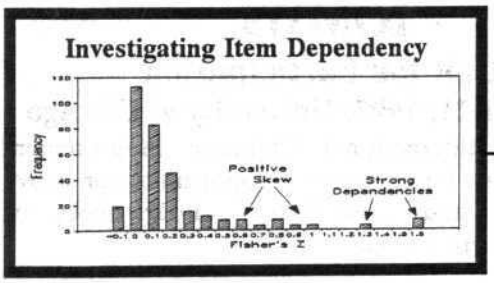
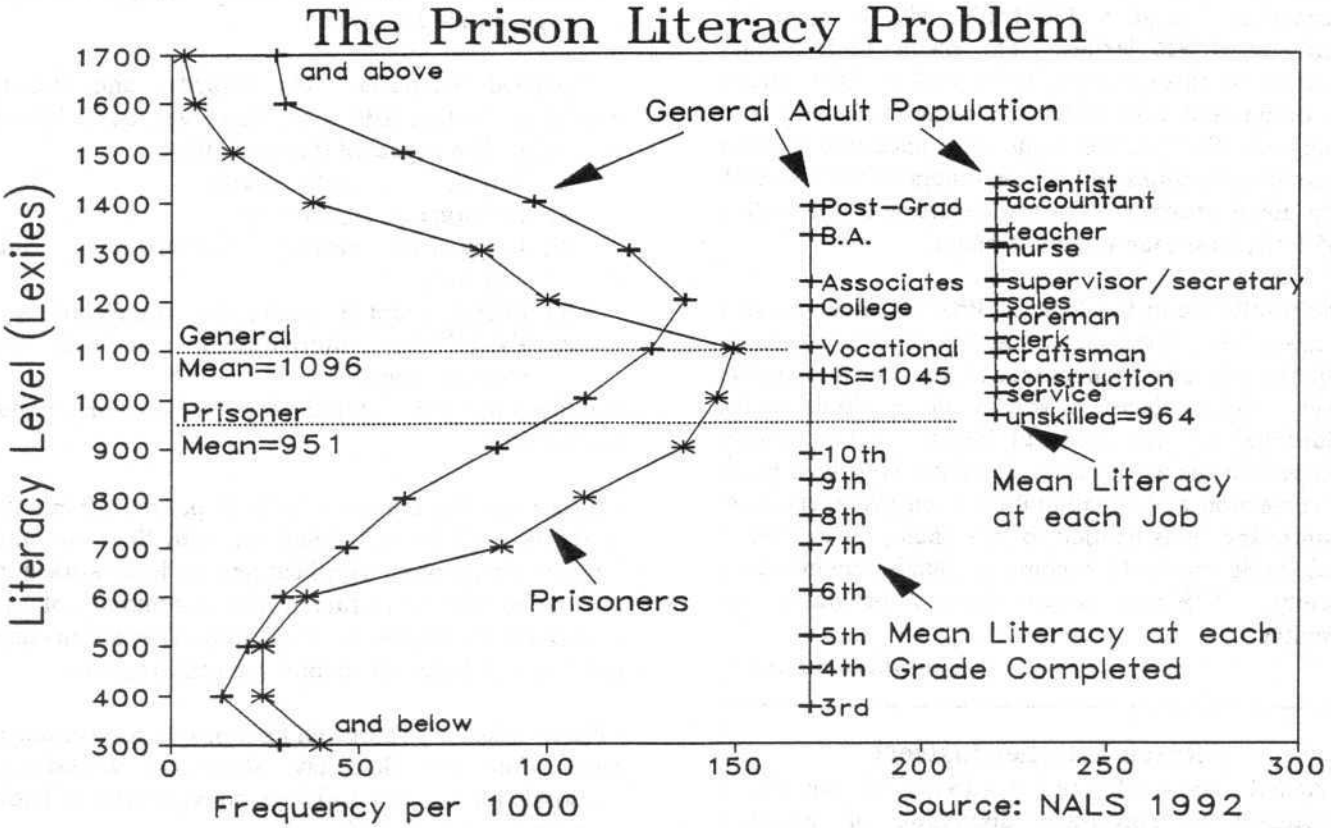


RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 10 No. 1 | Spring 1996 | ISSN 1051-0796



"The U.S. prison population has tripled since 1980 to a record 1.5 million. Another 3.5 million are on probation or parole. If this trend continues, the number of Americans under the control of the criminal justice system, including those in prison and on parole, will approach the number of full-time students enrolled in four-year colleges and universities. It is alarming that two-thirds don't have the literacy skills needed to function in society. An increasing number of states are reducing their support for education programs for prisoners. 'With so many of our young adults incarcerated, are we comfortable with their overall low levels of literacy? Most all will be released back into society. Should we let them remain so unprepared for employment and social responsibility?' stated Richard

Table of Contents

Coley." (From *ETS Developments* 41:2, 1995-96, p.8).

The 1992 National Adult Literacy Survey gives insight into the abysmal literacy of prisoners. Using sophisticated sampling techniques, the survey measured the literacy level of 1,103 prisoners and 23,617 adult members of the general population with a test of 184 literacy items.

Rasch analysis of the NALS data was conducted in two phases. First, the general adult population were measured and item calibrations obtained. By inspection, 1 logit = 180 lexiles with a mean test difficulty of 916 lexiles. The mean adult literacy measure on this scale was 1,096 lexiles. Adult means by occupation and education completed were also obtained. Second, the items were anchored at their general calibrations and the prisoners were measured. The mean prisoner literacy measure was 951 lexiles, 145 lexiles less than the adult mean.

The results are in the Figure. Prisoner mean literacy is lower than the mean of any occupational group. One third of the prisoners read at less than a 9th grade level. This places prisoners at a disadvantage additional to their criminal history. They cannot compete in the work-place. If prison is to be a place of correction and rehabilitation, then prisoners must acquire the skills needed to give them, on release, a reasonable chance to become productive members of society. Vigorous prison literacy programs are essential.

John Michael Linacre

Rasch Internet Listserv

ACER has established a LISTSERV on the internet to encourage discussion of Rasch measurement. If you would like to join the *Rasch Electronic Forum*, send to

mailserv@acer.edu.au

the following command as the text (not the Subject:) of your e-mail message:

subscribe rasch

You will receive a "welcome" message explaining briefly how the LISTSERV works. Then you will receive all messages posted to the Forum. You can read them and, if you wish, you can send a comment/reply of your own if you wish. The welcome message also explains how to unsubscribe.

Geoff Masters

IOMW9

Call for Participation

March 21-23, 1997, University of Chicago

The Ninth International Objective Measurement Workshop will be an exciting opportunity for active participation and networking. Four types of participation are invited:

· **Conventional Paper Presentations.** 12 Rasch-oriented papers will be scheduled for 20 minute presentations on Friday afternoon starting at 1:00 p.m. Paper proposals are invited.

· **Organized Symposia.** On Saturday and Sunday mornings, starting 8:30 a.m., there will be 5 1-hour symposia. The expected themes will be:

i) Setting Criterion Standards
ii) Performance Assessment
iii) Large-scale testing, measurement and reporting
iv) Measurement in rehabilitation medicine
v) Establishing international standards for measurement.

Applications are invited for symposium panel membership.

· **Discussion Workshops.** Up to 14 parallel 90 minute workshops will be scheduled on both Saturday and Sunday afternoons. Applications to lead workshop *discussions* (not to lecture!), with a summary of the material to be discussed, are invited. Discussions that build on the symposia themes are encouraged.

· **Poster Session.** There will be a poster session social hour after the Saturday afternoon workshops. Proposals for posters, software demonstrations, book signings, etc. are invited.

There will be social events at local restaurants on Friday and Saturday evenings.

Proposals by mail, FAX or e-mail, are due by **November 1, 1996.** Send a one-page summary of your presentation, symposium contribution, workshop theme or poster topic, along with your full name, mailing address, affiliation, telephone and FAX numbers, and e-mail address, to:

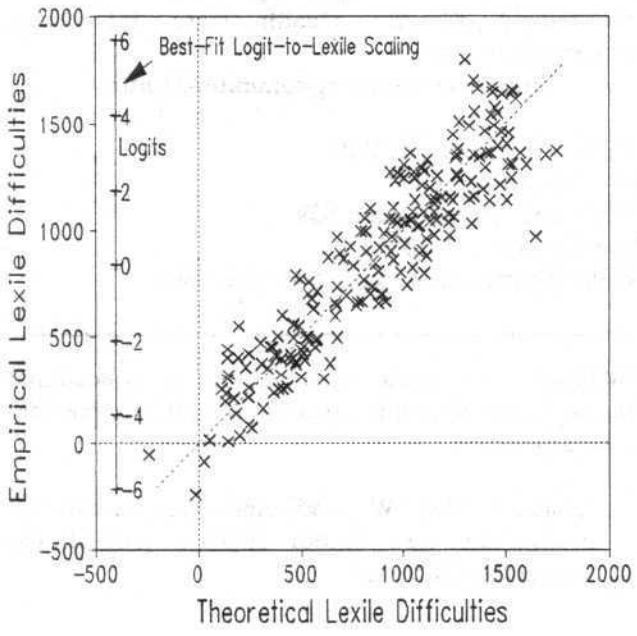
Ben Wright
IOMW9 Program — Tel. (312)702-1596
5835 S. Kimbark Ave. — FAX: (312)834-0326
Chicago IL 60637-1609 — MESA@uchicago.edu

Theoretical Prediction of Test Items

The Lexile theory of readability measurement is attempting for reading comprehension what Newtonian mechanics achieved for astronomy. Initially, the Ptolemaic system of eccentric circles and epicycles described the motions of the planets more precisely than Newtonian mechanics. Newtonian mechanics, however, embodied a strong predictive theory which identified discrepancies for investigation and explanation. Explanations included telescope misalignment, the influence of unobserved planets (Uranus, Neptune), and short-comings in Newtonian theory (orbit of Mercury, since explained by Einsteinian relativity). The Ptolemaic system of making the model fit the data would have discovered none of these. Once established, however, Newtonian mechanics went far beyond astronomy, transforming physics and enabling our modern world.

Lexiles simplify the complex, content-ridden process of reading comprehension into two abstract, content-free components: syntactic load (quantified by sentence length) and semantic demand (quantified by word frequency). Even though reading experts insist that lexiles are a hopeless over-simplification, empirical work with Lexiles demonstrates that *lexile theory-based* item calibrations produce reading comprehension measures as accurate as any *observational data-based* calibration.

The Figure shows the relationship between lexile-theory and data-based calibrations for 200 sentence completion test items. The data-calibrations are in logits, rescaled to follow an identity line. In this relationship (which gives equal weight to theoretical and empirical values), the standard deviation of empirical values around the theoretical values is 177 lexiles (about 1 logit). Though departures from the identity line invite investigation, their impact on practical measurement with tests of reasonable length is negligible (see Wright & Panchapakesan, 1969).



Because of findings like this, several U.S. States are using lexiles as a fair and practical method for equating nationally published and locally produced reading comprehension tests.

Hal Burdick, Jack Stenner
Metametrics Inc.
1100 Perimeter Park West, Suite 112
Morrisville NC 27560

Wright B.D. & Panchapakesan N. 1969. A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.

Questionnaire Item Difficulty

The wording of the widely-used Marlowe-Crowne Social Desirability Questionnaire provides useful guidance for interpreting constructs and writing items.

1) Moderate phrasing ("at times", "on occasion") is easier to agree with than extreme phrasing ("always", "never").

2) Admitting to a behavior ("I did it") is easier to agree with than correcting the behavior ("I took action about it")

3) Desire ("I want to give up smoking") is easier to agree with than commitment ("I will give up smoking").

4) Trying ("I've cut down smoking") is easier to agree with than succeeding ("I've given up smoking").

Fei Mo
Rush Cancer Institute
1725 West Harrison St.
Chicago IL 60612

Establishing Measurement Standards and Standard Measurements

The APA-NCME-AERA Standards for Educational and Psychological Testing (now under revision, *RMT* 9:3 p.440) are unenforceable procedural recommendations that perpetuate a proliferation of incommensurable numerical systems which fall far short of what physicists, merchants and carpenters require of measurement. A different organization, the American Society for Testing and Materials (ASTM), is concerned with enhancing communication by establishing quality-controlled universal metrics.

Measurement Standards Require Social Networks

Both science and commerce demand instrument-free linear (or ratio) measures. As was pointed out in *RMT*, 9(4) p.466-467, social scientists take the universal reproducibility of physical measuring units for granted, ignoring the global networks of technicians who establish and maintain these units. Social science lacks these networks and so is hobbled by a cacophony of different instruments each purporting to quantify an important variable. But each instrument defines its own variable with an indeterminate relationship to any other instrument's variable. Worse, the reported quantities are idiosyncratic ordinal units of indeterminate quality.

Establishing and maintaining universal metrics for medical outcomes measures, for instance, requires a network of practitioners among whom would circulate reference data and reference instruments for each variable. The network would evaluate different health care facilities' measurement results, and also test and certify different brands of instruments as measuring in the standard metric unit.

A Model Standards Network: ASTM

Models of, and a place for, such a network exist in the ASTM. ASTM was organized in 1898 and has grown into one of the largest voluntary standards development systems in the world. It provides a forum for producers, users and consumers to meet and write standards for materials, products, systems, and services. From the work of its 134 standards-writing committees, ASTM publishes more than 8,500 standards in 68 volumes of the "Annual Book of ASTM Standards".

What is a Standard?

"Standardization refers to the general acceptance of concepts, quantities, terms, rules, and definitions that serve as reference points for professionals in a given field. Standards provide criteria for a common language, ensuring reliable communication between all parties in the conduct of business. For example, U.S. electric current standards, which cover voltage, amperage, outlets, and plugs, are so widely accepted that manufacturers automatically incorporate them in their designs. These standards streamline product development for manufacturers and ensure consumers hassle-free usage regardless of where they live in the U.S. or what type of appliance they purchase."

From *Evolving with Technology: Information Systems, Standards, and Public Health*, published by the Joint Council of Governmental Public Health Agencies.

An Opportunity in Medical Standards

The ASTM electronic medical record subcommittee, E31.19, is interested in Rasch's models for scale-free measurement because they offer unique advantages for quality control, quantification, and outcomes comparison of rating scale data. The head of E31.19, Gretchen Murphy, who represents the American Health Information Management Association in ASTM, invites us to contribute our expertise to the development of quality and quantity standards for medical outcomes data. In order to coordinate these contributions, I was made head of a working group on health quality measurement standards of the Electronic Medical Records Committee, E31, of ASTM. I welcome your participation in this group.

For more information on ASTM point your Web browser to http://www.astm.org/. To go directly to information on the Health Care Informatics subcommittee, use

http://www.astm.org/commit/e-31.htm.

William P. Fisher, Jr., PhD
LSU Medical Center
1600 Canal Street, Room 809
New Orleans, LA 70112
wfishe@nomvs.lusmc.edu, (504) 568-6864

"Without clear goals, specification of educational standards and good measures of them, it is impossible to be productive."

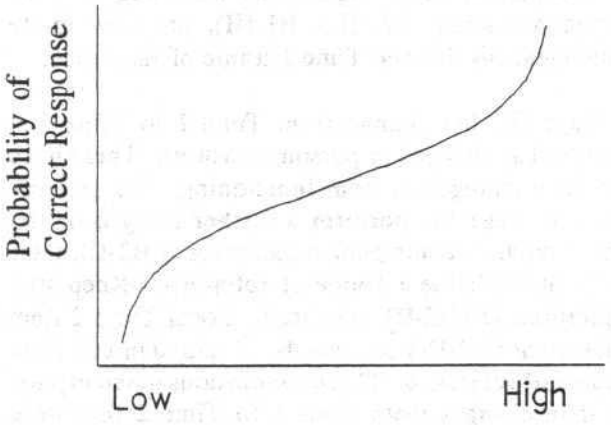
Odden A, Clune W (1995) Improving educational productivity and school finance. Educational Researcher 24(9) p.7

Post-Doctoral Fellowship

The Rehabilitation Institute of Chicago announces the availability of a Post-Doctoral Fellowship in its Rehabilitation Services Evaluation Unit. RIC is a national leader in rehabilitation and offers a comprehensive continuum of rehabilitation care. The Fellow will design and conduct health services research on medical rehabilitation topics. The position provides the opportunity to collaborate with a wide array of health service researchers and providers through the Medical School of Northwestern University. The principal responsibilities include developing and submitting research proposals on health services research topics across a continuum of settings, writing grant proposals, conducting statistical analyses, writing research reports, speaking at professional meetings, consulting with investigators about the dissemination of research presentations and reports, and providing lectures and consultation on research design and statistical analysis topics. The position requires a doctoral degree in public health, epidemiology, educational measurement, biostatistics or an allied health field with an emphasis on measurement and evaluation.

Send *Curriculum Vitae* and names of three references to:

Human Resources Department
Rehabilitation Institute of Chicago
345 East Superior Street
Chicago IL 60611



A typical item characteristic curve as depicted in Scott T. Meier (1994) *The Chronic Crisis in Psychological Measurement and Assessment: A Historical Survey*. San Diego, Ca.: Academic Press, p. 143, Fig. 28.

No wonder there's a crisis!

Mid-West Objective Measurement Seminar

Friday, May 17, 1996, Chicago

Defining the Collegiality of an Environment
Rita Bode, University of Illinois at Chicago

Taming Tasks with Facets
Greg Stone, Dental Assisting National Board

Relative Fit
Richard Smith, Research Foundation Inc.

Anxiety and Test Performance
Richard Gershon
Computer Adaptive Technologies

Reading Instruction in Chicago
Winifred Lopez
Malaysian Ministry of Education

Measuring Customer Satisfaction
Selwyn Becker & Kirk Becker
Graduate School of Business

Prison Literacy
John M. Linacre

Predicting the Unpredictable
Mark Moulton
MESA Psychometric Laboratory

Rasch Measurement with Fixed Guessing
Liru Zhang
Delaware Department of Public Instruction

Managing Interactions among Judges, Projects and Candidates
Mary Lunz
American Society of Clinical Pathologists

Discovering the Structure of Math
Julia Smith, University of Rochester

Is it Changing Children or Changing Opinions that Matters
Melissa Roderick & Susan Stone
School of Social Service Administration

Equating Quality of Life Instruments
David Cella & Chih-Hung Chang
Psychosocial Oncology, Rush Cancer Institute

Why Measures are Better than Raw Scores
Jin Shei Lai, University of Illinois at Chicago

5 Grading Points Can Work Better than 20.
Chew Chin Sing, Ngee Ann Polytechnic

The Future of Reading Comprehension Testing
Hal Burdick, Jack Stenner, Metametrics Inc.

Time 1 to Time 2 Comparison

Measurement of change presents a nasty challenge. We expect persons (patients, students, experimental subjects) to change from *Time 1* to *Time 2*. But the functioning of test items and rating scales may also change, even when identical data collection protocols are used. The challenge is to measure persons and items in the same clearly defined frame of reference encompassing both time points, so that measurements of change will have unambiguous numerical

Most analysts, including those misusing raw scores as measures, assume without verification that the functioning of test items and rating scales remains constant across time. The change-scores they report are spoiled by uncertain frames of reference.

Rasch analysts proceed at least to Stage I (see Figure). Here the Time 1 and Time 2 data are analyzed independently. This aids the detection and elimination of gross errors in data entry and test administration. It also permits a rough verification of the stability of the frame of reference by plotting the item difficulty calibrations at Time 2 (D2-I) against those at Time 1 (D1-I). A close fit to the identity line is reassuring. For each rating scale, cross-plotting key points on the *expected score ogives* for Time 1 and Time 2 (derived from F1-I and F2-I) and then observing fit to the identity line verifies scale stability. When these item and rating scale plots indicate stability, then the plot of ability measures for Time 2 (B2-I) against Time 1 (B1-I) provides a dependable picture of person changes.

Stage I, however, usually reveals problems. Some items are too far from the identity line. The rating scale structure is time dependent: upper categories may be rarely used at Time 1, lower categories at Time 2. The meaning of changes in person measures is now uncertain – further analysis is needed.

Stage II (see Figure) stacks the data vertically, so that each person appears twice (Time 1 and Time 2) and each item once. This matrix yields three findings:

a) Items that were away from the identity line in Stage I now show greater misfit than in the separate Stage I analyses. This confirms that these items function differently at the two time-points, and suggests that each such item might be split into two separate items: a Time 1 version and a Time 2 version. The column of item responses can be split into two columns (with missing data at the other time point) so that the two time-interacting versions of each original item are calibrated independently. Re-analysis should show an overall improvement in fit and an increase in person separation.

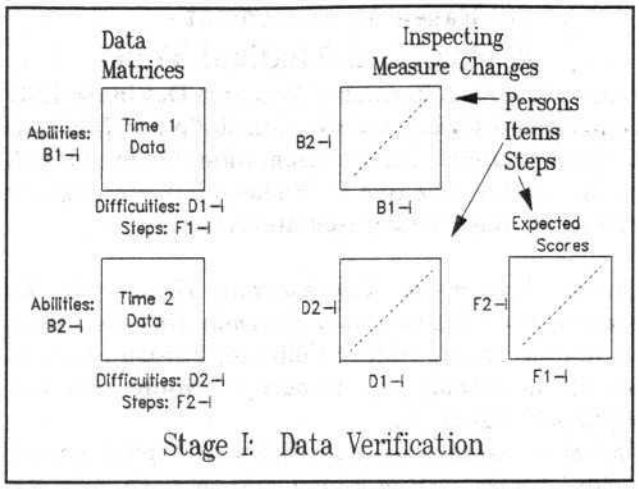
b) The rating scale calibrations used for the final item structure are those most consistent with both Time 1 and Time 2. These become the anchor calibrations (F1&2-II) for later analyses.

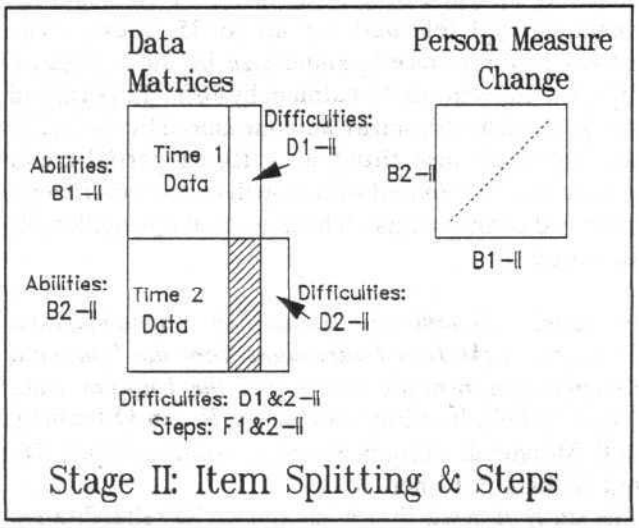
c) Plotting Time 2 abilities (B2-II) against Time 1 abilities (B1-II) at Stage II is more meaningful than Stage I. But even these measures are still in an intermediate frame of reference that reflects neither Time 1 nor Time 2 accurately.

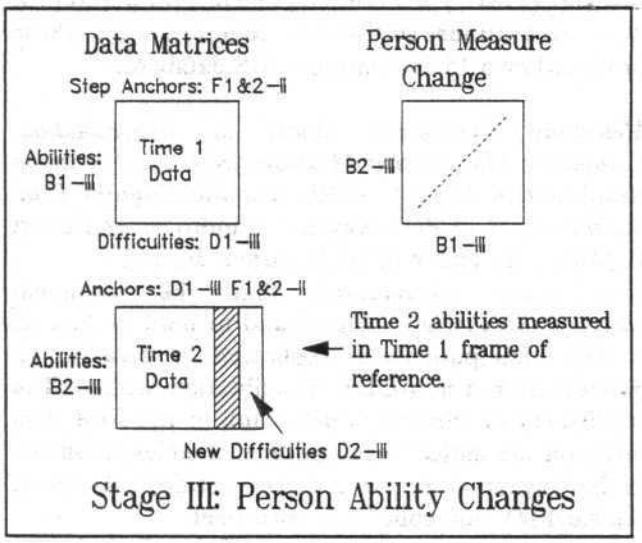
Stage III (see Figure) installs Time 1 as the benchmark. We measure change away from Time 1. (Time 2 can also be treated as a benchmark.) Benchmark item calibrations (D1-III) and person measures (B1-III) are obtained from the Time 1 data using the F1&2-II calibrations as step anchors. The D1-III and F1&2-II calibrations are now applied to the Time 2 data, except for split items. Split items are those which function differently at Time 1 and Time 2. Consequently, split items are not anchored at their Time 1 calibrations for the Time 2 data, but are calibrated from these data. Time 2 person measures (B2-III) and Time 2 calibrations for split items (D2-III) are now estimated in the Time 1 frame of reference. The same ruler has been applied at Time 1 and Time 2. The plot of B2-III against B1-III, along with the change measures (B2-III - B1-III), are now in an unambiguously defined Time 1 frame of reference.

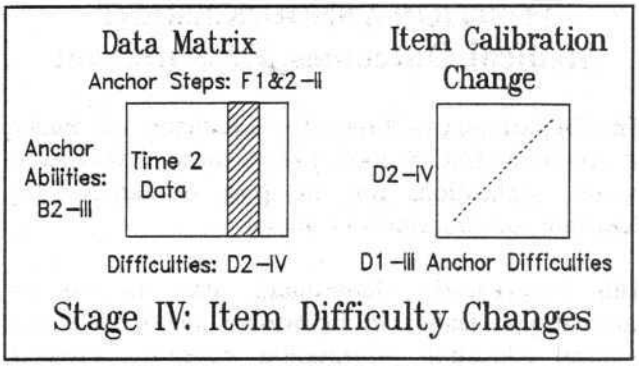
In Stage III, the change from Time 1 to Time 2 is expressed as changes in person measures. There have also been changes in item functioning. To examine these, in Stage IV, perform a further analysis of the Time 2 data. Anchor person measures at B2-III, their values in the Time 1 frame of reference. Keep step calibrations (F1&2-II) anchored. Local Time 2 item calibrations (D2-IV) can now be obtained in the Time 1 frame of reference. These calibrations make explicit the item changes from Time 1 to Time 2 that were implicit in the changes of person measures. A plot of D2-IV against D1-III (including split items) displays the changes in item difficulty across time, again in a clearly defined frame of reference.

Benjamin D. Wright









Disattenuating Correlation Coefficients

When two sets of measures, {x} and {y}, are correlated, measurement error lowers the correlation coefficient below the level it would have reached had the measures been precise. The *reliability* of a set of measures is the proportion of observed variance *not* due to measurement error, r_{xx} for set {x} and r_{yy} for set {y}. Measurement error can be removed from a correlation coefficient, r_{xy}, to estimate a correlation coefficients disattenuated of measurement error, ρ_{xy}, by the formula:

$$\rho_{xy} = r_{xy} / \sqrt{r_{xx}} \sqrt{r_{yy}}$$

Disattenuated values greater than 1.00 indicate that measurement error is not randomly distributed.

Muchinsky (1996) summarizes features of the disattenuated correlation coefficient:

1. Disattenuation does not change the quality of the measures or their predictive power.

2. Disattenuated correlations are not directly comparable with uncorrected correlations.

3. Disattenuated correlations are not suited to statistical hypothesis testing.

4. Disattenuation is not a substitute for precise measurement.

5. But, disattenuation tells us whether the correlation between two sets of measures is low because of measurement error or because the two sets are really uncorrelated.

Randall E. Schumacker

Muchinsky P.M. (1996) The correction for attenuation. Educational & Psychological Measurement 56:1, 63-75.

POSITION ANNOUNCEMENT
Medical Outcomes Measurement

The Department of Preventive Medicine and Public Health in the LSU School of Medicine at New Orleans invites applications for the post of tenure-track Assistant (or Associate) Professor.

This two-year-old department aims to be an international leader in innovative health care and medical education information systems. Projects involving Rasch measurement include computerized medical records access via networked, handheld pen devices; a state-wide telemedicine consultation system; an electronic emergency medicine data system; and educational projects involving computer-adaptive testing and performance assessments.

The successful candidate will have a Ph.D. in education, psychology, public health, or other relevant discipline, with a strong emphasis on quantitative methods employing Rasch models. Experience with medical outcomes, health status, customer/patient satisfaction, quality of life, or functional assessment research is not essential, but would be a plus. We are especially interested in skill in the application of Rasch measurement software to problems of item banking, instrument equating, multifaceted designs, and computer-adaptive item administration.

Salary will be commensurate with experience. Applications must be received by August 1, 1996. Starting date will be as soon as possible thereafter. Candidates should send a letter of interest stating relevant qualifications and achievements, a *curriculum vitae*, and the names, addresses, and telephone/fax numbers of three references to:

William P. Fisher, Jr.
Associate Professor
Department of Preventive Medicine and Public Health
LSU School of Medicine
1600 Canal Street, Suite 800
New Orleans, LA 70112

The LSU School of Medicine in New Orleans is an Equal Opportunity Employer, encourages applications by women, minorities, and persons with disabilities, and offers a smoke-free work environment.

Rasch Measurement -
A Practical Medical Tool

June 7, 1996, was Residents' Research Day in the LSU Department of Medicine's Section of Physical Medicine & Rehabilitation. Three presentations featured Rasch measurement. The one by Yadav, et al. was selected best of the day. The presentations were:

Growth Factors vs. Conventional Therapy in the Treatment of Chronic Lower Extremity Diabetic Ulcers, by Rajesh Yadav, Jeffery Filiberto, Furqan Siddiqui, Joseph J. Biundo, Jr., Robert C. Mipro, Jr., and William P. Fisher, Jr.
This study compared the efficacy of topical growth factor versus conventional therapy in chronic nonhealing diabetic foot ulcers. Fifteen subjects were randomly assigned to growth factor or conventional treatment and followed for up to 35 weeks. The growth factor reduced wound size by an average of 71%, the conventional treatment by 46%. Physical and psychosocial health status were measured by the SF-20 and the data were fitted to a Rasch partial credit model. Persons treated with growth factor experienced improved health status. Those treated conventionally worsened.

*Measuring Functional Status in Rehabilitation: Comparing FIM **Item Calibrations** from the Louisiana Rehabilitation Institute (LRI) and the Uniform Data System (UDS)*, by Maryam Qayum, Karen Ortenberg, Rolf Morstead, Furqan Siddiqui, Robert Mipro, Jr., and William P. Fisher, Jr.
This study showed that a sample of 70 rehabilitation patients measured with the Functional Independence Measure (FIM℠) at a non-UDS facility produced FIM item calibrations statistically identical with those produced by a 15,000-patient UDS database.

*Measuring Functional Status in Rehabilitation: Comparing FIM **Patient Measures** from the Louisiana Rehabilitation Institute (LRI) and the Uniform Data System (UDS)*, by Paul Mayes, Alejandro Perez, Robert C. Mipro, Jr., and William P. Fisher, Jr., PhD.
This study documented that the nonlinear score/measure relationships found to hold in data on 15,000 UDS patients is replicated in data on 70 patients from a non-UDS rehabilitation facility. The consistency of this relationship and of the FIM item order on the motor and cognitive variables, as shown in the Qayum, et al. presentation, justifies LRI use of the KeyFIM data collection worksheet.

William P. Fisher, Jr.

Calendar of Events

June 27,	1996 COMET (monthly), Chicago Niko Bezruczko (312) 684-8549
Dec 6,	1996 MOMS, Chicago Ben Wright (312) 702-1596
Mar 21-23,	1997 IOMW9, Chicago Ben Wright (312) 702-1596
Mar 24-28,	1997 AERA, Chicago AERA (202) 223-9485

AERA Annual Meeting

Call for Papers

March 24-28, 1997

Interest is mounting in discovering solutions to the challenge of new measurement problems. Help by participating in the 1997 AERA Meeting! Send your proposal in AERA format (see *Educational Researcher* May 1996 p. 33-46) to reach me by **August 15, 1996.** Though SIG sessions are well attended by our friends, we are especially interested in Rasch-related papers that will appeal to a wider audience. Your proposal will be reviewed anonymously. Presentation formats include papers, round-tables, posters, symposia and alternative session formats. If you are proposing a format other than a paper, explain that on your proposal. You may propose to present the same paper at both AERA and IOMW9, but plan to emphasize different aspects for the different audiences. In your proposal, **please include** adequate descriptions of data analyses and full citations for your references.

I would also like to hear from you, if you are willing to serve as a reviewer, session chair or discussant.

Richard M. Smith
Marianjoy RFI
26 West 171 Roosevelt
Box 675
Wheaton IL 60189
(708) 462-4102, FAX (708) 462-4547
After August 3rd: (630) 462-4102, FAX (630) 462-4547

Rasch Measurement SIG Officers

Geofferey N. Masters	Chair
Betty A. Bergstrom	Secretary
Richard M. Smith	Program Chair
Ben Wright	IOMW9 Chair
John Michael Linacre	Operations Manager

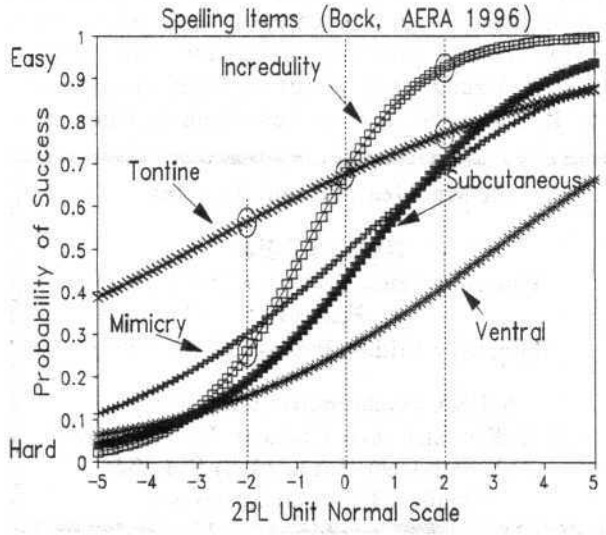
Construct Problems with Descriptive IRT

Items and constructs are in constant tension. Each item is intended to probe an underlying construct, but each item also has idiosyncracies. Only when items are coerced into a sufficient level of cooperation does a stable construct emerge. The process of coercion includes painstaking item writing, item selection, investigation of misfit statistics and careful consideration of the integrity of the evolving construct as it becomes defined by an emerging item hierarchy.

In its celebration of item idiosyncracies, IRT abandons the construct. At NCME 1996, Darrell Bock presented a Table of 2PL parameters for 100 spelling items. The 2PL model parameterizes, as "item discrimination", differential performance on items by high and low performing examinees. Since differential item performance disrupts construct stability, variation in "item discrimination" is identified as misfit in a Rasch analysis.

Allowing each item its own life muddles the construct. ICCs for 5 words from Bock's 100 spelling items are shown in the Figure. Bock reports that these words show good fit to his 2PL model. But the item hierarchy for low performers at -2 contradicts that for high performers at +2. For middle performers at 0, *tontine* and *incredulity* are equally easy to spell. But for low performers at -2, *tontine* was easier to spell than *incredulity*, while for high performers, *tontine* was harder to spell than *incredulity*. Since a clear construct is required in order to know what measures mean, descriptive IRT is not measurement.

Benjamin D. Wright



Communication Validity and Rating Scales

Test validity, the extent to which a test measures what it is intended to measure, is critical. Although many researchers review content, construct, and statistical aspects of validity, even conscientious researchers usually take for granted that respondents understood the tasks they were asked to perform and then performed them in a coherent way.

Despite the fact that rating scales and response formats are the media of communication with respondents, researchers ignore "communication validity". Did the rating scale categories perform as intended? Did respondents converse with the test developer in a common language free of idiosyncratic category usage, response sets, and ambiguous terminology? Were respondents able to distinguish the response levels of each rating scale? How did they order the levels? It is pointless to examine any other form of validity until we have established that we have listened carefully to what test respondents have told us about our variable.

We want our respondents to manifest a clear definition of the variable. We also want to locate them at separate locations along the variable. Their use of the rating categories is crucial. We need respondents to provide an unambiguous hierarchical ordering of our categories. Their response behavior may not concur with our original presentation of our response categories.

Rasch analysis provides a statistical method for ascertaining and verifying respondents' perceptions of the ordering of category meanings (*RMT* 9:3 450-451, 9:4 464-465). Categories labeled "Don't know", "No opinion", and "Does not apply" are prime candidates for misplacement in the category hierarchy. Such category labels provoke irrelevant and evasive responses. Usually they do not belong in the hierarchy at all. It is often better not to use them or, when used, to treat their selection as missing data.

Each category is intended to increase the discrimination of the rating scale and so to increase the information in all responses. But confrontation by too many response alternatives muddles respondents. Respondents rarely make stable discriminations among more than 6 levels. Sometimes 2 or 4 levels are all they can negotiate. Excess categories introduce more noise than information by forcing respondents to make their fine choices idiosyncratically, such as by preference for even or odd numbering.

Responses to excess categories can be combined with those of adjacent categories in a "collapsing" process. When we collapse adjacent categories, we construct new categorizations. Rasch analysis provides the opportunity to study how well these new categories function. The optimal categorization is that which
a) provides the best construct definition,
b) best separates respondents along the variable,
c) produces the best fit of data to model.
These criteria usually cooperate to identify an optimal scoring solution.

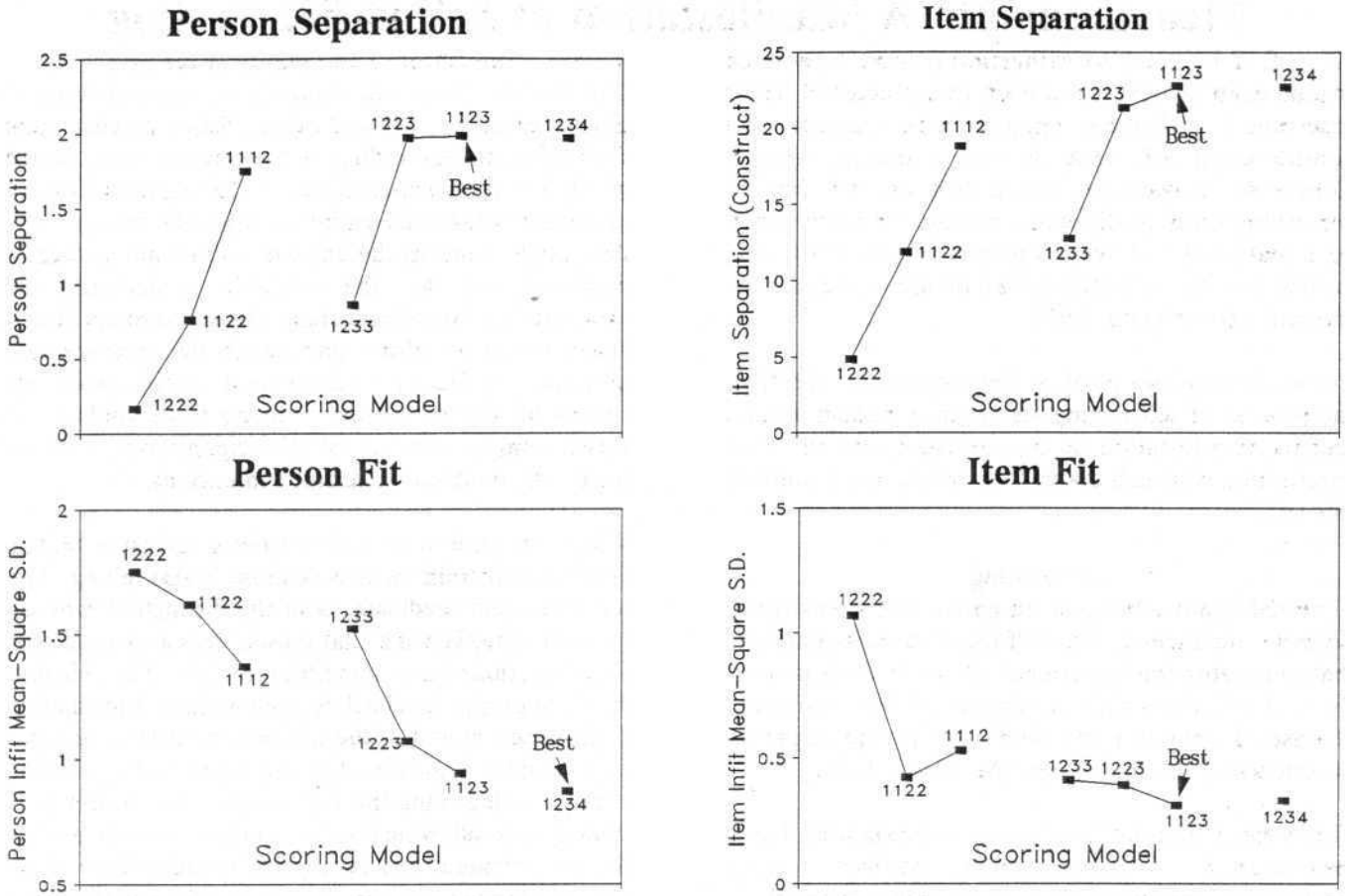
The Figures summarize different categorizations of the responses of teachers to 19 items about reading instruction. The printed rating scale was:

No Emphasis			Major Emphasis
1	2	3	4

This scale suffered from the common flaw of unlabelled (and hence not clearly defined) categories.

The Figures show the statistical implications of different collapsings. "1234" means the categories are assigned their printed ordering. "1222" means that original category "1" is retained as "1", but original categories "2", "3", and "4" are collapsed into one category "2". The statistics are almost unanimous in declaring that collapsing categories "1" and "2" provides the most informative categorization. Thus, our respondents tell us that they can only discriminate three levels of emphasis in this context. The most valid communication with our respondents is then not our printed scale of 4 theoretical categories, but their experiential scale of three empirical categories. It is on the basis of their scale that investigation of the other forms of validity is best pursued.

Winifred Lopez



When to Inspect!

K. Tang and J. Tang (1994, Design of screening procedures: a review. Journal of Quality Technology 26(3) p.209-226.) list some useful rules about when to inspect during manufacturing.

1. Inspect after procedures likely to produce nonconforming items.
   Rasch: After applying a scoring key, check for misfit and negative point-biserial correlations.

2. Inspect before costly procedures.
   Rasch: Inspect before analyzing an enormous sample. Inspect output on screen before printing on paper.

3. Inspect before procedures where nonconforming items may jam machines.
   Rasch: Misinterpretation of incorrect results "jams" the mind, vitiating analysis of correct results.

4. Inspect before procedures that cover up nonconforming items.
   Rasch: Scan maps and numerical details before counting on summary statistics.

5. Inspect before assembly procedures where rework is costly.
   Rasch: Inspect before using Rasch measures for statistical analysis or substantive decisions. Errant pass-fail decisions are hard to reverse.

Rasch Measurement Transactions
5835 S. Kimbark Ave, Chicago IL 60637-1609
Tel. (312) 288-5650 FAX (312) 834-0326
E-mail: Rasch@uchicago.edu
Editor: John Michael Linacre
Associate Editor: Benjamin D. Wright

Published quarterly: June, Sept, Dec, March
Annual dues: $8, renewable in March

Text on ERIC/AE at *gopher.cua.edu*

Measurement: A Manifestation of Self-Other Testing

"A self, if it is not to wither away, must forever be testing itself against the nonself in a process of active assertion... Testing implies both respect and consideration for what we test ourselves against. Otherwise it becomes not a test of self, but of something entirely different, perhaps of brute force. As a matter of fact, what a person selects as a testing ground is most indicative of the nature and quality of the self" (Bettelheim, 1967).

Let us develop two parts of Bettelheim's insight: first, the process of self-testing, its origin in human nature, and its manifestation in measurement; second, what measurement reveals about 'the nature and quality of the self.'

Self-Testing

"The self is not a thing or an entity; it is a concept; a symbolic abstraction. The self refers to the uniqueness that separates the experience of an individual from those of all others while, at the same time, conferring a sense of cohesion and continuity to the disparate experience of the individual" (M. Basch, 1983).

This definition of self has much in common with Rasch measurement. Measurement involves testing expectations for workable fictions called "variables". Like the self, variables are not things or entities. They are concepts. Measurement separates one variable from all others. Measures describe continuity in a variable. Fit statistics indicate cohesion.

Jane Loevinger (1976) argues that the 'self' reasons, judges, evaluates in order to make sense of the world. Rasch measurement relies on a model that functions to make sense of the world. The model takes into account the actors and their tasks in the world. It makes predictions. It allows us to evaluate and judge the "world" on the basis of those predictions.

Many developmental psychologists consider the development of self to occupy only a short time span. They suppose that while newborn infants do not separate 'self' from 'other', by 24 months children have made the distinction. Bettelheim suggests, however, that testing the distinction between self and other never ends. Measurement, too, is a never-ending process: every attempt at measurement confirms or challenges the measurement process and the underlying variable – we can never stop.

The Nature and Quality of the Self

The way we choose to conduct measurement reveals our beliefs about self and other. Rasch measurement defines an understanding of the way the world works which leads to expectations. The expectations are examined using data gathered from the world. The data might confirm the expectations or might suggest improvements to the variable underlying the expectations. Most important, the tested sample has a voice, a way to inform and enrich the measurement process. A Rasch measurement sample does not consist of victims, voiceless objects to be studied. A Rasch sample consists of participants who provide feedback about the measurer's intentions.

Whom we choose to test ourselves against is telling. How we constrain those we choose is also telling. The self that seeks feedback about the distinction between self and other is a self that grows. It is also a self that believes others have something to say. The self that does not permit feedback is a self without information, a self whose model of the distinction between self and other cannot improve. On the other hand, allowing others to determine the self, rather than inform it, is analogous to allowing the data to determine the model. No improvement can take place because there is no self to improve, no variable to enrich. Zaner (1981) writes that testing is "a continually ongoing, internally rhythmed, and always precarious mutuality." This mutuality is the key to self-other testing. It is also the key to Rasch measurement.

Self-other testing is one of our first tasks in development. Rasch measurement mirrors this natural and essential tendency. In the first stage of constructing measures, our understanding and expectations for the world are organized into a variable. This variable is based on our experience in the world, our understanding of ourselves and our position, and our observation of and interaction with others. In a second stage, Rasch measurement compares expectations based on this variable with a sample. The resulting comparison informs and enriches the variable. In Rasch measurement both self and other participate in exploring relationships. It is this essential ongoing, dynamic mutuality that sets Rasch measurement apart.

Kathy Aldred

(References opposite)

Quantifying Item Dependency

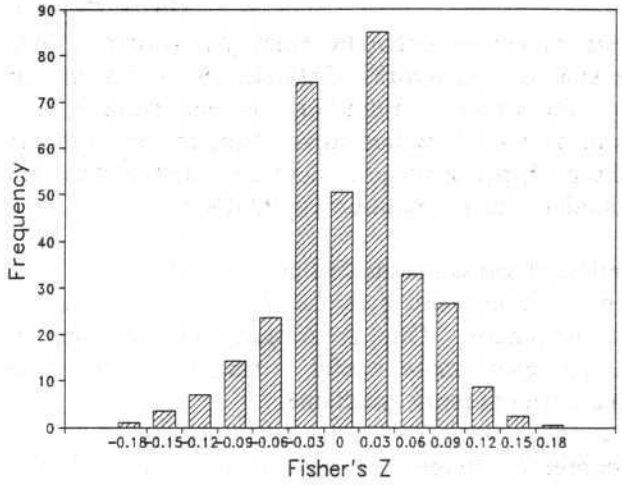
Some items are more closely related than others. Comprehension items that address the same written text and arithmetic items requiring the same operation are examples. In Rasch analysis, this is no major source for concern provided there is no extreme systematic dependency between two or a few items. When such dependency does exist, combining the separate items into one partial-credit item may be useful. Low values of the Rasch INFIT statistic signal dependency. When a diagnostically specific investigation of inter-item dependency is required, Fisher's Z is useful.

1) Obtain the Rasch-based standardized residuals for each observation of Person n on item i:

$$z_{ni} = \frac{Observed_{ni} - Expected_{ni}}{Standard\ Error_{ni}}$$

2) Correlate the standardized residuals for all pairs of items i, j across all N persons:

$$r_{ij} = \frac{\sum_n z_{ni} z_{nj} - \sum_n z_{ni} \sum_n z_{nj} / N}{N\ SD(z_{ni})\ SD(z_{nj})}$$



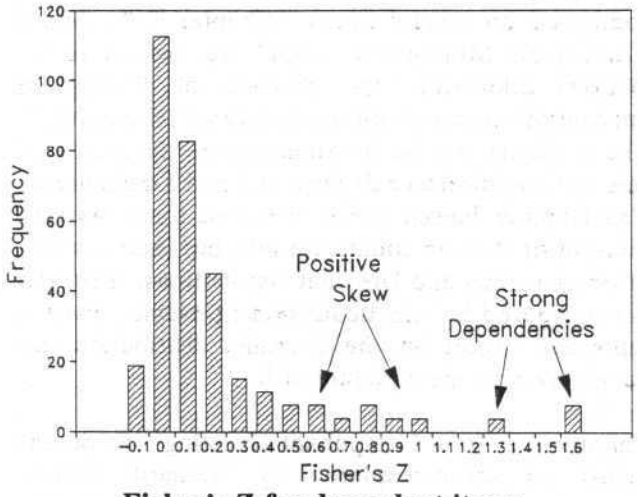
Fisher's Z for independent items

3) Use Fisher's Z to linearize Pearson correlations:

$$Z_{ij} = \frac{1}{2}\log\left(\frac{1 + r_{ij}}{1 - r_{ij}}\right)$$

4) Draw a histogram of the Fisher Z values. Noticeable positive skew (see Figures) indicates the presence of sets of dependent items. A cluster analysis of the Z matrix will identify the sets. Outlying positive Z values flag pairs or sets of highly dependent items for further investigation.

Linjun Shen
National Board of Osteopathic Medical Examiners
2700 River Road, Suite 407
Des Plaines IL 60018



Fisher's Z for dependent items

Self-Other Testing

Basch, Michael Franz. The Concept of "Self": An Operational Definition, 1983. In Lee, B. & Noam, G. (Eds.) Developmental Approaches to the Self. New York: Plenum Press.

Bettelheim B 1967. The Empty Fortress: Infantile Autism and the Birth of the Self. New York: The Free Press, p. 81. Quoted in Fisher, 1991.

Fisher, William Jr. 1991. Bettelheim's Test. Rasch Measurement Transactions 5:3 p.164-5.

Loevinger, J. 1969. Ego Development. San Francisco: Jossey-Bass.

Zaner R 1981. The Context of Self: A Phenomenological Inquiry Using Medicine as a Clue. Athens, Ohio: Ohio University Press. p. 188. Quoted in Fisher, 1991.

Objective Measurement: Theory into Practice, Volume 3 as a Provocation to Thought

This volume, published by Ablex (Norwood NJ, 1996) emanated from IOMW7 (Atlanta, 1993). Its Preface, by editors George Engelhard, Jr. and Mark Wilson, summarizes 22 authoritative chapters (35 authors). These chapters suggest ideas and approaches that stimulate our preparation for IOMW9.

Philosophical concerns are not central in Vol. 3 ,as they were in Vols. 1 & 2. It is practical concerns that dominate. One is to make the outcome of measurement more useful. Another is to make measurement more adaptable.

Chapter 7, "Judge Performance Reports: Media and Message" (J. Stahl & M. Lunz) addresses the pivotal issue in performance assessment: How can we monitor, diagnose, control and improve judge rating behavior? The first step is to treat raters as intelligent humans (rather than rating machines). The second is to give them feed-back they can understand and use to modify their own behavior. Pages 120-121 show quality control charts that raters act on, and W.E. Deming would be proud of. What is the third step?

Chapter 8, "Examining Changes in the Home Environment..." (J. Monsaas & G. Engelhard, Jr.) intrigues us with a variety of graphical devices for presenting results. The juxtaposition on p. 132 of a Table (of reliability coefficients) and a Figure (depicting time effects) convinces the reader that while a Figure is memorable, a Table is forgotten, even as it is read.

The struggle to make measurement more flexible is conducted on several fronts. Chapter 9, "... Mixed Coefficients Multinomial Logit" (R. Adams & M. Wilson) addresses "the problem of finding an appropriate model to suit the structure of the context." The challenge will be to communicate the results of this mathematical *tour de force* and make them useful. Van Duijn & Jansen (1995) overcome some awkward features of Poisson counts adroitly, but their solution involves gamma and Dirichlet distributions. They too are hampered by communication problems. Even a renowned expert on the Dirichlet distribution was unable to draw me a picture of it.

Chapter 15, "Item Component Equating" (R. Smith) builds on ideas explored by Gerhard Fischer. Measuring the parts can be more useful than measuring the whole. Practicality is the problem. Can designs be developed that allow parts to be embedded in different contexts and then used to link diverse wholes? This chapter evokes a new approach to test equating based on the components of test items, situations, tasks, judges(?). Who will put it to use?

Chapter 18, "Constructing Questionnaires.." (E. Roskam, N. Broers) presents one way in which items can be designed around component parts with the intention of measuring and learning from the parts rather than the items. This suggests that instead of trying to decompose existing items into parts, it will be more fruitful to construct items from parts. Chapter 18 also demonstrates how failure of parts to predict the difficulty of wholes stimulates further investigation into the nature of the variable.

At first glance, Chapter 22, "...Selection Methods for Optimal Test Design" (M. Berger, W. Veerkamp) appears anachronistic in an age of computer-adaptive testing and performance assessment. But the test designs on p. 440 have a marked similarity to judging plans. Paper-and-pencil tests could be expensive, but performance assessment is far more so. Optimal large-scale, minimum-cost *judging* plans are now demanded by education administrators. Can the techniques of this and other Chapters provide these plans?

Since Ben Wright first formulated Chapter 13, "Composition Analysis", I have been struck by how often we frail humans use the wrong approach to solve problems. As p. 250 illustrates, when a problem is hard for a group to solve, the group should resort to "pack" work - everyone trying to come up with a solution independently. Instead, in difficult times the theme is always "unity" - we walk in lock-step in a futile attempt to preserve what we have. On the other hand, when a problem is easy, then "team" work, consensus, is most effective. Instead, we say "That's easy! I don't need anyone else's advice." - and proceed to blunder. This chapter illustrates how, as we understand measurement, we understand ourselves.

Vol. 3 is an excellent example of pack work!

John Michael Linacre

Van Duijn M.A.J., Jansen M.G.H. (1995) Modeling repeated count data: some extensions of the Rasch Poisson Counts model. JEBS 20:3, 241-258.

Journal of Outcome Measurement
Contributor Information

JOM publishes scholarly work from all academic disciplines relating to outcome measurement, the measurement of the result of any intervention designed to alter the physical or mental state of an individual. *JOM* will consider both theoretical and applied articles on measurement models, scale development, applications and demonstrations. *JOM*'s multi-disciplinary activity is directed by two broad-base editorial boards, one in the Health Sciences, the other in Social Sciences.

Manuscript Submission: Authors are responsible for all statements made in their work and for obtaining copyright permission to reprint tables, figures or quotations of 500 words or more. Copies of permissions and credit lines must be submitted. Prepare manuscripts according to the *Publications Manual of the American Psychological Association* (4th ed., 1994). Limit manuscripts to 25 pages of text, exclusive of tables and figures. Double space manuscripts, including title page, abstract, text, quotes, acknowledgements, references, and appendices. List author name(s), affiliation(s), address(es), telephone number(s), and e-mail address(es) on a cover page. Supply a 100-150 word abstract on the second page. Place each table on a separate page. Include photocopies of all figures. Number all pages consecutively. State, in a cover letter, that the manuscript contains original material, not previously published and not under review elsewhere.

Submit four manuscript copies. Prepare three of these for peer review by removing all references to author(s) and institution(s). Manuscripts are peer-reviewed anonymously by two experts appropriate to topic. Review takes three months. When manuscripts are accepted, authors submit a final printed copy of the manuscript, camera-ready figures, a disk copy in WordPerfect format on a 3½" MS-DOS disk, and sign a copyright-transfer agreement. Manuscripts are copy-edited and composed into page proofs, which authors may review before publication.

Submit manuscripts to:

Richard M. Smith
Editor, *Journal of Outcome Measurement*
Rehabilitation Foundation, Inc.
P.O. Box 675
Wheaton IL 60189

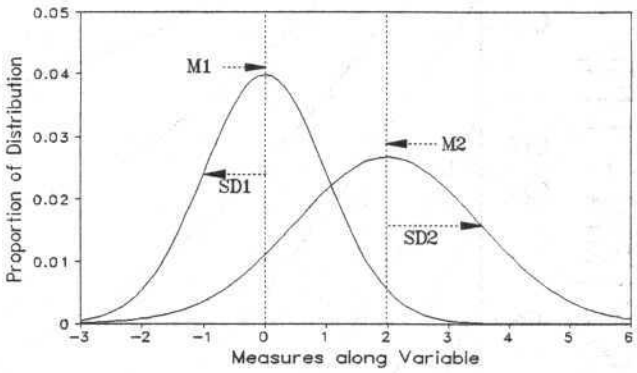
Overlapping Normal Distributions

Are men taller than women? Are women more patient than men? We answer such questions by comparing putatively normal distributions. A typical analysis computes a *t*-test of differences between sample means. When samples are large, any difference between means will be declared "statistically significant". Before drawing substantive conclusions from a "significant" result, however, it is wise to discover how much the distributions overlap, contradicting any finding of clear "difference".

Consider two normal distributions, N_1 and N_2, with the same sample sizes but different means (M_1 and M_2) and standard deviations (SD_1 and SD_2). Number the samples so that N_1 has the smaller standard deviation, SD_1 (see Figure). To discover the expected percentage by which each sample distribution overlaps the other, consult the nomogram overleaf. For the x-coordinate, compute a standardized absolute distance between the means, $|M_2-M_1|/SD_1$. For the y-coordinate, compute the ratio of the standard deviations, SD_2/SD_1. Interpolate by eye between contours to estimate the percent, p%, of each distribution that is in common with the other. The unique amount of each distribution is then 100-p%.

If $M_1=2.4$, $SD_1=1.6$, and $M_2=3.2$, $SD_2=2.0$, then $|M_2-M_1|/SD_1=0.5$, $SD_2/SD_1=1.25$, and, by reference to the nomogram, about 80% of each distribution overlaps the other.

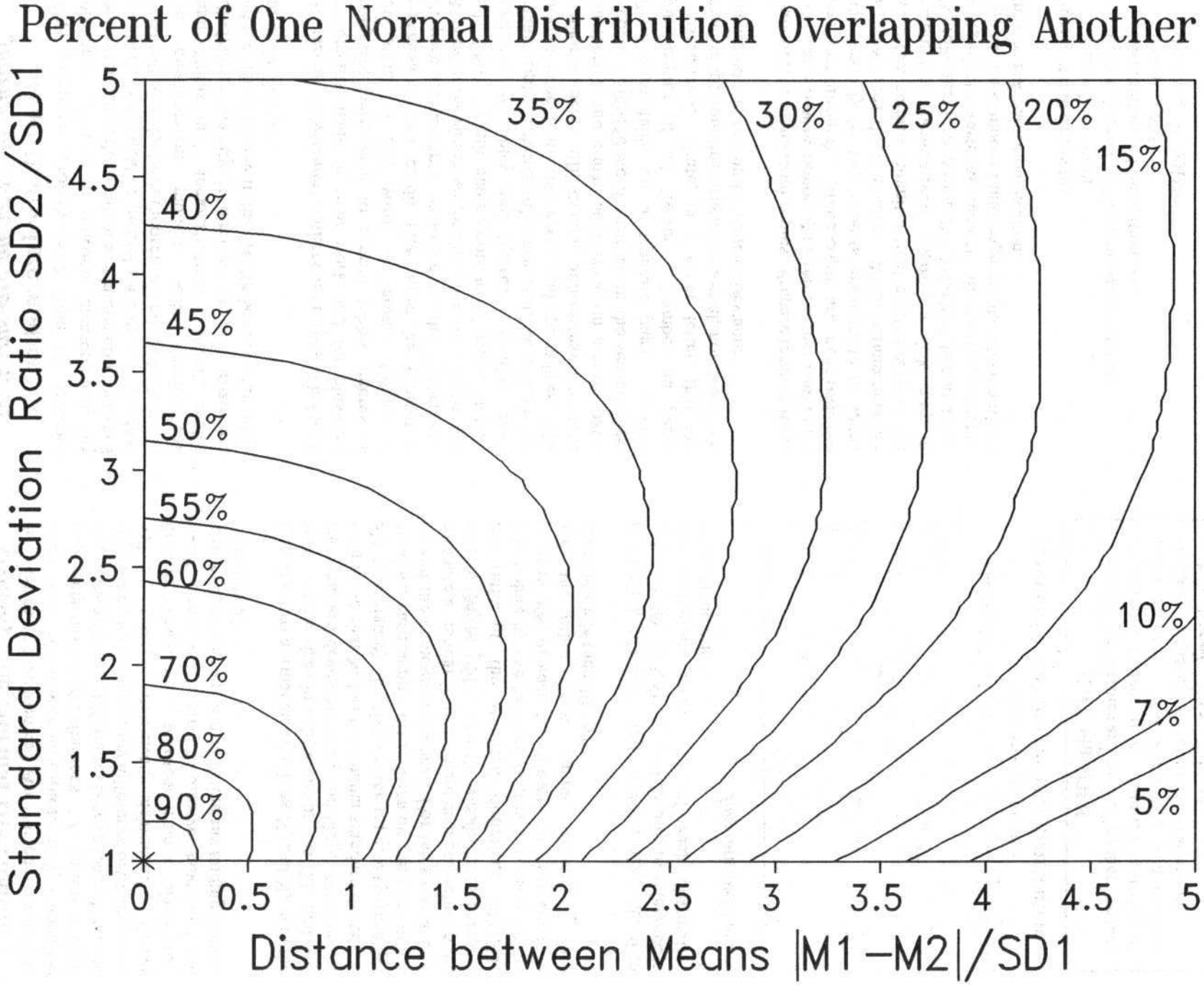
John Michael Linacre



Comparison of two normal distributions

RMT via *WWW*

The *RMT* back-issues are on the "gopher" system at Catholic University of America, courtesy of Larry Rudner. The world-wide web browser command is *http://www.cua.edu/www/eric_ae/* – then click on "Essays" button.


Percent of One Normal Distribution Overlapping Another
Standard Deviation Ratio SD2/SD1
Distance between Means |M1−M2|/SD1
35%
30%
25%
20%
15%
10%
7%
5%
40%
45%
50%
55%
60%
70%
80%
90%
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5