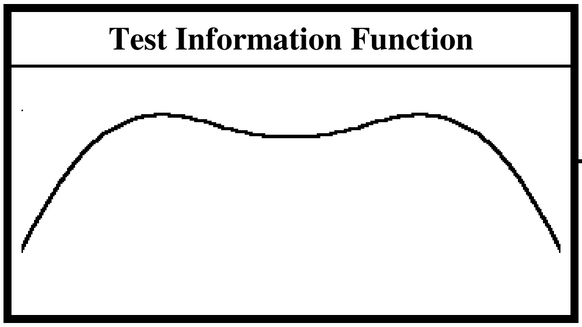


RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 17 No. 1 Summer 2003 ISSN 1051-0796

Rack and Stack: Time 1 vs. Time 2

Measures of the same persons are often obtained at two time-points, or under two conditions, with the intention of investigating changes.

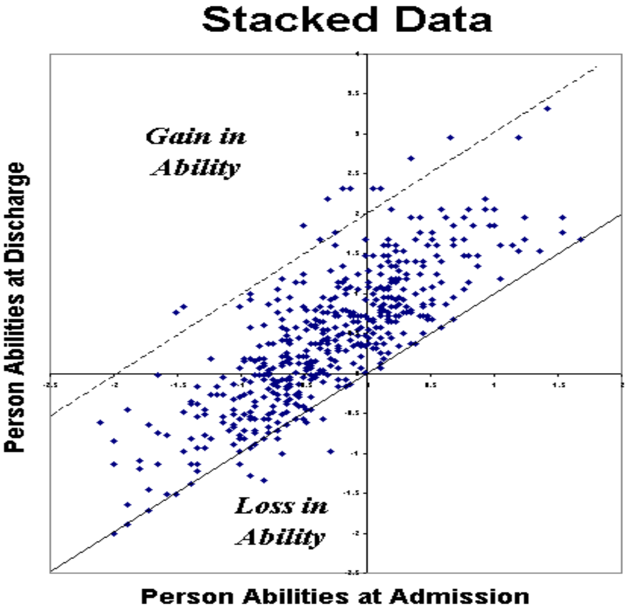
Stacking Data

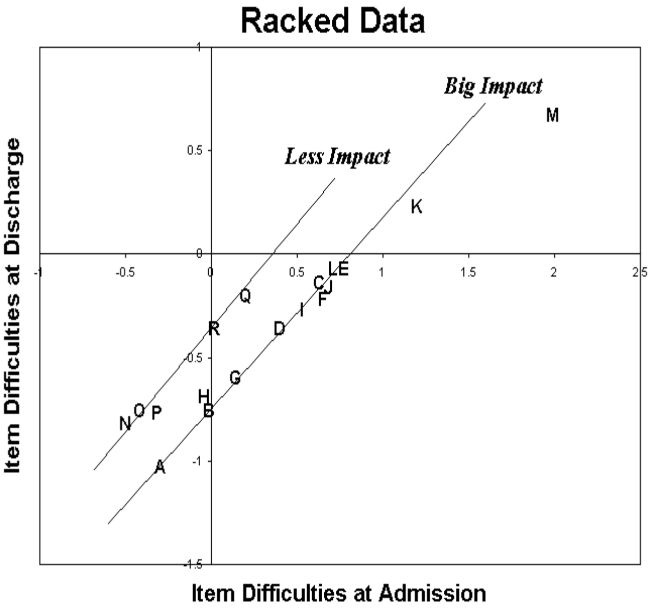
Consider patients being assessed for level of independent functioning on entering and leaving rehabilitation. Each patient has two sets of observations. A useful approach is to convert each set of observations for each patient into a measure, with all measures in the same frame of reference. This would be exactly the same as measuring all their heights at admission and discharge. Then a patient's change in level of functioning would simply be the difference between the admission and discharge measures.

This can be done by "stacking" the data. Each set of observations for each patient is appended to the data file as a further case. Finally, the data file contains twice as many cases as there are patients. Measures are constructed on all cases simultaneously.

Patient Dependency?

A question often raised is "Doesn't putting the same patients in twice introduce dependency?" It probably does in a small way. But let's think about the situation:



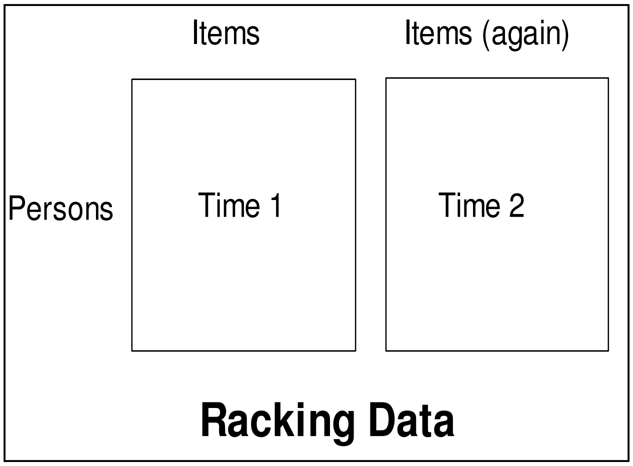


1. The patients are not identical patients. They have changed.

2. There are many sources of dependency within the data. The dependency among patients with similar diagnoses at Time 1 may be greater than the dependency between the same patients at Time 1 and Time 2.

What is the effect of dependency on Rasch measurement? The data are no longer as random in the way that the Rasch

Table of Contents

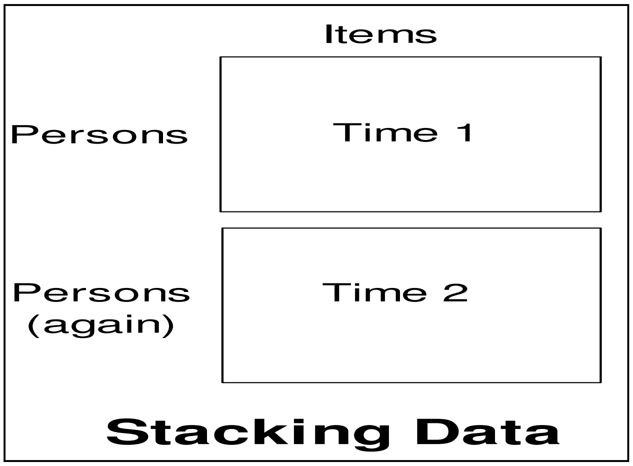


model predicts. Lack of randomness can increase misfit (if the dependency is generally in observations that are unexpected according to model predictions). It can lessen misfit (if the dependency is generally in observations close to model-predictions). Increased misfit reduces sample reliability and separation, making differences smaller in terms of logits. Decreased misfit increases sample reliability and separation, making differences larger in terms of logits.

In practice, dependency between Time 1 and Time 2 is difficult to identify for individuals, except those grossly misfitting at both time points. Dependencies across time points often are present, but they cluster across patients within items. This brings us to ...

Racking Data

In the physical sciences, great effort is exerted to prevent the measuring device from changing. But often, in the social sciences this is not possible. The construct hierarchy changes between Time 1 and Time 2. Physicists would despair, but social scientists can rejoice because this provides a special insight into what has changed.



Between Time 1 and Time some intervention has occurred or some other change has happened. It is unlikely to affect the responses to all items equally. Some items will relate directly to the therapy, teaching or intervention, others will not.

Imagine that the patients have not changed, but the effect of the intervention is to change the items. “I’m still the same person, but now climbing stairs is easier!” Then each person is entered once into the data, but each item twice: once for Time 1 and once for Time 2. This is “racking” the data.

In this racked analysis, the item difficulties are of greater interest than the person measures. Items with the biggest change in measure are those on which the intervention has had the greatest effect.

A Practical Example

The Functional Independence Measure, FIM™ was administered to 500 stroke patients at admission to, and discharge from, rehabilitation. The stacked data of 1000 FIM administrations was analyzed. The Figure shows each person’s ability at discharge plotted against that person’s ability at admission. The patients improved by 0.75 logits on average. Some toward the top left have gained more than 2 logits (above the dashed line). A few have regressed (below the solid identity line).

In the racked data, 500 patients were administered the 18 item FIM twice, so there are 36 items. The plot shows how the change in patient status is reflected in the measures of the items at the two time points. Rehabilitation has had the biggest impact of 0.75 logits on the motor items, A – M. There has been less impact, 0.35 logits, on the mental items, N-R.

Stacking the data, we see **who** has changed. Racking the data, we see **what** has changed.

Benjamin D. Wright

Constructing Scientific Measurement Models

The aim of the scientific process is, in some sense, to predict the future. It may be a future-in-the-past, for instance an eclipse of the sun that occurred in Ireland in 688 A.D., or it may a future-yet-to-happen. Scientific models deliberately embody simplified, but manageable, versions of reality. Henry David Thoreau wrote a universal truth in another context: "Our life is frittered away by detail … Simplify , simplify." If we attempt to include every possible detail into our analysis, we exhaust ourselves and obtain results that are so specific as to become merely restatements of the original details.

Thus the scientific challenge is to formulate models general enough to encompass the scope of situations usually encountered, but specific enough to give practical and useable guidance in the outcomes to be expected in those situations. Thus the scientific model embodies a theory about the relationships that generate the data. Of course, the predicted outcomes only approximate the actual ones. "Empirical problems are frequently solved because, for problem solving purposes, we do not require an exact, but only an approximate, resemblance between theoretical results and experimental ones." (Laudan, 1977). Indeed "in many aspects of statistics it is necessary to assume a mathematical model to make progress." (Draper and Smith, 1966).

There are an infinity of possible models that generate outcomes which approximate the data, so which ones to choose? There is no absolute or correct answer, but there is the answer of utility. "All science is only a refinement of everyday thinking" (Einstein, 1936). The more generally applicable the model, and the more useable the results, the more it is likely to meet practical needs and form the basis for scientific progress. William of Ockham suggests that "What can be accounted for by fewer assumptions is explained in vain by more." Scientists are also generally comfortable performing arithmetical operations. "Measurement is primarily a device which enables us to use the laws of arithmetic to solve problems relating to phenomenal events" (Guild, 1938). Accordingly, a good starting point would be to look for models with as few parameters as possible within a framework that can be manipulated by arithmetical operations.

Classical test theory (CTT) appears to meet these requirements. In fact, it is almost ubiquitously used for summarizing and reporting the results of scoreable tests. Its strength is that the outcome of a test for an examinee can be expressed as one number which has at least the arithmetical properties of rank order, and often approximates linearity. CTT fails when results must be compared across tests, or there is missing data, or score differences within a test need to be compared, or when ... Rasch's insight was that a simple logistic transformation overcomes the obvious predictive flaws of CTT. The logistic transformation is mathematically tractable, and yet, as Derek de Solla Price observed, it underlies a multitude of natural process. Under many circumstances, merely replacing a reported percent with

$$\text{Measure} = 50 + 25 * \text{Log}_{10} (\%\text{Right} / \%\text{Wrong})$$

will approximate linearity will enough.

John Michael Linacre

Draper, N. R., & Smith, H., Jr. (1966) *Applied Regression Analysis.* New York: Wiley.

Einstein, A. (1936) *Physics and reality.* Journal of the Franklin Institute, 221. Translated by Syllabus Division, University of Chicago.

Guild, J. (1938) *Are Sensation Intensities Measurable?* Report of the 108th Annual Meeting of the British Association for the Advancement of Science, Cambridge.

Laudan, L. (1977) *Progress and its Problems.* Berkeley, CA: University of California Press.

Price, D. J. de Solla *(1986) Little Science, Big Science ... and Beyond.* New York: Columbia University Press.

IOMW XII
Cairns, Australia
June 28 - July 3, 2004

Ready for a little temptation? Has IOMW ever offered such lush location? Try this:

www.soe.jcu.edu.au/IOMW2004/

the website for *IOMW Down Under!* None of the photos on the IOMW site are promo glossies; just snapshots I took one weekend. IOMW will have fun and substance! Many key folk in the Rasch measurement have undertaken to participate. The program already includes paper sessions, software workshops, networking opportunities, and SCUBA diving. This is the one you'll tell your grand-kids about!

Joseph Indorato, Operations Manager, at Quest Marlin Cove *www.quest-inns.com.au,*is waiting to help you with your accommodation requirements and is ready to take those early bookings:

jindorato@questapartments.com.au

Joe is willing to help you share an apartment if you give him the appropriate information. Remember to mention the magic code *IOMW* when booking/enquiring. Quest Inns also has accommodation all round Australia for those who want to make the most of the visit.

Trevor Bond, Chair IOMW XII Committee

Conference Report: “Epistemology of Measurement”

On May 6, 2003 a thematic ‘Conference’ on the **Epistemology of Measurement in the Social Sciences : Contemporary Perspectives** was organized at the *Institute of History and Philosophy of Science and Technique, Paris, France.* At IHPST the concerns with the epistemology of social sciences are multifaceted. The May 6 conference was concerned with the family of measurement models familiar to the readers of *Rasch Measurement Transactions.*

David Andrich (Murdoch University, Australia) gave a paper on *Recognizing problems after they are solved in the construction of models of measurement in the social sciences.* He showed us, in reference to Kuhn’s work, that various measurement problems in the social sciences were recognized as problems after their solution was derived. This required insights and recognition of the implications of the models, but the models were not constructed with the solution to those problems in mind. Indeed, He suggested that the problems could not have been solved if it was set out to solve them.

Joel Michell (University of Sydney University, Australia) gave a paper on *The theory of additive conjoint measurement and the Rasch model.* His point was that although the relationship between the Rasch model for psychometric measurement and the theory of additive conjoint measurement has been recognized since the 1970s, surprisingly little attention has been given to the issue of testing the hierarchy of conjoint measurement cancellation conditions in the Rasch context and in particular the contrast between the following three issues: (1) the information that each of these cancellation conditions supplies about the structure of the ability & difficulty attributes, (2) the empirical content they each possess, and (3) the a priori probability of falsifying them in the Rasch context.

William Fisher (Metametrics Inc., Durham, NC, USA) gave a paper on *The metaphysics of measurement : Toward a hermeneutic-mathematical methodological continuum.* He suggested that we could understand the role of measuring instruments in the social sciences as some kind of text production. From this he presented his own theory of measurement’s objectivity taking into account some of the post modernist criticism, in reference to Latour, while enabling to understand the character both conventional and objective of measurement in its relation to experiment. He also outlined some metaphysical implications of this understanding.

Trevor Bond (James Cook University, Australia) setting the tone for the discussion, gave a paper on *The Rasch model and the progress of science.* He provided wide ranging perspectives on the uses of the family of Rasch models in the social sciences and their consequences along with insights on the difficulties and snags of their applications and dissemination.

Philippe Lemoigne (CNRS-Paris 5 University, France) Opened the discussion with the following question: why is measurement, on the one hand, very much developed in the social sciences and on the other hand somewhat blind to questions related to the metrical character of the data ? He analyzed several examples from clinical psychology and psychiatry where he thinks this dilemma is especially salient.

Alain Leplège
Institut d’Histoire et de Philosophie des Sciences et des Techniques, Paris & Département de Philosophie, Université d’Amiens, France.

Benjamin D. Wright
Lifetime Achievement Award

On the afternoon of Friday, April 25, 2003, the day before his *Festchrift* Conference, Mary Lunz, a member of the Board of the *Institute for Objective Measurement* (IOM) presented Ben Wright with the “IOM Lifetime Achievement Award.”

The presentation took place in Ben’s back-garden in front of an audience of Ben’s colleagues and former students. Ben Wright made a short speech in response to Mary’s account of Ben’s achievements and her personal reminiscences. Here are summary jottings of Ben’s speech, made at the time:

> “Thank you. I value your optimism and friendship. When I met Georg he spoke the same way. It was wonderful. It was absorbing. It was exhausting. Our conversations sometimes lasted two days. His wife insisted we be social, which was itself sometimes exhausting! I’ve enjoyed being social with each of you.
>
> I began this years ago, and the [journey continues, but now you must make it.]
>
> You will always count, but not all the same!
>
> I wish Georg was here. He is in a way.”

Ed Bouchard then presented Ben with a pre-publication copy of Ben’s latest book, *“Directing Observations, Inventing Constructs, Crafting Yardsticks, and Examining Fit”,* B. D. Wright & Mark H. Stone, Chicago: The Phaneron Press, 2003, (shortly to be available through Amazon.com). Ben autographed copies for those resourceful enough to obtain them on the spot.

A Celebration of the Career and Contributions of Benjamin D. Wright

The Ben Wright "Festschrift"

Presentations, Roundtables, and Posters: Chicago, April 26-27, 2003

Reflections on Ben Wright pre- and post-Rasch. *Herbert J. Walberg*

Ben Wright's Kinesthetic Ventures. *Ed Bouchard*

Ben Wright: "Idiosyncrasies of autobiography and personality" in taking up the Rasch measurement paradigm. *David Andrich*

Ben Wright's Contributions to Test Interpretation. *Dick Woodcock*

Dr. Benjamin D. Wright: A multi-facet analysis. *Mary E. Lunz*

The Multiple Careers of Benjamin D. Wright. *Mark H. Stone*

Expected Values of Items Within a Measure May be Used to Monitor Patient Care. *Carl V. Granger*

Number-Freed Reporting: Grandchild of KIDMAP. R*onald Mead*

The influence of the Pearsonian chi-square approach to assessing the fit of the data. *Richard M. Smith and Ronald Mead*

Ben Wright, Fundamental Measurement, and Cognitive Psychology. *Ryan Patrick Bowles & Karen M. Schmidt*

Measurement of political participation. *Filemon Cerda*

Measurement Properties of the Chinese Version of Cognitive Failures Questionnaire: A Rasch Model Rating Scale Analysis. *Raymond Chan*

Measurement in Evaluation Or Is It Evolution? Creating a New Measure. *Ken Conrad*

Fortuitous Advancements in Rasch Analysis Software Packages. *Tracy L. Kline & Karen M. Schmidt*

A Historical View of the Development the Assessment of Motor and Process Skills (AMPS). *Brenda K. Merritt & Anne G. Fisher*

Wright: A Standard of Human Expression. *Gregory Stone*

Use of the 'Ensemble Interpretation' When Fitting Construct Theories to Item Difficulties. *Jack Stenner*

Healthy Measurement. *Alan Tennant*

Application of the polytomous Saltus model to stage-like data. *Karen Draney*

Paired Comparison Matrices of Partial Credit Rasch Models. *Mary Garner*

Mapping Student Progress on Multiple Variables. *Cathleen Kennedy & Karen Draney*

RASCHLAB: Using Computer Simulation to Teach Objective Measurement. *Rense Lange*

A Bootstrap Approach To Rating Scale Optimization. *Eric Van Lente & George Karabatsos*

Reliability in Rasch Measurement: Avoiding the Rubber Ruler. *Randy Schumacker & Ev Smith*

Ben Wright's Role in Bringing Substantive Meaning to Variables. *Geoff Masters*

From Theoretical Research to For-Profit Ventures: My Journey with Ben Wright. *Richard Gershon*

The Influence on Ben Wright of some Friends and Acquaintances. *John M. Linacre*

A Rasch Scale of Perceived Seriousness of Boundary Crossings and Violations. *Michael Lamport Commons, Patrice Marie Miller, & Thomas G. Gutheil*

A multifaceted analysis of seven asthma severity indicators from a multi-site clinical trial with multiple physician assessors and multiple visits. *T. Joseph Sheehan, Judith Fifield, and Joseph Burleson*

Incorporating Rasch Measurement Ideas into Descriptions of Achievement based on Multiple Domains. *Matt Schulz*

Development of Measurement Tool for Estimating of Educators Attitude to Unified Graduation Exam. *Anatoliy Maslak & Tatiana S. Anisimova*

Measuring Social-Emotional and Self-Help/Adaptive Development in Young Children: An Application of Rasch Measurement. *Futoshi Yumoto & Gregory R. Anderson*

Optimal Categorization Construction: A Rasch View. *Weimo Zhu*

Measurement as struggle. *Mark Wilson*

Thurstone, Guttman, and Rasch: A comparison of their perspectives on invariant measurement. *George Engelhard, Jr.*

Provoking Professional Development. *William Fisher*

Information: When Gaps Can Be Bridged

Fit of data to the Rasch model always makes us feel comfortable. While fit provides evidence of accuracy of the measurement model of the variable of interest, precision, however, is a different issue, namely the issue of targeting. (Strictly speaking, mistargeting also affects the power of the test of fit.) If items are operational in a range of the latent dimension but most of the respondents are located in a different range, person (and item) parameter estimates lack precision and standard errors are large. The reason, of course, is that the further apart an item and a person is, the less information the item provides about the location of the person, with item information being calculated as P*(1-P) (Fischer, 1974, p.294). Over the items in a test, individual item information adds up to test information (Fischer, 1974, p.296). The test information curve typically but not necessarily is bell-shaped with its maximum at 0 (provided the scale has been defined by the sum of items equal to 0).

An interesting variant of mistargeting occurs when items are clustered in terms of their location and most or many persons are between the clusters (see Figure 1).

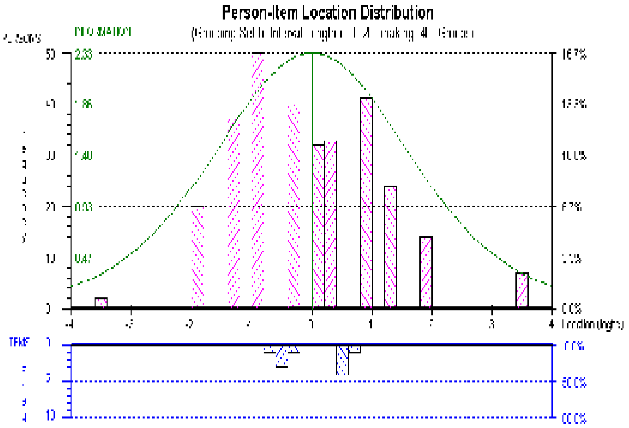


Figure 1: Distribution of Person and Item Locations (Produced by RUMM2020 (Andrich et al., 2003), simulated data)

Let us assume that, for simplicity, in a ten item test, five dichotomous items are located exactly at -0.5 and five other items are located at 0.5 on a logit scale. (Assuming that there are no further items below -.5 and above +.5 does not have a substantial impact on the conclusions drawn.) In other words, there is a gap between these item clusters. Intuitively, one might think that information is higher at -0.5 and at 0.5, respectively, than *between* -0.5 and 0.5 due to the gap between item locations. However, in most cases, the gap between items does *not* imply less information. On the contrary, information peaks out right in the middle of the gap, i.e. at 0 (with 2.35 of total test information in the given example). At the centers of either item cluster, information only amounts to 2.23 (see Figure 2).

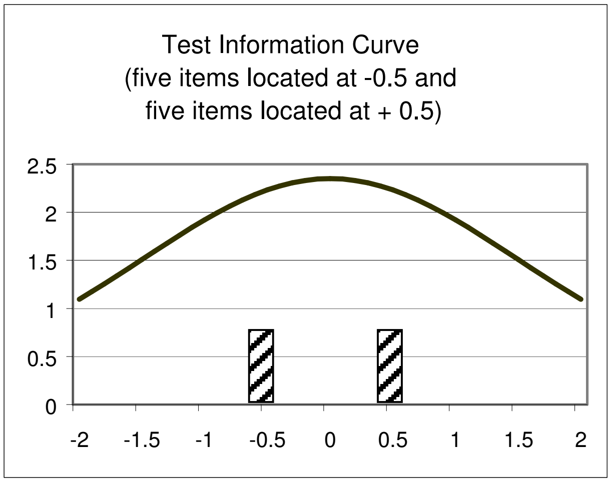


Figure 2: Test Information Curve for five items located at -0.5 and five items located at + .5.

The explanation is that, though, information reaches its maximum at ξ (person ability) = δ (item difficulty) with .5 * (1 - .5) = .25, it remains rather high over a relatively wide range. At |ξ – δ| = .5 it still is .235 and at |ξ – δ| = 1 it amounts to .197. While information is reduced when moving from ξ = δ to |ξ – δ| = 1 by .053 points, the rate of decrease almost doubles for a further logit unit between x and d with information being only .105 at |ξ – δ| = 2. Consequently, in our example, when moving from -.5 into the gap between -.5 and +.5 we loose information provided by the items clustered at -.5. But at the same time we gain information by the items clustered at +.5. For a person location of, e.g., -.2 we loose .006 per item clustered at -.5 whereas we gain .025 from items located at +.5, i.e. four times as much. In other words, total information increases. Thus, we get the same bell-shaped information curve even though there is a gap. It should be noted that in terms of interpretation of person locations within the gap, the lack of items located in this range is certainly not satisfactory.

It is interesting to compare the situation with a Guttman scale. If the items would function perfectly all persons within the gap would get all the five items at -.5 right and all the five items at +.5 wrong. There would be no way to tell whether a particular person is closer to -.5 or to +.5. So, it is the stochastic uncertainty under the Rasch model, the "imperfect nature" so to speak, that yields information and, as a consequence, precision.

However, not all gaps can be bridged. If the distance between the item clusters is too large, the loss of information from "near" items when moving into the center of the gap is not compensated for by the gain of information from "distant" items. The result is a double-peaked total information curve (see Figures 3 and 4).

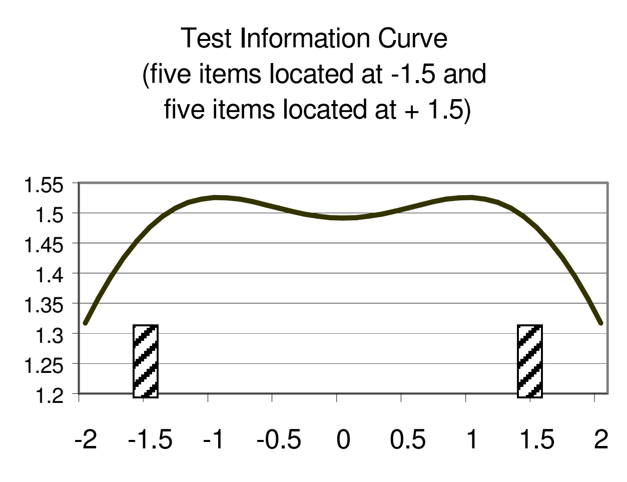


Figure 3: Test Information Curve for five items located at -1.5 and five items located at + .1.5.

AERA Annual Meeting
San Diego, California
April 12-16, 2004
Call for Papers

Proposals for papers, symposia and other presentations are invited via the on-line submissions system of the American Educational Research Association.

www.aera.net/meeting/

All proposals must be submitted by 11:59 p.m. (Pacific Time) on **August 1, 2003.** The Rasch SIG program chair, Randall E. Schumacker, invites members to volunteer as session chairs, discussants and reviewers through the on-line submission system.

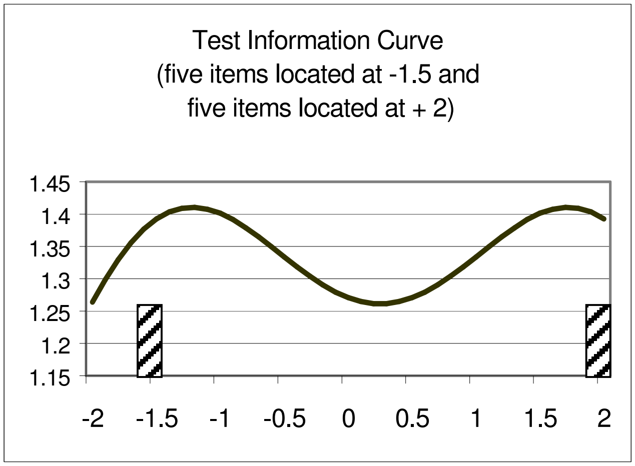


Figure 4: Test Information Curve for five items located at -1.5 and five items located at + 2.0.

Thomas Salzberger
Vienna University of Economics
and Business Administration, Austria

The author wants to thank Alan Tennant (University of Leeds) and David Andrich (Murdoch University) for the stimulating input they provided.

Andrich, D.; Lyne, A.; Sheridan, B.; Luo, G. (2003): RUMM2020, v. 4.0, Rumm Laboratory.

Fischer, Gerhard (1974): Einführung in die Theorie psychologischer Tests [Introduction to the Theory of Psychological Tests], Huber, Bern.

On-line Rasch Course for Beginners

James Cook University, Australia will be offering an introductory Rasch measurement course **on-line** starting **July 28, 2003,** or soon afterwards**.**

ED5191 Rasch Measurement for the Social Sciences

Rasch measurement provides for the construction and monitoring of data collection, analysis and interpretation for key educational, developmental and psychological variables. Topics include Rasch's simple logistic model and its extensions including the Partial Credit Model, the Rating Scale Model and the Many-facets Rasch model. Data analysis and interpretation will focus on the theory-practice nexus.

Trevor Bond will instruct a group of Australian and international students in an on-line one-semester Masters-level course based around the successful Bond & Fox text *Applying the Rasch model (Erlbaum, 2001)*. Supporting audio-visual materials, on-line discussion, data sets and (at least) weekly feedback are features of this hands-on approach to learning Rasch measurement basics.

www.jcu.edu.au/courses/handbooks/2003/subjects/ed5191.html

Five Steps to Science:
Observing, Scoring, Measuring, Analyzing, and Applying

Science progresses by dialogue between the inner world of abstract ideas and the outer word of concrete experience. Ideas are hypothetical guidelines. Experience brings them to life.

Experience is tangible. But it needs ideas to become useful. Raw experience is chaotic. Guidelines are necessary to organize perceptions of reality, to make them recognizable, and to make some sense of them.

Science is the conversation between ideas and experience. We require ideas to recognize reality. We require experience to nudge ideas into new shapes. This dialogue is the relationship between ideas and experience out of which all knowledge evolves. This is the method of science. Science is not facts but method, not data but interpretation of data, not just experience but also idea.

The scientific tasks of psychometrics are pursued by working through a dialogue of five successive procedural models. Articulating these models enables focus and reduces confusion. Explicating a model for each step focuses attention on the specific problem addressed at that step.

1. The **Observing Model** sets the standards for data production and established the first level of quality control. When Isherwood (1939) imagines, "I am a camera with its shutter open, quite passive, recording, not thinking," he supposes that we are able to observe without thinking. But to observe and record is to think about what to observe and what to record. Isherwood's camera is pointed in a direction, has a focal length, and responds to a particular wave band. We are unable merely to observe.

To make observations is to select what to attend. Even more important, to observe is to select what to not attend. Data does not exist of itself. It must be conceived and produced. To produce data requires deciding what to address, when and how. Data is not found. It is anticipated in imagination and constructed in action. Like any product, its manufacture needs quality control.

Quality control over data production means continuous monitoring. The Observing Model nominates the qualities to be sought and recorded. It specifies what is to be looked for and what is to be counted.

In psychometrics, the response requirements for collecting data influence item difficulty, e.g. double negatives, reversals conditionals. Sometimes item intention is so overwhelmed by an ambiguous response process that we do not obtain a reproducible item difficulty, only a variety of local difficulties provoked by interactions between the eccentricities of respondents and the peculiarities of the item.

2. The **Scoring Model** addresses the observed data as though it were nothing but comparisons of ordered alternatives like 0/1 for dichotomous responses and 0/1/2/3 … for ranks and rating scales. The stochastic model by which these data are given inferential meaning is based on defining transition odds for successive categories in terms of a few conjointly additive parameters. To proceed in this direction it is necessary to determine which of the various ordinal scorings the particular categories offer is most useful for measurement, which scoring format provides the most information. The Scoring Model chosen specifies the *more* and *less* comparisons used to infer measurement from the observations (Wright & Stone, 1996).

3. All useful **Measuring Models**, however complicated appearing, reduce to a Rasch formulation, an equation connecting logs of category transition odds for observable events to conjointly additive parameters designed to explain these odds (Rasch, 1960, 1980; Wright & Stone, 1979). The stochastic model interprets the data as instances of probabilities.

The simplest Rasch model identifies parameters B_n for person ability and D_i for item difficulty. Their difference ($B_n - D_i$) is defined to govern the probability of what is expected to happen when person n uses their ability B_n against the difficulty D_i of item i. The data are interpreted as independent of the distribution of the other B_n, and the measures of B_n are independent of the distributions of D_i. The log-odds function establishes a linear scale and the

MODEL STEP	NUMERICAL LEVEL	PRIMARY ACTIVITY
1. Observing	Nominal	Determining what to observe and what to overlook
2. Scoring	Ordinal	Determining which ordinal scoring of the observations categories provides the most informative comparisons.
3. Measuring	Interval	Calibrating items, measuring persons, and evaluating fit, and so defining the construct.
4. Analyzing	Relational	Investigating the relationships among measures and tracking processes.
5. Applying	Practical	Applying results back to the initial problems and forward to new ones.

parameter separation establishes generality (Wright & Stone 1979).

a) Person separation indicates the ability of the items to separate measures of these persons.

b) Item separation indicates the degree to which a variable has been defined by these persons.

c) Item fit evaluates the relevance of each item to the conjoint variable.

d) Person fit evaluates the validity of each person measure and directs response diagnosis.

4. The **Analyzing Model** cannot do its work until we have satisfied the requirements of the first three models. Premature analysis of data mistaken for measures without taking into account and using the models for observing, scoring and measuring only confounds results with an inextricable mass of unidentified and uncontrolled interactions.

In the Analyzing Model we study process and relation. We determine the implications of the measures we derive from our observations and investigate how these measures relate to other variables also measured by applying the observing, scoring and measuring models.

5. The **Applying Model** follows analysis of the measures constructed from observations. In this step, we apply the results obtained to the problems that initially provoked our investigations and also to new situations. The Applying Model brings the prior steps into focus and use. It orients the prior models to an outcome.

The scientific productivity of the five models depends on the vitality of their stepwise reconciliation of idea and experience. The models articulate a dialogue proceeding forwards and backwards as we apply what has been clarified by one model to expediting the tasks of another. Quality control and continuous monitoring is essential.

The organizer that integrates the five models is the **MAP of the Variable**. The MAP begins an idea about experience, an expectation, and a plan. The results from applying the models are incorporated in the MAP. The MAP coordinates and explains the idea by illustration, conceptually and experientially. The MAP portrays the status of results achieved, pictures what has been accomplished and identifies what remains to be done. Successful mapping brings ideas and experience together in a visual manifestation and synthesis of the dialectic process.

Benjamin D. Wright & Mark H. Stone
May 2003 (original paper 1996)

Isherwood, D. (1939) *I am a camera*. New York: Random House.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press. (Original work published in 1960)

Wright, B. D. & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Wright, B. D. & Stone, M. H. (1996). *Measurement essentials*. Wilmington, DE: Wide Range, Inc.

Rasch Measurement SIG News

Notes from the SIG Business Meeting at AERA

George Karabatsos, outgoing SIG Program Chair, presented reports from SIG Chair, Trevor Bond, and SIG Secretary/Treasurer, Edward W. Wolfe, to the SIG. Those reports indicated that (a) the Rasch SIG finances are coming under control - the SIG is in the black and the bank account is holding steady, (b) distributing the newsletter via email and raising dues from $8 to $10 annually would offset the recent losses in SIG funds, (c) the SIG should revisit the Paper Proposal policy. In light of these indications, RMT will only be available to new members via email distribution (only 52% of the members currently receive their newsletters via email), the dues will be raised to $10 per year, and Randy Schumacker will be SIG Program Chair for AERA, San Diego, April 12-16, 2004.

Membership

The membership of the Rasch SIG increased from 178 to 231 between August of 2002 and March of 2003 while the official membership (those counted toward AERA paper sessions) increased from 85 to 144. If you would like to become a member of the Rasch SIG, you are encouraged to become an "official" member by joining the SIG at the same time you become a member to AERA (or renew you membership). The cost is $10 per year. You can also become a member of the Rasch SIG by sending a check for $10 (US) per year to Edward Wolfe, made out to AERA SIG with Rasch SIG in the memo line.

Electronic Newsletters

If you are able to download and read PDF documents, please contact Ed Wolfe to let him know that you would like to receive your newsletter via email. Not only would you save the SIG about $12 per year, but you would also receive your newsletter about 2 weeks ahead of those who subscribe via hardcopy.

In Pursuit of Rasch Measurement - Explorations Following Michell

Michell (2002) concluded that "... in no area of traditional psychometrics is there yet any evidence that the relevant attributes are quantitative" - no data have yet been shown to be consistent with either the monotonic or Rasch theories.

Michell's argument that leads to this conclusion can be summarized as follows:

1) Premise. One can not measure a psychological attribute using any specific procedure (conjoint measurement theory or otherwise) until one **first** obtains satisfactory evidence that the attribute we wish to measure is measurable, i.e., quantitative.

2) Premise. If measurement is to be done according to conjoint measurement and item response theory principles, then satisfactory evidence that the attribute is quantitative requires evidence that a data matrix satisfies all orders of cancellation conditions. (He points out that satisfying only single cancellation, which is entailed by monotonic but crossing item characteristic curves, ICC's, is evidence for ordinal theory; satisfying single, double and all higher cancellation conditions, which is entailed by monotonic non-crossing ICC's, is evidence for monotonic theory and for a quantitative attribute; satisfying all cancellation conditions plus having logistic ICC's is evidence for the Rasch theory.)

3) Evidence that a data matrix satisfies all cancellation conditions must include statistical tests that can be shown to be sensitive to higher order cancellation conditions ("that such tests are capable of discriminating between monotone theory and ordinal theory").

4) Michell claims that no such evidence that any data structure satisfies more than single cancellation conditions has yet been presented.

5) Therefore, Michell concludes, no psychological attribute has yet been shown to be quantitative and measurable.

In fact, an even stronger conclusion can be drawn if in place of 4) and 5) we substitute:

4b) Any proposed measurement procedure will fail fit tests of the cancellation conditions if a statistical test of sufficient power is provided (i.e., if a large enough data matrix is provided) since realistically any point null hypothesis is false. For example, there is no reason to expect that any behavior is affected by only one trait with the influence of all other traits exactly zero for all individuals across all test items.

5b) Therefore, no psychological attribute can ever be shown to be measurable.

Thus, Michell's premises lead not just to refutation of claims that quantitative measurement of psychological attributes has been achieved - they entail a refutation of the possibility that such measurement can ever be achieved. On first reading, Michell's argument seems to mean 'Abandon hope all ye Rasch believers trying to enter the Hell of psychological traits'. While angels may wisely fear to tread here, I see two possible paths.

I. We compromise our high principles in order to get into Hell by following the "good enough" approach (cf. Serlin & Lapsley, 1993, based on work of Lakatos).

In this approach we accept that attributes (data generators) may not be perfectly quantitative, and/or that we can not perfectly measure them, but we propose that they are quantitative and measurable enough for practical purposes. In this case we modify Michell's second premise to allow a wider range of "satisfactory evidence that an attribute is quantitative."

Consider the following. Newtonian mechanics and optics theories have been shown to fail fit tests - observations deviate from the predictions of these theories when sufficiently powerful tests are conducted. Nevertheless, for most practical purposes these theories predict a close approximation to data - and quite useful bridges and telescopes can be built using them.

One could quite reasonably hold that the relation between cognitive structure, including attributes, and behavior is sufficiently complex that no theory relating the two is likely to be complete enough that real data will satisfy all possible fit tests even if the attributes are measurable. The task of psychologists is to steadily improve understanding of the cognitive structure/behavior relationship - to be able to build useful bridges and telescopes. Some reasonable amount of misfit, e.g., within interval null hypotheses, is accepted - a theory can be held until the data diverge "too much" for practical utility or until a better theory is proposed. Only if the data deviate more than this criterion would the theory of quantitative measurement of an attribute be rejected. Michell's requirement that tests of higher order cancellation conditions be provided is still relevant, but some amount of misfit must be accepted. Studies of the implications of particular amounts or types of misfit on quality of measurement would also be relevant.

II. We give up trying to enter Hell and instead try to create a Heaven by following the model approach.

In this approach we do not attempt to directly measure psychological attributes and do not postulate that the attribute as it "really exists" is strictly quantitative. Instead we simply **define** a quantitative attribute, an ideal attribute, to exist in our model of the individual. In this approach the quantitative nature of this variable is not an empirical question - it is not open to disconfirmation by data - and Michell's entire argument does not apply.

Next we propose a particular measurement procedure, e.g., a test and Rasch analysis. A Rasch analysis of data then provides quantitative measures of the ideal attribute. Again, the issue of whether these measurements are quantitative or not can not be empirically challenged. The analysis necessarily, by virtue of the characteristics of the Rasch model, provides a quantitative measure. The empirically testable questions concern how good the ideal attribute and measurement thereof are. For example, we could set a criterion such as that the idealized model must explain 90% of the real data variance. We are not concerned whether the 10% misfit also involves failure of higher order cancellation conditions since we did not originally assert that the data generator was strictly quantitative, only that it is a close enough approximation that the analysis can extract a quantitative component.

This process would be analogous to proposing that some physical process, say the vibration of piano strings, could be modeled as a pure sine wave. The actual generator may have non-linear components, but we go ahead and fit the noisy data to a sine wave. The sine wave frequency measurements are quantitative. If we can say that, 90% of the vibration energy in a set of data can be accounted for as sine waves, then our idealized model of strings may be useful (even though we could disprove the theory that the real piano strings generate only pure sine waves). With time we may be able improve our model by explaining the remaining 10% in terms of other quantitative or non-quantitative (e.g., non-linear) components.

In summary, one can accept Michell's claim that no psychological attribute has yet been proven to be measured and even perhaps, in the absolute sense, that no attribute can in principle be proven to be measurable. This would accord with the general view that no scientific theory can be proven, only disproved. Thus, when a Rasch analysis is said to satisfy fit criteria, we should be sensitive to Michell's argument and not uncritically conclude that 'the results indicate that the attribute is quantitative and can be measured by the test in question'. Instead we could say, for example, either (Hell, approach 1) 'within Z degree of accuracy, these data are consistent with the predictions of the theory that attribute X is quantitative and measurable by test Y with Rasch scaling'; or (Heaven, approach 2) 'attribute X **in our quantitative model** as measured by test Y with Rasch scaling was able to account for Z amount of the variance in behavior'. In either case, the more important issues will concern how well these putative measurements of the attribute, whether conceptualized as real or modeled, relate to other behaviors and other theories.

Roger. E. Graves
University of Victoria

Michell, J. (2002). Conjoint Measurement & the Rasch Model: Quantitative versus Ordinal Structure. Paper presented at the International Objective Measurement Workshop, New Orleans, LA, 6 April 2002.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In: G. Keren & C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: methodological issues. Hillsdale, NJ: Erlbaum

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 © 2003 Rasch Measurement SIG
Permission to copy is granted.
SIG Chair: Trevor Bond   SIG Secretary: Ed Wolfe

Journal of Applied Measurement

Volume 4, Number 2. Summer 2003

Maximum Information Approach to Scale Description for Affective Measures Based on the Rasch Model. *Huynh Huynh and J. Patrick Meyer, p. 101-110*

Measuring Client Satisfaction with Public Education I: Meeting Competing Demands in Establishing State-wide Benchmarks. *John A. King and Trevor G. Bond, p. 111-123*

Developing an Initial Physical Function Item Bank from Existing Sources. *Rita K. Bode, David Cella, Jin-shei Lai, and Allen W. Heinemann, p. 124-136*

Breakthrough Measuring Neighborhoods. *Nikolaus Bezruczko, p. 137-152*

Rasch Fit Statistics as a Test of the Invariance of Item Parameter Estimates. *Richard M. Smith and Kyunghee K. Suh, p. 153-163*

Measuring Attitudes and Behaviors to Studying and Learning for University Students: A Rasch Measurement Model Analysis. *Russell Waugh, p. 164-180*

Understanding Rasch Measurement: Rasch Techniques for Detecting Bias in Performance Assessment: An Example Comparing the Performance of Native and Non-Native Speakers on a Test of Academic English. *Catherine Elder, Tim McNamara, and Peter Congdon, p. 181-197*

Guidelines for Manuscripts. *Richard M. Smith, John M. Linacre, and Everett V. Smith, Jr., 199-203.*

This issue is the second issue in an expanded format. There is a larger page size and a double column format to make it easier to read, as well as seven articles per issue. Please visit our website for a complete list of published articles. Sample copies are available from the Editor. *Recommend JAM to your librarian!*

Richard M. Smith, Editor
Journal of Applied Measurement
P.O. Box 1283, Maple Grove, MN 55311
JAM web site: ***www.jampress.org***

All Measuring the Same Thing

Luigi Tesio (2003, Measuring Behaviours and Perceptions: Rasch Analysis as a Tool for Rehabilitation Research, Journal of Rehabilitation Medicine, 2003; 35: 105–115) presents a remarkable summary of the progress made in the last 15 years. His Table III (reproduced here) supports the contention that instruments developed at different times for different purposes may actually measuring the same thing. Historically, this parallels a major development in thermometry.

Source scale	Item	Logit measure 0-100 transform	Fit mean-square
McGill	Aching	60	0.71
Oswestry	Lifting	57	1.13
McGill	Tiring	56	0.93
FASQ	Sitting	55	1.07
FASQ	Standing	53	1.30
Oswestry	Traveling	48	0.79
FASQ	Low sit (raising from)	47	0.92
Oswestry	Walking	42	0.83
Oswestry	Personal care	34	0.86

Table III. Rasch measure (logit transformed into 0–100 units) and fit mean square of the BACKILL scale for low-back pain syndromes, derived from existing scales (14). McGILL:McGill Pain Questionnaire—short form; OSWESTRY: Oswestry Low Back Pain Disability Questionnaire; FASQ: Functional Assessment Screening Questionnaire. Items made up a linear measure of "back illness". Fit indexes revealed good fit to the Rasch model.

"The development of common constructs can also contribute to a cohesive knowledge core and further enhance theoretical understanding."

Michael J. Feuer, Lisa Towne, and Richard J. Shavelson,
in "Scientific Culture and Educational Research,"
Educational Researcher, 2002, 31, 8, p. 11

"The Spirit of the Age and Fashion in Data Analysis Methods"

"What today is a state-of-the-art research approach is tomorrow already like yesterday's snow."

Then, 1980?	Now, 2000:
Categorical data	Metrical data
Explanation	Description
Operationalization	Measurement
Model	Method
Classification	Quantification
Statistical significance	Effect size

Jürgen Rost. *(2002) Zeitgeist und Moden empirischer Analysemethoden. ZUMA - Nachrichten Spezial Band 8, Von Generation zu Generation.*

Psychometrics is at a Crossroads

I've come to believe that Psychometrics has become

1. **too inbred** - too many of us are paying attention only to ourselves. We have forgotten our users - psychologists and other social scientists.

2. **too driven by mathematics** without regard to practical usefulness.

3. **too focused on minutiae** - often mathematically tweaking existing work to provide a minute new feature that makes little (if any) useful difference.

4. to focused on methods and models that have gotten way **more complicated than users can understand** – and than we can communicate to anyone other than ourselves.

5. **insufficiently** driven by its original purpose of trying to provide mathematical tools **useful** to Psychologists.

The common denominator of these points is:

we are too focused on mathematics at the expense of usefulness.

I believe that Psychometrics is in need of **rejuvenation,** and that this can be done by expanding the purview of Psychometrics by working on

1. the development and use of **Psychological tools** that are useful in strengthening the scientific basis of the practice of Statistics; while at the same time strengthening the traditional focus on

2. the development and use of **Statistical tools** useful in strengthening the scientific basis of the practice of Psychology.

I encourage the field of Psychometrics to encompass and encourage the use of Psychological tools that are useful in strengthening the scientific basis of the practice of Statistics. These tools include the theories and methods of **cognitive and perceptual Psychology,** which are already being used to improve the scientific basis of the practice of Statistics - especially for **Graphics** and Software design.

Excerpted and edited from: Forrest W. Young, L.L. Thurstone Psychometric Laboratory, University of North Carolina, (1996) *New Directions in Psychometrics.* Invited presentation at the annual meeting of the Psychometric Society, Banff, Canada

"... for only by varied iteration can alien conceptions be forced on reluctant minds."

Herbert Spencer, in the preface to The Data of Ethics, 1881. *Courtesy of Lise DeShea*

Bad Things Can Happen to a Good Field!

William Fisher points out that fields of scientific research share many features, good and bad. He identifies "The Emperor's New Methods" (Spence et al., 2003) as a cautionary tale for us all. It describes how research decisions are made in the field of genetics. Four thematic pitfalls are identified. Unwittingly, our field may fall into those same traps.

Theme 1: The Most Popular Approach Being Taken as the Only Acceptable One.
This is prone to happen when many new researchers are entering a field. They each ask, "What is the appropriate method?" They are told the most familiar one. An example was the analysis of rating scales. For some 20 years after the introduction of unidimensional polytomous Rasch models, many researchers continued to routinely analyze rating scales by dichotomizing them.

Theme 2: Scientific Practice Based on Myth Rather Than Evidence.
Rasch analysis has its own share of myths. A widely circulated one is that a large sample size is necessary. Another is the supposedly deleterious effect of "significant" misfit, leading to model rejection. Ben Wright's advice was to analyze all the data, then put the noticeably misfitting portion of the data to one side, reanalyze and compare the findings. Rarely was there any noticeable difference. Model fit is not the same as substantive impact.

Theme 3: Willingness to Establish Standards without the Protections of Rigorous Testing.
Spence at al. remark "end users, in general, know little about whether methods are accurately implemented in [computer] programs or how to recognize when the program has failed to give the correct answer. These shoddy standards for validation and calibration of tools almost certainly contribute to a climate in which it is extremely difficult to decide which methods are working and which are not. This deprives us in part of the single most important protective facet of empirical work: the proof should be in the pudding! However, what if one has no definition of what constitutes a palatable pudding?"

We have encountered sometimes humorous examples of this at Conferences across the years. A Presenter would show us an item hierarchy but without any indication of which direction corresponded to "more of the latent variable". Even the Presenter didn't know! Soon the audience would divide into two camps, "The top is more of the variable!" vs. "The bottom is more!", each with good supporting rationalizations. Finally, someone would notice that Appendix 3 of the Paper included a fragment of the original survey instrument. The dispute would be settled, but the audience was left bemused.

Rough prediction of results in advance of analysis is a powerful cross-check on software functioning. Are the sample expected to exhibit much or little of the latent variable? Are they expected to be homogeneous or diverse? What will the general form of the item hierarchy be? Are the rating categories intended to correspond to wide or narrow slices of the variable?

Theme 4: The Unfortunate Development of a "Cult of Personality"
"Reliance of an entire field on the recommendations or prejudices of a handful of individuals has, in the history of science as a whole, proved to be a very poor method of moving closer to the truth." (Spence et al.).

It is annoying to read published papers advocating, but misrepresenting, Rasch methodology. But this is far better than reading a succession of papers parroting the "party line". What is perceived to be a misrepresentation may be a deeper insight or a different perspective. Perhaps even the first step towards the next break-through. Georg Rasch himself perceived progress to lie in a certain direction: "It is to be hoped, however, that ... contributions from others will gradually enlarge the field where fruitful models can be established" (Rasch, 1980, xxi). Happily, this hope continues to be fulfilled. But areas he merely touches upon, such as investigation of construct validity and systematic diagnosis of local misfit, are now prime reasons for the adoption of Rasch techniques. Indeed, it may be that the philosophy of Rasch measurement has greater impact than its mathematics – a phenomenon already witnessed in the work of Newton and Einstein.

Spence M.A., Greenberg D.A. Hodge S.E., Vieland V.J. (2003) The Emperor's New Methods. *Am. J. Hum. Genet. 72:1084–1087, 2003.*

Rejecting Best Items?

"I would suggest rethinking your reliance on Rasch fit statistics as a criterion for item rejection .... In many case, the best (most highly discriminating) items would be rejected if one relied on Infit and Outfit statistics."

NCME reviewer, as reported by Ryan Bowles.

Conventional wisdom says "When items correlate highly with one another, those with the highest average correlations are the best items" (Jm Nunnally, *Psychometric Theory,* 1967, p. 261). But it is well-established that there can be too much of a good thing ... inter-item correlations can become too high:

> "Other things being equal, interdependent items tend to decrease the reliability of a test. ... For the tendency becomes to answer neither item or both items and thereby produces an effect equivalent to reducing the number of items in a test." (Percival M. Symonds, *Factors influencing test reliability,* Journal of Educational Psychology, 1928, 19, 73-87. Italics his.)

Rasch Infit and Outfit statistics flag items to which responses are overly predictable, an indication that, in some way, they are interdependent with other items.

Size vs. Significance: Standardized Chi-Square Fit Statistic

"The first of the distributions characteristic of modern tests of significance, though originating with F.R. Helmert [1875], was rediscovered by Karl Pearson in 1900, for the measure of discrepancy between observation and hypothesis, known as χ^2 [chi-square].. ... It supplies an exact and objective measure of the joint discrepancy from their expectations of a number of normally distributed ... variates" (R. A. Fisher, Statistical Methods for Research Workers.)

It is the χ^2 distribution which underlies many Rasch-model fit statistics. Even those based on the likelihood of the data capitalize on the fact that –2 log (likelihood) is asymptotically χ^2.

A χ^2 statistic with *k* degrees of freedom, d.f., is the sum of the squares of *k* random unit-normal deviates. Therefore its expected value is *k,* and its model variance is *2k.* This provides the convenient feature that the expected value of a mean-square statistic, i.e., a χ^2 statistic divided by its d.f. is 1.0. But the model variance of a mean-square statistic is *2/k.* Thus, as the number of degrees of freedom, i.e., the sample size, increases, the power to detect small divergences increases, and ever smaller departures of the mean-square from 1.0 become statistically "significant", i.e., surprising, if the data are indeed as modeled.

The relationship between the size and significance of mean-square statistics is shown in the Figure. The statistical significance is expressed as the value of the corresponding value on a unit normal distribution. For 2-sided *t*-tests, 1.96 corresponds to p=.05.

Test of Perfect Fit

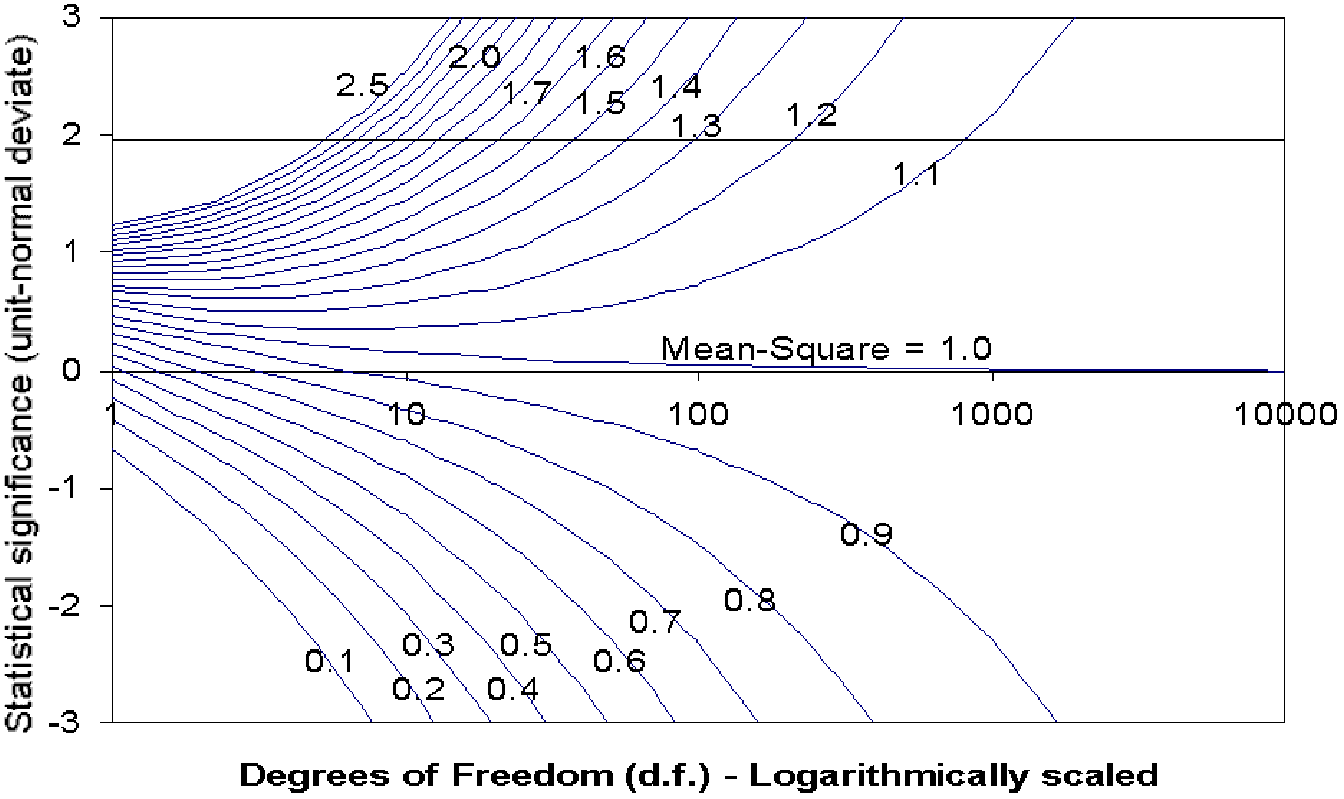
The null hypothesis for a significance test of "perfect" fit of these data would be "Mean-square=1.0". Since the Rasch model is a mathematical ideal, like a Pythagorean triangle, we never expect to encounter empirical data that match it exactly. So this is an instance in which we know, *a priori,* that the null hypothesis cannot be accepted.

A mean-square of 1.2 means 1 unit of modeled information and .2 of unmodeled noise. The plot indicates that items with as little misfit as this would be flagged as significantly misfitting if observed in samples of over 200 persons. On the other hand, grossly noisy items, with more unmodeled noise than modeled information, i.e., with mean-squares of 2.0 or more, are not flagged in samples of less than 10. Overall, useful sample sizes for standardized fit statistics appear to be in the range 50-250 data points for the "perfect fit" null hypothesis.

Test of Useful Fit

A null hypothesis of "useful" fit could be "mean-square = 1.5 or less" (e.g., RMT 14:2, p. 743). This would give a one-sided *t*-test. As the sample size (d.f.) increase beyond 30, there is increasing certainty as to whether these data are productive (mean-square ≤ 1.5) or unproductive (mean-square>1.5).

John Michael Linacre


Statistical significance (unit-normal deviate)
Mean-Square = 1.0
Degrees of Freedom (d.f.) - Logarithmically scaled