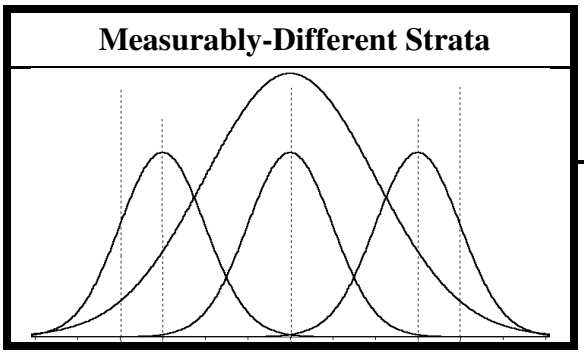


RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

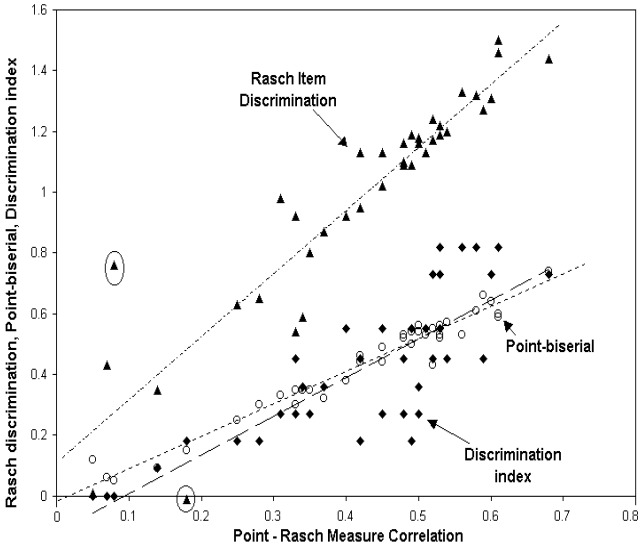
Vol. 16 No. 3 Winter 2002 ISSN 1051-0796

Item Discrimination Indices

Item discrimination indicates the extent to which success on an item corresponds to success on the whole test. Since all items in a test are intended to cooperate to generate an overall test score, any item with negative or zero discrimination undermines the test. Positive item discrimination is generally productive, unless it is so high that the item merely repeats the information provided by other items on the test. This is the "attenuation paradox."

The *Discrimination Index* (D) is computed from equal-sized high and low scoring groups on the test. Subtract the number of successes by the low group on the item from the number of successes by the high group, and divide this difference by the size of a group. The range of this index is +1 to –1. Using Truman Kelley's "27% of sample" group size, values of 0.4 and above are regarded as high and less than 0.2 as low by Ebel (1954, "Procedures...", *Educational and Psychological Measurement, 14,* 352-364).

The *Point-biserial Correlation* is the Pearson correlation between responses to a particular item and scores on the total test (with or without that item). The *Biserial Correlation* models the responses to the item to represent stratification of a normal distribution and computes the correlation accordingly. Again the ranges are +1 to –1. The biserial is always more extreme than the point-biserial. Jm Nunnally (*Psychometric Theory*, 1967, p. 123) states that "to use the biserial is to paint a faulty picture of the actual size of the correlations obtainable from existing data." A convenient substitute for these correlations, particularly when data are missing, is the correlation between the Rasch person measures and their responses to the item, the *point-measure correlation*.



The 2-PL model parameterizes item discrimination in the model and uses it to estimate person ability. A 2-PL model can be written:

$$\log\left(\frac{P_{ni}}{1-P_{ni}}\right) \equiv a_i(\theta_n - b_i) \quad (1)$$

At its core, the estimation process is:

$$\sum_i \hat{a}_i \, X_{ni} \Rightarrow \hat{\theta}_n \quad (2)$$

$$\sum_n \hat{\theta}_n \, X_{ni} \Rightarrow \hat{a}_i \quad (3)$$

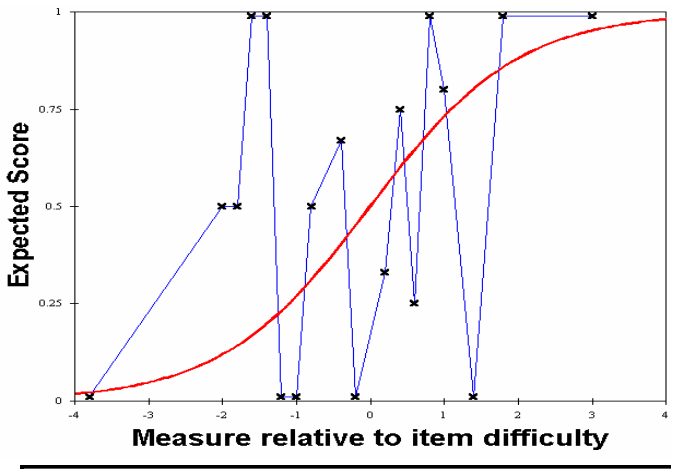
Unconstrained, this produces a feed-back loop. In (2), success on highly discriminating items (X_{ni}=1) raises the person measure, failure (X_{ni}=0) lowers the person measure. In (3), success on an item by those with high measures, coupled with failure by those with low measures, raises the item discrimination. This raised discrimination then feeds back into (2) to increase the measure difference between the successful and unsuccessful, which then, in (3), increases the item discrimination, *ad infinitum.* To avoid this, 2-PL estimation programs introduce constraints such as a maximum limit on item discrimination estimates, and a pre-

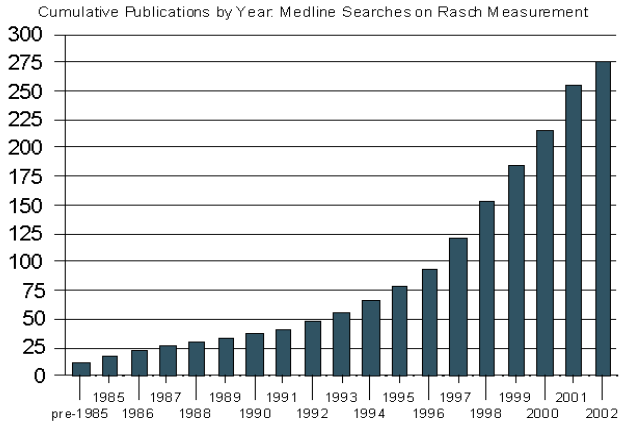
Table of Contents

set person distribution. Rasch models have pre-set item discriminations, so feedback does not occur. Person and item measures can be estimated in (2) because a_i is set to 1 at this step. Then those measures can be used in (3) to estimate item discriminations. There is no return to (2).

The plot shows the relationship between these indices. It reports item discrimination indices for dichotomous data reported in W J. Micheels and M. R. Karnes (1950, *Measuring Educational Achievement,* p. 478-9). For these data, the biserial correlations sometimes exceed 1, so that index is contra-indicated. The Discrimination Index (D) has been computed with the top 27% of the person sample in the high group and the bottom 27% in the low group. The trendlines (.... point-biserial, - - - discrimination index) show that all indices give similar information. The item ringed on left side of the plot has a low correlation but high Rasch item discrimination. It is an easy item with a few misfitting incorrect responses. The item ringed in the bottom left of the plot has negative Rasch discrimination. Its model and empirical ICCs are shown here. These results suggest that the point-measure or, for complete data, the point-biserial correlation capture the useful item discrimination information.

John M. Linacre





The growth of Rasch publications in medical literature. Its logistic shape (predicted by Derek de Solla Price) suggests a ceiling of 500 articles per year, at which point Rasch will be regarded as routine. *Courtesy of William Fisher.*

Midwestern Objective Measurement Seminar

Sponsored by UIC and
the Institute for Objective Measurement
Friday, December 13, 2002
University of Illinois at Chicago

Assessing change in depression and suicidal ideation in Mexican-American school children following a problem solving intervention. *Sarah Ailey, Julia Cowell, Diane McNaughton, and Louis Fogg, Rush University College of Nursing*

Testing of the URICA in a PTSD treatment group. *Cynthia W. Kelly and Mohamed Aziz, Northern Kentucky University*

Using the Rasch Model for Assigning Course Grades. *Robert J. Belloto Jr., Nevada College of Pharmacy*

Psychometric properties of the Knox's Cube Test – Revised. *Mark H. Stone, Adler Institute*

Examination of Rater Behavior in Grading Histology Practicals. *Johnna Gueorguieva, American Society of Clinical Pathologists*

Impact of item redundancy on Rasch estimates. *Everett V. Smith Jr., University of Illinois at Chicago*

Analysis of Illinois State Board of Education ISAT data: Investigating the impact of differential item functioning. *Lidia Dobria, University of Illinois at Chicago*

NO item misfit but ZERO person separation. *Rita K. Bode, Rehabilitation Institute of Chicago*

Cross-cultural and validation of the Functional Assessment of Cancer Therapy for patients receiving biological response modifiers (FACT-BRM). *Stacie A. Hudgens, Elizabeth A. Hahn, Alastair Glendenning, Ari Gnanasakthy, Center on Outcomes, Research and Education, Evanston Northwestern Healthcare*

Running Winsteps 10 Times Faster: The use of SAS ® (or SPSS ®). *Kazuaki Uekawa, The University of Chicago*

A Financial Disability and Victimization Questionnaire: Creating a new measure. *Ken Conrad, University of Illinois at Chicago*

Rasch analysis of Firefighters EMS Tasks. *Karen Conrad, University of Illinois at Chicago*

FESTSCHRIFT in honor of Ben Wright

25-27 April 2003 (weekend after AERA)

Rehabilitation Institute of Chicago

We, a committee organizing a *Festschrift* in honor of Ben Wright, invite presentation proposals addressing some aspect of the theme: **"Access, Provocation, and the Development of Professional Identity: Celebrating the Careers of Benjamin D. Wright."** Though the choice of the specific topics addressed is for you to make, we hope that you will take up an issue that involves or builds on Ben's extensive contributions to making measurement more accessible and to the fundamental foundations of measurement, his reputation as an irascible provocateur, his selfless support for others' professional development, and/or his multiple careers, as explained below.

We will provide a forum at the Rehabilitation Institute of Chicago, the weekend of Friday through Sunday, April 25-7 (immediately after AERA), in support of 1) platform presentations; 2) poster presentations, roundtables, and "artifact" displays; 3) software demonstrations (Friday afternoon, 25 April), and 4) a social event (Saturday evening, 26 April). The conference will close by early afternoon on Sunday, 27 April.

Presentation abstracts of 500-1,000 words should be submitted via e-mail before February 15, 2003 to Mark Wilson at *mrwilson@socrates.berkeley.edu* before February 15, 2003. Abstracts should take up one or more of the following themes from either a historical or a state-of-the-art perspective: Access to Measurement (including data applications in any field), Foundations of Measurement, Provocation of and Development of Professional Identity, Multiple Careers. More detail on these possibilities is provided below.

All presentations will be eligible for publication in the conference proceedings, to be edited by Mark Wilson and George Engelhard. If you want your work considered for the book, please indicate that you plan to submit a paper at the conference, and provide three copies of the paper to Mark Wilson or George Engelhard. at the conference.

Access to Measurement: simpler, faster estimation (PROX, UCON); software that works; models for more kinds of data; error, reliability, and fit statistic development; applications to tests, surveys, and assessments in dozens of fields; publishing (MESA Press, RMT, support for OM:TiP, JOM, JAM, PM); associations (the SIG, IOM); meetings (MOMS, AERA/SIG, IOMW); and constant improvement to all of that via substantive interactions with students and colleagues.

Foundations of Measurement: measurement as a scientific enterprise, relation to scientific revolutions, relation to foundational ideas such as specific objectivity and additive conjoint measurement, relation to foundational work of figures such as Thurstone, Guttman and Rasch.

Provocation and Development of Professional Identity: Ben is well-known for strongly challenging and even abruptly dismissing anything that strikes him as irrelevant, foolish, or half-baked, and he seems to have had explicit reasons for behaving in this manner, reasons stemming from his work on identity development with Bruno Bettelheim. Personal accounts of Ben's successes and failures in this regard are of particular interest.

Multiple Careers: In addition to his work in measurement theory and practice, Ben worked as a physicist, and then as a psychologist and factor analyst. He taught a course on the psychology of becoming a teacher for many years, and continued working in this area long after most people associated him primarily with Rasch measurement. Even within the area of measurement alone, Ben's early work on estimation, models, fit, error, reliability, and software stands in considerable contrast with his later emphases on applications, organizations, and publishing. Papers touching on more than one of these careers will be of special interest.

See you in April in Chicago!

William Fisher (chair), David Andrich, Kendon Conrad, George Engelhard, Allen Heinemann, Mary Lunz, Geoff Masters, Alan Tennant, Ev Smith, Mark Wilson

April 2003, Chicago

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

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

April 25-27, Friday-Sunday
Ben Wright *Festschrift*

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

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

Bettelheim's Test Revisited

In an autobiographical account, Ben Wright (RMT 2(3):25-32) elaborates on how his early "career led to an identity confusion." After pursuing a PhD in physics doing almost nothing but measuring, Ben decided to seek out something livelier, more human. He explored possibilities in English and history, but wound up in the 1950s in psychology, and consulted doing factor analyses for Chicago marketing firms. In Ben's account, the contrast between the stable, interpretable results of measurement in physics and the unstable, uninterpretable results of factor analysis made him feel "like a crook." After some time in this awkward position, he met Georg Rasch, and joked that he could then "stop going to the psychoanalyst to have [his] schizophrenia mended week by week."

Ben does not mention them in his account, but his work in psychology included considerable time with Bruno Bettelheim. Their two co-authored publications focus on extending the lesson of identity development learned from autistic children into the domain of professional identity development. These publications, his book *Hero, Villain, Saint* (Wright & Yonke, Peter Lang Publishing, 1989), and the course on the *psychology of becoming a teacher*, that Ben taught for many years, explore the ways in which professionals emerge as independent thinkers and actors from a process that includes a decisive break with a key mentor.

In Ben's (RMT 2(3): 27) own account, he made an ineffectual step in this direction in 1964 when he contradicted Rasch by incorporating an item discrimination parameter into software he was writing with Bruce Choppin. He made a cleaner break a few years later with the development of the UCON estimation algorithm, which Rasch also opposed but which retained a connection with parameter separation and sufficient statistics in a way that the earlier 2p program did not. In Wright's own words, his UCON work

> "was an important point in our [Ben and Rasch's] relationship because at that moment he and I separated a little bit. Up until then, as far as he was concerned, I was doing everything exactly the way he told me. But UCON was a new something that I did on my own, not to his liking, which seemed to me plainly convenient, practical and useful. So it was a point in our work where I was becoming myself, in spite of, indeed, against his wishes. We continued to be good friends. But from that summer of 1967, there was that bit of difference between us." *(1981 Interview of Ben Wright by David Andrich, www.rasch.org/rmt/rmt0.htm)*

By taking this step, Ben took responsibility for advancing his own ideas and innovations in a direction not specifically foreseen or supported by his teacher. As Ben already well knew from his work with Bettelheim, this meant he had completed a significant stage in the development of his own identity as a professional. His explicit awareness of the importance of this step raises the question as to whether he might have tried deliberately to provoke others into taking it.

Something that has rarely, if ever, been appreciated about Ben is his way of alternating between, on the one hand, improved access to measurement and, on the other, provocations to measure better and think more clearly. Rasch's models abstract information about individuals, but also integrate that information with that of the populations to which they belong. Similarly, Wright simultaneously supported the professional development of both individuals and populations by making measurement more accessible, and by provoking others into overtly testing and asserting the validity of their own measurement innovations and contributions.

In my own case, for instance, I had the great fortune of discovering on my first day in Ben's classroom concepts and tools that I had previously thought I was going to have to invent. But this revelation of open access to what I recognized to be of great value was soon (within 2 or 3 weeks) countered by Ben's flat dismissal of my approach to the language of measurement theory. That really made me mad, so I wrote an impassioned paper explaining my position, and Ben warmly embraced my point of view, adding with it respect for pushing back at him in an assertion of my independent identity.

A question we need to raise is how we as individuals and as a field are now to respond to the access and provocations of Ben's work. How have others historically risen to the challenge of Ben's one-two punch? How have some failed to rise to the task, or even failed to recognize that there was one? And, with the fairly recent realization of the vital role in measurement played by metrological networks of instruments traceable to reference standard metrics, we can now also articulate the question as to the extent to which Ben's combination of access and provocation reaches beyond the development of individual professional identities to the development of professions' identities.

After all, to what extent is psychology, sociology, or any other *-ology* actually fulfilling its mission as an effective

Rasch Measurement Transactions
P.O. Box 811322, Chicago IL 60681-1322
Tel. & FAX (312) 264-2352
rmt@rasch.org www.rasch.org/rmt/
Editor: John Michael Linacre

SIG Chair: Trevor Bond SIG Secretary: Ed Wolfe

manner of expressing a particular field of meanings if its *logos* remains blatantly dependent on the particular persons and phrasings of the questions and answers embodying the conversation? In other words, to what extent does a field of study actually have a professional identity if its objects and subjects are not clearly expressed and distinct from those of other fields? There are many expressions of the opinion that fields of study are as scientific as they are mathematical, but mathematical means quantitative far less than it implies a rigorous independence of figure (numeric, geometric, metaphoric, dramatic) from meaning. Rasch's separability theorem provides the basis for tests of that independence, and thereby becomes the basis for the development of professions' identities.

Does not Ben's work amount to a repetition and extension of Socrates' similarly simultaneously enacted roles of midwife and gadfly? And in the same way that, first, harmonic and geometric studies, and later, the modern sciences, emerged from Socrates' tests of ideas as hypotheses, so, too, today we are witnessing the conception and birth of new forms of understanding relevant to mathematical structures accessible in large part to the inspiration and perspiration of Benjamin Drake Wright.

A conference scheduled for April 25-27, 2003, at the Rehabilitation Institute of Chicago, immediately after AERA, will present further elaborations on these themes. See the call for presentation proposals elsewhere in this issue of RMT for more information.

William Fisher

Reviewers Reviewed ....

"I would suggest rethinking your reliance on Rasch fit statistics as a criterion for item rejection .... In many cases, the best (most highly discriminating) items would be rejected if one relied on Winsteps' 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 [correctly] 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. ☸

"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 (2002) Scientific Culture and Educational Research. *Educational Researcher, 31, 8*, 11

2nd International Conference

on Measurement in Health, Education, Psychology and Marketing: Developments with Rasch models

Perth and Fremantle, Western Australia

January 20 –22, 2004

Exciting developments in the theory and practice of measurement in health, education, psychology and marketing provide an opportunity to review the state of the art in measurement science, learn from the experts in an extensive pre-conference program, and enjoy the delights of summer in Western Australia.

Topics for the conference:

- Epistemology, fundamental measurement and Rasch models
- Cumulative models for attitude and trait measurement – dichotomous and ordered category models.
- Unfolding models for preference and choice – folding the Rasch models
- Rasch model applications in education (e.g., large scale test equating, benchmarking)
- Applications in psychology (e.g., intelligence testing, linking quantitative and stage developmental data)
- Applications in marketing (e.g., pairwise designs for preference and choice studies)
- Applications in health care (e.g., cross-cultural validity)
- Item banking
- Computer adaptive testing
- Using simulation studies for clarifying methodological issues (e.g., tests of fit)
- Developments in Rasch modeling (e.g., differential item functioning)
- Understanding response processes compatible with the Rasch models
- History and philosophy of measurement and Rasch models

Abstracts are invited by July 31, 2003. Further information regarding the conference (Scientific Committee, Web Site, Registration, Costs, etc, and courses), will be made available. If you wish to be put on the mailing list for this information, please send a message to Angelina Chillino, email ***chillino@murdoch.edu.au*** and make the subject: *RaschSummer2004*.

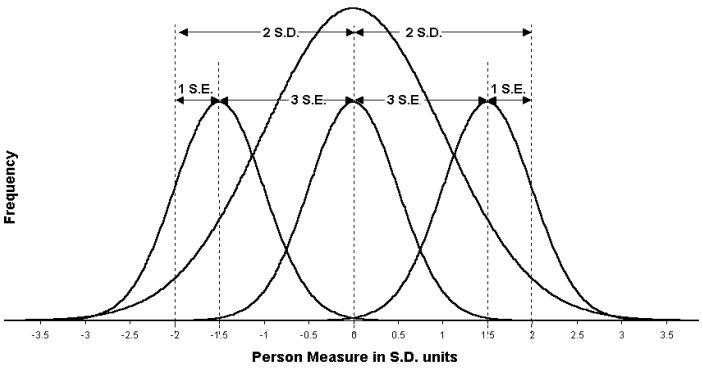
Number of Person or Item Strata

Wright and Masters (*Rating Scale Analysis,* 1982, pp. 92, 106) write: "... if we define statistically distinct levels of item difficulty as **difficulty strata** with centers three calibration errors apart, then this separation index G can be translated into the **number of item strata** defined by the test H=..." (emphasis mine) and similarly for persons.

$$H = (4G + 1)/3$$

$$\text{where } \mathbf{G} = \frac{\text{"True" standard deviation}}{\text{Average measurement error}}$$

The "average measurement error" is the statistical average, i.e., the root-mean-square, of the standard errors of the measures of items of persons. The "True" standard deviation of the item or person measures is obtained from: "True" standard deviation^2 = Observed standard deviation^2 - average measurement error^2.



What does this mean? The plot shows the relationship between sample distribution and measurement error when H=3, i.e., when 3 strata are observed. This is when G=2, so that the "true" sample standard deviation, S.D., is twice the average measurement error, S.E. We see that the relevant range of the "true" distribution (assumed normal) is two standard deviations away from the mean, encompassing over 95% of the distribution.

The strata are defined as statistically distinct measures. They are located 3 S.E. apart, because this is conveniently more than 1.96 * sqrt (2) = 2.77 S.E., the distance corresponding to .05 significance. The centers of the extreme strata are also positioned 1 S.E. within the boundaries of the sample distribution. Thus 84% of even the most extreme strata lie within the 4 S.D. range.

G itself is a more conservative "Separation Index" than H. For instance, suppose that the "true" standard deviation of a sample is the same as the average measurement error. Then G=1, and the test reliability is 0.5, warning us that we don't know whether observed differences within the sample are real differences or merely measurement error . H is (4+1)/3, i.e., roughly 2. This indicates that the opposite ends of the "true" distribution are measurably different, implying that, if the observed measures are sufficiently far apart, they probably reflect real differences. *JML*

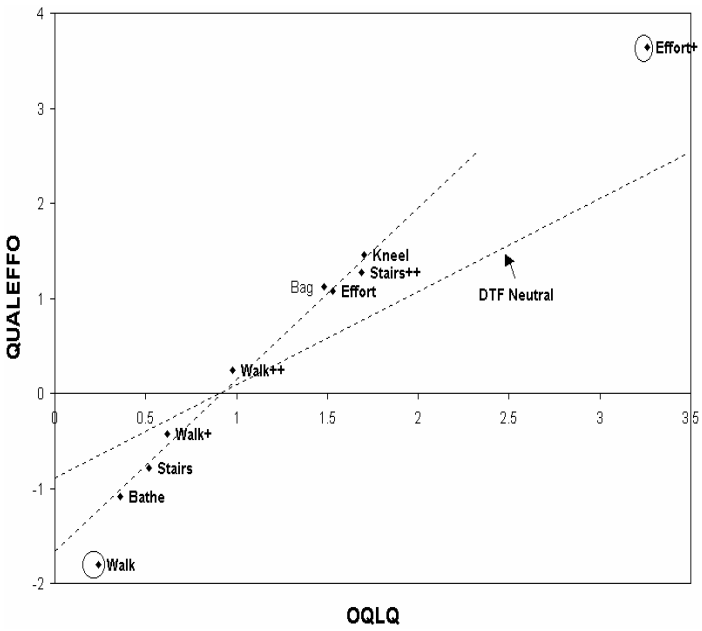
Differential Item and Test Functioning (DIF & DTF)

Do test items function in different ways for different groups of test-takers? Item functioning is intended to be invariant with respect to irrelevant aspects of the test-takers, such as gender, ethnicity and socio-economic status. But item functioning is expected to be altered by interventions targeted at those items, for instance, the use of calculators in arithmetic tests or the use of assistive devices on mobility items.

Differential Item Functioning (DIF) investigates the items in a test, one at a time, for signs of interactions with sample characteristics. In the widely used Mantel-Haenszel procedure (1959, *www.rasch.org/memo39.htm),* reference and focal groups are identified which differ in a discernible way. These groups are stratified into matching ability levels and their relative performance on each item is quantified. The ability levels are usually determined by the total scores on the test. In this way, the DIF analysis for one item is as independent as possible of the DIF analyses of the other items. But a consequence is that the overall impact of item DIF, accumulated across the whole test, is unclear.

Differential Test Functioning (DTF) compares the functioning of sets of items. Wright and Stone (1979, p. 93) compare the difficulty measures of 14 items obtained from two separate analyses. This technique has been extended to separate analyses of the test responses by reference and focal groups. The effect of separate analyses is that two separate item hierarchies are defined, and the measures of the two groups are obtained in the context of their own hierarchies.

Badia, Prieto et al. (2002) provide a good example of Differential Test Functioning. Two instruments, OQLQ and QUALEFFO, were assigned to randomly equivalent groups of patients. Then the difficulty measures of the common items (originally from the SF-36) were cross-plotted. The Figure shows the results (with abbreviated item labels).



What has happened? The overall item difficulty order has been maintained, but the relationship between the difficulties is QUALEFFO = -1.7 + 1.8 OQLQ. There are also two somewhat off-diagonal items. "Walk" and "Effort+", which are easier for the QUALEFFO sample.

An overall difficulty shift is expected. The independent analyses of the two instruments result in each having its own zero-difficulty point set at the mean difficulty of its own items. The common items are the harder items on the OQLQ, but span the difficulty range on the QUALEFFO. This produces the difficulty shift of 1.7 logits.

The slope is explained in a different way. The paper reports the mean-square fit statistics for the common items. Their mean-squares on the QUALEFFO distribute around the expected 1.0, with an average of 1.03, but their mean-squares on the OQLQ are in the range 0.5 to 1.0, with an average of 0.65. This implies that other items in the OQLQ, such as "cut nails", "care for plants", "buy clothes", are less predictable, and so force the more general SF-36 items to overfit. If the common items were calibrated independently of the other items in the OQLQ, their own logit range would be approximately
1/(average mean-square) = 1 / 0.65 = 1.5 times wider
(www.rasch.org/rmt/rmt142n.htm). This would roughly match the 1.8 times wider observed in the QUALEFFO.

The two circled off-diagonal item, "Walk" and "Effort+" are relatively easier for the QUALEFFO sample, This may be explained by sample differences. Despite the intention of having randomly equivalent samples, the paper's demographic Table reports the QUALEFFO sample to have better general health, more vitality and better physical functioning than the OQLQ. Even though each of these items has a standard error on each instrument, it is not possible to make precise DIF tests because of the changes of scale and uncontrolled interactions with items unique to each instrument.

Further investigation of "Walk" and "Effort+" requires a definitive examination of Differential Item Functioning across the two samples. A joint analysis of only the common items would remove the distorting effects of the items unique to each instrument. Further, the analyses would then share the same logit metric. The interactions between sample differences and item difficulties could then be precisely determined. *JML*

Badia X, Prieto L, Roset M, Díez-Pérez A, Herdman M (2002) Development of a short osteoporosis quality of life questionnaire by equating items from two existing instruments. *Journal of Clinical Epidemiology, 55,* 32–40.

"Linear" Rescaling vs. Linear Measurement

"Rescaling Ordinal Data to Interval Data" (Harwell MR & Gatti GG, *Review of Educational Research, 2001, 71:1,* 105-31) is honest, but misleading. The paper itself comprises two parts. First, a useful survey of the prevalence of, and problems with, using ordinal data in quantitative research. Second, two examples using IRT to rescale ordinal data, one an analysis of a real dichotomous dataset using BILOG, the other an analysis of simulated Graded Response data using MULTILOG.

The paper begins

> "Many statistical procedures used in educational research are described as requiring that dependent variables follow a **normal distribution, implying an interval scale of measurement.** The advantage of an interval scale is that relative differences among values composing the scale are **assumed** to be equal in terms of what is measured, allowing arithmetic operations (e.g., addition, multiplication) to be used unambiguously" (Emphasis mine).

Certainly, normality requires linearity. A normal distribution only makes analytical sense if it is based on an underlying linear frame of reference. The paper helpfully provides supporting references to this (Guilford, 1954, p. 17; Gaito, 1959; Lord & Novick, 1968, p. 22). But the paper leaves the mistaken impression that *linearity implies normality.* Normality may be hypothesized to exist. But linearity itself is independent of any particular sample distribution.

A second misconception follows. Linearity is not a property that can be safely "assumed". No physicist, carpenter or cook would be so foolhardy as to merely "assume" the linearity of a measuring instrument. Usually there is evidence that a manufacturer has taken pains to construct linearity. Then the instrument must be used in such a way as to maintain its linearity. If linearity is in doubt, as when an instrument is damaged or of unknown provenance, its linearity is checked before it is used. Thus a linear scale must be constructed and then it can be tested. A further complication is that a scale that is linear for one purpose, e.g., time as expressing duration, may be non-linear for another, e.g., time as expressing running or swimming prowess.

Most IRT models concur with the implication that "normality implies linearity". The sample is assumed, or rather asserted, to have a normal distribution. This assertion is then imposed on the analysis, and the resulting scale scores are declared to be "linear". The paper honestly admits the difficulty of demonstrating that such scale scores are, in fact, linear.

> "Clearly, additional work is needed to demonstrate that the estimated proficiencies for a variety of IRT models and item types show an interval scale. One option is to follow Fischer's (1995) approach in which proficiencies under the Rasch model were proved to possess an interval scale . This is the most attractive approach, but such proofs are difficult beyond the case of the Rasch model for dichotomous responses. Alternatively, computer simulation studies could be performed ...." (p.127).

There are proofs for linear scaling with polytomous and other Rasch models (Andrich, 1977; Fischer, 1995; Linacre, 1989). A basic property of all Rasch models is separability of parameters, which is manifested statistically by each parameter having a sufficient statistic. From this basis, linearity can be constructed. But there are no proofs of linearity for non-Rasch IRT models, i.e., those without separability of parameters. And no amount of computer simulation will "turn a sow's ear into a silk purse!"

Harwell & Gatti's BILOG "Rasch" Example

The paper's idiosyncratic analysis of a real dichotomous dataset prompts a comment. 1,000 4th-grade students responded to 30 dichotomous items. Our authors must be congratulated for choosing to perform a Rasch analysis, even if their motivation lacks conviction: "we had no reason to believe that the items varied in discrimination or that guessing needed to be modeled." Thus BILOG was instructed to perform a "Rasch" analysis.

The reported results make most sense when interpreted with a local scaling of 1 BILOG unit = 0.7 logits. But the paper's Figure 1 (reproduced here) shows a score range (4-20) that fails to include all of those in its Table 3 (6-

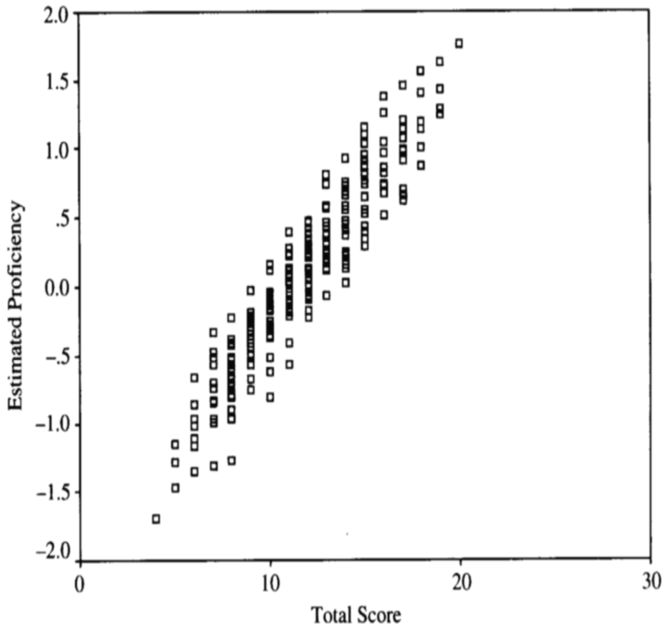


FIGURE 1. *Plot of total scores and estimated proficiencies for the Rasch model.*

27). Figure 1 implies that most item p-values are less than 0.5, but its Table 2 reports that 28 out of 30 p-values exceed 0.5. Further, Figure 1 and its accompanying text explain how different response patterns, for the same raw score, yield different person measures. This accords with IRT scaling philosophy, but contradicts a basic tenet of Rasch measurement –

> "we may conclude that *as far as the model goes [measures] should be estimated from the marginals ... only, while any further details about the structure of [the response matrix] is irrelevant for estimation – but of course not for controlling the model."* (Rasch, 1980, p. 177. Italics his.)

Ben Wright (1977) remarked that "Progress marches on the invention of simple ways to handle complicated situations." As it stands, this paper makes the linearization of ordinal data, a complex but manageable problem, unintelligible.

JML

Andrich D (1977) Summary Equations on Notes for a Rasch Model for Likert Scales. Paper presented at AERA. *www.rasch.org/memos/memo48.htm*

Fischer G.H. (1995) Derivations of the Rasch model. & The derivation of polytomous Rasch models. In G.H. Fischer and I.W. Molenaar (Eds.) Rasch Models. New York: Spring-Verlag.

Gaito J (1959) Non-parametric methods in psychological research. Psychological Reports, 5, 115-125.

Guilford JP (1954) Psychometric Methods. 2nd Ed. New York: McGraw-Hill.

Linacre JM (1989) Many-facet Rasch Measurement. Chicago: MESA Press.

Lord FM & Novick MR (1968) Statistical theories of mental test scores. Reading MA: Addison-Wesley.

Wright BD (1977) Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14(2),* 97-116.

Bond & Fox (2001) with Winsteps-Ministep

Chapters 2-7 of "Applying the Rasch Model" (Bond & Fox, 2001, Mahwah NJ: Lawrence Erlbaum Assoc., available from *www.rasch.org/books.htm)* include practical examples of the Rasch model. Step-by-step instructions for these analyses using Winsteps or the free Ministep software are at

www.winsteps.com/bondfox.htm.

Expected *A Posteriori* (EAP) Measures

Under Rasch model conditions, there is some probability that a person will succeed or fail on any item, no matter how easy or hard. This means that there is some probability that any person could produce any response string. Even the most able person could fail on every item.

The measure estimated for a person is usually that for which the observed response string is most likely, or that for which the response string best fits a Rasch model. We may, however, have some rough idea about a person's ability measure (or an item's difficulty) prior to the current data collection and wish to incorporate this idea into the newly estimated measure. To do this, we calibrate the test items in the usual way. Then we combine the item calibrations, our prior rough idea, and the observed responses to obtain an improved, *a posteriori*, person measure. Mislevy and Stocking (1989) recommend this approach for IRT models. John Uebersax (1993 and on his website) outlines a general procedure for this.

The technique capitalizes on an insight of Thomas Bayes:

Prior Probability x Data Probability ⇨
Posterior Probability

which implies that

Prob (B' given {X}) =
Prob (B') x Prob ({X} given B') /
Sum over all B [Prob (B) x Prob ({X} given B)]

where B' is a particular value of the person measure, and the sum is over all possible values of our rough idea, B. {X} is the person's response string. The EAP estimate of the person measure is the expected value of this:
EAP estimate = Sum over all B [B x Prob (B given {X})].

Thus, suppose that our rough idea, the prior distribution of B, φ(B), is a convenient distribution, such as N(μ,σ). The test consists i=1,L items. P_{Xni} is the probability of person *n* of ability B scoring X_{ni} on item *i*.

Then

$$EAP(B_n) = \frac{\int_{B=-\infty}^{\infty} B\,\varphi(B) \prod_{i=1}^{L} P_{X_{ni}}\, dB}{\int_{B=-\infty}^{\infty} \varphi(B) \prod_{i=1}^{L} P_{X_{ni}}\, dB}$$

This can be evaluated using numeric quadrature to approximate the integrals.

JML

Mislevy RJ & Stocking ML (1989) A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57-75.

Uebersax JS (1993) Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association, 88,* 421-427.

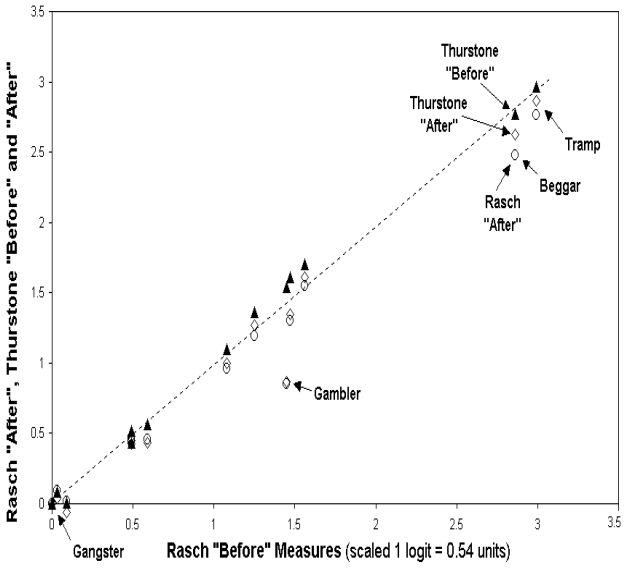
Thurstone Case 5, Rasch, DTF and DIF

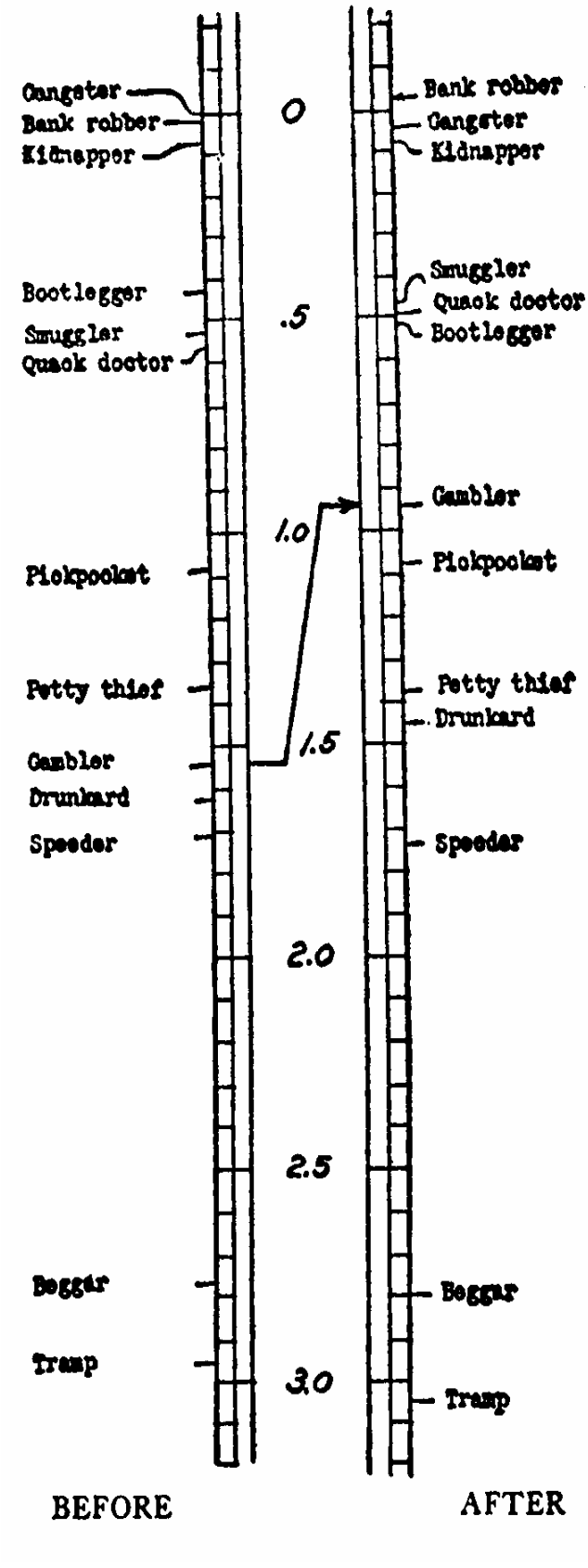
Differential Test Functioning (DTF) and Differential Item Functioning (DIF) were investigated by Louis L. Thurstone using Case 5 of his "Law of Comparative Judgment" which is based on "discriminal differences" *(Psychological Review, 1927, 34,* 273-286).

240 children in Mendota, Illinois, were each asked to compare pairs of 13 delinquent activities and indicate which should be punished more severely. About a week later, they were shown a movie about the life of a gambler and then instructed to do their comparisons again. The results are shown here from "Influence of motion pictures on children's attitudes" *(Journal of Social Psychology, 1931, 2,* 291-305).

Thurstone's sample exhibited DTF. The spread of scale values on the second occasion was .95 of the first. This was attributed to increased boredom or indifference on the second occasion. Accordingly, in the Figure, Thurstone inflated the scale values for the second occasion by 1.05. "Gambler" exhibits noticeable DIF. The movie about the gambling lifestyle affected attitudes, at least that day.

Thurstone published the data for his analysis. This was reanalyzed with a Bradley-Terry (Rasch) model using *Facets*. The plot below (with 1 logit = 0.54 Thurstone units) shows the collinearity of scale values. The placement of the black triangles above the dotted trend line indicates the slight non-linearity of Thurstone "Before" relative to Rasch "Before" values. The "After" values are not adjusted for DTF so their narrower range is evident. Thurstone and Rasch coincide in their placement of "Gambler" in the "After" analyses. Rasch standard errors are about .05 units, so a DIF study shows the shift in "Gambler" to be highly significant. "Gangster" and "Beggar" show noticeable misfit "Before". None misfit "After". *JML*





Thurstone, 1931, Fig. 20. Seriousness of delinquencies.