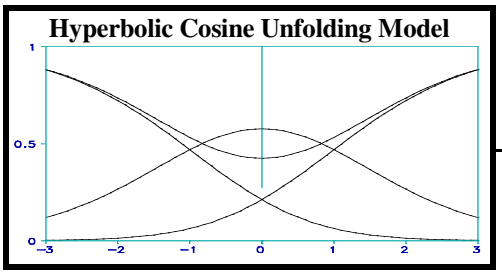


RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

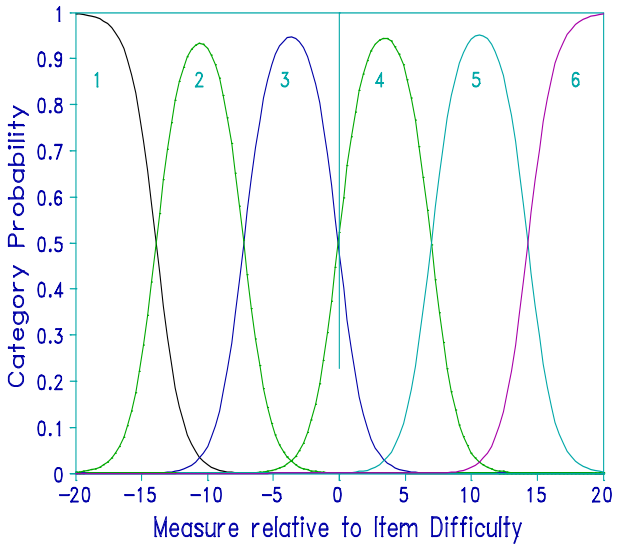
Vol. 16 No. 1 Summer 2002 ISSN 1051-0796

Judge Ratings with Forced Agreement

Many performance assessments have each piece of work rated by a pair of judges, supposedly rating independently. But a commonly applied rule is that, whenever the ratings awarded by the pair of judges differ by more than one category, that piece of work is rated by a third rater whose rating replaces that of the more discrepant of the original pair. Raters who are deemed discrepant too frequently are retrained and may be dismissed. The result is pressure on the judges to be "consistent", i.e., to conform to an imaginary consensus. The consequence of this pressure is a dataset in which the ratings of pairs of judges do not differ by more than one score-point for any piece of work. What are the measurement implications of this?

It is straightforward to construct a data matrix that accords with this intent. You can do it yourself. Imagine 7 pieces of work of increasing quality. These are the columns of the data matrix. Each is rated on a 1-6 rating scale. Each row of the data matrix is a judge, assigning ratings to each piece of work, but in such a way that the ratings of each piece of work (i.e., in each column) do not differ by more than one score-point. Your data matrix will look something like this:

1123456
1234566
2133456
1234455
1123456
1234566
1123456
1234566
1123456
1234566



A Rasch analysis reveals the measurement implications of this forced agreement. The Figure depicts the category probability curves for the rating scale. The category curves display very little overlap with curves other than their immediate neighbors. For my dataset, the range of the scale is around 40 logits. This accords with the ranges of over 30 logits sometimes reported for assessments using this type of judging procedure.

What has happened? The attempt to increase reliability by forcing judge agreement has not worked as intended. Reliability is an ordinal or even, in the case of Cohen's *Kappa,* a nominal index. If the two judges were perfectly reliable, they would be like machines, always producing identical ratings. So they would act as one judge. We have here a variant of the "attenuation paradox" of raw-score test theory, or of what the legal profession "wood-shedding".

Table of Contents

From the measurement perspective, each rating is expected to provide independent information about the location of the performance on the latent trait. It is the accumulation of that information, not the ratings themselves, that is decisive. Ratings which contradict the accumulated information certainly merit investigation, but are not automatically rejected. In the situation described here, the attempt to increase inter-rater reliability has actually reduced the independence of the judges, and so degraded the validity of the measures as measures.

John M. Linacre

MIDWEST OBJECTIVE MEASUREMENT SEMINAR

University of Illinois at Chicago, May 10, 2002

Housing quality: What does this mean for U.S. veterans
who live in an institutional setting of a psychiatric medical center?
Edward Clark, University of Illinois-Chicago

Psychometric evaluation of the Center for Epidemiological Studies - Depression (CES-D) Scale
in stroke patients
Mehul Dalal, University of Illinois-Chicago

Scaling indicators of frequency
Tom O'Neill, University of Illinois-Chicago

A comparison of the separation ratio and coefficient alpha in the creation of minimum item sets
Trudy Mallinson, Northwestern University

The direction and meaning of depression in adults with Down Syndrome
Sarah Ailey, University of Illinois-Chicago

Eliminating disconnected subsets in FACETS
Patrick Fisher, Measurement Research Associates, Inc.

Collapsing a rating scale: Meanings and implications
James Houston, Measurement Research Associates, Inc.

Evaluating student-teacher trust in the Chicago public school system
Lidia Dobria, University of Illinois-Chicago

Describing NAEP achievement levels with multiple domain scores
Matthew Schulz, ACT, Inc.

Measuring functional outcome one-year post severe brain injury
Theresa Louise-Bender, Northwestern University

Test structure and item parameters: The effect of basal and ceiling rules
Kirk Becker, Riverside Publishing, Inc.

Medical technologist satisfaction with organization/management
Johnna Gueorguieva, American Society for Clinical Pathologists

Measuring student engagement level in mathematics and science classrooms:
Using the Experience Sampling Method (ESM) and Rasch model analysis
Kazuaki Uekawa, The University of Chicago

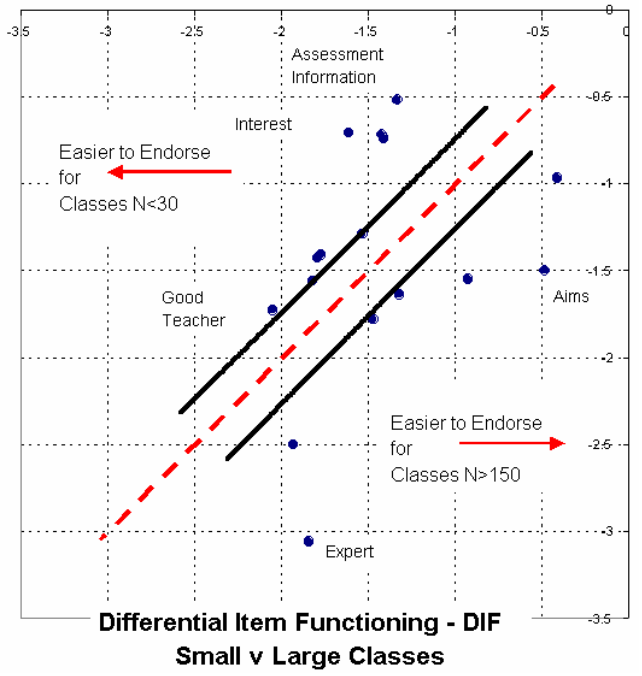
Does familiarity breed contempt?

Student evaluation of faculty teaching performance is fast becoming a high-stakes game, but few are interested in measuring attributes influencing good teaching - most settle for some sort of statistical interpretation of counts.

The SFT (Student Feedback about Teaching) scheme at James Cook University in Australia is based on Rasch modelled estimates of responses to a Likert-type scale. In order to help to address a perennial question about how difficult it is to teach large classes, responses for classes smaller than N=30 were pooled, as were responses for classes larger than N=150 to generate the DIF plot above. While about half of the items located close to identically for both groups (with suitable allowance for measurement error), it was easier for students in small classes to report satisfaction with the `Interest' their teachers showed and the `Assessment Information' they provided. While teachers of large classes were reported as doing better at achieving the subject's `Aims', the largest DIF effect was for the easiest to endorse item: the teacher in the large lecture theater is much more readily seen as being an `Expert' in the field. (Disclaimer of personal interest; the author teaches classes of 450 first year teacher education students about developmental psychology.)

Trevor Bond, Trevor.Bond@jcu.edu.au

From: Trevor G. Bond, "Accountability in the Academy: Rasch measurement of student surveys", Survey Research in Education SIG, AERA 2002.



Comparisons vs. Preferences

It is important to distinguish between "comparisons" and "preferences". The person makes a **comparison** as to which of the pair of stimuli has more of the property, e.g., comparing cups of coffee as to which is sweeter. This comparison should essentially not be a function of the person's liking for sugar in coffee. A sweeter cup should, in general, taste sweeter to everyone (and, when we have relative similar amounts of sugar, we get proportions rather than perfection in the numbers deciding one way or the other.)

In the case of **preferences,** the person parameter ("ideal point" in the language of Clyde Coombs) plays a central role. Taking the sweetness of coffee example again, the person is asked which of each pair of cups of coffee the person prefers as to sweetness, not which is sweeter (irrespective of the person's preference). In this case, the person will prefer the cup of coffee which is closest to the person's ideal amount of sweetness relative to those cups that are both less and more sweet.

It is very important to distinguish the **instructions** that are given to people and to consider which model is the most appropriate (i.e., has the correct properties). It is convenient to use the word "comparison" when the person's location **is not** supposed to be involved, and the word "preference" when the person's location **is** supposed to be involved.

This is confused in the literature. For instance, there is Luce's so-called *choice axiom* (1959). This essentially states that, when there are several alternatives available, the probability of the preferred option is independent of the sequence of decisions. When this axiom is expressed algebraically, no person parameter is specified. But preferences are obviously decision-maker dependent. The same flaw is evident in the algebraic formulation of the *Shepard-Luce choice rule*, which can be expressed: "Choice probability increases with strength of evidence that an object belongs to a category." These have added further to the confusion of terminology, models and response processes.

David Andrich, dandrich@murdoch.edu.au

Luce RD (1959) Individual Choice Behavior. New York: Wiley.

Shepard, Roger N. (1964), "On Subjectively Optimum Selection among Multi-Attribute Alternatives," in Human Judgments and Optimality, eds. M.W. Shelley and G. L. Bryan, N.Y.: John Wiley.

Fifth New England Objective Measurement Workshop
Boston, April 25, 2002

Welcome and Introduction to Principles of Rasch Measurement
Larry Ludlow (Boston College)

A Proposal for the Construction of a Rasch-based Summative Assessment to Measure Chemistry Achievement
James Cheng (Boston College)

Examining Performance Assessment with a Many-Facet Rasch Model
Jere Turner (Boston College)

Measuring Bilingual Students' Vocabulary Knowledge
Andrea Rolla San Francisco (Harvard)

A One Parameter Model Look at a Published and Popular Developmental Math Test: The Test of Early Mathematics Abilities (TEMA-2)
Amy Warren (Harvard)

The Psychometric Structure of the Conformity to Feminine Norms Inventory
Camelia Rosca (Boston College)

The Utility of Rasch Residuals to Reveal Gender Differences on the MCAS
Kathy Rhoades (Boston College)

Results of a National Teacher Survey: Perspectives of the Practitioner. How State-Mandated Testing Impacts Teaching and Learning
Lisa Abrams (Boston College)

MCAS Scale Invariance for LEP vs. non-LEP 10th Grade Math Students
Helena Miranda (Boston College)

Institute for Objective Measurement Session organizer: Dr. Larry Ludlow; Chair: Jere Turner

"When schemes are laid in advance,it is surprising how often the circumstances fit in with them."
Sir William Osler (1849-1919)

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
Associate Editor: Benjamin D. Wright
Quarterly

Rasch Measurement SIG Officers
Trevor Bond ..... Chair
Edward Wolfe..... Secretary
George Karabatsos ..... Program Chair
Join the SIG with your AERA renewal!

Calendar of Events
Jul 10-12, 2002 ASEESA, Namibia
www.polytechnic.edu.na/aseesa/
Jul 22 - Nov 4, 2002 Coursework, External study
David Andrich, *andrich@murdoch.edu*
Apr 21-25, 2003 AERA, Chicago
AERA, *www.aera.net*

Chi-Square Local Independence Meets the Rasch Model

Hambleton et al. (1991) suggest using a chi-square test to identify local independence between two items. The procedure consists of constructing a 2x2 table for two items using the correct and incorrect answers of persons at the same level of ability (i.e., with the same raw scores on the test):

	Item X		
Item Y	Right	Wrong	Total
Right	A	B	Y_R = A+B
Wrong	C	D	N-Y_R = C+D
Total	X_R = A+C	N-X_R = B+D	N = A+B+C+D

This yields a chi-square statistic with 1 degree of freedom:

$$\chi^2 = \frac{(A+B+C+D)(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)}$$

This is minimized to zero, and local independence is apparently assured in this dataset at this ability level for this pair of items, when

$$AD = BC$$

i.e., when

$$A/N = X_R/N * Y_R/N$$

as it would be when the data exactly conform to the Rasch model, or any other model in which the two items are conditionally independent.

In conventional usage, however, when $\chi^2 < 3.84$, there is a presumption of local independence with 95% confidence. Then "local independence" is declared!

Let us conduct an experiment based on this convention. Let us constrain the possible outcomes so that, if a subject succeeds on item X, that subject cannot fail on item Y. The data matrix becomes:

	Item X		
Item Y	Right	Wrong	Total
Right	A	B	Y_R = A+B
Wrong	0	D	N-Y_R = D
Total	X_R = A	N-X_R = B+D	N = A+B+D

The chi-square test apparently becomes:

$$\chi^2 = \frac{(A+B+D)(AD)}{(A+B)(B+D)} = N\left(\frac{X_R}{N-X_R}\right)\left(\frac{N-Y_R}{Y_R}\right)$$

Then we can determine the relationship between successes on items X and Y for any value of chi-square, say N*k, where k is a constant. Accordingly,

$$\left(\frac{X_R/N}{1-X_R/N}\right)\left(\frac{1-Y_R/N}{Y_R/N}\right) = k$$

Rewriting this in terms of probabilities, where P_X is the probability of success on item X under these conditions, etc., and taking logarithms, such that K=log(k):

$$\log\left(\frac{1-P_Y}{P_Y}\right) - \log\left(\frac{1-P_x}{P_X}\right) = K$$

It is seen that the relationship between the items is expressed in terms of log-odds difficulties, in accord with the Rasch model, despite that fact that the data do not accord with Rasch model conditions. Item Y is always easier than item X.

Further, for any particular chi-square "significance" value, e.g., 3.84, items with paired log-odds difficulties differing by more than log (3.84 / N) might be unwittingly declared "locally independent".

Here is an example:

	Item X		
Item Y	Right	Wrong	Total
Right	20	67	87
Wrong	0	13	13
Total	20	80	100

The log-odds difficulties of the two items are log(13/87) = -1.9, and log(80/20) = 1.4 logits, differing by -3.3. The criterion value is log(3.84/100) = -3.3.

Agustin Tristan

Hambleton R.K., Swaminathan H., Rogers H.J. (1991) "Fundamentals of item response theory", Sage publications Inc. London, Chapter 2, pp 9-12 and Exercise 6, pp 29-31

DNA and the Origins of the Jewish Ethnic Groups

A long-standing debate concerns the ethnic origins of the Ashkenazi (Eastern European) Jews. Are they of predominately the same ancestry as the Sephardic (Southern) and Oriental Jews? Or are they, to a large extent, descended from a different origin, perhaps the Khazar (Turkic) converts to Judaism in the 8th and 9th Century? A parallel debate concerns the genetic origins of the Lemba, a Bantu tribe in southern Africa who claim Jewish paternal ancestry. Hammer et al. (2000) published an exploratory paper with this conclusion "The results support the hypothesis that the paternal gene pools of Jewish communities from Europe, North Africa, and the Middle East descended from a common Middle Eastern ancestral population, and suggest that most Jewish communities have remained relatively isolated from neighboring non-Jewish communities during and after the Diaspora."

Table 1: Y-Chromosome haplotype frequency %
(from Hammer et al., 2000, Table 1)

4S	1R	Med	1Ha	1U	1C	1L	1D	Other	Ethnic Group
16	0	45	5	7	0	20	5	2	J-Roman
20	2	42	18	7	4	2	4	0	J-North African
13	3	28	6	16	31	3	0	0	J-Near Eastern
8	4	44	6	18	8	8	4	0	J-Kurdish
17	0	43	0	7	10	17	3	3	J-Yemenite
45	0	10	0	5	0	0	0	40	J-Ethiopian
24	12	31	2	3	4	10	9	3	?-Ashkenazi
6	6	26	0	32	0	0	0	29	?-Lemba
19	5	51	3	7	1	8	0	5	G-Palestinians
10	3	57	3	6	1	9	9	0	G-Syrians
29	13	46	0	4	0	0	4	4	G-Lebanese
19	0	38	19	19	5	0	0	0	G-Druze
5	5	33	5	24	5	0	19	5	G-Saudi Arabians
15	16	11	6	2	1	37	11	1	G-Europeans
50	1	26	0	4	1	4	1	12	G-North Africans
6	0	1	0	1	0	0	0	93	G-Sub-Saharan
6	12	26	9	12	3	20	5	6	G-Turks

Table 1 excerpts haplotype percentages from Hammer. Ethnic groups commencing "J" are recognized to have Jewish ethnicity. Those commencing "G" to be non-Jewish. "?" indicates the two groups that are the focus of the study. Hammer et al. investigate with various descriptive statistics. What insights does a measurement approach provide?

The first step is to orient the 9 haplotype variables with the construct of interest, Jewish ethnicity. At this point, the Ashkenazi and Lemba are omitted from the analysis, so as not skew their own placement within the construct. The "J" groups are anchored at 2 logits. The "G" groups at 0 logits. The difference between the two anchor values is chosen to be big enough to make a clear distinction in the output. An analysis in which each haplotype "item" is allowed to define its own rating scale ("partial credit") is then performed. Since the meaning of percentage gaps is not clear, the unobserved percentages are treated as structural zeros and are dropped out of the rating scales. This analysis indicates that haplotypes 1D, 1R and "Other" are negatively oriented (large negative correlations), so that larger percentages indicate less Jewish Ethnicity. "Med", 1L and 1Ha have almost zero correlations.

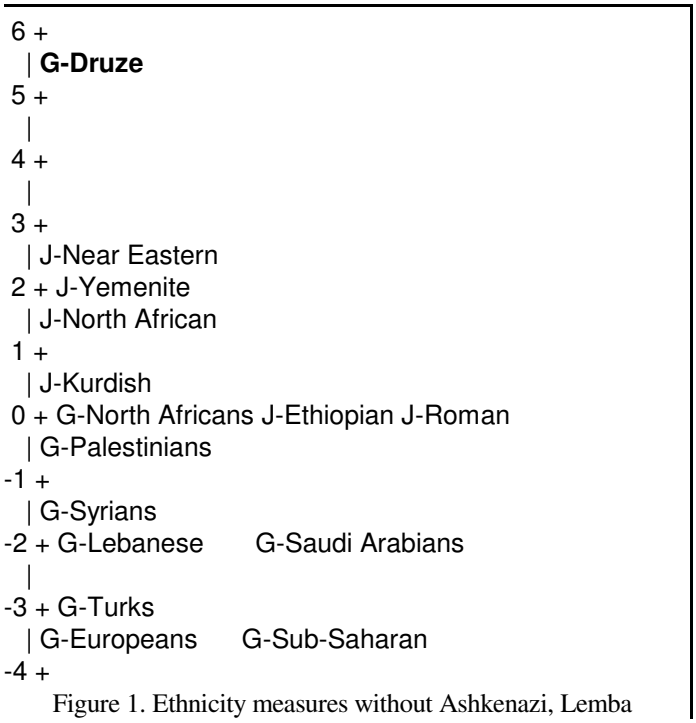


Figure 1. Ethnicity measures without Ashkenazi, Lemba

The second step is to reverse-code the negatively-oriented percentages and drop the haplotypes with very low correlations. The analysis is rerun without anchoring and still without the focus ethnicities. Figure 1 plots the resulting ethnicity measures. Logits have been scaled by 10, and the origin located at a meaningful point. It is seen that there is a neat stratification of "J" and "G" ethnicities, except for the Druze. The Druze number about 1 million and are located in southern Syria and Lebanon, and northern Israel. They originated from an Islamic reform movement in Egypt in the 11th Century. Exactly what this implies for their ethnicity is not clear. Perhaps, if this study had included "Egyptians" we would be better informed. In the present instance, however, we do not want to evaluate the "Druze-ness" of the focus ethnicities, and so the Druze are dropped from the study. Repeating the first step without the Druze, indicates

that 1Ha now has a positive correlation and so is kept in the analysis. The results already shown in Figure 1, but now without the Druze, emerge again.

The focus ethnicities are now introduced into the data, and measures constructed. Figure 2 depicts the results. The basic "J"-"G" contrast remains. The "Ashkenazi" are positioned between the "J" and the Turks, suggesting the possibility that they have both Jewish and Turkic descent. The Lemba are positioned between the "J" and the Sub-Saharan, also suggesting they may have mixed descent.

John M. Linacre

M. F. Hammer, A. J. Redd, et al. (2000) Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes . *Proc. Natl. Acad. Sci. USA*, Vol. 97, Issue 12, 6769-6774, June 6, 2000

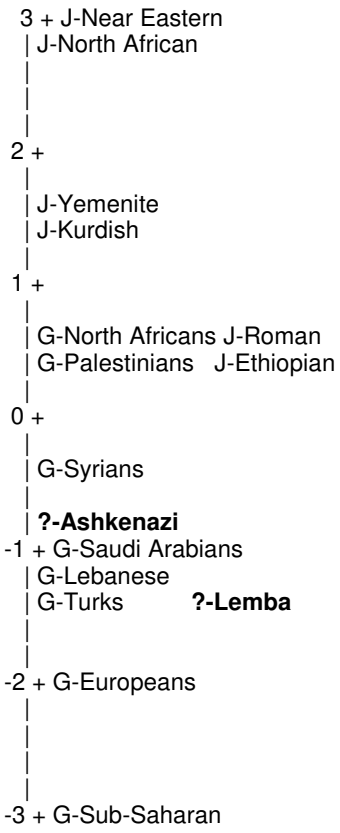


Figure 2. "Jewishness" of Ethnic Groups.

From the Toledo Blade, April 29, 2002

UT, BGSU students share $300 prize for work on different topics

Doctoral research of 2 honored

Mona Amer used what some say is a "revolutionary technique" to determine the best way to survey Arab-Americans about the push and pull of American culture.

Alana Raber gave children diaries to get closer to learning why some of them are overweight.

The two doctoral students were rewarded for their efforts. They shared a $300 prize given during the Ninth Annual Symposium on Research in Psychiatry, Psychology, and Behavioral Science yesterday at the Medical College of Ohio. The University of Toledo, Bowling Green State University, and MCO sponsor the conference.

"Both of these studies involved very careful looks in areas where people often do not look carefully." said Dr. Robert Elliot, the University of Toledo's clinical psychology professor who chaired this year's conference and served as a judge.

Ms. Amer, a doctoral student at UT, used a statistical technique called the Rasch model to measure cultural stress and family function among second-generation Arab-Americans.

Without the Rasch analysis, surveys can be inaccurate and misleading, said Dr. Christine Fox, the University of Toledo professor who co-wrote Ms. Amer's work. Dr. Fox, an associate professor of education research, is one of the nation's leading proponents of the Rasch model, and author of a book on the subject.

By employing this system, Ms. Amer, a Detroit native and second-generation Egyptian-American, arrived at a standardized way to collect data.

"You need to speak the same language," Dr. Fox said. "In psychology, it's the Tower of Babel. Everyone's making up their own systems, instead of speaking in a universal metric."

By Jenni Laidman, Blade science writer

"If" or "When" to Assess?

"No single assessment or combination of assessments offers the whole picture. The greater question to me is not **if** we use a given assessment tool (be it testing or other) but **when** to do so." (John Roope)

True. We never get the whole picture, but multiple methods shine different kinds of light from different angles, and can provide **real** illumination. But what does it really mean that "No single assessment or combination of assessments offers the whole picture." It means that any sample of test questions or assessment criteria are inherently incomplete. We can never imagine and pose every conceivable kind of problem that a student could possibly some day encounter. The assessment problem is then one of sampling from an infinite universe of possible questions and criteria, making sure that the ones chosen actually belong to that common universe (typically called a population), that they adequately represent it, and then calibrating the ones chosen so that they measure in a quantitative unit that any other similar sample from the same population will also measure in.

This is what Rasch measurement is all about, this is what gets people like me excited, because we have seen this work in practice over and over. Traditional methods focus on counts of right answers, or on sums of ratings, but those are only a preliminary step in the process.

Statistical models are usually chosen so as to describe the raw data. But what is the point of describing data that will never happen again in their specific detail? The raw data are inherently dependent on the particular questions asked, and cannot be generalized so as to be comparable with the scores likely to be obtained by the students on another, or even the same, sample of questions. We are then deliberately and prematurely restricting ourselves to the limited picture we have in hand, without even checking to see if a better picture can be brought into focus.

Rasch models are chosen so as to obtain generalizable comparability across samples of items and/or examinees/respondents. Instead of fitting models to data, and rejecting models that don't describe the data well enough, Rasch models prescribe data structures general enough to think with, and non-constructive data are rejected. This process much more closely approximates what has historically worked in experimental science.

If we can't generalize from our data, no amount of statistical *hocus pocus* is going to construct meaningful results. But if we start from a strong sense of how meaningful results are constructed, and we carefully monitor the process, we stand a pretty good chance of mediating past and future. By that last phrase, I mean that the past data we have in hand are only something we can learn from to help manage the future. We can't do anything about the past. Those data are history. But maybe we can extract a general structure from them that we see applies over and over again across data sets. So we try to take advantage of that structure by building it into a measuring instrument that will tell us what is going on with a child, in detailed quantitative and qualitative terms, at the very moment that the measure is made.

And that brings us to John's second point: "The greater question to me is not **if** we use a given assessment tool (be it testing or other) but **when** to do so." Right here is the crux of my passion. I start from the observation that everyday conversational language is an assessment tool that precedes formal testing, and it has a lot still to teach us about what testing could be.

Before getting to the *if* and *when,* think again about what assessment is supposed to do. It is supposed to let us know how we're doing, right? So how do we know when we know? One criterion for whether someone knows something for themselves is whether they can explain it in their own words. As Albert Einstein is reported to have said, "You do not really understand something unless you can explain it to your grandmother."

Well then, for different assessments and tests, intended to measure the same thing, to be shown to do so, and to do so in the same amount, is just another way of showing that we know what we're talking about, right? And because we, by and large, have hardly even started to assess experimentally the quality of our assessments and tests, we don't really have a clue as to whether we know what we're talking about, do we?

We might start by further consideration of where we're coming from. The ancient Greek term, *ta mathemata,* (the mathematical,) was used to refer to anything teachable and learnable. Mathematical thinking is characterized by its abstract generality and lack of dependence on concrete particulars, not primarily by an association with number. So conversation can function as a test that has qualitatively

ΜΗΔΕΙΣ ΑΓΕΩΜΕΤΡΗΤΟΣ ΕΙΣΙΤΩ
ΠΛΑΤΩΝ

Plato's Dictum: **"Let no one a-geometric enter!"**
From above the door of Geoffrey Opat, late Professor of Physics at the University of Melbourne, Australia.

mathematical consequences in the form of what is found to be teachable and learnable. In fact, it seems that in order to read or hear, and to learn from what is read or heard, one must have implicitly in mind the question to which the thing learned is an answer.

Developmentally, don't infants and kids also test themselves against the things they come up against, learning at the level they're at by means of the challenges posed by their environments? Then shouldn't their tests and assessments be sampling that same population of questions in determining whether children know something or not, or in assessing their developmental readiness?

And now we're at the *if/when* issue. The current fashion of high stakes tests poses hoops that are jumped through just one time and then are left behind forever. Teachers try to help their students get through those hoops by fashioning their own hoops to as nearly the same specifications as the high stakes ones. And then the results are used to decide whether the kid advances a grade, or gets into college, or if the school gets praise or money or a new principal.

But didn't we just see that actual daily life is a process of constant testing? Life doesn't pose rare high stakes tests, and it doesn't prepare us for them by posing other tests that are as similar as possible to the "real" one. Instead we have a constant random sampling of problems from each particular domain or construct. Some kinds of problems come up fairly frequently, and we develop routines for dealing with them.

But the point is, wouldn't a superior testing and assessment environment be built the way life tests us? We would want at least a small sample of problems to be posed daily, as, in fact, they already are in many textbooks and classrooms. The key difference is to calibrate these problems so that 1) the teacher, the student, the parent, and anyone who cares to look can see that the challenge posed is relevant to the student's llevel of ability, and 2) success on a new more difficult challenge can be immediately related to a likely measure on the high stakes assessment. In fact, given frequent and reproducible daily measures, there might be no further need for the high stakes assessments.

Rasch models help us implement tests and assessments patterned after the way that life itself challenges and promotes growth and development. Rasch's models are the tools we need to check the offspring of our assessments and tests for viability, because tests of sample- and scale-independence are what they provide. Because they are probabilistic, they support adaptive administration, meaning that short tests could be administered daily, and results obtained in a standard uniform metric that would inform the student and teacher as to the most relevant point of entry in the curriculum. These processes could help to integrate teaching and testing in a way that brings back to life the ancient Greek connection between the curriculum and the mathematical in *ta mathemata.*

The biggest obstacle to actually implementing an approach like this is that we have not yet created the measurement-friendly environment, the ecological niche, in which Rasch-born constructs can thrive. Many carefully designed instruments have been closely studied and precisely calibrated, but exist in complete isolation from 1) other similar instruments that in all likelihood could measure in the same unit and could also throw considerable new light on the theory of the construct, and from 2) the communities of practitioners and researchers who could benefit from the sharing of common languages for exchanging qualitative and quantitative value. Let us push our own work in the direction of creating these niches.

William P. Fisher, Jr.

"If the facts don't fit the theory, change the facts."

Attributed to Albert Einstein (1879-1955)

April 2003, Chicago

April 19-20, Saturday-Sunday
An Introduction To Rasch Measurement: Theory And Applications. UIC.
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,** *www.winsteps.com*

April 30-May 1, Monday-Tuesday
***Winsteps* Workshop**, *www.winsteps.com*

Residuals and Rating Scales

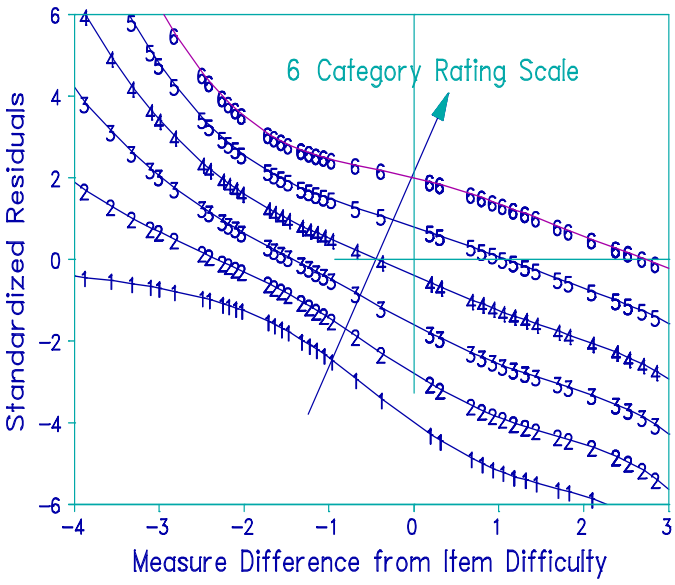
Measures are located on infinite, continuous variables, but they are observed as observations on finite discrete rating (or other) scales. Stephen Humphry (West Australia Writing Assessment, 2002) notices the distinctive patterns of residuals that result.

Figure 1 shows the standardized residuals for observations on a typical 6 category rating scale plotted against measures (relative to the mean difficulty of the item). A residual is the difference between the observation and its expected value for a respondent of a particular ability on that item. The Rasch model also predicts the distribution of observations around their expected value. The residual is divided by the standard deviation of this "model" distribution to obtain the standardized residual.

Each of the distinctive striations in the plot corresponds to the standardized residuals for one of the categories of the rating scale. At each measure, the striations are located at equal vertical intervals, equal in size to the "model" standard deviation. The curvature of the lines is indicative of changes in the item information along the latent variable.

When the standardized residuals for the responses to two items by the same persons are cross-plotted, then distinctive patterns emerge. These are shown in Figure 2. The pairs of responses to the 3-category items are shown on the plot as "response to easy item"+"response to hard item". For each traceline, the lower, left end is generated by *high* abilities, and the higher, right end by *low* abilities.

Only 3 instances of the unlikely pairing, "0"+"1" (low



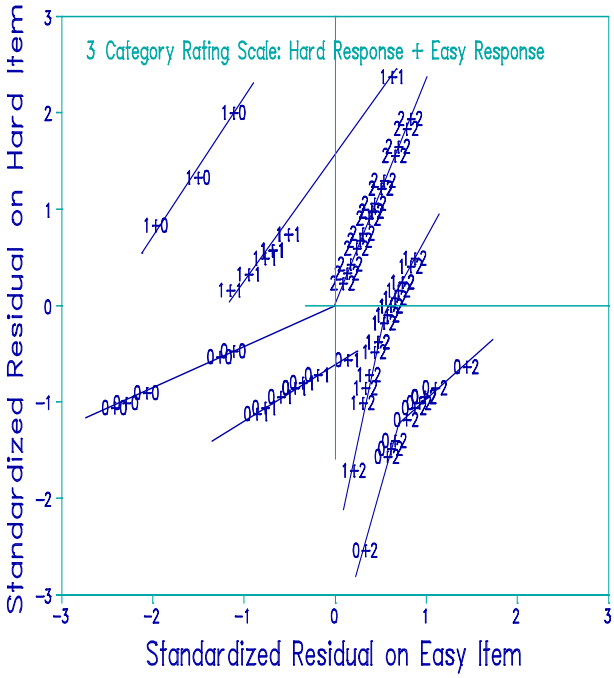


Fig. 2. Standardized residuals for two 3-category items.

rating on the easy item, middle rating on the hard item), are observed. They are in the upper left corner. The most unlikely response pairing, "0"+"2", is not observed here, but would be found in the extreme upper-left corner. The "0"+"0" and "2"+"2" trace lines approach the origin for persons toward the extreme low and extreme high ends, respectively, of the latent variable.

John M. Linacre

Too Many Factors?

"Therefore, one might expect the emergence of only one factor when a factor analysis would be performed on all newly defined subsets [of unidimensional items]. However, factor analysis of the newly defined subsets yielded two factors. Further inspection of the factor plot showed that the emergence of a second factor could be considered as an artefact due to the skewness of the subset scores."

Van der Ven, A.H.G.S., & Ellis, J.L. (2000). A Rasch Analysis of Raven's Standard Progressive Matrices. Personality and Individual Differences, 29 (1), 45-64.

Manual Estimation of "Partial Credit" Item Difficulties

Computer programs are routinely used to estimate Rasch measures. But is your program functioning correctly? Are the estimates based on the intended data? Are they based on the model you think you specified? A quick manual check of the computer's results can reassure you, or point you to a discrepancy.

Estimates for partial credit items, i.e., items specified to have unique rating scales, can be awkward to verify. Here is one approach. Imagine a sample of students who respond to both a dichotomous item and a polytomous "partial credit" items. We take the reported difficulty of the dichotomy, item i, as the benchmark difficulty. From this we estimate the difficulty of the dichotomy, D_{jk}, between each pair of two adjacent categories, k-1 and k, of the partial credit item, j.

From their respective Rasch models,

$$N_{0k} \approx \sum_n P_{ni0} \quad P_{njk}$$

$$= \sum_n \frac{P_{ni1}}{e^{B_n - D_i}} \frac{P_{nj(k-1)}}{e^{-B_n + D_{jk}}}$$

$$= e^{D_i - D_{jk}} \sum_n P_{ni1} \quad P_{nj(k-1)}$$

$$\approx e^{D_i - D_{jk}} \quad N_{1(k-1)}$$

So that

$$D_{jk} \approx D_i + \log \left(N_{1(k-1)} / N_{0k} \right)$$

We can use this equivalence to estimate the difficulty of all the $\{D_{jk}\}$. The mean item difficulty is

$$D_j = \sum_{k=1}^{m} D_{jk} \quad / \quad m \qquad \text{for } k = 1,m$$

So that, within the item, the adjacent category dichotomies, F_{jk}, are located at

$$F_{jk} = D_j - D_{jk}$$

Here is an example from Items 12 and 14 of the "Liking for Science" data (Wright and Masters, 1982), specified as "partial credit".

Paired Frequencies
for "Liking for Science" items 12 and 14.
(Item 12 recoded 0,1)

N_{00}	N_{01}	N_{02}	N_{10}	N_{11}	N_{12}
5	6	2	9	28	25

According to an analysis, the difficulty of the dichotomous item i=12, is -1.14. So that for item j=14,

$$D_j 1 \approx D_i + \log \left(N_{10} \ / \ N_{01} \right)$$

$$= -1.14 + \log (6/9) = -1.55$$

$$D_j 2 \approx D_i + \log \left(N_{11} \ / \ N_{02} \right)$$

$$= -1.14 + \log (28/2) = 1.50$$

In this example, the reported estimates are -.71 and 1.30, somewhat more central than the values, -1.55 and 1.50, given by our approximation.

Andrew Stephanou
Australian Council for Educational Research

Estimating Item Discriminations

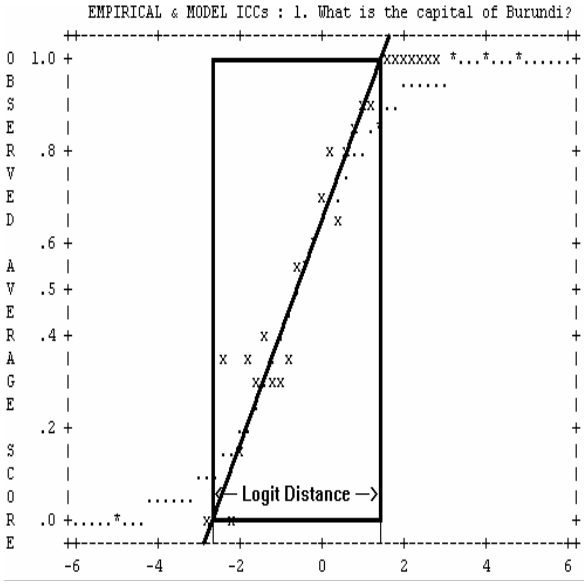
Plots of empirical item characteristic curves (ICCs) enable one to estimate the empirical item discrimination, at least as well as a 2-PL IRT computer program. This is because 2-PL discrimination estimation is degraded by the imputation of a person distribution and by arbitrary constraints placed on discrimination values. It is also skewed by accidental outliers which the eye can disregard.

On a plot such as Figure 1 draw in the line that, to your eye, matches the central slope of the empirical item characteristic curve (ICC). In Figure 1, the dots indicate the Rasch-model-predicted ICC. The points marked `x' indicate the empirical values, obtained by averaging the observed responses, `0' and `1', for persons whose measures are estimated to be in the .1 logit interval centered on the `x'. When a dot and an `x' coincide, a `*' is shown. The line follows the trail of `x' and `*'.

Estimate the logit (x-axis) distance from where the line intercepts the .0 score value to where it intercepts the 1.0 score value (for dichotomies). The logit distance here is about 4.0 logits.

Use the nomogram in Figure 2 to estimate empirical item discrimination. In this nomogram, looking at the middle traceline, a logit distance of 4.0 logits, corresponds to a logit discrimination of 1.0, in accordance with model prediction. Shorter distances, i.e., steeper slopes, correspond to higher discriminations.

In Figure 2, there are 3 lines in order to simplify comparison and reporting of results in an IRT context. 2-PL programs



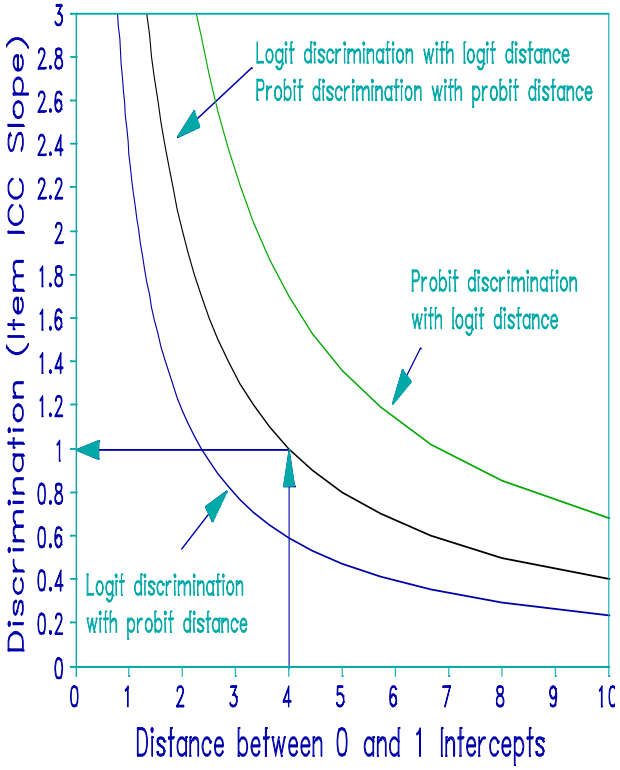


Figure 2. Nomogram for item discrimination.

generally report their results in probits. So the probit discrimination for logits distances is given by the upper line. That corresponding to a 4-logit distance is 1.7. On the other hand, if one is looking at plot scaled in probits, instead of logits, one would look at the lower line to find the logit discrimination.

John M. Linacre

"The most valuable contribution to the area of tests of fit for Rasch models in recent years has been the recognition by some psychometricians that there is no such thing as a final `fit' of data to the model and hence that no one test is ever likely to be complete. Appreciation of this point still needs to be given much wider circulation among workers in the field. Then there will be less of a tendency to reject data sets (or the model) outright, simply because one test failed to show `fit'. Implicit in this perspective is the assumption that there is as much to be learnt about a data set from the responses which misfit as there is from those which do fit."

Graham Douglas (1982) Issues in the fit of data to psychometric models. Education Research and Perspectives, 9:1, p. 43.

BICAL Item Discrimination Index

Question: An item "Discrimination Index" is shown in Wright & Stone's book, *Best Test Design* (1979, p. 52), but is not described in detail. What is it? Is it useful?

K. L., Hong Kong

Answer: This Discrimination Index was a feature of the Rasch computer program, BICAL. It is described in MESA Research Memorandum, 23B, by B.D. Wright, R.J. Mead, & S.R. Bell, June 1979. Here is what is written on pages 15 and 16:

If the data ... produced characteristic curves which varied in slope, then the person-item logit could be expressed as

$$L_{ni} = \alpha_i (\beta_n - \delta_i) + \varepsilon_{ni} \qquad (21)$$

or

$$\gamma_{ni} = (\alpha_i - 1)(\beta_n - \delta_i) + \varepsilon_{ni} \qquad (22)$$

where

$$\gamma_{ni} = L_{ni} - (B_n - D_i)$$

In this form, γ_{ni} is the difference between a "true" logit response and the Rasch model with which we have attempted to explain it. If the Rasch model adequately accounts for the data, the regression in Equation 22 should have a slope of zero.

In terms of estimates, this expression can be written as

$$y_{ni} = (a_i - 1)(b_n - d_i) + e_{ni} \qquad (23)$$

[but y_{ni} can be approximated from the Rasch ICC as]

$$y_{ni} = (x_{ni} - P_{ni}) / [P_{ni} (1 - P_{ni})]$$

Therefore an indication of α_i can be calculated as

$$a_i = 1 + \sum_n (y_{ni} - y_{.i})(b_n - d_i - m_{.i}) / \sum_n (b_n - d_i - m_{.i})^2$$

[where $y_{.i}$ is the mean of the residuals for item *i,* and $m_{.i}$ is the mean of the person-item logits, b_n-d_i.]

This is the residual index given in BICAL output.

Comment: Item discrimination is conceptualized as a property of the central slope of the item characteristic curve (ICC). This index, however, is strongly influenced by outliers (e.g., lucky guesses, careless mistakes). This drawback might be overcome by dropping or down-weighting off-target responses. On the other hand, Wright & Stone (p. 53) claim this index to be less affected by sample targeting and dispersion than the point-biserial correlation.

Now online - Frank Baker's "Basics of Item Response Theory"

Dennis Roberts writes: *"http://ericae.net/irt/baker/* is a link to the full text of Frank Baker's classic little book on *Basics of Item Response Theory* (1985) ... with his old software included! This is provided by the ERIC Clearinghouse on Assessment and Evaluation."

The complete text of the revised and updated 2001 edition can be studied one screen at a time, or the entire book can be downloaded as one *pdf* file.

This book views the "Rasch or One-Parameter, Logistic Model" as a special case of Birnbaum's 3-PL model. Here is what is stated on page 25:

"The next model of interest was first published by the Danish mathematician Georg Rasch in the 1960s. Rasch approached the analysis of test data from a probability theory point of view. Although he started from a very different frame of reference, the resultant item characteristic curve model was a logistic model. ... Under this model, the discrimination parameter of the two-parameter logistic model is fixed at a value of $a = 1.0$ for all items; only the difficulty parameter can take on different values. Because of this, the Rasch model is often referred to as the one-parameter logistic model.

"The equation for the Rasch model is given by the following:

$$P (\theta) = \frac{1}{1 + e^{-1 (\theta - b)}}$$

"where: *b* is the difficulty parameter and θ is the ability level.

"It should be noted that a discrimination parameter ["1"] was used in [the] equation, but because it always has a value of 1.0, it usually is not shown in the formula."

[Thus Baker's presentation of the Rasch model follows IRT conventions, but somewhat idiosyncratically.]

The Hyperbolic Cosine Unfolding Quasi-Rasch Model

The Hyperbolic Cosine Model (HCM, Andrich & Luo, 1993) for dichotomous unfolding responses is derived from the Rasch model for 3-ordered-category responses, but is not, itself, a Rasch model. Consider a dichotomous preference item: "I owe a lot to my parents" (Agree/Disagree). The meaning of Agree seems obvious. But what does Disagree mean? I owe little to my parents? I owe everything to my parents? Thus, in constructing the HCM, the Disagree response is resolved into two latent components. One component is, *Disagree below,* "I owe a little ..." The other is *Disagree above,* "I owe everything ..."

A Rasch model for three ordered categories is:

$$\log\left(\frac{P_{nik}}{P_{ni(k-1)}}\right) = B_n - D_i - F_{ik} \qquad k = 1,2$$

where k=0 is "Disagree below", k=1 is "Agree", and k=2 is "Disagree above", and $\Sigma F_{ik}=0$. The HCM function for a Disagree response, P_{niD}, is the sum of the probabilities of the "Disagree below" and "Disagree above" categories. The Agree category remains as P_{niA}. Summing,

$$\left(\frac{P_{niA}}{P_{niD}}\right) = \frac{e^{-F_{i1}}}{e^{B_n - D_i} + e^{-(B_n - D_i)}}$$

A convenient identity for the hyperbolic cosine is:

$$\cosh(x) = (e^{x} + e^{-x}) / 2$$

So that, after reparameterization (Andrich, 1996; Luo, 1998), the HCM can be expressed more elegantly as:

$$\left(\frac{P_{niA}}{P_{niD}}\right) = \frac{\cosh(\rho_i)}{\cosh(B_n - D_i)}$$

where

$$\rho_i = \cosh^{-1}\left(0.5\ e^{-F_{i1}}\right) \quad \text{and} \quad F_{i1} < -0.69$$

It is seen that $P_{niA} = P_{niD} = 0.5$ when $\rho_i = B_n - D_i$ or, because cosh(x) = cosh(-x), when $-\rho_i = B_n - D_i$. Thus the new parameter, ρ_i, is half the distance between the two crossing points of the Agree and Disagree response curves. This characterizes the *latitude of acceptance,* an important concept in attitude measurement. Note that this model requires that the probability of observing Agree reach .5 at some point along the latent variable. The Figure shows the HCM functions and the corresponding Rasch model for 3 categories (in dotted lines).

Why is HCM not a Rasch model? Rasch models require parameter separability or, in statistical terms, sufficient statistics. HCM does not have these.

Choosing a Response Model

If the data follow the Rasch (or other cumulative) model, responses are positively correlated across items. If the data follow the HCM (or other unfolding model), an item has positive correlations with nearby items, but negative correlations with distant items. The HCM equation has been expanded into a general form for dichotomous unfolding responses (Luo, 1998), and then into a general form for polytomous unfolding responses (Luo, 2001).

Guanzhong Luo, Murdoch University, Australia

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: reconciling Thurstone and Likert methodologies. British Journal of Mathematical and Statistical Psychology, 49, 347-365.

Andrich, D. & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. Applied Psychological Measurement, 17, 253-276.

Luo, G. (1998). A general formulation of unidimensional unfolding and pairwise preference models: making explicit the latitude of acceptance. Journal of Mathematical Psychology. 42, 400-417.

Luo, G. (2001). A class of probabilistic unfolding models for polytomous responses. Journal of Mathematical Psychology. 45, 224-248.