). The ROI of human capital: Measuring the economic value of employee performance. New York, New York: AMACOM.

Hawken, P., Lovins, A., & Lovins, H. L. (1999). Natural capitalism: Creating the next industrial revolution. New York, New York: Little, Brown, and Co.

Heidegger, M. (1977). Modern science, metaphysics, and mathematics. In D. F. Krell (Ed.), Basic writings (pp. 243-282). New York, New York: Harper & Row.

Latour, B. (1987). Science in action: How to follow scientists and engineers through society. New York, New York: Cambridge University 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: An International Journal, 3(1), 54-63.

Mandel, J. (1977, March). The analysis of interlaboratory test data. ASTM Standardization News, 5, 17-20, 56.

Mandel, J. (1978, December). Interlaboratory testing. ASTM Standardization News, 6, 11-12.

Plato. (1961). Theaetetus (F. M. Cornford, Trans.). In E. Hamilton & H. Cairns (Eds.), The Collected Dialogues of Plato, including the Letters (pp. 845-919). Bollingen Series LXXI. Princeton, New Jersey: Princeton University Press.

Putnam, R. D. (1993). Making democracy work: Civic traditions in modern Italy. Princeton, New Jersey: Princeton University Press.

Ricoeur, P. (1992). Oneself as another. Chicago, Illinois: University of Chicago Press.

Rogoff, B., Matusov, E., & White, C. (1996). Models of teaching and learning: Participation in a community of learners. In D. R. Olson & N. Torrance (Eds.), The handbook of education and human development: New models of learning, teaching and schooling (pp. 388-414). Oxford, England: Basil Blackwell Publishers, Inc.

Smith, R. R. (1998). Using Lexile reading measures to improve literacy. Rasch Measurement Transactions, 12(3), 644

Stenner, A. J., & Burdick, D. S. (1997, January 3). The objective measurement of reading comprehension. www.lexile.com/about_lex/tech-papers/documents/ObjectveMS.pdf (visited 10 March 2003) (Ed.), [Response to technical questions raised by the California Department of Education Technical Study Group], Durham, North Carolina: MetaMetrics, Inc.

Wernimont, G. (1977, March). Ruggedness evaluation of test procedures. ASTM Standardization News, 5, 13-16.

Wernimont, G. (1978, December). Careful intralaboratory study must come first. ASTM Standardization News, 6, 11-12.

Wright, B. D., Stenner, A. J., Vanezky, R. (1995, Winter). Reading in America: Stenner's Lexiles confirmed. Rasch Measurement Transactions, 8(4), 387-388

Zaner, R. (1981). The context of self: A phenomenological inquiry using medicine as a clue. Athens, Ohio: Ohio University Press.

"There's a real resistance to standardization in the healthcare community. People who go into medicine often prefer autonomy, so this type of cultural change takes time."

JCAHO Protocol to Eliminate Wrong-Site, Wrong-Procedure, Wrong-Patient Surgery: A Newsmaker Interview With Rick Croteau, MD by Laurie Barclay, MD, Dec. 5, 2003, *Medscape Medical News*

Plato's Separability Theorem

What does Gadamer (1989, p. 412) mean when he says that "We see that it is not word but number that is the real paradigm of the noetic"? The noetic, from the ancient Greek noesis, is the entire abstract population of all things that can be understood. Gadamer is addressing the same issue taken up by Descartes (1961, p. 8) when he says that "...in seeking the correct path to truth we should be concerned with nothing about which we cannot have a certainty equal to that of the demonstrations of arithmetic and geometry."

Mathematical thinking is too often assumed to be inherently numerical and quantitative (Michell 1990, 1999). The mere use of numbers in many fields is deemed sufficient indication of mathematical thinking, though the numbers may only rarely express anything substantively meaningful.

So how could Gadamer, the quintessential hermeneutic philosopher and anti-methodologist, seriously consider number to be the paradigmatic model of understandable meaning? He gives a vital clue when he acknowledges that "numerical signs [are coordinated] with particular numbers, and they are the most ideal signs because their position in the order completely exhausts them" (Gadamer 1989, p. 413), and he elsewhere gives extensive consideration to the "overall structural parallel between number and logos" (Gadamer 1980, p. 149; also see Gadamer 1979), but for those familiar with Rasch's separability theorem, perhaps no one illustrates the crux of the situation better than Ballard (1983).

To see the value in Ballard's treatment, first recall Rasch's (1961, p. 325) statement of the separability theorem: "On the basis of [one of the equations in the model] we may estimate the item parameters independently of the personal parameters, the latter having been replaced by something observable, namely, by the individual total number of correct answers. Furthermore, on the basis of [the next equation] we may estimate the personal parameters without knowing the item parameters which have been replaced by the total number of correct answers per item. Finally, [the third equation] allows for checks on the model [another equation] which are independent of all the parameters, relying only on the observations."

Ballard is concerned with Plato's examination of the concepts of small and large in the Republic (523D-525D). Plato shows that these concepts are insufficient for rigorous comparison due to the ambiguity of having things that can simultaneously be smaller than some things and larger than others. Ballard (1983, pp. 135-6) points out that Plato has Socrates "show that the confusion engendered by a finger being both large and small may be cleared up by the art of quantitative measurement. In order to execute this measurement, we first separate each finger from one of its properties, in this instance a quantitative property, its length, and replace each finger by its (abstract) length. So now the length of each finger can be measured by some equal and common unit of length; then one of these fingers, the middle, will appear to be larger by so much (i.e., by so many units of length) than another, and smaller by so much ...."

Later, Ballard (1983, pp. 136) points out that Plato also sees to it that "an analogous technique is brought to bear upon the puzzling aspect of the unit and the techniques of measure," so that "a still greater clarification can be achieved." So in the same way that Socrates separates the concrete instance of the thing to be measured from its quantitative property, and also separates the unit and the techniques of measurement from their concrete expressions, Rasch separates observations from parameters for both person measures and item calibrations.

Gadamer (1980, p. 150) takes the matter still further, pointing out that a "characteristic of a proportion is that its mathematical value is independent of the given factors in it, provided that they keep the same proportion to one another. The same relation can exist even when the numbers in it are changed. The universality of the relationship as such transcends its components."

Gadamer here identifies in proportionality one of the key features of Thurstone's and Rasch's measurement models, the capacity through which different items can provoke different responses from different people but still remain consistent with one another and provide comparable measures, as in computer adaptive testing (Lunz, Bergstrom, & Gershon 1994). Gadamer (1980, p. 149) also understands that "the real problem in the logos lies in its being the unity of an opinion composed of factors or items which are distinct from the opinion itself. Now, as we know, logos is a mathematical term that means 'proportion.'"

These comments are quite reminiscent of Thurstone (1928, p. 228): "If the scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help to construct it."

In other words, for measures to represent the logos of an object of discourse, the factors or items instrumental to that representation must remain in constant proportion to one another, so as to be separable and distinct from the opinion itself. Thurstone, writing 40 years before Gadamer's article first appeared in German, appears to have applied a principle of reason fundamental to science since Plato.

As is well known, Thurstone (1928, p. 228) characterizes checks on the extent to which scale values are or are not affected by the opinions of the people who help to

calibrate the tool as a "crucial experimental test." This test is rarely employed by psychologists, who apparently find vulnerability to falsification more of a fault than a virtue (Michell 1990, p. 130). The test is, however, routinely implemented in Rasch measurement (Andrich 1978; Smith 2000).

It is remarkable then that Gadamer goes so far as to pinpoint the crux of what Andrich (2002) calls "resistance to the data-model relationship in Rasch's paradigm," saying that "the test which is to be applied in respect to the eidos [the logos of a particular idea] is a test of the immanent, internal coherence of all that is intrinsic to it. One should go no further until one is clear about what the assumption of the eidos means and what it does not mean. It should be noted that consequently the hypothesis is not to be tested against presumed empirical consequences, but conversely the empirical consequences are to be tested against the hypothesis, i.e., that from the start everything empirical or accidental which the eidos does not mean and imply is to be excluded from consideration. This means above all that the particular which participates in an eidos is of importance in an argument only in regard to that in which it may be said to participate, i.e., only in regard to its eidetic content." Gadamer (1980: 33-4; 1986: 101-2).

This passage conveys the essential importance of instrument calibration as the isolation of a particular thing, a variable or construct, that dominates a repeatably identifiable object of discourse. As was repeatedly stressed by Messick (for instance, 1975) in his work on construct validity, making measures inherently assumes that responses to questions embody a certain internal coherence, and so measures certainly should not be subjected to statistical comparisons until we are clear about what they mean and do not mean.

Unfortunately, IRT and classical test theory (CTT) begin from the position that the hypothetical model of what is being measured (referred to by Gadamer as the hypothesis of the eidos) is tested against the data (the presumed empirical consequences). In this paradigm, the model that best describes the data is taken as the basis for instrument calibration, even when that model explicitly (in the IRT case) or implicitly (in the CTT case) incorporates parameters that make it impossible to separate the particular factors or items involved in a unitary opinion from that opinion.

But as Gadamer says in the sentence immediately preceding the passage just quoted, "Such a procedure would be totally absurd in respect to a postulated eidos: that which constitutes being a horse could never be proved or disproved by a particular horse." It is common practice, however, in the implementation of IRT models with multiple item parameters, to decide that that which constitutes reading ability or moral development is proved or disproved by particular items or factors that are not distinct from the particular abilities of the persons measured.

Rasch models, in contrast, test empirical consequences against the hypothesized construct, holding, precisely in accord with Gadamer, that any test or survey question in particular is important only to the extent that it actually participates in and contributes to the generalizable measurement of the object of interest by being separable from it.

Ever since Kuhn's 1962 extension of the linguistic concept of the paradigm to the history of science, we have come to a fuller appreciation of the fact that "...reason has insight only into that which it produces after a plan of its own, and that it must not allow itself to be kept, as it were, in nature's leading-strings, but must itself show the way with principles of judgment based upon fixed laws, constraining nature to give answer to questions of reason's own determining. Accidental observations, made in obedience to no previously thought-out plan, can never be made to yield a necessary law, which alone reason is concerned to discover. Reason, holding in one hand its principles, according to which alone concordant appearances can be admitted as equivalent to laws, and in the other hand the experiment which it has devised in conformity with these principles, must approach nature in order to be taught by it. It must not, however, do so in the character of a pupil who listens to everything that the teacher chooses to say, but of an appointed judge who compels the witnesses to answer questions which he himself has formulated. ... It is thus that the study of nature has entered on the secure path of a science, after having for so many centuries been nothing but a process of merely random groping" (Kant 1965: 20-1).

As long as IRT and CTT dominate test- and survey-based measurement, we can expect nothing but continued random groping from the human sciences, since "the road from scientific law to scientific measurement can rarely be traveled in the reverse direction" (Kuhn 1977, p. 219). Rasch models specify the structure of scientific laws (Rasch 1960, p. 110-5) and so provide a framework in which reason can have insight through the projection of a plan of its own, showing the way with principles of judgment based on necessary and sufficient lawful relations.

For those seriously interested in pursuing this line of thought, I strongly recommend close and repeated reading of Heidegger's (1967, 1977) book, What is a thing? Fuller treatments of these ideas are taken up in my own recent work (Fisher 2003a, 2003b, 2003c, 2004).

William P. Fisher, Jr.

Andrich, D. A. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. Applied Psychological Measurement, 2, 449-460.

Andrich, D. A. (2002). Understanding Rasch measurement: Understanding resistance to the data-model relationship in Rasch's paradigm: A reflection for the next generation. Journal of Applied Measurement, 3(3), 325-59.

Descartes, R. (1961). Rules for the direction of the mind. Indianapolis: Bobbs-Merrill.

Gadamer, H.-G. (1979). Historical transformations of reason. In T. F. Geraets (Ed.), Rationality today (pp. 3-14). Ottawa, Canada: University of Ottawa Press.

Gadamer, H.-G. (1980). Dialogue and dialectic: Eight hermeneutical studies on Plato (P. C. Smith, Trans.). New Haven: Yale University Press.

Gadamer, H.-G. (1986). The idea of the good in Platonic-Aristotelian philosophy (P. C. Smith, Trans.). New Haven: Yale University Press.

Gadamer, H.-G. (1989). Truth and method (J. Weinsheimer & D. G. Marshall, Trans.) (Rev. ed.). New York, New York: Crossroad (Original work published 1960).

Heidegger, M. (1967). What is a thing? (W. B. Barton, Jr. & V. Deutsch, Trans.). South Bend, Indiana: Regnery/Gateway.

Heidegger, M. (1977). Modern science, metaphysics, and mathematics (W. B. Barton, & V. Deutsch, Trans.). In D. F. Krell, (Ed.). Basic writings (pp. 243-282). New York, New York: Harper & Row. Rpt. from M. Heidegger, What is a Thing? South Bend, Indiana: Regnery/Gateway, pp. 66-108.

Kant, I. (1929/1965). Critique of pure reason (N. K. Smith, Trans.) (Unabridged). New York, New York: St. Martin's 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. Chicago, Illinois: University of Chicago Press (pp. 178-224). Reprinted from T. S. Kuhn, (1961), The function of measurement in modern physical science. Isis, 52(168), 161-193.

Lunz, M. E., Bergstrom, B. A., & Gershon, R. C. (1994). Computer adaptive testing. International Journal of Educational Research, 21(6), 623-634.

Messick, S. (1975, October). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Michell, J. (1990). An introduction to the logic of psychological measurement. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In Proceedings of the fourth Berkeley symposium on mathematical statistics and probability (pp. 321-333). Berkeley, California: University of California Press.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. Journal of Applied Measurement, 1(2), 199-218.

Thurstone, L. L. (1928). Attitudes can be measured. American Journal of Sociology, XXXIII, 529-544. Reprinted in L. L. Thurstone, The Measurement of Values. Midway Reprint Series. Chicago, Illinois: University of Chicago Press, 1959, pp. 215-233.

Data Variance: Explained, Modeled and Empirical

How much of the variance in my data do the Rasch measures explain? This is a crucial question, but its answer is far from obvious and can only be known approximately.

Here are three sources of variance in the data:

i) People differ in ability and items differ in difficulty. These cause different responses, and it is these differences that the Rasch measures are intended to reflect.

ii) People respond in an apparently random way, but still in accord with Rasch model predictions.

iii) People respond in a way that conflicts with Rasch model predictions.

Suppose that N people respond to L dichotomous items, scored 0, 1. The response by person *n* to item *i* is scored X_{ni} (using the notation of Wright & Masters, 1982, p. 100). Then the overall average response, A, is

$$A = \sum_{n=1}^{N} \sum_{i=1}^{L} X_{ni} / NL .$$

So, conceptualizing the scored observations to be linear, as is typically done, the observed variance sum-of-squares, OV, in the data is

$$OV = \sum_{n=1}^{N} \sum_{i=1}^{L} (X_{ni} - A)^2 .$$

This includes (i), (ii) and (iii) above.

Once the Rasch ability measures $\{B_n\}$ and difficulty measures $\{D_i\}$ have been estimated, there is an expected value, E_{ni}, for each X_{ni}. The variance explained by the Rasch measures, RV, can then be expressed as:

$$RV = \sum_{n=1}^{N} \sum_{i=1}^{L} (E_{ni} - A)^2$$

corresponding to (i) above.

Associated with each E_{ni} is its Rasch-predicted model variance W_{ni}. Thus the variance not explained by the measures, but predicted by the Rasch model, MV, is

$$MV = \sum_{n=1}^{N} \sum_{i=1}^{L} W_{ni}$$

corresponding to (ii) above. The total variance in the data, TV, is predicted to be

TV = RV + MV.

When the data fit the Rasch model, then OV ≈ TV.

Empirically, the unexplained variance, UV, is

$$UV = \sum_{n=1}^{N} \sum_{i=1}^{L} (X_{ni} - E_{ni})^2$$

corresponding to (ii) + (iii) above. Then, since fit to the model is never perfect, the variance actually explained, AV, a shown in Table T1, becomes

AV = OV – UV.

T1: **Raw Score** Sum-of-squares Variance components	Empirical conceptualization	Rasch model prediction
Explained by measures (i)	AV = OV – UV	$RV = \sum_{n=1}^{N} \sum_{i=1}^{L} (E_{ni} - A)^2$
Predicted Unexplained (ii)	$UV = \sum_{n=1}^{N} \sum_{i=1}^{L} (X_{ni} - E_{ni})^2$	$MV = \sum_{n=1}^{N} \sum_{i=1}^{L} W_{ni}$
Unpredicted Unexplained (iii)		0
Total = Explained + Unexplained	$OV = \sum_{n=1}^{N} \sum_{i=1}^{L} (X_{ni} - A)^2$	TV = RV + MV
Proportion of variance explained	AV/OV	RV/TV

These variance computations can be extended to allow for missing data and polytomies by adjusting the summations.

When the data approximate the Rasch model, the proportion of variance explained is about equal for the two conceptualizations. When the data grossly misfit the model, the empirical variance explained by the measures, AV, can become negative. On the other hand, with anchored measures, the empirically unexplained variance can become less than the Rasch predicted variance, indicating overfit of the current data to the measures. Tables T1 and T2 show the algebraic components and also their values for the "Liking for Science" data.

T2: **Raw Score** Variance components in the "Liking for Science" data	Empirical conceptualization	Rasch model prediction
Explained by measures	AV = OV – UV = 564.63	RV = 562.70
Unexplained	UV = 543.92	MV = 546.48
Total = Explained + Unexplained	OV = 1108.55	TV = RV + MV = 1109.18
Proportion of variance explained	AV/OV = 51%	RV/TV = 51%

Variance in Standardized Units

An alternative conceptualization is in standardized units. Here each *response* is modeled to contribute one unit of statistical information. Consequently, the summations are in unit normal deviates rather than in raw scores. This is summarized in Table T3.

It is followed by Table T4, a practical example for data noticeably contradicting the Rasch model. In this example of an MCQ test, 4 of 20 items have negative point-biserial correlations, i.e., are oriented in opposition to the Rasch dimension. This has reduced the variance explained by the Rasch dimension to half what would be expected were these data to fit the model.

T3: **Standardized** Variance components	Empirical conceptualization	Rasch model prediction
Explained by measures	AV = OV – UV	$RV = \sum_{n=1}^{N} \sum_{i=1}^{L} (E_{ni} - A)^2 / W_{ni}$
Unexplained	$UV = \sum_{n=1}^{N} \sum_{i=1}^{L} (X_{ni} - E_{ni})^2 / W_{ni}$	$MV = \sum_{n=1}^{N} \sum_{i=1}^{L} W_{ni} / W_{ni}$
Total = Explained + Unexplained	$OV = \sum_{n=1}^{N} \sum_{i=1}^{L} (X_{ni} - A)^2 / W_{ni}$	TV = RV + MV
Proportion of variance explained	AV/OV	RV/TV

T4: **Standardized** Variance component in Winsteps Example 10A data	Empirical conceptualization	Rasch model prediction
Explained by measures	AV = OV – UV = 113.41	RV = 220.04
Unexplained	UV = 400.08	MV = 240.00
Total = Explained + Unexplained	OV = 513.49	TV = RV + MV = 460.04
Proportion of variance explained	AV/OV = 22%	RV/TV = 48%

Relationship to
Principal Components Analysis of Residuals (PCAR)

The variance "explained by the measures" corresponds to the Rasch dimension. The "unexplained" variance corresponds to all other dimensions and random noise. PCAR attempts to partition the unexplained variance based on factors representing other dimensions. This is done by decomposing the matrix of inter-item (or inter-person) correlations of residuals. In this matrix, each diagonal element is set to 1, indicating that there is one unit of residual variance contributed by each item (or person). Thus the total amount of variance to be explained by the PCAR, i.e., the sum of the factor eigenvalues, equals the number of items (or persons).

The "unexplained" variances in the Tables are in summed raw score or standardized units with little immediate meaning, so it is convenient to rescale them into eigenvalue units such that the Unexplained variance corresponds to the sum of the eigenvalues to be explained by the PCAR. This is shown in Table T5 using the Liking for Science data comprising 25 items.

The strength of the Rasch dimension, 50.8, can then be compared directly with the strength of the biggest secondary dimension, 4.3, indicating that, for most practical purposes, the Liking for Science data can be treated as unidimensional.

John M. Linacre

T5: **Standardized** Variance *Liking for Science*	Empirical Eigenvalue units
Total = Explained + Unexplained	50.8 rescaled
Explained	25.8 rescaled
Unexplained	25.0 rescaled = PCAR
Explained by PCAR: 1st Factor 2nd Factor 3rd Factor	 4.3 eigenvalue 2.9 eigenvalue 2.3 eigenvalue

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: Trevor Bond SIG Secretary: Ed Wolfe

Guttman Parameterization of Rating Scale

"A reparameterised form of thresholds into their principal components is the method of estimation operationalised in RUMM2020. This notion of principal components is used in the sense of Guttman (1950), who rearranged ordered categories into successive principal components, beginning with the usual linear one. They are analogous to the use of orthogonal polynomials in regression where the independent variable is ordered. The term does not refer to the common *principal components analysis* in which a matrix of correlation coefficients is decomposed."

Excerpted from *www.rummlab.com.au*

A convenient logit-linear expression of a typical form of the Rasch polytomous model is

$$\log(P_{nix} / P_{ni(x-1)}) \equiv B_n - D_i - F_x$$

where F_x is the centralized (Andrich, Rasch) threshold (also called step calibration) corresponding to the point on the latent variable where categories x-1 and x are predicted to be equally likely to be observed. Categories are numbered from 0 to m.

A Guttman-parameterized version of this same model, derived from Andrich and Luo (2003, eqn. 13) is:

D_i = the item difficulty, which is also the rating scale *location*

F_x = $(2x - m - 1)\,\theta$ where θ is the rating scale *dispersion* or *unit*

$+ (6(x - 1)(x - m) + (m - 1)(m - 2))\,\eta$ where η is the *skewness*

$+ (20x^3 - 30x^2(m + 1) + 2x(6m^2 + 15m + 11) - m^3 - 6m^2 - 11m - 6)\,\zeta$ where ζ is the *kurtosis*.

This enables the Rasch threshold parameters, $\{F_x\}$, to be computed directly from the Guttman parameters, θ, η, ζ, when they are known. The numerical values of the multipliers for m = 2, 10 are shown in the Table.

Direct computation of θ, η, and ζ from the $\{F_x\}$ can usually be performed by means of linear regression, solving the *m* equations of the form above, with the $\{F_x\}$ as the independent variables, the values in the Table as the dependent variables, and θ, η, ζ as the coefficients to be estimated.

Example 1: Item 14 in the *RUMM2020 runAll* example is a 4-category item, so m = 3. On *www.rummlab.com.au,* the reported estimates are $\theta = 2.445$ and $\eta = -0.160$. Thus, by computation,

$F_1 = -2 * 2.445 + 2 * -0.160 = -5.210$

$F_2 = 0 * 2.445 + -4 * -0.160 = 0.640$

$F_3 = 2 * 2.445 + 2* -0.160 = 4.570$

The estimates reported for the $\{F_x\}$ on *www.rummlab.com.au* are: -5.231, .641, 4.590, indicating a close match between theoretical and empirical results.

Example 2: An m=6 rating scale has category frequencies: 96, 88, 101, 168, 210, 146, 101, The $\{F_x\}$ are estimated by *Winsteps* at -2.30, -1.75, -1.34, 0.08, 2.08, 3.23. *Excel* regression analysis reports θ = 0.5794, η = 0.02786, ζ = -0.002241. According to Andrich and Luo (2003, p. 209) these values have greater stability than the $\{F_x\}$. The consequent smoothed values of $\{F_x\}$ are -2.21, -2.04, -1.13, 0.24, 1.82, 3.32.

John Michael Linacre

Guttman Principal Component Multipliers

m	x	θ	η	ζ	m	x	θ	η	ζ
2	1	-1			8	1	-7	42	-210
	2	1				2	-5	6	150
3	1	-2	2			3	-3	-18	210
	2	0	-4			4	-1	-30	90
	3	2	2			5	1	-30	-90
4	1	-3	6	-6		6	3	-18	-210
	2	-1	-6	18		7	5	6	-150
	3	1	-6	-18		8	7	42	210
	4	3	6	6	9	1	-8	56	-336
5	1	-4	12	-24		2	-6	14	168
	2	-2	-6	48		3	-4	-16	312
	3	0	-12	0		4	-2	-34	216
	4	2	-6	-48		5	0	-40	0
	5	4	12	24		6	2	-34	-216
6	1	-5	20	-60		7	4	-16	-312
	2	-3	-4	84		8	6	14	-168
	3	-1	-16	48		9	8	56	336
	4	1	-16	-48	10	1	-9	72	-504
	5	3	-4	-84		2	-7	24	168
	6	5	20	60		3	-5	-12	420
7	1	-6	30	-120		4	-3	-36	372
	2	-4	0	120		5	-1	-48	144
	3	-2	-18	120		6	1	-48	-144
	4	0	-24	0		7	3	-36	-372
	5	2	-18	-120		8	5	-12	-420
	6	4	0	-120		9	7	24	-168
	7	6	30	120		10	9	72	504

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

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