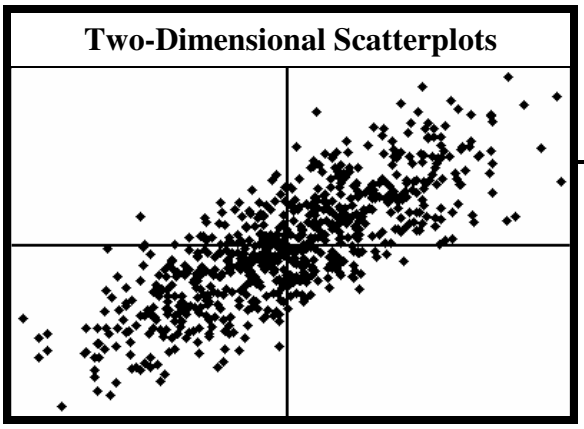


RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 18 No. 3 Winter 2004 ISSN 1051-0796

The Rasch Model and the Quest for Perfection

"The Rasch model is too rigid." "It demands too much from our data." "It throws out too many items." Perhaps the Rasch model is a mathematical *Maud:*

Perfectly beautiful: let it be granted her:
where is the fault?
Faultily faultless, icily regular, splendidly null.
Dead perfection, no more.

Tennyson

Are Rasch analysts wrong to seek perfection? The book, "The Customer is the Key" (M.M. Lee with J.N. Sheth, Wiley, 1991) describes the six characteristics of businesses identified as "winners". They are summarized thus:

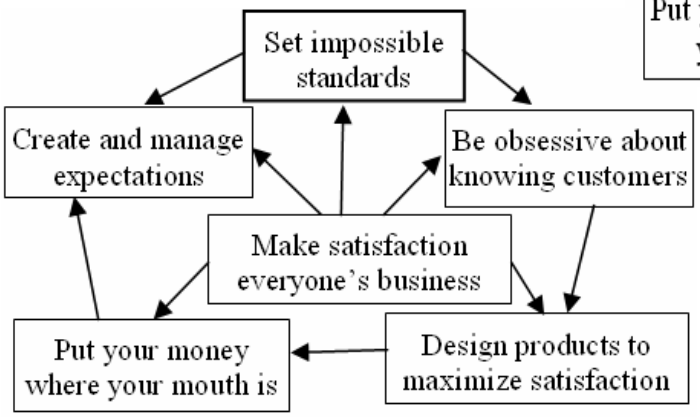


Fig. 1. Six key characteristics of "winning" businesses. (redrawn from Fig. 3.1, Lele & Sheth, 1991).

Notice the extreme language, "impossible", "obsessive", "everyone", "maximize." Surely no business can actually achieve these characteristics? Lele & Sheth admit that they can't, but "these companies realize that performance often falls short of expectations. Therefore, in order to deliver merely good results, they must set their sights on impossible goals. .... knowing that they are likely to achieve something less than what they aim for is the most compelling reason that they can give for aiming for the best" (p. 61). This suggests that Fig. 1 also applies to "winning" measurement projects, see Fig.2:

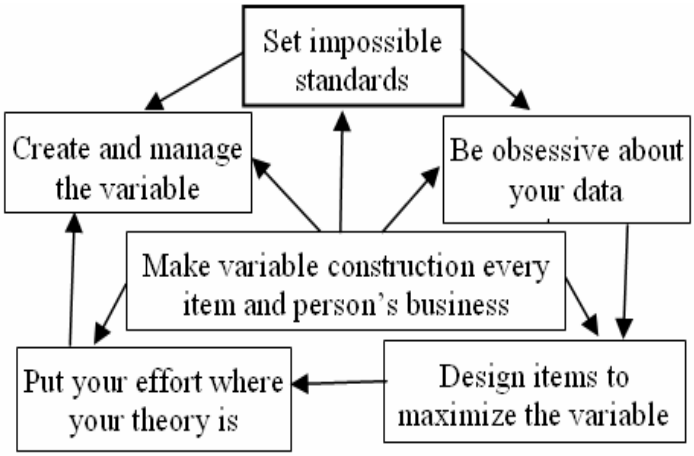


Fig. 2. Six key characteristics of "winning" measurement projects. (after Lele & Sheth, 1991).

"Science must begin with myths." *(Alexander Pope)*

"Yours is a distinctive vision, you believe. And it's also ideal. After all, you want to set a new standard of perfection, beauty or excellence. You want to be a model for others. Yours is an ideal and unique image of the future" *(J.M Kouzes & B.Z. Posner, The Leadership Challenge, Jossey-Bass, 1995, p. 96)*

Let our myth be that perfection is attainable, and let us resist the shame of having mediocrity thrust upon us *(Joseph Heller, Catch-22).*

John Michael Linacre

Table of Contents

Book Review: "Constructing Measures: An Item Response Modeling Approach"

Mark Wilson (2004) Mahwah NJ: Lawrence Erlbaum Associates

The plot-line of "Constructing Measures" is straight-forward. The first chapter presents "A Constructivist Approach to Measurement" outlining four building blocks. These four blocks are the topics of the next four chapters: construct maps, item design, outcome space (i.e., the data and its structure), and the measurement model. The next three chapters focus on specific aspects of instrument quality: model-data fit, measurement error and evidence of test validity. The last chapter connects the topics in the book to the wider worlds of psychology, statistics and assessment.

In total, this book is an excellent text for those desiring to construct, apply and benefit from valid test instruments for measuring educational or psychological traits. The book contrasts strongly with the typical text on educational and psychological testing and measurement. Such texts contain a mass of technical detail and jargon, but lack the careful guidance the neophyte requires to successfully construct a test. A parallel is a book of cooking recipes. The novice cook sees what the final result should look like from the picture and also sees the list of ingredients, but cannot get successfully from the one to the other. "Constructing Measures" guides the reader safely along the rocky path.

The power of the practical, yet deeply philosophical, test-development model central to "Constructing Measures" is obvious when compared with models presented in conventional textbooks. "Constructing Measures" presents a development cycle, Fig. 1, with clearly specified actions to be taken at each stage. The process is one of incremental improvement of every block. The cycle is repeated until the desired results are obtained. A conventional model, such as that in Fig.2, implies a development cycle, but one in which only the test items change, all else is fixed. It operationalizes the "pile-up" theory of respondent performance. The test is a check-list, and respondent success is counted up from "none" to "all". At any moment, each respondent has a certain accumulation, but there is little information about what should be accumulated next, or whether there are holes in the pile.. In contrast, Fig. 1 supports a "ladder" theory of respondent success. At any moment each respondent is at a certain height. We know what is above the respondent and what is below. We can identify the respondent's special strengths and deficits. Further, Fig 1 contains the fundamental insight that, during the test development process, our understanding of the instructional objectives and activities, i.e., the formal statement of the construct, will change. Gaps, obscurities and ambiguities will be encountered. The formal test probes the respondents, but Fig 1. also asserts that their responses probe the construct.

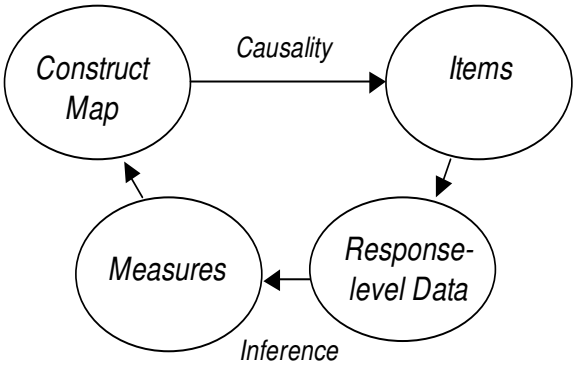


Fig. 1. The 4 building blocks of Wilson, 2004, Fig. 1.9

Accompanying the book is the computer program *GradeMap.* This features graphical representations of the construct, the functioning of the items, and the response patterns of individual respondents. At the time of this review, *GradeMap* produced conceptually interesting output, but was somewhat slow.

A couple of slight improvements to the text: first, Wright & Masters (1981) should be (1982). More fundamentally, page 4, "1.1 What is Measurement?", contains the enigmatic phrase, "those numbers [measures] have certain properties." It is not until we get to page 92 (in my reading of the text) that the essential property is revealed: "the *difference* between them is what matters" (Emphasis author's), i.e., the numbers must have linear scaling. In fact, a central purpose of the book is to instruct the reader how to design tests that locate the performance of each person and the difficulty of each item on a shared linear measurement ruler.

This title is currently on offer at $29.95. Its contents complement "Applying the Rasch Model" (Bond & Fox, 2001, Erlbaum).

John Michael Linacre

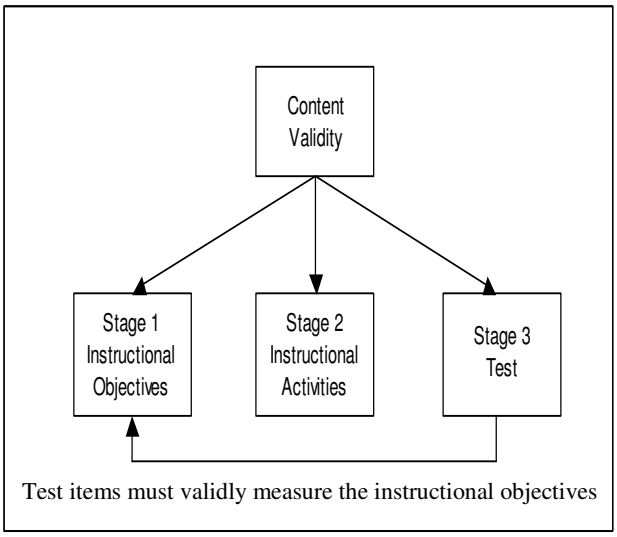


Fig. 2. The three-stage classroom measurement model of Kubiszyn and Borich, 2000, Fig. 4.1

Comment: Mixed Methods Research: A Research Paradigm whose Time has Come

R.B. Johnson & A.J. Onwuegbuzie, Educational Researcher (2004) 33:7, 14-26.

Yet again, Benjamin D. Wright was ahead of the wave. Qualitative or quantitative methodology? Ben advocated using both simultaneously. Now so do our authors.

Fig.1 is our authors' flowchart of their recommended research methodology, the "Mixed research process model."

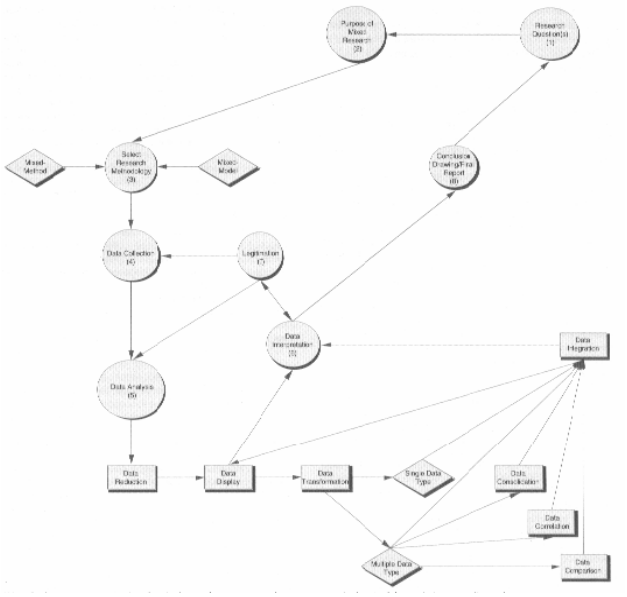


Fig. 1. Mixed research process model (reduced size). Johnson & Onwuegbuzie, 2004

Let's streamline and redraw their flowchart:

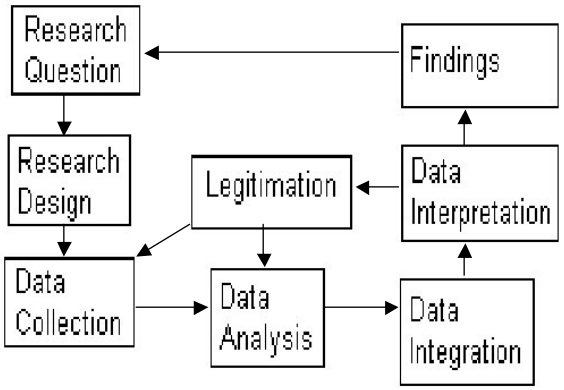


Fig. 2. Mixed research process model (streamlined and redrawn)

This is now seen to correspond closely to Ben's Fig. 3 in "The Road to Reason", *Wright B. D. (1998) Rasch Measurement Transactions, 11:4 p. 589.* Ben wrote:

> "There is no contradiction or conflict between the qualitative and the quantitative. The qualitative is complex, inscrutable, unique. But to learn from it, utilize it, manipulate it, it must be made simple, obvious, general. The leap from qualitative to quantitative is based on this organizing principle. We want to leave behind the contradiction, chaos and idiosyncrasy of the impractical concrete. We want to build an artificial world based on the practical abstract."

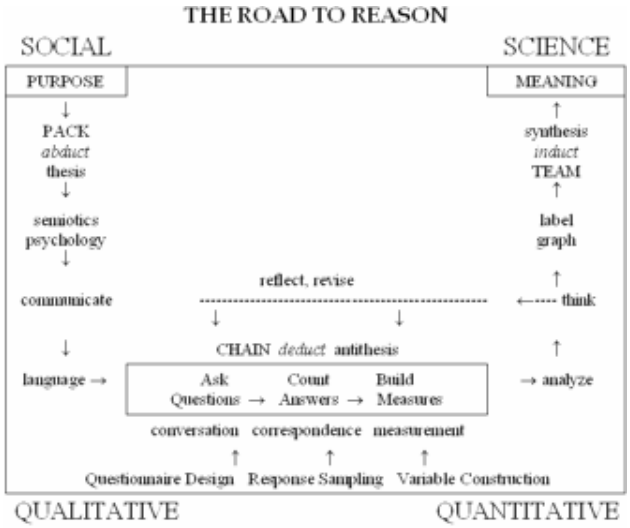


Fig. 3. Ben Wright's (1998) "The Road to Reason"

Ricoeur's Kluge Prize and Its Relevance to Rasch

On December 8, 2004, the philosopher Paul Ricoeur shared, with the historian, Jaroslav Pelikan, the US$1 million Kluge Prize awarded by the U.S. Library of Congress. Ricoeur's prize honors his lifetime of achievements in philosophy, many of which were realized in the twenty years he taught at the University of Chicago. The Kluge is equivalent monetarily with the Nobel Prizes, and intends to be considered equivalent in intellectual prestige, as well. For more information, see the Library of Congress web site at *www.loc.gov/loc/kluge/* .

Qualitative Objectivity

Ricoeur's work overlaps significantly with Rasch measurement in the area of epistemology, which is the logic of the way we speak and write (for in-depth accounts of the overlaps, see Fisher, 2003a, 2004). An obvious place to begin is from Rasch's recognition that "even in physics observations may be qualitative . . . as in the last analysis they always are! (e.g. as reading off a point as located between two marks on a measuring rod)" (Rasch, 1977, p. 68; original parenthetical insertion).

Ricoeur (1981a, p. 210) similarly offers a paradigm of reading that takes the text as a basis for a form of objectivity that owes nothing to a positivist world of facts, but which is "congenial" to this kind of objectivity. Non-vicious Circularity

A second overlap is suggested by Rasch's (1960, p. 110) remarks on the non-vicious circularity through which measures and calibrations are mutually constituted. The dialectical interaction of questions and answers is a classic example of the hermeneutic circle, one of Ricoeur's major areas of investigation. Rasch's remarks in this overtly interpretive, qualitative vein are significant for leading off a passage that explores the mathematical similarities between his model for reading ability measurement and Maxwell's 1876 analysis of the relations of mass, force, and acceleration.

Textual Independence

A third overlap pertains to the specific details of Ricoeur's and Rasch's epistemological claims. Ricoeur (1977, p. 293) makes the strong assertion that "No philosophical discourse would be possible, not even a discourse of deconstruction, if we ceased to assume what Derrida justly holds to be 'the sole thesis of philosophy,' namely 'that the meaning aimed at through these figures [of metaphor] is an essence rigorously independent of that which carries it over.'" Ricoeur's (1981a) elaboration of the four traits characteristic of textual objectivity comprise criteria for recognizing when and where a text's inherent metaphoricity achieves a status of rigorous independence from its meaning.

Providing the measurement analogue, Rasch (Rasch 1961, p. 325; 1960, p. 122), of course, is known for his separability theorem, in which, to be meaningful, in Rasch's sense of specifically objective, the measurement and calibration parameters estimated must be rigorously independent from one another, as well as from the model itself.

Textual Vitality

Ricoeur (1981b, pp. 159, 162) also suggests a direction for measurement practice that takes a step beyond Rasch's thinking and beyond the typical state of the art in psychosocial measurement theory and practice. Ricoeur considers the reading of a text, which we construe to include test, assessment, and survey instruments, to be meaningful when the interpretation is more "an objective act of the text" than it is "a subjective act on the text." What Ricoeur means by this is that texts have lives of their own evident in the way that they compel certain interpretive invariances across samples of readers.

Different readers bring different sets of presupposed and explicit questions to a text, but the text nonetheless still persists in showing itself as itself, insofar as it has been understood. The same kind of thing happens in measurement when different instruments intended to measure the same construct are administered to different samples at different times and places but still give rise to the same order of things (Fisher, 1997, 2004).

Invariance and Self-Identification

New understandings, of course, may well provoke whole new kinds of invariance, and this leads into Ricoeur's later work on identity, time, and narrative. It is of interest in this context to note that, linguistically, we separate the identities of fields of study by naming them according to their relevant type of logos, or proportionate rationality. Thus we have psychology, sociology, biology, etc. It then appears that the professional identities of communities of inquiry and their members are constituted through the questions they pursue and the things they measure.

Might not we then achieve firmer, more coherent, and more meaningful professional senses of ourselves to the extent that we achieve more objective, transparent, and universally uniform measurement of the things we investigate? There is reason to hope that the overlap of Ricoeur's theories of interpretation and identity with Rasch measurement will lead to yet greater things.

William P. Fisher, Jr.

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

Fisher, W. P., Jr. (2003a, December). Mathematics, measurement, metaphor, metaphysics: Part II. Accounting for Galileo's "fateful omission." Theory & Psychology, 13(6), 791-828.

Fisher, W. P., Jr. (2004, October). Meaning and method in the social sciences. Human Studies: A Journal for Philosophy and the Social Sciences, 27(4), in press.

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

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

Ricoeur, P. (1977). The rule of metaphor: Multi-disciplinary studies of the creation of meaning in language (R. Czerny, Trans.). Toronto: University of Toronto Press.

Ricoeur, P. (1981a). The model of the text: Meaningful action considered as a text. In J. B. Thompson (Ed.), Hermeneutics and the human sciences: Essays on language, action and interpretation (pp. 197-221). Cambridge, England: Cambridge University Press.

Ricoeur, P. (1981b). What is a text? Explanation and understanding. In J. B. Thompson (Ed.), Hermeneutics and the human sciences: Essays on language, action and interpretation (pp. 145-64). Cambridge, England: Cambridge University Press.

Model Overfit in Theology

"In short, our tendency has been to fight our fiercest battles at the theological periphery of evangelicalism, where we believe the limits of tolerance have been exceeded *[i.e., where data no longer fit our theological model]*. We rarely ask who in our midst may be equally misguided (and possibly even more dangerous) because they have drawn the boundaries too narrowly rather than too broadly *[i.e., fit the theological model too well]*. As Arland Hultgren's survey of the earliest eras of church history reminds us, one can become heretical by being either too broad-minded or too narrow-minded. It would be a salutary exercise to survey the history of the *Evangelical Theological Society* to see if we have ever addressed the second of these categories *[too narrow, i.e., theological overfit],* having obviously addressed the first numerous times *[too broad, i.e., theological underfit]*. It would be even more salutary as we currently wrestle with definitions of orthodoxy more generally to make sure that we address both extremes."

Craig L. Blomberg, *The New Testament Definition of Heresy.*

Arland J. Hultgren (1994) The Rise of Normative Christianity. Minneapolis: Fortress Press

Midwest Objective Measurement Seminar

Friday, December 3, 2004
Rehabilitation Institute of Chicago
Institute for Objective Measurement
Moderators: Mary E. Lunz and Allen Heinemann

Reliability for Performance Examinations. *Mary E. Lunz and Lidia Martinez, Measurement Research Associates.*

A comparison of Rasch and KR-20 reliability. *Kirk Becker, Promissor, Inc.*

Factor Analysis of Survey Responses to Physics associated with Medical Diagnostic Sonography. *Timothy Sares, American Registry of Diagnostic Medical Sonographers*

Translating Job Task Analysis: Data to Test Blueprints An Automated Interactive Procedure. *John Stahl and Kirk Becker, Promissor, Inc.*

Rasch Model Measures Developmental Change. *Nikolaus Bezruczko, Measurement and Evaluation Consulting, Chicago, Illinois*

Monitoring sources of variability within the English writing competency examinations. *Lidia Dobria, MESA University of Illinois at Chicago*

Ethnicity and Teacher Placement. *Patricia Garza, George Karabatsos, and Josh Radinsky, University of Illinois at Chicago*

Functional Caregiving: Empirical Evidence for a New Construct of Mothers' Caregiving for Adult Children with Intellectual Disabilities. *Chen, Shu-Pi C., St. Xavier University, Chicago; Ryan-Henry, Sheila, Seguin Retarded Citizens' Association*

Measuring the Accessibility of Trails and Walking Paths for Persons with Disabilities: A Many Faceted Rasch Analysis. *Barth Riley, Ph.D., Department of Disability and Human Development, University of Illinois-Chicago*

Measuring Health Barriers for Chinese Americans Using Rasch Measurement Model. *Cuiqing Huang, Kendon Conrad, Terri Morris, University of Illinois at Chicago*

Analyzing the Substance Problems Scale with Winsteps and Facets. *Michael Dennis, Chestnut Health Center; Ken Conrad, University of Illinois at Chicago and Hines VA Hospital; Chris Scott, Lighthouse Institute Rod Funk, Chestnut Health System; Carol Myford, University of Illinois at Chicago*

Statistics and Truth

"Statistical models are sometimes misunderstood in epidemiology. Statistical models for data are *never true.* The question of whether a model is true is irrelevant. A more appropriate question is whether we obtain the correct scientific conclusion if we pretend that the process under study behaves according to a particular statistical model."

Scott L. Zeger, American Journal of Epidemiology 1991, 134 (10), 1062. *Courtesy of William Fisher*

Mapping Multi-Dimensionality

The physical universe is imagined to exist in 11 dimensions (and time) according to "string" or "(mem)brane" theory. Only three of these dimensions, length, depth, width, do we encounter directly. Very few can understand this complexity. But the mental universe of everyday experience is of a far higher dimensionality, and we encounter many of its dimensions simultaneously intertwined. Consider Figure 1 (excerpted from Figure 1 in De Mars C.E., "Measuring Higher Education Outcomes with a Multidimensional Rasch Model", Journal of Applied Measurement, 5, 4, 350-361, 2004).

Figure 1 shows student measures on two dimensions. Dimension 1 depicts the measures for the students on "American Experience" objectives and Dimension 2 on "Global Experience" objectives. Items 1-40 focus on American Experience and 41-78 on Global Experience.

The Paper reports the reliabilities of the student measures in the two dimensions of "Experience" to be .86 and .76, and the correlation between the student measures from separate analyses of the dimensions to be .64. This implies that the correlation between the dimensions, when disattenuated for measurement error, is .64 /√(.86*.76) = 0.78. This approximates 0.79, the correlation between the two dimensions in the joint multidimensional analysis. Thus analyzing the two Experiences together increases overall estimation accuracy. However, the accuracy for those students with atypical relationships between the dimensions may be reduced.

In wrestling with dimensional concepts, it is useful to consider parallel situations in physics. The two

Figure 1. Experience map in one dimension (De Mars, 2004)

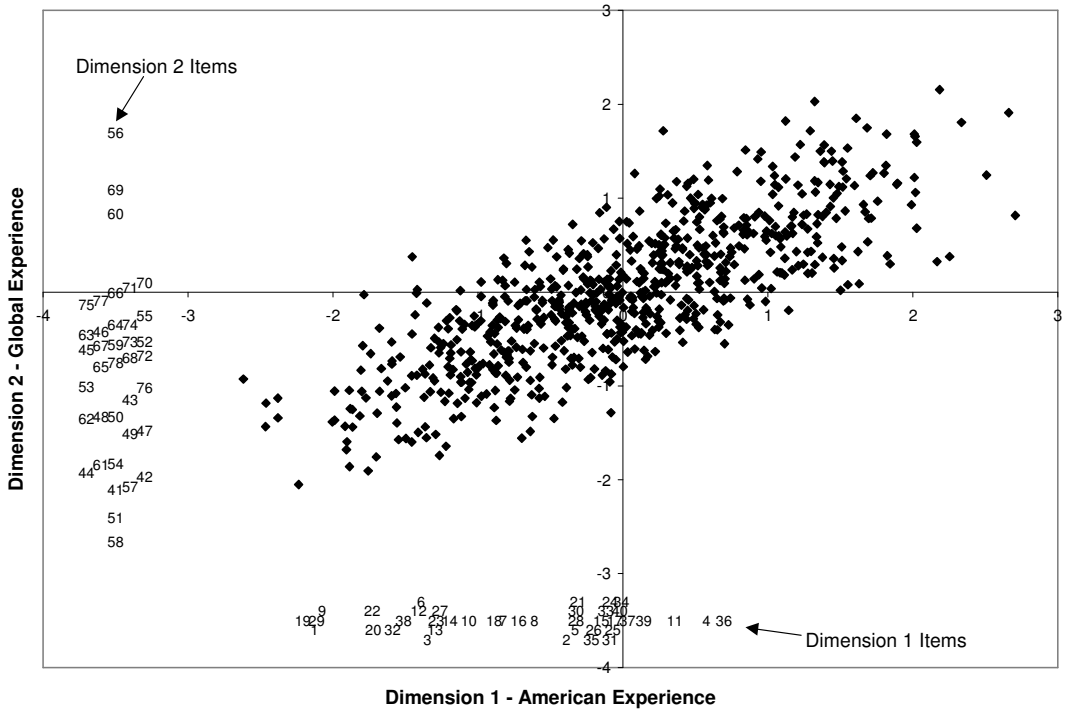


Figure 2. Experience map in two dimensions

"Experience" variables are correlated .79. This is roughly the correlation of height and weight for growing children. So, looking at Figure 1 we could understand Dimension 1 to be height, and Dimension 2 to be weight. And items 1-40 to be marks on a height scale, and items 41-78 to be marks on a weight scale. When we don't know a child's weight, or measure it too sloppily, we can infer that weight from the child's height, and vice-versa. In fact, no matter how precisely we estimate a child's weight, the child's height estimate would indicate whether our weight estimate is on the heavy side (if the child is relatively short) or on the light side (if the child is relatively tall) and similarly for the child's height estimate. Further, our height and weight scales may be crude or uneven. In which case, the correlation between height and weight enables us to smooth out some of the irregularity in the scale markings. But, though these manipulations produce an overall improvement in accuracy, the accuracy for any particular child might be lessened.

On the map, however, these operational considerations, and the non-orthogonality of height and weight in our sample, must be left behind. There are two dimensions, so, to understand these data, a two dimensional map is needed. Figure 2 is Figure 1 redrawn with student locations simulated to match the distributions in Figure 1.

On viewing Figure 2, an immediate question is, "Which students have substantively or statistically significantly different measures on the two dimensions?" Substantively, in many educational situations, one logit approximates one year of growth. Statistically, in the multi-dimensional analysis, the two sets of measures (and their standard errors) are not independent, so drawing confidence bands on Figure 2 is awkward. However those students whose two measures differ by more than a logit are statistically, and probably substantively, irregular.

John Michael Linacre

Does Item Order or Context Matter?

Tom Snider-Lotz asked, and some of the answers were ...

David Andrich: It is always an empirical matter with every particular data set (collection of items and persons) whether the item order is independent in the sense that is required by the model, and therefore that order will not matter. However, there are good testing reasons, reflected by the model, for having items independent. For example, we do not want one item imply the answer to another and so on. If students are to do a set of items, it is not helpful to independence and good testing to put the most difficult items first. It will upset the students and they will not be able to do the ones that they could do later in the test had they been earlier in the test.

Jack Stenner: We have conducted a number of [in-house] studies over the last decade on the effects of context on reading item calibrations. Context includes variation due to person sample, placement on the test, and resolution of location indeterminacy via a text analysis of all items on the test. We have found a rather consistent context effect of slightly more than .40 logits. Of course, the effect on the mean item difficulty, which is what matters most when making person measures, is reduced proportional to the square root of the number of items. Thus "ambient noise", which is what we call this irreducible variation in item difficulty, would contribute on average only .40/7 logits of error to the centering on a 49 item test.

Bryce Reeve: Lynne Steinberg did some analyses .... and found a context effect.

> "Question order effects [in questionnaires] have been found to reliably influence an item's item-total correlation (Knowles, 1988), item-trait correlation (Steinberg, 1994), slope parameter ..., and reliability (Knowles & Byers, 1996)." (Assessing Performance: Investigation of the Influence of Item Context using Item Response Theory Methods. Kuang, D.C., & Steinberg, L., 2004 Annual Meeting of the Society of Industrial and Organizational Psychology, Chicago, IL.)

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. Journal of Personality and Social Psychology, 57, 351-357.

Knowles, E. S. & Byers, B. (1996). Reliability shift in measurement reactivity. Driven by content engagement or self-engagement? Journal of Personality and Social Psychology, 70, 1080-1090.

Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. Journal of Personality and Social Psychology, 66, 341-349.

SBIR RFA 211: Developing Item Response Theory Software for Outcomes and Behavioral Measurement: Bryce Reeve, the NIH Project Officer, reports that several responses have been received to the RFA. These will be reviewed in early 2005, with notification of the granting of awards to follow in mid 2005.

My Best Items Don't Fit!

"I have a instrument of 13 items. I used the Partial Credit model. There are 3 items whose INFIT "t" statistic (ZSTD) is outside -9.9, and so significantly misfitting my Rasch model. But these items' score correlations are greater than 0.83. The discrimination of these three items' is higher than the other 10 items, so in classical test theory and much of IRT, these are my best three items. Should I say that the Rasch model is not suitable for this instrument, and maybe a Generalized (2-parameter) Partial Credit model analysis is indicated?

Desperate Test Constructor

t-statistics (ZSTD) are tests of the hypothesis "these data fit the model perfectly." In statistics this is called a "false null hypothesis", because it can never be true! No empirical data fit the Rasch model perfectly. So a more crucial question is, "Do the data fit the model usefully?" "Do they distort the measures more than they contribute to measurement accuracy and precision?"

Here are some steps to take in your investigation:
1. Those three items are over-discriminating from a Rasch perspective. Are these items really good items or are they substantively flawed? Look at the content of the items and refer to Geoff Masters (1988) "Item discrimination: when more is worse", Journal of Educational Measurement, 25:1, 15-29, and www.rasch.org/rmt/rmt72f.htm - RMT 7:2, 289.

2. Don't be intimidated by the statistics. What is your sample size? Is it making the hypothesis test too sensitive? See www.rasch.org/rmt/rmt171n.htm - RMT 17:1, p. 918. Are the mean-squares (chi-squares divided by their degrees of freedom) so close to their expectations that the differences have no substantive implications, despite being significantly unexpected?

3. Are the three items contributing to accurate measurement, or are they distorting measurement? The usual way to check this is to measure the persons with and without these 3 suspect items and cross-plot the person measures. Who is off the diagonal? Which set of measures better represent the abilities of your sample?

4. If these 3 items really are substantively "bad", changing the analytical model will not make them "good". A different model will merely hide the symptoms. So omitting the items is preferable to changing the analytical model.

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: Randy Schumacker, Secretary: Steve Stemler

Equating Constants with Mixed Item Types

Question: I am doing non-equivalent-groups common-item equating of two tests using the Rasch model. My common items are 20 multiple-choice items (worth 1 point) and 1 partial credit (essay) item worth 4 points. How do I compute an equating constant?

Response: Let's assume a scatter-plot of the difficulties of the 21 items on the two tests indicates that their pattern approximates an identity line. (If it does not, you need to investigate whether the "common items" really are common.) You now need to compute a defensible equating constant.

Check if the structure of the partial credit item has changed between tests. The default "item difficulty" for a partial credit item is the point at which top and bottom categories are equally probable. The bottom category of a long scale tends to be relatively rarely and idiosyncratically used. So, for equating purposes, it may be more robust to define the item difficulty to the be point at which the two most frequent categories (across the two tests) are equally probable. This provides a more stable, and statistically more secure, "item difficulty" for the polytomous item.

If your software does not report one overall partial-credit item difficulty in each analysis, but instead a set of threshold difficulties, then average the threshold difficulties for an overall difficulty.

A. The examination board may assume that the one 4-point partial credit item is equivalent to 4 dichotomous items. If so the equating constant between the two tests is:
((Sum of MCQ common-item difficulty differences) + 4*(partial credit difficulty difference)) / (20 + 4)

B. The examination board may want the one 4-point partial credit item to have the same influence as all 20 dichotomous items. If so the equating constant is:
((Sum of MCQ common-item difficulty differences) + 20*(partial credit difficulty difference)) / (20 + 20)

C. An option beloved of statisticians is "information-weighting" (i.e., weighting by the inverse of the item-difficulty-standard-error-squared). This will give the 4-point partial credit item about 6 times the influence of a 1-point dichotomous item.

D. Inverse-standard-error weighting ("effect-size" weighting) of the items is more consistent when combining subtests, see www.rasch.org/rmt/rmt83f.htm RMT 8:3, p. 376. This will give the 4-point partial credit item about 2.5 times the influence of a 1-point dichotomous item.

In this example, to specify that for 1 partial-credit item = 4 dichotomous items would also be to opt for a compromise between the two statistical viewpoints.

When does a Gap between Measures Matter?

When two item difficulty measures (or two person measures) are located along a latent variable, how big must the gap be for it to be important?

Gaps based on probabilities: for dichotomies, these correspond to what differential chance of success would matter. If a 60% chance of success is thought to be importantly different from a 50% chance, then the logit difference is 0.4 logits, so a gap of 0.4 logits matters. For polytomies, this calculation tends is more complex.

Gaps based on substance: these usually correspond to "what is the smallest difference that an informed observer would see to be definitely different"? In many educational situations a gap that matters is about 0.5 logits, roughly half a grade level at school.

Gaps based on statistical significance: these are computed from the standard errors of the individual measures. The more data usually the smaller the standard errors. So for .15 logits to represent a statistically significant gap (using a two-sided .05 *t*-test) between two measures, the individual measure standard errors must be about .05 logits, corresponding to about 250 dichotomous responses underlying each measure.

Gaps based on effect-size: these are used in education, e.g., where students whose abilities are 2 S.D.s above the sample mean ability are in a higher performing group.

For polytomies (rating scales, partial credit, etc.): The math is more complicated and probabilistic implications hard to explain, so it usually comes down to substance. Lai & Eton (2002, RMT 15:4, 850) report 0.5 logits to be a clinically meaningful gap for one instrument.

Equating/Linking with Anchors

Using pre-set "anchor" values to fix the measures of items (or persons) in order to equate the results of the current analysis to those of other analyses is a form of "common item" (or "common person") equating. Unlike common-item equating methods in which all datasets contribute to determining the difficulties of the linking items, the current anchored dataset has no influence on those values. Typically, the use of anchored items (or persons) does not require the computation of equating or linking constants. During an anchored analysis, the person measures are computed from the anchored item values. Those person measures are used to compute item difficulties for all non-anchored items. Then all non-anchored item and person measures are fine-tuned until the best possible overall set of measures is obtained. Discrepancies between the anchor values and the values that would have been estimated from the current data can be reported as displacements. The standard errors associated with the displacements can be used to compute approximate t-statistics to test the hypothesis that the displacements are merely due to measurement error.

Rasch with an Error Term

Question: Regression models include an explicit error term, why don't Rasch models?

Answer: The Rasch model is usually presented in a way which emphasizes its unique statistical properties, but it can be written to conform with a "general linear hypothesis" as :

$$X = E \pm \sqrt{W}$$

where X is the empirical observation and E is the expected value of the observation according to the relevant Rasch model. W is the error variance, specific to this observation, i.e., modeled as heteroscedastic, in contrast to the typical regression model in which the error variance is averaged across all observations, i.e., assumed to be homoscedastic.

The algebraic expressions for E and W are shown on p. 100 of *Rating Scale Analysis* (Wright & Masters, 1982). For dichotomous data they are

$$E_{ni} = P_{ni} = \frac{e^{B_n - D_i}}{1 + e^{B_n - D_i}}$$

and

$$W_{ni} = P_{ni}(1-P_{ni})$$

Rasch Workshops

March 21-22, 2005 – Monday-Tuesday, Chicago IL
July 25-26, 2005 – Monday-Tuesday, Chicago IL
Introduction to IRT/Rasch measurement using *Winsteps*
conducted by Ken Conrad & Nick Bezruczko
www.winsteps.com/workshop.htm

April 9-10, 2005 – Sat. -Sun., Montreal QU (pre-AERA)
An Introduction to Rasch Measurement: Theory and Applications
conducted by Richard M. Smith and Everett Smith
www.jampress.org

May 24-26, 2005 – Tuesday-Thursday, Dallas TX
Winsteps workshops
May 31-June 2, 2005 – Tuesday-Thursday, Dallas TX
Facets workshop
conducted by Mike Linacre
www.winsteps.com/seminar.htm

June 20, 2005 - Monday, Kuala Lumpur, Malaysia
Winsteps and Facets workshop
conducted by Mike Linacre
www.iiu.edu.my/proms&isme2005

July 27-28, 2005 – Wed.-Thursday, Chicago IL
Introduction to Many-Facet Rasch Measurement using *Facets*
conducted by Carol Myford & Lidia Dobria
www.winsteps.com/workshop.htm

Comparing Rasch variables?

Question: "I want to compare each individual students' math and science achievements on the same scale. I want to be able say if this student did better on math than on science (after taking into account the different level of test difficulties)."

Response: In this type of situation it is always helpful to think of what you would do in a similar practical physical situation. Pretend your two tests are "weight" and "height" of children. How would you proceed? You would have to make an assertion about the relationship between height and weight for your students.

So, for your math and science tests, you need to make an assertion (assumption) about their relationship. Common assertions include:
a. The test items are equally difficult, on average, for both samples (with equal item difficulty dispersion).
b. The samples are equally able, on average, on both tests (with equal person measure dispersion).
c. Particular items on the math test have the same difficulty as particular items on the science test.
d. Particular persons or groups of persons on the math test have the same ability as particular (perhaps the same) persons or groups of persons on the science test.

An attractive short-cut might be to do a joint calibration of the math and science items. But imagine we are comparing the weight and height of children. If we tried to force them both into the same numerical variable, it would skew results for both. So what we might do instead is to match the mean and standard deviations of the sample's weights and heights in order to make weight/height comparisons.

Measure the math ability of each of the students. Measure the science ability of each of the students. Implement your assertion as to how the two ability distributions relate. Then you can report individual relative performances on math and science.

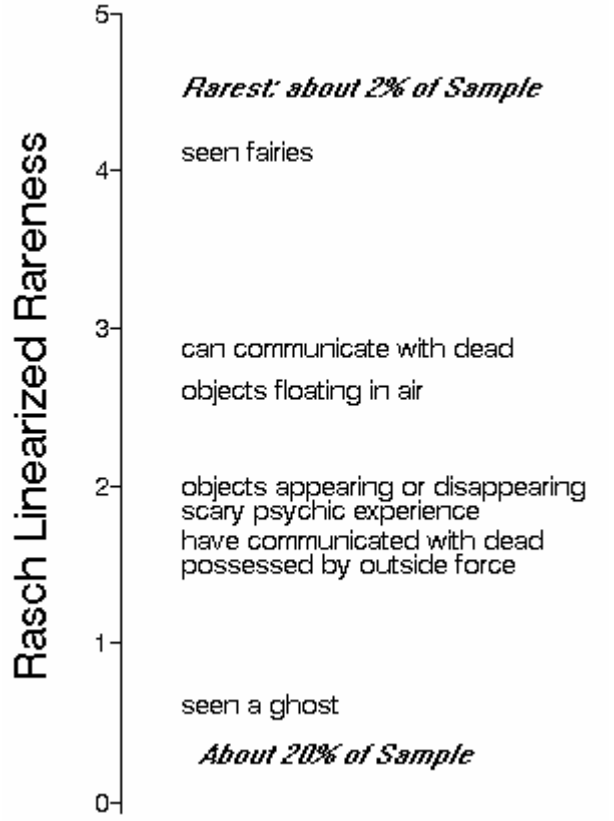
Computations with Rasch Measures

Question: My school is thinking of using the AGS Reading-Level Indicator to measure the progress of our reading program (student growth). We have the W-ability (Rasch) table that accompanies this measure. We would like to know what kind of statistical manipulations would be appropriate to use with these numbers (t-tests, ANOVA, etc.).

Answer: Wonderful! Rasch measures are designed and intended for all those arithmetical and statistical operations which expect their numbers to obey the rules of addition and subtraction, such as the computations of means and standard deviations. In the classification of S. S. Stevens, Rasch measures are "interval".

A Haunting Hierarchy

A first step to making sense of the apparently irrational is to organize it. That step is taken by James Houran and Rense Lange, "A Rasch Hierarchy of Haunt and Poltergeist Experiences", Journal of Parapsychology, 65, 41-58, 2001. Here is a map drawn from their Table 1. Experts in the field perceive this hierarchy to have construct validity.



Hard Science Sometimes Somewhat Soft!

"There can be personal bias in reading non-digital instruments or estimating certain quantities. This is referred to as the 'personal equation'. One example is a mercury thermometer. Mercury has a lower surface tension in contact with glass, than in contact with air, so a meniscus forms on top of the mercury column in a glass tube. This results in a subjective reading, notably in estimating values between the marks on the scale (1). Another example is the estimation of cloudiness. Apparently many people prefer reporting cloudiness as 1 or 3 or 7 oktas rather than other values (2). Sometimes a discontinuity in a time series of station temperature, cloudiness, or other meteorological variable can be attributed to a staff change. The personal equation is an important factor in other sciences such as anthropometry, demography, geography, and physics (1)."

E.T. Linacre [no relation to Editor of RMT]

(1) Cox, N.J. 1991. Human factors. Nature 353, 597.
(2) Linacre, E.T. 1992. Climate Data & Resources. Routledge.