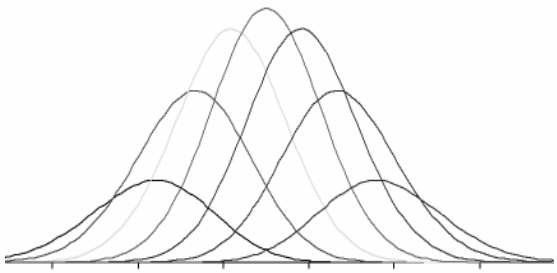
Likelihood of a Raw Score



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 18 No. 2 Autumn 2004 ISSN 1051-0796

Raw Score Nonlinearity Obscures Growth

Historians and philosophers of science generally agree that measuring linear change lies at the foundations of modern science separating it from Aristotelian physics in the 17th century (Burtt, 1924). Ever since Galileo described laws of terrestrial motion and Newton generalized them to celestial bodies, empirical measures of linear change have advanced theoretical science. Not surprisingly, contemporary social theory suffers desperately from a profound inability to measure linear change with dismal implications for social science.

Cronbach is widely recognized for describing raw score problems measuring change but abandoned this challenge to improve social research methodology by advising researchers to "frame their questions in other ways" (Cronbach & Furby, 1970, p. 80). Virtually all commentaries and methods proposed since Cronbach have attempted to dismiss the limitations of measuring change with raw scores by offering feeble substitutes. In general, these approaches are complicated (Cohen & Cohen, 1975), as well as controversial (see Ragosa et al., 1982), while Collins and Horn (1991) suggest only analyzing change not measuring it.

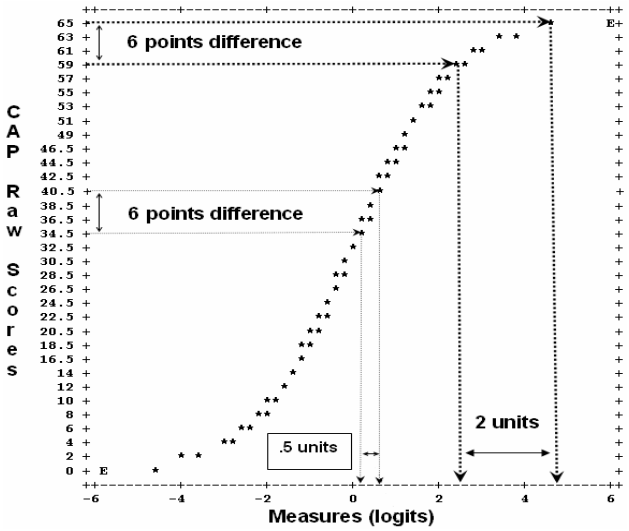


Figure 1. Raw score to Linear Measure relationship

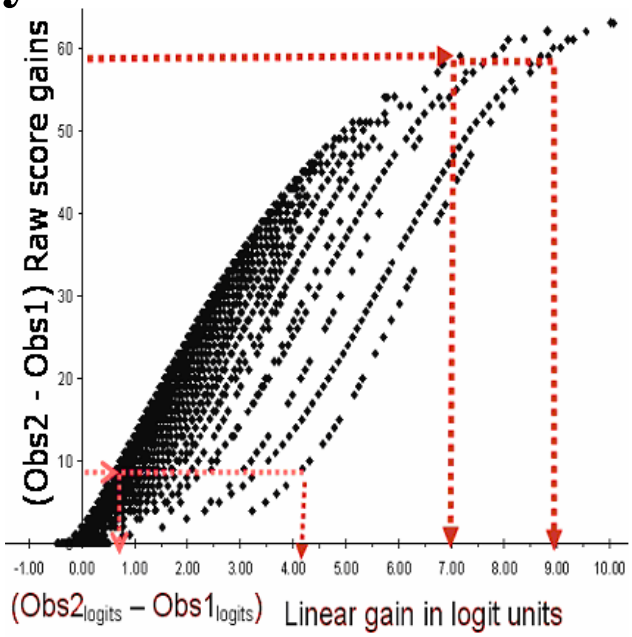


Figure 2. Raw score gains vs. Linear gains.

While the literature is full of discussions about raw score problems in measuring change such as low reliability, spurious negative correlations with initial status, and lack of constant meaning (see Embretson & Reise, 2000), few researchers understand why raw scores are fundamentally flawed. Consequently, the purpose of this report is to present an example of measuring change with both raw scores and linear units that may help clarify this problem.

The data are 25,000 CAP (Child Assessment Profile; Chicago Public Schools, 1993-2002) raw score records that were collected by Chicago preschool teachers in

Table of Contents

September (Observation 1) and the following June (Observation 2). CAP consists of 65 items that represent five domains of early childhood learning sampled across a hierarchical construct. Teachers dichotomously (0/1) score each child and higher summed scores indicate higher CAP growth. Separation reliability is high (> .95) and person/item fit excellent . Typical CAP items are "Count to 10", "Names colors", and "Writes own name".

Figure 1 shows initial transformation of CAP raw scores to linear logit measures, and, as expected, upper and lower tails show substantial raw score distortions. A six point raw score difference in upper tail is four times greater when represented with logits. While disturbing, these distortions commonly occur in raw score analyses.

Figure 2 presents a less well known relationship between raw scores and linear measures with important implications for understanding meaningful gain measurement. On the vertical axis appears CAP raw score change between Observations 2 and 1 (Obs2 raw scores – Obs1 raw scores). Then CAP raw scores were transformed to logits and their differences (Obs2 logits – Obs1 logits) appear along the horizontal axis. Consequently, Figure 2 shows ordinal raw score differences matched with corresponding linear differences. Some prominent results are:

1. Every linear difference (Obs2 logits – Obs1 logits) corresponds to a range of raw score differences. The raw score range differs depending on Observation 1 initial status (see Figure 1).

2. The range of raw scores corresponding to logits is enormous. For example, CAP gain measuring four logits corresponds to a raw score range between 10 and 45 points! Conversely, a 10 point raw score gain corresponds to a logit range between .7 and 4 units.

3. As logit values increase, raw score range diminishes. CAP gain of 7 logits corresponds to a raw score range between 40 and 60 points, while a gain of 9 logits only corresponds to a range between 55 and 60 points.

These results show that raw score change is virtually impossible to interpret with meaning and accuracy because position on the measurement construct is confounded with a nonuniform metric. A 10-point raw score change, for example, appears to be the same for both high and low ability children but thee corresponding linear values show the change for high ability children to be four times greater. A practical result of raw score interpretation is conflation of growth with nonlinearity.

The effect of nonlinearity on growth interpretation in this example may be considered relatively benign because CAP is not high stakes and all children advance to kindergarten. But most children start lower down where raw score distortion tends to inflate growth while underestimating child growth near the top. Moreover, these raw score results suggest that children are much more homogenous than linear units would indicate and this distortion is further obscured when child scores are aggregated by preschool centers.

Although program evaluators will find almost everyone improving on this assessment, raw score convolutions are too complicated to establish useful normative growth expectations or isolate child abnormalities. The common strategy of aggregating raw scores into summary statistics only obscures the underlying problem of unequal scale intervals and eliminates an opportunity to understand individual growth.

Nikolaus Bezruczko

Burtt, E. A. (1924). The Metaphysical Foundations of Modern Science. New York: Doubleday Anchor.

Cohen, J., Cohen, P. (1975). Applied Multiple Regression/Correlation Analysis for Behavioral Sciences. Hillsdale, NJ: Erlbaum.

Collins, L. M., Horn, J. L. (1991). Best Methods for Analyzing Change. Washington, DC: American Psychological Association.

Cronbach, L. J., Furby, L. (1970). How should we measure "change" - or should we? Psychological Bulletin, 74, 68-80.

Embretson, S. E., Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, NJ: Erlbaum.

Ragosa, D., Brandt, D., Zimowski, M. (1982). A growth curve approach to measurement of change. Psychological Bulletin, 92, 726-748.

Opening Remarks: International Objective Measurement Workshop XII

June 30, 2004, Cairns, Australia

Prof. Bernard Moulden:

Good morning everybody, and welcome to the Twelfth International Objective Measurement Workshop, hosted this year by James Cook University in Cairns.

Because we are geographically a little off the beaten track in this little corner of paradise, you will understand that I can't resist the opportunity to give you a 60-second burst of bragging about the university that of which I am privileged to be the Vice Chancellor and President.

A few years ago James Cook University committed itself to the goal of becoming one of the top five universities of the world enhancing life in the tropics through education and research. At the time that might have seemed a bit of a stretch target for some, but recent objective evidence shows that in fact we are well on the way to achieving it.

I know you are all interested in evidence and here is a piece of evidence that I like a lot. The recent survey by researchers at Shanghai's Jiao Tong university identified the top 500 universities of the world in terms of their research performance. That survey placed Harvard, Stanford, Caltech, and UC Berkeley at the top of the list. In fact it showed that the USA was home to 160 of the world top 500; Germany and the UK have about 40 each, and that Australia has just 13 universities in the world Top 500. Now of course I wouldn't be telling you this if it wasn't for the fact that James Cook University is one of those top 13, one of only three to be located outside of a capital city, and one of only two in Queensland - but wait, there's more.

Obviously a big university will nearly always produce more than a small one - but if you measure not total output but research intensity, by dividing output by the number of staff - then you discover that JCU ranks number three in Australia, behind ANU and Macquarie and, and with a research intensity score almost double that of the University of Queensland.

Other evidence shows that if we look just at the universities located in the tropical regions JCU ranks in the top dozen in the world, and what is more, it shows that in some disciplines the impact measures of our scientist's research - the number of times their work is cited by others - puts us in the top three or four in the world.

So there you are - I bet you didn't know that before, and I bet you feel a lot better now that you do. It certainly makes me feel good.

Once upon a time - half a lifetime ago - I was a Professor of Psychology. I worked at what some of my colleagues called the "hard" end of the discipline, on the neurophysiology of vision. They worked at what *I* called the "'soft" end, in what seemed to me to be a context of intrinsically untestable theory and either, on the one hand, a complete absence of quantitative data or, on the other, a wealth of data of indeterminate validity and an interpretability status that I could only charitably describe as astrological. Needless to say, we didn't talk much.

Until around 1970, the advance of science had generally been assumed to be smoothly cumulative. Then Thomas Kuhn published his remarkable book "The Structure of Scientific Revolutions" and established the notion that science proceeds in punctate steps, as one paradigm of thought replaces a previous one. Many people believe that Rasch analysis, or perhaps more generally Item Response Theory, constitutes a significant enough change in thought and approach to social sciences to merit the status of a genuine paradigm shift.

Indeed, in 2003 Mark Blais, of Harvard Medical School, wrote a book review entitled ***"Have you heard we're having a revolution? The coming of modern test theory"*** *[Journal of Personality Assessment, 80, 2: 208-210].* The book in question was of course Bond & Fox's ambitiously titled ***"Applying the Rasch model: Fundamental Measurement in Human Science".*** Ambitious it may have been, but Blais was clearly converted: "This is a great book", he said, "and reading it...might just make you part of the quiet revolution in test development." (Trevor Bond can make the usual commission payments to the Vice Chancellor's special account.)

Having seen the briefing notes for your conference I'm in no doubt that a genuine revolution *has* occurred, and I suspect that it is well on the way to robbing the 'hard science/soft science' dimension of any reality that it may ever have had. I envy you the exciting sense of redefining the frontiers that you must all be enjoying, and I wish you well in your enthusiastic development of the new paradigm. From what I said at the outset I have no doubt you will find that James Cook University provides the ideal intellectual environment and context for your scholarly activities.

Colleagues, I apologize for not being there in person to greet you, and I can't even use pressure of work as an excuse because in June I shall be on recreational leave in Europe. I hope that northern Queensland is living up to its reputation as being glorious one day and perfect the next, and round about now I shall be thinking of you with envy and probably longing to be home. Even from the Loire Valley I shall be envying you your immersion in stochastic Guttman ordering, conjoint additivity, Campbell concatenation, sufficiency, and infinite divisibility.

Thank you for listening, enjoy yourselves, and welcome.

Plausible Values

Plausible values were first developed for the analyses of 1983-84 NAEP (National Assessment of Educational Progress) data, by Mislevy, Sheehan, Beaton and Johnson, based on Rubin's work on multiple imputations. Plausible values were used in all subsequent NAEP surveys, TIMSS and now PISA.

What Plausible Values Are

The simplest way to describe plausible values is to say that plausible values are *some kind* of student ability estimates. There are some differences between plausible values and the θ (student ability parameter) as in the usual 1, 2 or 3-PL item response models. Instead of directly estimating a student's θ, we now estimate a probability distribution for a student's θ. That is, instead of obtaining a point-estimate for θ, we now come up with a range of possible values for a student's θ, with associated likelihood of each of these values. Plausible values are random draws from this (estimated) distribution for a student's θ (I will call this distribution "the posterior distribution").

Mathematically, we can describe the process as follows: Given an item response pattern **x**, and ability θ, let $f(\mathbf{x}/\theta)$ be the item response probability, $f(\mathbf{x}/\theta)$ could be 1, 2 or 3-PL model, for example). Further, we assume that θ comes from a normal distribution $g(\theta) \sim N(\mu, \sigma^2)$. (In our terminology, we often call $f(\mathbf{x}/\theta)$ the item response model, and g (θ) the population model). It can be shown that, the posterior distribution, $h(\theta/\mathbf{x})$, is given by

$$h(\theta/\mathbf{x}) = \frac{f(\mathbf{x}/\theta)g(\theta)}{\int f(\mathbf{x}/\theta)g(\theta)d\theta}$$

That is, if a student's item response pattern is **x**, then the student's posterior (θ) distribution is given by $h(\theta/\mathbf{x})$. Plausible values for a student with item response pattern **x** are random draws from the probability distribution with density $h(\theta/\mathbf{x})$. Therefore, plausible values provide not only information about a student's "ability estimate", but also the uncertainty associated with this estimate.

If we draw many plausible values from a student's posterior distribution $h(\theta/\mathbf{x})$, these plausible values will form an empirical distribution for $h(\theta/\mathbf{x})$ (as plausible values are observations from $h(\theta/\mathbf{x})$). So if a data analyst is given a number of plausible values for each student, an empirical distribution of $h(\theta/\mathbf{x})$ can be built for that student. This is done because there is no nice closed form for $h(\theta/\mathbf{x})$ to give to data analysts, except for through the empirical way (plausible values) (unless you have ConQuest). Typically, 5 plausible values are generated for each student (and deemed sufficient to build an empirical distribution!)

As plausible values are random draws from a student's posterior distribution, plausible values are not appropriate to be used as individual student scores for reporting back to the students. Suppose two students have the same raw score on a test, their plausible values are likely to be different as these are random draws from the posterior distribution. Imagine the outcry if we ever give two students different ability scores when they have the same raw score. However, plausible values are used to estimate population characteristics, and they do a better job than a set of point estimates of abilities. I will give more details about this in the next section. In NAEP, TIMSS and PISA, we do not report individual scores. We only estimate population parameters such as mean, variance and percentiles.

Why We Need Plausible Values

So why are plausible values used?

(1) Some population estimates are biased when point estimates are used to construct population characteristics.

(2) Secondary data analysts can use "standard" techniques (e.g., SPSS, SAS) to analyze achievement data provided in the form of plausible values.

(3) Plausible values facilitate the computation of standard errors of estimates for complex sample designs.

Plausible Values versus Point Estimates

If we are interested in the population characteristics of some ability, Θ, one way to go about it is to compute an estimate for each student, $\hat{\theta}_n$, and then compute the mean, variance, percentiles, etc. from these $\hat{\theta}_n$.

Consider two possible estimates for $\hat{\theta}_n$: the Maximum Likelihood Estimate (MLE) and the Expected A-Posteriori estimate (EAP). In the case of the 1-parameter (Rasch) model, MLEs are ability estimates that maximise the likelihood function of response patterns

$$\prod_n \prod_i \frac{\exp[x_{in}(\theta_n - \delta_i)]}{1+\exp(\theta_n - \delta_i)},$$

where i is the index over items, and n is the index over people, and x_{in} is the item response (0 or 1) for person n on item i. We use the formula for dichotomous items, for simplicity. That is, MLE estimates only involve the item response model. There is no assumption about the population model.

Mean and Variance

It can be shown that if $\hat{\theta}_n$ s are MLEs, the mean of $\hat{\theta}_n$ is an unbiased estimate of μ, the population mean of Θ. But the variance of $\hat{\theta}_n$ is an **over-estimate** of σ^2, the population variance. But if our $\hat{\theta}_n$ s are EAPs (e.g., ability estimates from Marginal Maximum Likelihood MML models), where we assume an item response model, e.g., $f(\mathbf{x}/\theta)$, it can be shown that the mean of EAPs is an unbiased estimate of the population mean, μ, but the variance of the EAPs is an **under-estimate** of σ^2. In both

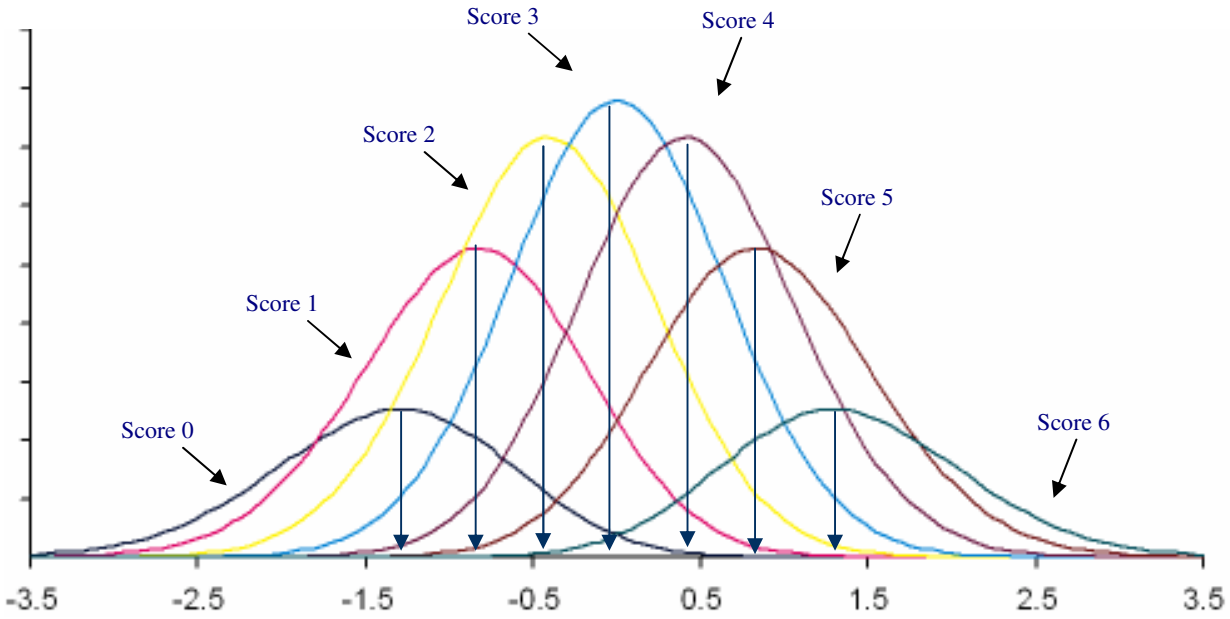


Figure: Proficiency on Logit Scale

the MLE and EAP cases, the bias does not go away when the sample size increases. The bias is reduced when the number of items increases, and can be removed by a mathematical disattenuation.

One way to overcome the vairance bias problem is to use the MML and directly estimate μ and σ^2 without going through the step of computing individual $\hat{\theta}_n$. This is possible with MML because we can integrate out the ability parameter θ in the likelihood equation:

$$f(x) = \int f(x/\theta) g(\theta) \, d\theta$$

so that the parameters to estimate are only δ_i (item difficulties), and μ and σ^2 (population parameters). Such direct estimation method will give unbiased results for μ and σ^2. This is what ConQuest does. But most data analysts do not have ConQuest or other similar software that will enable them to do this direct estimation easily. Data analysts have available to them general statistical software such as SPSS and SAS. To allow the data analysts to compute the correct estimates of population parameters, plausible values are provided.

Recall that plausible values are random draws from each student's posterior distribution. The collection of posterior distributions for all students, put together, gives us an estimate of the population distribution, $g(\theta)$. Therefore, we can regard the collection of plausible values (over all students) as a sampling distribution from $g(\theta)$. This is an important statement, and some results follow from this statement:

(1) Population characteristics (e.g., mean, variance, percentiles) can be constructed using plausible values.

(2) Supppose we generate 5 plausible values for each student, and form 5 sets of plausible values (set 1 contains the first plausible value for each student; set 2 contains the second plausible value for each student, etc.). Then each set is equally as good for estimating population characteristics, as each set forms a sampling distribution of $g(\theta)$. It follows that, even if we only use one plausible value per student to estimate population characteristics, we still have unbiased estimates, in contrast to using each student's EAP estimates (mean of plausible values for each student) and getting biased estimates. So the apparent paradox is that using one random draw (PV) from the posterior distribution is better than using the mean of the posterior, in terms of getting unbiased estimates.

Percent Below Cutpoint and Percentiles

The following example shows why point estimates are not the best for estimating percent in bands or percentiles. Suppose we have a 6-item test, so students' test scores range from 0 to 6. The Figure above shows the 7 (weighted) posterior distributions, corresponding to the 7 possible scores, and the corresponding EAP estimates (shown by the black vertical lines).

Averaged over 10 replications:	MLE	EAP	PV1	PV2	PV3	PV4	PV5	Direct Estimate	Generating value
Ability variance	-0.05	-0.05	-0.05	-0.04	-0.06	-0.04	-0.05	-0.05	0
Ability mean	1.45	0.78	1.01	0.99	1.01	1.00	1.01	1.00	1

Suppose we are interested in the proportion of students below a cutpoint, say –1.0. If we use EAP estimates, then the proportion of people below –1.0 is the proportion of people obtaining a score of 0. In fact, for any cutpoint between EAP_0 and EAP_1, we obtain this same proportion because the (EAP) ability estimates are discrete, not continuous. In contrast, if we look at the area of the curves of the posterior distributions that is below –1.0, we see that this is a continuous function, and that this area contains contributions from all posterior distributions (corresponding to all scores).

As plausible values are random draws from the posterior distributions, the proportion of plausible values below a cutpoint gives us an estimate of the area of the posterior distributions below that cutpoint. By using plausible values, we are able to overcome the problems associated with the discrete nature of point estimates. Similarly, for percentiles, using plausible values will overcome the problem of having to interpolate between discrete ability estimates.

Some Simulation Results

Some simple simulation results are shown in the Table. A data file containing student responses was generated for a 20-item test with 300 students whose abilities were sampled from N(0,1). MLE, EAP and 5 PVs were computed for each student, and the sample mean and variance (across students) were computed for each of these estimates. This process was repeated 10 times (10 replications). Plausible values (and direct estimation) do a better job for estimating the population variance. That is, had we provided data analysts with students' MLE (or EAP) ability estimates, they would not be able to recover the variance (and other statistics such as percentiles) correctly.

Margaret Wu,
Australian Council for Educational Research

Beaton, A. E. & Gonzalez, E. (1995). NAEP Primer. Chestnut Hill, MA, Boston College.
Journal of Educational Statistics (Summer 1992) Special Issue: NAEP.
Journal of Educational Measurement (Summer 1992) Special Issue: NAEP

"The significant problems we face cannot be solved at the same level of thinking we were at when we created them."
Albert Einstein (1879-1955)

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

Developing Item Response Theory Software for Outcomes and Behavioral Measurement

Solicitation of the Public Health Service for Small Business Innovation Research contract proposals #211.
Proposal receipt date: **November 5, 2004**

http://grants.nih.gov/grants/funding/sbir.htm

Excerpts from the Solicitation:

The goals of this topic are to develop and/or adapt software that employs both traditional and modern measurement methods [i.e., item response theory (IRT) modeling] to respond to the needs of cancer outcomes, health surveillance, and behavioral researchers. Software should be user-friendly, flexible, and inclusive of a variety of IRT models for both dichotomous and polytomous response data, with sophisticated graphic capabilities, tests of model fit, and extensions of the software for multi-dimensional modeling, testing for differential item functioning, linking questionnaires, and computerized-adaptive testing.

Phase I Activities and Expected Deliverables: The contractor should consult with both leading psychometricians who have experience in IRT modeling and health outcomes, health surveillance, and behavioral researchers who have a range of training in measurement to help shape the functionality and presentation of the software and literature to be developed in Phase II. Deliverables should include: (1) a complete program design and specification, (2) an outline of the manual and primer, and (3) a prototype of the software that responds to the minimal changes recommended in this proposal. Offerors may request a one year Phase I.

Phase II Activities and Expected Deliverables: Develop the full IRT software and supporting documents based on Phase I findings including beta-testing of the software on a variety of datasets among healthcare researchers with a variety of measurement backgrounds. Also, develop a curriculum, evaluation measures, and other educational materials designed to integrate this software into the healthcare community. Deliverables will include: (1) the software, (2) the manual, primer, and other educational materials, and (3) at least one article describing the development and evaluation of the program that is suitable for publication in appropriate scientific journals and/or books.

Phase I awards will be firm fixed price contracts. Normally, Phase II awards will be cost-plus-fixed-fee contracts. Normally, Phase I contracts may not exceed **$100,000.** Phase II contracts normally may not exceed **$750,000.** Work must be performed by a Small Business in the USA.

Relevant contact: Bryce Reeve, reeveb@mail.nih.gov

Book: Introduction to Rasch Measurement: Theory, Models, and Applications

Edited By Everett V. Smith, Jr. and Richard M. Smith

24 chapters (700 pages, $47), written by the leading experts in Rasch measurement. More details are at www.jampress.org

I. Foundations

1. An Overview of the Family of Rasch Measurement Models. *Benjamin Wright, University of Chicago Magdalena Mo Ching Mok, Hong Kong Institute of Education*
2. Estimation Methods for Rasch Measures. *J. Michael Linacre, University of Sunshine Coast*
3. Rasch Model Estimation: Further Topics. *J. Michael Linacre, University of Sunshine Coast*
4. Fit Analysis in Latent Trait Measurement Models. *Richard Smith, Data Recognition Corporation*
5. Evidence of the Reliability of Measures and Validity of Measure Interpretation: A Rasch Measurement Perspective. *Everett Smith, University of Illinois at Chicago*
6. On Choosing a Model for Measuring. *Mark Wilson, University of California, Berkeley*
7. Controversy and the Rasch Model: A Characteristic of Incompatible Paradigms? *David Andrich, Murdoch University*
8. Understanding Resistance to the Data-Model Relationship in Rasch's Paradigm: A Reflection for the Next Generation. *David Andrich, Murdoch University*

II. Basic Applications

9. Substantive Scale Construction. *Mark H. Stone, Adler School of Professional Psychology*
10. Rasch Measurement: The Dichotomous Model. *Randall Schumacker, University of North Texas*
11. Optimizing Rating Scale Category Effectiveness. *J. Michael Linacre, University of Sunshine Coast*
12. Partial Credit Model and Pivot Anchoring. *Rita Bode, Rehabilitation Institute of Chicago*
13. Construction of Measures from Many-Facet Data. *J. Michael Linacre, University of Sunshine Coast Benjamin Wright, University of Chicago*
14. An Introduction to Multidimensional Measurement using Rasch Models. *Derek C. Briggs, University of Colorado Mark Wilson, University of California, Berkeley*
15. Metric Development and Score Reporting in Rasch Measurement. *Everett Smith, University of Illinois at Chicago*

III. Advanced Applications

16. Equating and Item Banking with the Rasch Model. *Edward Wolfe, Michigan State University*
17. Detecting Item Bias with the Rasch Model. *Richard Smith, Data Recognition Corporation*
18. Rasch Techniques for Detecting Bias in Performance Assessments: An Example Comparing the Performance Native and Non-Native Speakers on a Test of Academic English. *Catherine Elder, University of Auckland Tim McNamara, University of Melbourne Peter Congdon, Victorian Curriculum and Assessment Authority*
19. Objective Standard Setting (or Truth in Advertising). *Gregory Stone, University of Toledo*
20. Detected and Measuring Rater Effects using Many-Facet Rasch Measurement: Part I.
21. Detected and Measuring Rater Effects using Many-Facet Rasch Measurement :Part II. *Carol Myford, University of Illinois at Chicago Edward Wolfe, Michigan State University*
22. Detecting and Evaluating the Impact of Multidimensionality using Item Fit Statistics and Principal Component Analysis of Residuals. *Everett Smith, University of Illinois at Chicago*
23. Computer Adaptive Testing. *Richard Gershon, Northwestern University*
24. The Rasch Model, Additive Conjoint Measurement, and New Models of Probabilistic Measurement Theory. *George Karabatsos, University of Illinois at Chicago*

Why Not Estimate Ability by Merely Adding Up Item Difficulties?

Question: "One thing that I struggle with is why one doesn't add up the item difficulties for the items endorsed/answered correctly in order to estimate a person's ability level?"

Response: Yes, this can be perplexing. Let's look at the simplest case. Suppose that we have 100 equally difficult items. which are at the local origin of the scale, and so are of 0 logit difficulty. A high performer would get most of them, say 90, correct. A low performer might get 10 correct. What ability levels do these performances indicate? No matter how we add up, subtract or multiply the item difficulties, the high performer would be reported with an ability of 0 logits, exactly the same as the low performer. But this is the type of test which E. L. Thorndike (1904) stated to be ideal. Indeed, the Rasch model can be derived from Thorndike's criteria (RMT 14:3, p. 763). In fact, this simple case is an example of Bernoulli binomial trials. The logit ability of the high performer is the average item difficulty, 0, plus the log-success-to-failure-ratio, log(90/10) = 2.2 logits. The logit ability of the low performer is the item difficulty, 0, plus the log-success-to-failure-ratio, log(10/90) = -2.2 logits.

From this example, we can see that more is required to estimate ability than merely adding up item difficulties.

John M. Linacre

Thorndike, E.L. (1904). An introduction to the theory of mental and social measurements. New York: Teacher's College.

Ordinal vs. Ratio Revisited Again

Roberts (1994, pp. 625-6) points out that, in measuring:
"...one seeks to assign numbers to objects so that *a* is judged louder than **b** if and only if the number assigned to *a* is greater than the number assigned to *b.* Such a mapping from objects to numbers is called a homomorphism from the observed relation to the numerical relation. In measurement theory, scales are identified with homomorphisms. Formally, an admissible transformation of a scale is then a transformation of the numbers assigned so that one gets another homomorphism."

Thus, Roberts (p. 626) continues:
"One of the goals of research in the theory of measurement is to develop a collection of tools which can be used to determine what assertions can meaningfully be made and what conclusions can meaningfully be drawn, using scales of measurement. A statement involving scales of measurement is called meaningful if its truth value is unchanged whenever every scale in the statement is modified by an admissible transformation. This definition goes back to Suppes (1959) and Suppes & Zinnes (1963). (While it is not mentioned explicitly in the work of Stevens, it is inherent in his treatment of admissible transformations; see Mundy (1986).)"

Roberts (p. 627) concludes his summary of measurement theory saying:
"The notion of meaningfulness is concerned with which assertions it makes sense to make, and which ones are just artifacts of the particular version of the scale of measurement that happens to be in use. This notion of meaningfulness is closely related to the concept of invariance in classical geometry."

"The definitions we have given are reasonably well accepted, at least to the extent that it is widely agreed that 'invariance' is a desirable condition and that it is implied by 'meaningfulness'."

The Consequences of Ordinal Status for Measurement
Roberts (pp. 628-30) then gives a list of meaningful and meaningless statements, and shows the logical fallacy involved in averaging and comparing raw scores. One example shows that we have the strong tendency to treat ordinal scales as interval, contrary to the empirical fact that the spacing between the categories is unknown. When this fact of unknown and likely variable spacing is recognized, we see that the categories may be acceptably scored by any algorithm that maintains their order, no matter how different the spacing between them. That is, ordinal homomorphisms do not restrict the spacing of the categories, but only their order, since the spacing is unknown.

Roberts gives the example of two groups of three individuals each rated once on a five-point scale, scored, as is commonly deemed the natural way of proceeding, as 1, 2, 3, 4, 5. Group 1 scores 4, 4, and 4; group 2 scores 5, 4, and 1. Group 1's mean of 4 is higher than group 2's mean of 3.33.

Now, given that we recognize and accept our scale's status as ordinal, the ratings may be transformed in any way that invariantly preserves their order. A logical and scientific way of proceeding to test the hypothesis of the group difference would then require that we try out different admissible transformations of the scale to see if we obtain the same result. Roberts accordingly rescores 5s as 200, 4s as 100, 3s as 50, 2s as 20, and 1s as 3. Now group 1 has a mean of 100, and group 2 has a mean of 101.

The change in the ordering of the groups in the context of an admissible transformation of the raw scores renders any test of a hypothetical average difference between the groups undecidable; the failure of invariance makes any statement about the groups' order meaningless. Roberts notes that comparing the group medians would be meaningful, since the order would always be preserved across admissible transformations.

Though Roberts does not go into it, we see in this example why ordinal comparisons are commonly justified within the context of normal distributions and similar standard deviations. The two groups of scores in Roberts' example have significantly different standard deviations (Group 1 SD = 0; Group 2, 2.08). Were the scores in Group 1 more dispersed, or those in Group 2 less so, the original scoring's order would more likely be preserved across permissible transformations.

Even though similar and normally distributed variation across groups can aid in preventing meaningless assertions, ones that "are just artifacts of the particular version of the scale of measurement that happens to be in use," a number of other problems dog ordinal scores (Wright & Linacre, 1989). As was recognized by Wilson (1971):

> "The ordinal level of measurement prohibits all but the weakest inferences concerning the fit between data and a theoretical model formulated in terms of interval variables.... The task of developing valid, reliable interval measurement is not a technical detail that can be postponed indefinitely while the main efforts in sociological research are devoted to substantive theory construction; rather it is the central theoretical and methodological problem in scientifically-oriented sociology."

It is in this context that one sees the real truth and value of an opinion widely held among natural scientists and often attributed (Wise, 1995, p. 11) to Ernest Rutherford, winner of the 1908 Nobel Prize in Chemistry, namely,

that if your experiment requires statistics, then you should have designed a better experiment. This opinion is expressed by Feinstein (1995), the long-time editor of the Journal of Clinical Epidemiology, in his critical examination of meta-analytic methodology. The implication is that when measurement is realized, it provides all the relevant information needed to make informed judgments about more and less.

Roberts provides another 30 pages of analyses concerning the kinds of conclusions that may be logically drawn from different scales of measurement in different contexts. He does not take up the problem of how interval/ratio scales might be calibrated on the basis of ordinal observations.

Ordinal to Ratio

To take up this question ourselves requires first of all recognizing that the rating scale is simply a generic way of labeling observations that we suppose involve some increasing amount of something. At the start of a new investigation into a new construct, we do not know how much increase is represented by any transition across categories, or even whether any increase at all is represented by these transitions. It may, after all, turn out that the construct cannot be quantified, or that the items and/or people brought together to explore the construct's quantitative status do not work well together, and so falsify the quantitative hypothesis.

Accordingly, how the categories are labeled is irrelevant. The labels are there only so that we can unambiguously distinguish them from one another and place them in ascending qualitative order according to some construct theory. The object of our interest is how many observations are labeled by each category. When that is ascertained, then we can estimate the log-odds that any respondent will reply in any one of the categories relative to any other category, for any item or group of items. The numbering of the rating scale categories is merely a convenience to facilitate thinking and to simplify the log-odds estimation procedure.

As is demonstrated in numerous developments in Rasch measurement theory and applications (Andrich, 1978a, 1978b; Linacre, 1999, 2002; Wright & Masters, 1982), this analysis reveals whether the rating categories are in fact ordered as hypothesized, and, if so, what their actual spacing is. Each numeric unit increase in the measures homomorphically maps the observed relation onto the numeric relation. The log-odds unit provides a ratio scale in the sense that any meaningful difference between two ratings, two items, two respondents, a respondent and an item, or a respondent and a category on an item could be identified as the smallest meaningful unit of measurement, and all other differences could be scaled in that unit. In other words, any magnitude difference can be divided up into any number of smaller ratio-unit differences, or divided into any number of larger ones, with no change in either the order or the proportionate spacing of any individual measures or group averages.

Admissible transformations for ratio scales are then those that preserve both the order of the relations as well as the magnitude of their proportionate spacing. Had the measures given in Roberts' example of the meaninglessness of averaged ordinal scores been ratio, all permissible transformations would have invariantly maintained the same proportionate difference between the individual measures and between the groups' average measures.

The Structure of Scientific Laws

Roberts closes his article with speculations based in Luce's (1959, 1990) classic article on the ratio form of scientific laws in general. When both independent and dependent variables are ratio scales, scientific laws are power laws. In Ohm's Law, for instance, voltage is proportional to current when resistance is fixed.

These comments echo similar observations made by Rasch (1960, pp. 110-5) concerning the identical form shared by his model for reading measurement and Maxwell's model for the relations of mass, force, and acceleration. Just as force is proportional to acceleration when mass is fixed, so, too, is reading ability proportional to reading comprehension when the reading difficulty of the text is fixed.

Roberts (p. 664) points out that researchers have been able to establish psychological laws that conform with Luce's method "only in rather limited circumstances." This conclusion would seem to clash with the widespread applicability to an enormous variety of data types enjoyed by Rasch's models. Rasch software routinely 1) scales both the independent and dependent variables in ratio form, and 2) assesses and isolates failures of invariance via fit analysis, overcoming both of the major barriers to identifying and testing scientific power laws.

Perhaps because construct theory continues to be underdeveloped, the value of the laws established by means of Rasch scaling remains under-appreciated. The invariant stability of the qualitative relations quantified in Rasch measurement constitutes a fundamental form of capital. But much remains to be done before the human and economic value of that capital is leveraged in practical applications.

William P. Fisher, Jr.

Andrich, D. A. (1978a). A binomial latent trait model for the study of Likert-style attitude questionnaires. British Journal of Mathematical and Statistical Psychology, 31, 84-98.

Andrich, D. A. (1978b). A rating formulation for ordered response categories. Psychometrika, 43, 357-374.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. Journal of Clinical Epidemiology, 48(1), 71-79.

Linacre, J. M. (1999). Investigating rating scale category utility. Journal of Outcome Measurement, 3(2), 103-22.

Linacre, J. M. (2002). Understanding Rasch measurement: Optimizing rating scale category effectiveness. Journal of Applied Measurement, 3(1), 85-106.

Luce, R. D. (1959). On the possible psychophysical laws. Psychological Review, 66, 81-95.

Luce, R. D. (1990). "On the possible psychophysical laws" revisited: Remarks on cross-modal matching. Psychological Review, 97, 66-77.

Mundy, B. (1986). On the general theory of meaningful representation. Synthese, 67, 391-437.

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).

Roberts, F. S. (1994). Limitations on conclusions using scales of measurement. In A. Barnett, S. Pollock & M. Rothkopf (Eds.), Operations research and the public sector (pp. 621-671). Amsterdam: Elsevier.

Suppes, P. (1959). Measurement, empirical meaningfulness, and three-valued logic. In C. W. Churchman & P. Ratoosh (Eds.), Measurement: Definitions and theories (pp. 129-143). New York, New York: Wiley.

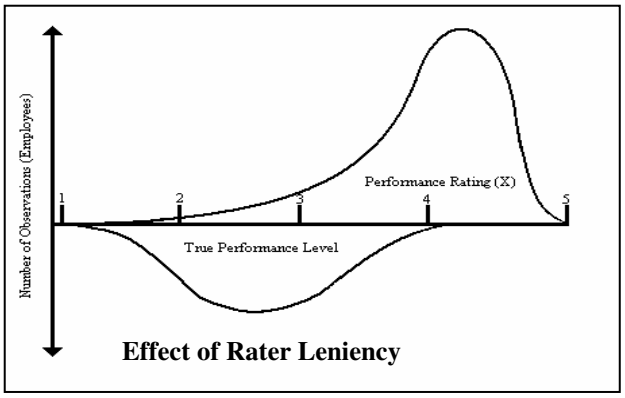
Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), Handbook of mathematical psychology (pp. 1-76). New York, New York: Wiley.

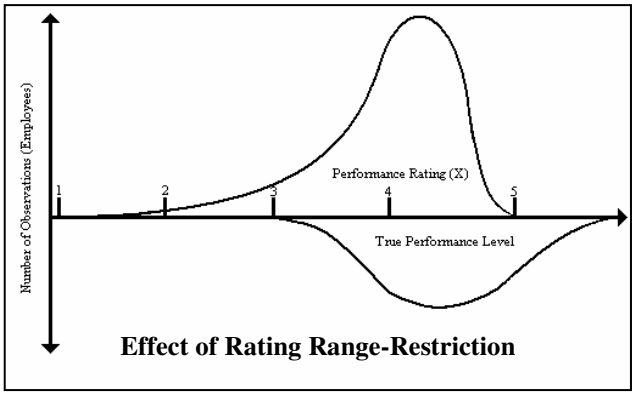
Wilson, T. P. (1971). Critique of ordinal variables. Social Forces, 49, 432-444.

Wise, M. N. (Ed.). (1995). The values of precision. Princeton, New Jersey: Princeton University Press.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. Archives of Physical Medicine and Rehabilitation, 70(12), 857-867. www.rasch.org/memo44.htm

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis: Rasch measurement. Chicago: MESA Press.





Investigation of 360-Degree Instrumentation Effects: Application of the Rasch Model, John T. Kulas (2004) Society for Industrial and Organizational Psychology (Poster)

An Introduction to Rasch Measurement: Theory and Applications
November 6-7, 2004 Chicago, IL

The purpose of this training session is to introduce participants to the theory and applications of Rasch measurement and provide hands-on experience using Rasch calibration programs to scale ordinal data. This session will provide participants with the necessary tools to become effective consumers of research employing Rasch measurement and the skills necessary to solve practical measurement problems. Instructional material will be based on four Rasch measurement models: dichotomous, rating scale, partial credit, and many-facet data. Participants will have the opportunity to use current Rasch software. *Directors:* Everett V. Smith Jr. and Richard M. Smith.

The format will consist of eight self-contained units. The units are: **Introduction to Rasch Measurement; Item and Person Calibration; Dichotomous and Polytomous Data; Performance and Judged Data; Applications of Rasch Measurement I and II; Examples of Rasch Analyses; and Analysis of Participants Data.** The co-directors will divide the topics in each session to maximize individual strengths. The instructional format will combine lecture, question and answer, and small group instruction.

Registration includes the full 2-day workshop, a continental and bagel bar breakfast each morning, over 800 pages of handouts and tutorial material, a copy *of Introduction to Rasch Measurement* (a 698 page book) and a one-year subscription to the *Journal of Applied Measurement.* More details are at www.jampress.org

Audience: Anyone interested in learning about the practical aspects of Rasch measurement. Previous training in measurement is recommended, but not necessary.

Location: Chicago Circle Center (CCC) building on the campus of the University of Illinois at Chicago. The CCC is located at 750 South Halstead Street, Chicago.

Agenda: *Saturday, November 6, 2004*

Session I – Introduction to Rasch Measurement: What is Measurement, Rasch Measurement Models, True Score vs. Rasch Measurement Models.

Session II – Item and Person Calibration: Testing the Fit of Data, Dimensionality and PC Analysis of Residuals.

Session III – Dichotomous and Polytomous Data: WINSTEPS Control Language, Example of dichotomous data analysis, Example of polytomous data analysis.

Session IV – Performance and Judged Data: FACETS Control Language & example of facets analysis with nested data (ratings of conference proposals), Example of facets analysis and G-Theory (ratings of student performance).

Group dinner (optional)

Sunday, November 7, 2004

Session V – Applications of Rasch Measurement: Score Reporting, Standard Setting, Item Bias.

Session VI – Applications of Rasch Measurement: Test Equating and Item Banking, Computer Adaptive Testing, Rasch vs. Multi-Parameter IRT Models.

Session VII – Examples of Rasch Analyses: Rating Scale Data, Partial Credit Data.

Session VIII – Analysis of Participants: Running WINSTEPS and FACETS, Other Rasch software: RUMM, Conquest, MULTIRA, WINMIRA, and LPCM-WIN, Your turn to analyze data.

COMET and IOM Chicago Chapter at UIC

3:30 PM, Thursday, September 23, 2004

UIC Department of Education
1040 W. Harrison St., Third Floor, Room 3427
Chicago IL

E. Matthew Schulz: Map-mark Standard Setting

In conjunction with a contract between ACT Inc., and the National Assessment Governing Board (NAGB), a new standard setting method, map-mark, has been developed as a possible procedure for recommending cut scores for the 2005 National Assessment of Educational Progress (NAEP) Grade 12 mathematics assessment.

Map-mark has been designed as an augmentation of the bookmark standard setting method with item maps and content domain scores. The item map number line represents the relative difficulty of test items and the location of cut scores. Student performance data is further organized into content domains covering a wide range of difficulty. Tables and plots show expected percentage correct scores on domains as a function of achievement. Panelists recommend cut scores directly in terms of scores on the achievement scale.

The map-mark procedure was implemented and modified through a series of two field trials of 10 panelists each, one study involving 20 panelists, and a pilot study involving 20 panelists for each of two methods: map-mark and an Angoff-based, item-rating procedure similar to the one used to set standards on the 1998 NAEP Civics Assessment. This presentation will describe the map-mark standard setting method and present results from the series of studies described above. Results will address the reliability of the method and the comparability of results to past cut scores (produced in 1992 using an Angoff-based procedure) and to cut scores set concurrently using an Angoff-based procedure.

. *Book:* Constructing Measures: An Item Response Modeling Approach

Mark Wilson, (Lawrence Erlbaum Associates, 2004) www.erlbaum.com

Part I: *A Constructive Approach to Measurement.*

Part II: *The Four Building Blocks.* Construct Maps. The Items Design. The Outcome Space. The Measurement Model.

Part III: *Quality Control Methods.* Choosing and Evaluating a Measurement Model. Reliability. Validity.

Part IV: *A Beginning Rather Than a Conclusion.* Next Steps in Measuring. The Cases Archive. GradeMap CD.

Publisher's description: Constructing Measures introduces a way to understand the advantages and disadvantages of measurement instruments, how to use such instruments, and how to apply these methods to develop new instruments and/or adapt old ones. The author believes that the best way to learn is by doing. It is therefore recommended that the reader review the book while in the process of actually constructing an instrument.

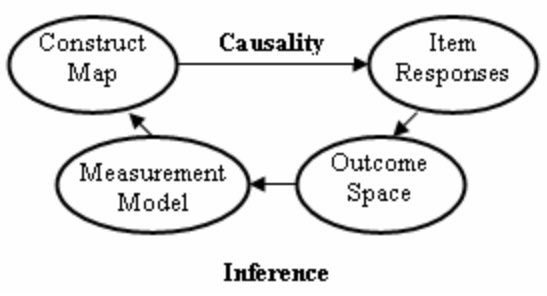
The book is organized around the steps taken while constructing an instrument. It opens with a summary of the constructive steps involved. Each step is then expanded on in the next four chapters. These chapters develop the "building blocks" that make up an instrument--the construct map, the design plan for the items, the outcome space, and the statistical measurement model. The next three chapters focus on quality control. They rely heavily on the calibrated construct map and review how to check if scores are operating consistently and how to evaluate the reliability and validity evidence. The book introduces a variety of item formats, including multiple-choice, open-ended, and performance items; projects; portfolios; Likert and Guttman items; behavioral observations; and interview protocols.

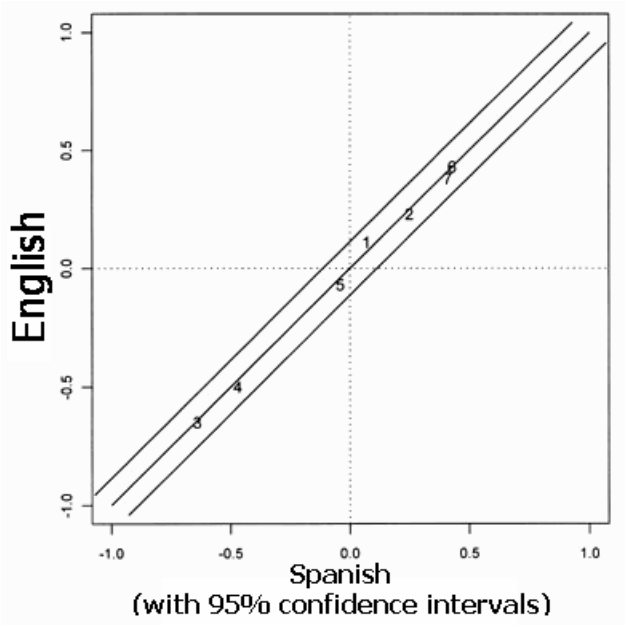
Each chapter includes several features to help the reader: a chapter overview provides the key concepts, related resources provide details for further investigation of certain topics, and exercises and activities provide an opportunity to apply the chapter's concepts. Some chapters feature appendices that describe parts of the instrument development process in more detail, numerical manipulations used in the text, and/or data results of computer analyses. A variety of examples from the behavioral and social sciences and education, including achievement and performance testing; attitude measures; health measures, such as quality of life, and general sociological scales demonstrate the application of the material.

An accompanying CD features: "GradeMap" software with control files, output, and a data set to allow readers to compute all of the text's exercises and examples, and create and explore new analyses; and Case archives based on the book's examples so the reader can work through the entire development of an instrument and gain a greater understanding of the ways the approach varies depending on the circumstances.

Constructing Measures is intended to serve as an advanced text or supplement in courses on item, test, or instrument development, measurement, item response theory, or Rasch analysis taught in a variety of departments, including education and psychology. The book also appeals to those who develop instruments, including industrial/organizational, educational, and school psychologists, health outcomes researchers, program evaluators, and sociological measurers. Knowledge of basic descriptive statistics and elementary regression is recommended. *(Price: $29.95 and up)*

Mark Wilson's Development Cycle





Successful equivalence of language versions.

Hernández L. et al. (2000) Development and Validation of the "Satisfaction with Pharmacist" Scale. Pharmacotherapy 20(7): 837-843, 2000