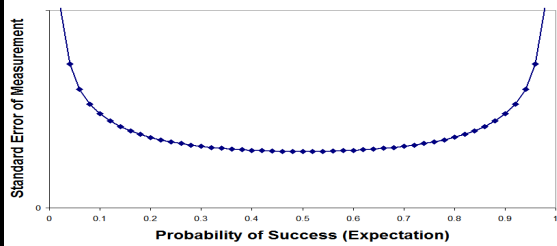


## Standard Error of Measurement



# RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG  
American Educational Research Association

Vol. 20 No. 3

Winter 2006

ISSN 1051-0796

## Bernoulli, Fisher, Shannon and Rasch

Jakob Bernoulli (1654-1705) wrote the founding treatise on mathematical probability, *Ars Conjectandi*, published posthumously in 1713. This discusses binomial trials, known to us as dichotomous items.

Bernoulli explains that, if the probability of passing a dichotomous item is  $p$ , this is also the expected value of the trial when a success is scored 1 and a failure 0. After  $n$  identical trials, we expect to have  $n \cdot p$  successes and  $n \cdot (1-p)$  failures. The sum-of-squares of these around their expectation,  $p$ , is  $n \cdot p \cdot (1-p)^2$  for the successes and  $n \cdot (1-p) \cdot (0-p)^2$  for the failures. Thus total sum-of-squares  

$$= n \cdot p \cdot (1-p)^2 + n \cdot (1-p) \cdot (0-p)^2 = n \cdot p \cdot (1-p)$$
 so the sum-of-squares for a single trial, its Bernoulli variance, is  $n \cdot p \cdot (1-p) / n = p \cdot (1-p)$ .

The curve of Bernoulli variance against probability of success is plotted here in Figure 1.

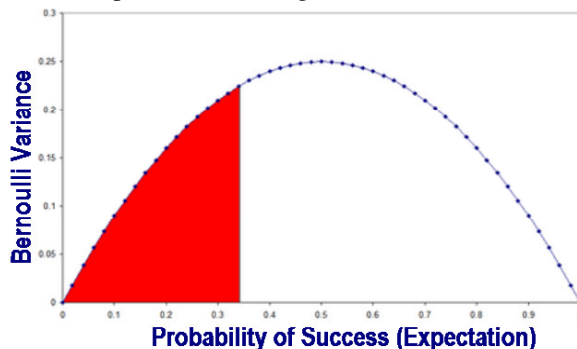


Figure 1. Bernoulli variance of one binomial trial.

The reciprocal of this curve  $I = 1 / (p \cdot (1-p))$  is Ronald Fisher's (1925) "statistical information" function for a binomial response, shown in Figure 2. Fisher defined statistical "information" to be the "intrinsic accuracy of the error curve". The square-root of the Fisher Information is the standard error of the Rasch measure.

Suppose that the probability is a function of a variable,  $x$ , so that  $p = p(x)$ . This function has the property that it is the accumulation (integral) of the Bernoulli variance (indicated by the red area in Figure 1). Then, since the variance is always positive,  $p(x)$  always increases as  $x$  increases.  $p(x)$  increases most rapidly with  $x$  when the Bernoulli variance is at its maximum where  $p=0.5$  and

$p(x) = 0.25$ . What is the function? It must satisfy the integration:

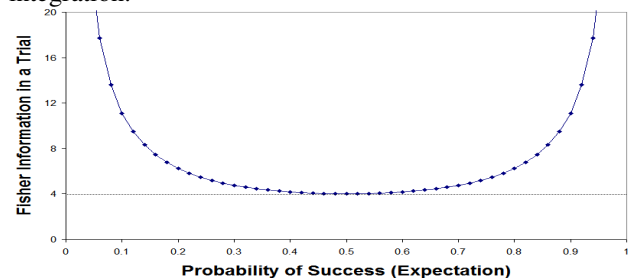


Figure 2. Fisher Information in one binomial trial.

$$\int (p(x) \cdot (1-p(x))) dx = p(x)$$

or, written as a differential equation,

$$dp(x) / dx = p(x) \cdot (1-p(x))$$

for which the solution is

$$p(x) = \exp(x) / (1 + \exp(x))$$

This is the logistic ogive of the Rasch model, with  $x$  being the difference between the person ability and the item difficulty for a dichotomous item. So  $p(x)$ , the Rasch logistic ogive, has the property that its rate of change, its slope, is the Bernoulli variance, the inverse of the Fisher information.

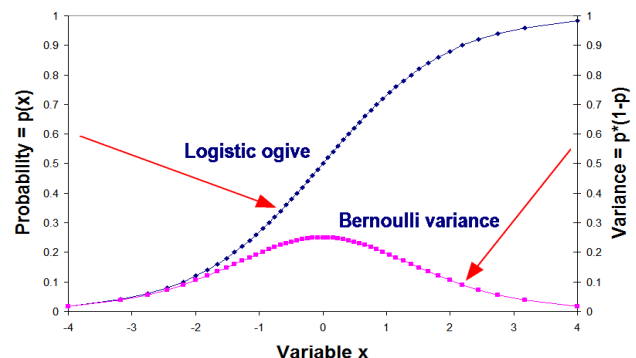


Figure 3. Rasch logistic ogive and Bernoulli variance.

## Table of Contents

|                                                 |      |
|-------------------------------------------------|------|
| Adjustment for Sample Size in DIF Analysis..... | 1070 |
| Bernoulli, Fisher, Shannon and Rasch.....       | 1063 |
| Consequences for Measurement of Mindsets .....  | 1066 |
| Rasch: Too Complicated or Too Simple?.....      | 1065 |
| Survey Design Recommendations .....             | 1072 |

Figure 3 shows the logistic ogive and the Bernoulli variance of Figure 1, now redrawn in terms of the independent variable  $x$ , which we can interpret as the Rasch measure-difference. It can be seen that the slope of the logistic ogive curve is given by the variance curve, and that the accumulation of the variance curve is the logistic ogive.

When we re-express the Rasch logistic ogive in terms of  $p$ , then we have the log-odds form of the Rasch dichotomous model:

$$x = \log(p / (1-p))$$

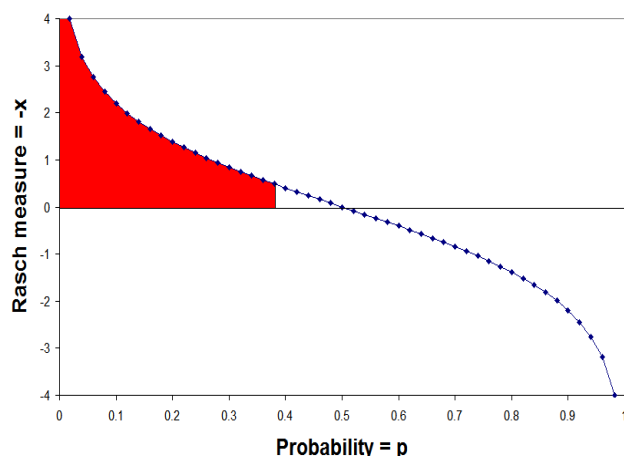


Figure 4. Reoriented Rasch logistic ogive.

### UK Rasch Users' Group

The *Cambridge Assessment Network* hosted the 2nd meeting of the *UK Rasch Users' Group* at Hughes Hall in Cambridge, England on Monday 5th February 2007. The purpose of the group is to provide a forum for Rasch practitioners in the UK working in different fields to get together to share ideas and present research.

The 38 delegates heard presentations covering a wide range of practical applications of Rasch measurement - from investigating misfit in classroom mathematics tests to measuring rehabilitation outcomes in brain-injured patients. During the lunch break there were demonstrations of developments from Cambridge ESOL in item banking, item analysis and computer adaptive testing, and of Cambridge Assessment's TSA (Thinking Skills Assessment) online test for university admissions.

The program for the day, and the abstracts and slides for the presentations will soon be available at [www.assessnet.org.uk/mod/resource/view.php?id=142](http://www.assessnet.org.uk/mod/resource/view.php?id=142)

Tom Bramley  
Assistant Director, Research Division  
Assessment Research & Development  
Cambridge Assessment  
1 Regent Street, Cambridge, CB2 1GG  
Direct Dial: 01223 553985  
[www.cambridgeassessment.org.uk](http://www.cambridgeassessment.org.uk)

Redrawing Figure 3 in terms of  $p$  on the x-axis, and  $-x$  on the y-axis, we have Figure 4, an ogive with the vertical orientation originally used by Francis Galton (1875), except that he had the tall men stand on the right side, not the left. The area under this curve (shown in red) is an indication of the entropy in the curve. It can be expressed by  $H$  where

$$H = \int (-x)dp = -(p \log(p) + (1-p) \log(1-p))$$

This is equivalent to Claude Shannon's (1948) "binary entropy function" which he identifies as related to the communication information in a binary observation, and also to the entropy function in Boltzmann's (1872) statistical mechanics. Shannon expressed his function in terms of  $\log_2(p)$ , which is  $1.44 * \log_e(p)$ . Shannon's information function is plotted here in Figure 5. Its maximum value is 1.0.

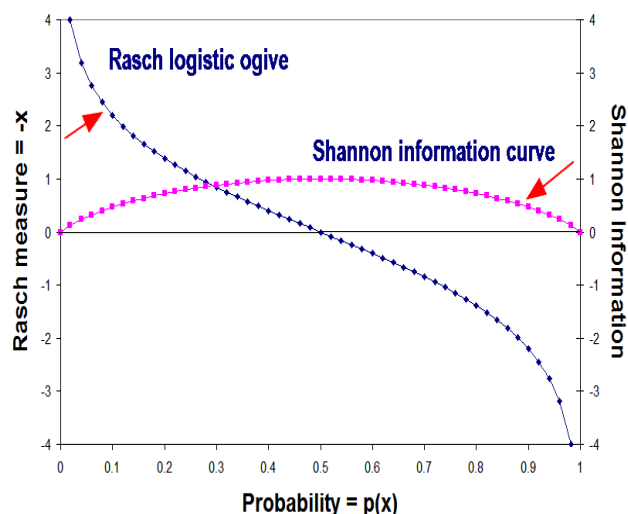


Figure 5. Shannon information and the Rasch ogive.

The Rasch ogive is seen to tie together neatly Fisher information and Shannon information. In the literature, these are presented as contradictory formulations. An example is "... a theory based upon Fisher information [may be] less powerful than one based on Shannon information" (*Wikipedia*, art. "Physical Information"). In fact, Fisher Information and Shannon Information are different ways of saying the same thing.

John Michael Linacre

Shannon, C. E. (1948). A mathematical theory of communication (parts I and II). *Bell System Technical Journal*, XXVII:379-423.

Boltzmann L. (1872) Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen, *Wiener Berichte*, 66: 275-370 ("Further studies on heat equilibrium among gas molecules")

Fisher, R. A. (1925). Theory of statistical estimation, *Proc. Camb. Phil. Soc.*, 22: 700-725.

Galton, F. (1875) Statistics by intercomparison, with remarks on the law of frequency of error. *Philosophical Magazine*, 4th series, 49: 33-46.

## Rasch: Too Complicated or Too Simple?

*A colleague writes:*

"I've proposed a Rasch analysis, not a Classical analysis, but people here are not familiar with it and they are not very open to other perspectives than Classical Test Theory. I've try to explain to them all the advantages of Rasch, but they said: "This is to complicated and people won't understand it. We need a Classical Analysis." But on another project, people were very critical about Rasch and proposed instead a 3-PL model."

*Comment:* Yes, familiar methods are difficult to unseat. The astrolabe was still being used 150 years after Isaac Newton demonstrated better methods of locating the planets. Classical methods usually survive until a situation arises they can't deal with, such as missing data, adaptive tests, maintaining criterion-based pass-fail points, test-equating with small samples ...

Advocating the 3-PL model means that they expect their students to guess answers at random - so that the student measures will have a chaotic component, lowering the predictive validity of the test. They also expect their items to differ widely in discrimination - which means they expect their item writers to write deficient items, lowering the construct validity of the test.

### *MetaMetrics Workshop Series on Psychometrics*

#### **An Introduction to Rasch Measurement: Theory and Application**

*by David Andrich*

**March 26-29, 2007 - Monday-Friday  
Durham , North Carolina**

*You are invited to a free four-day workshop* introducing the theory and applications of Rasch measurement and providing hands-on experience with RUMM2020 data analysis software. The workshop will combine lecture, question-and-answer and small-group instruction. You will have opportunities to analyze your own data.

We will study principles of the Rasch models from the perspective of the Item Characteristic Curve and Differential Item Functioning. RUMM2020 will be used to demonstrate concepts and teach you how to analyze data in a flexible way. Case studies will be used that bring together the professional understanding of the variable of assessment, item construction, and the use of statistical indices in determining the validity of item sets.

Instructional material will apply Rasch models to dichotomous (multiple choice) and polytomous (rating scale and partial credit) data. Familiarity with Microsoft Excel, basic statistics and the Windows platform is a plus.

More information and registration details at:

<http://www.lexile.com/DesktopDefault.aspx?view=re&tabindex=3&tabid=92>

The 3-P logistic model was designed to model messy data. According to Martha Stocking, an advocate of 3-PL, "Building statistical models is just like this. You take a real situation with real data, messy as this is, and build a model that works to explain the behavior of real data." *New York Times*, 2-10-2000.

Instead of designing a descriptive model, such as 3-PL, to explain messy data, Georg Rasch designed a model that demands useful data and points out where the data is messy. Messy data produces messy findings. Useful data produces useful findings. But most statisticians believe that the data, however messy, always tells the truth. The problem is that the messy data's "truth" may not be the truth that we need for decision-making.

### ***Rasch Workshop***

#### **Hands-on Introduction to IRT/Rasch Measurement Using Winsteps**

*by Ken Conrad & Barth Riley*

**March 26-27, 2007 - Monday-Tuesday  
University of Illinois - Chicago**

Social scientists have great need for the development of valid measures, e.g., of the quantity and quality of health services and of the outcomes of those services. Many researchers are frustrated when existing instruments are not well tailored to the task, since they then cannot expect sensitive, accurate, or valid findings. This workshop presents the theory and practice of classical test theory, the traditional approach. It then provides an overview of modern measurement as practiced using item response theory with a focus on Rasch measurement. Rasch analysis provides the social sciences with the kind of measurement that characterizes the natural sciences. Since Rasch focuses on the items and the persons rather than the test score, the synthesis of quantitative analysis with qualitative issues is experienced in a way that is rare in social science. Ultimately, Rasch measurement can facilitate more efficient, reliable, and valid assessment while improving privacy and convenience to users. The Workshop is useful for anyone who wants to understand the role of modern measurement in research.

*Attendees will learn hands-on:*

- \* Differences between Classical Test Theory and Rasch
- \* Why and how Rasch creates linear, interval measures
- \* The inner workings of the Rasch model
- \* How to run Winsteps analyses
- \* Interpretation of Rasch/Winsteps output

You need a recent Winsteps running on a lap-top computer. We provide Winsteps free, but time-limited.

For more details and registration:

[www.winsteps.com/workshop.htm](http://www.winsteps.com/workshop.htm)

## The Consequences for Measurement of Mindsets

Dweck's (2000, 2006) social-cognitive model of motivation, personality and development leads to the realization that individuals' behaviors and beliefs are often influenced by implicit theories concerning the malleability of intelligence. For example, implicit self-theories about intelligence have been shown to influence

- individuals' responses to achievement challenges (Dweck & Leggett, 1988; Henderson & Dweck, 1990),
- their conceptions of morality as duty-based versus human-rights based, and their overall assumptions about the kind of person someone is and how they will behave (Chiu, Dweck, Tong, & Fu, 1997),
- differences in degree of social stereo-typing (Levy, Stroessner, & Dweck, 1998), and
- individuals' predictions of the future behavior of another individual (Chiu, Hong, & Dweck, 1997).

In the course of conducting research aimed at the quantification of Dweck's constructs, it became apparent that the different ways self-theories influence responses to achievement challenges suggests an illuminating approach to understanding the history of measurement in psychology.

The implicit self-theory research demonstrates that an individual usually responds to life experiences through the unconscious lens of one of two theories about intelligence, fixed or malleable (Dweck, 2000, 2006). Dweck and colleagues (2000, 2006) carefully articulate that, though the fixed "Entity" and malleable "Growth" mindsets present different ways of approaching a variety of experiences, neither is wrong or right—they are just different in their consequences.

The fixed theory sees intelligence as a quantity received as a function of biologic inheritance that cannot be enlarged by experience. In other words, a person's ability is set in stone at birth. Although one may learn new things, the amount of available intelligence remains the same. The malleable theory (referred to as the "growth theory" here) sees intelligence as a function of experience and effort. A person's ability over the course of a lifetime may be increased through effort and experience. One can increase intelligence through trial-and-error problem solving in the course of learning new things. Research has shown that holding one or the other of these theories has specific consequences when the individual is faced with an achievement challenge.

### Self-Theories of Intelligence

Fixed theorists usually exhibit a need to protect their perception of the amount of intelligence they possess. When given a choice, fixed theorists select achievement challenges they perceive as not likely to threaten their image of themselves as successful. So, they rarely choose to take a risk in order to learn. Instead, they prefer to repeat a challenge known to be within their capacity to succeed. Fixed theorists center their concerns on how they

appear as they perform; they are more likely to be concerned with looking smart than with entertaining the possibility of failure in order to learn something new.

Growth theorists usually associate their intelligence with the amount of effort they invest in achievement challenges. The consequences of holding this theory means that when given a choice they more often select achievement challenges that offer the opportunities to learn something new instead of one that repeats a previous success. They see trial and error problem solving as a confirmation of intelligence. When faced with achievement challenges, growth theorists' concerns center on matching effort with the challenge. Big challenges require big efforts.

Dweck and colleagues (Dweck 2000, 2006) find differences in the way these two implicit self-theories influence individual's explanations of, and responses, to failure. Fixed theorists are likely to explain failure in terms of external circumstances beyond their control. Typical comments in the face of failure might be "I've never been good at that," or "I'm just not that smart," or "That teacher is just too demanding," blaming failure on their fixed amount of intelligence or the level of the challenge and exhibit a "helpless" response. Growth theorists on the other hand typically say, "Okay, so that way doesn't work," or "If I had just worked a little harder," suggesting that failure could be averted by moving beyond what doesn't work to calculate another approach predicting success as a function of effort.

Fixed theorists are known to develop maladaptive behaviors when confronted with achievement challenges that keep them from solving problems they had previously thought soluble (Dweck, 2000, 2006). These maladaptive behaviors bear striking resemblances to the responses of some theorists to the problems of psychological measurement.

### Measurement in the Human Sciences: The View from Two Mindsets

Psychology and the human sciences faced a particularly salient measurement puzzle when fundamental measurement as defined by Campbell (1920) seemed impossible for these fields.

Without fundamental measurement there could be no derived measurement and so, according to Campbell's theory, psychology was without measurement. What is more, the task of locating analogues of numerical addition pertinent to psychological measurement appeared grim. (Michell, 1990)

How might implicit self-theories of intelligence help us understand the ways in which these problems were addressed in the history of science? In other words, construing the situation as an "achievement challenge," how might the historical approaches taken by different individuals to the problem of psychological measurement

be interpreted in light of the implicit self-theory constructs? Let's look at two such individuals.

### **S. S. Stevens: The Road Most Traveled**

The problem of measurement filled many in the field of psychology with consternation, and none more than S. S. Stevens. Stevens attacked the problem and constructed the road most traveled by human scientists since. His approach was to redefine measurement operationally.

In an autobiographical article Stevens wrote, "My own central problem throughout the 1930's was measurement, because the quantification of the sensory attributes seemed impossible unless the nature of measurement could be properly understood" (1974, p. 409). The theory of measurement that he came to propose owed much to Campbell's, but it also owed a lot to what was then a fledgling philosophical movement, *operationalism*. (Michell, 1990, p. 15)

In drawing from operationalism, Stevens' response to the measurement challenge exhibited the characteristics of a "fixed theory of intelligence"—a fixed mindset.

Remember, the fixed mindset looks for a way to appear smart or rigorous with the least risk. Failure would be a flagrant denial of innate intelligence, or, in this case, a flagrant denial of the value of a putative science. So, Stevens' approach to the achievement challenge was to retreat to a challenge posing a diluted risk of failure.

Thus, in his 1935 article, "The operational definition of psychological terms," in the journal, *Psychological Review*, Stevens set the stage for success in measuring by establishing new guidelines. His proposal set forth a program that said as long as you define your operations and keep true to those definitions then you have a "measurement." Fundamental measurement as defined by Campbell was abandoned due to circumstances beyond the control of psychologists, in favor of a solution that guarantees the appearance of success.

As in the case of Dweck's mindset research, the consequence of making the choice in favor of the immediately solvable problem unnecessarily limits the learning boundaries of those involved. This is seen in the way research that reports results defined through Stevens' model consistently restricts itself to descriptive statistics with context specific applications. The consequences of this mindset in human science research means that the utility of reported findings is restricted to instrument- and sample-specific results. The alternative, although a road less traveled, yields quite different results.

### **G. Rasch: The Road Less Traveled**

Rasch (1960) was also concerned with the problem of measurement in the human sciences, particularly as it affected the measurement of individuals. In the preface to his *Probabilistic Models for Some Intelligence and Attainment Tests*, he presents the problem as one of requiring models that demand that the result of an encounter between an instrument and an individual

depend only on the individual's ability and the instrument's difficulty.

Symmetrically, it ought to be possible to compare stimuli belonging to the same class – "measuring the same thing" – independent of which particular individuals within a class considered were instrumental for comparison.

This is a huge challenge, but once the problem has been formulated it does seem possible to meet it. (Rasch, 1960, p. xx). In other words, he advocated for models that maintained the requirements of fundamental measurement, that the "calibration of the measuring

### **Rasch-related Coming Events: 2007**

Feb. 12-16, 2007, Mon.-Fri. Item Response Modeling With ConQuest (Ray Adams & Margaret Wu), Australia  
[www.edfac.unimelb.edu.au](http://www.edfac.unimelb.edu.au)

Feb. 16 - Mar. 16, 2007, Fri.-Fri. Practical Rasch Measurement (Winsteps) online course (Mike Linacre)  
[www.statistics.com/courses/rasch](http://www.statistics.com/courses/rasch) (Sorry! Fully booked.)

March 2007 - Dec 2008 3-day Rasch courses, Leeds, UK  
[http://home.btconnect.com/Psylab\\_at\\_Leeds/](http://home.btconnect.com/Psylab_at_Leeds/)

March 26-27, 2007, Mon.-Tues. Introduction to IRT/Rasch Measurement Using Winsteps (Conrad & Bezruczko), Chicago [www.winsteps.com/workshop.htm](http://www.winsteps.com/workshop.htm)

March 26-29, 2007, Mon.-Thurs. MetaMetrics Workshop Series in Psychometrics – Introduction to Rasch Measurement: Theory and Application (David Andrich) (free!), Durham, North Carolina [www.lexile.com](http://www.lexile.com)

Apr. 7-8, 2007, Sat.-Sun. Introduction to Rasch Measurement: Theory and Applications, Chicago IL (Smith & Smith) [www.jampress.org](http://www.jampress.org)

Apr. 9-13, 2007, Mon.-Fri. AERA Annual Meeting, Chicago [www.aera.net](http://www.aera.net)

May 4 - June 1, 2007, Fri.-Fri. Facets online course (Mike Linacre) [www.statistics.com/courses/facets](http://www.statistics.com/courses/facets)

June 21 - July 1, 2007, Thur.-Sun. 3rd Summer School Measurement of Latent Variables (Rasch), Russia  
June 22, 2007, Fri. Workshop: Theory and Practice of Measurement of Latent Variables, Russia  
[www.rasch.org/russia.pdf](http://www.rasch.org/russia.pdf)

July 16, 2007, Mon. ConQuest Workshop (Margaret Wu), Taiwan

July 16, 2007, Mon. Winsteps Workshop (Mike Linacre), Taiwan

July 17-19, 2007, Tues.-Thurs.  
Pacific Rim Measurement Symposium PROMS Taiwan  
<http://210.60.0.152/PROMS2007TAIWAN/>

Aug. 3 - Aug. 31, 2007, Fri.-Fri. Practical Rasch Measurement with Winsteps online course (Mike Linacre)  
[www.statistics.com/courses/rasch](http://www.statistics.com/courses/rasch)

instruments must be independent of those objects that happen to be used for calibration. Second, the measurement of objects must be independent of the instrument that happens to be used for measuring,” (Wright and Stone, 1979, p. xii). And it is apparent that Rasch did so with the full knowledge of the scope of the challenge.

Embracing the challenge, like a growth mindset theorist, he saw the solution as one of effort, a trial and error endeavor. Unlike Stevens, Rasch engaged the challenge at its heart, rather than change the nature of the problem to something less formidable. Rasch’s effort produced a set of probabilistic models making human science variables amenable to fundamental measurement requirements. Results can be reported in invariant units creating a common language among interested parties. This releases results from context specific applications and is crucial to meaningful, linked conversations among various interested parties, such as teachers, students, parents, administrators, researchers, accreditors, etc.

In this context, the historian of science, Bruno Latour, remarks, “Every time you hear about a successful application of science, look for the progressive extension of a network” (Latour, 1987, p. 249). Choosing methods capable of supporting expanding networks of people communicating in common languages about the same things breaks the silence of non-connected networks enforced by analyses that contextually imprison research results.

When a common language is mobilized within a network of shared signification, with its terms and symbols everywhere recognized and accepted by those trained in reading them, meaningful communication is achieved, shared understandings and histories are more easily accumulated, and collective productivity is markedly enhanced. (Fisher, 2003, p. 801)

Those following Rasch’s model (and other models like his) open learning boundaries. They extend what is possible for communities or networks of human science researchers to accomplish. They participate in the fascinating power of shared knowledge by way of preferring effort over retreat—a road too long less traveled.

Sharon G. Solloway

Campbell, N. R. (1920). *Physics, the elements*. Cambridge: Cambridge University Press.

Chiu, C., Dweck, C. S., Tong, J. Y., & Fu, J. H. (1997). Implicit theories and conceptions of morality. *Journal of Personality and Social Psychology*, 73(5), 923-940.

Chiu, C., Hong, Y., Dweck, C.S. (1997). Lay dispositionism and implicit theories of personality. *Journal of Personality and Social Psychology*, 73(1), 19-30.

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development*. Philadelphia,

PA: Psychology Press.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York: Random House.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95(2), 256-273.

Fisher, W. P., Jr. (2003, December). Mathematics, measurement, metaphor, metaphysics: Part II. Accounting for Galileo’s “fateful omission.” *Theory & Psychology*, 13(6), 790-828.

Henderson, V., & Dweck, c. s. (1990). Motivation and achievement. In S. S. Feldman & G. R. Elliott (Eds.). *At the threshold: The developing adolescent* (pp. 308-329). Cambridge, MA: Harvard University Press.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. New York: Cambridge University Press.

Levy, S. R., Stroessner, S. J., & Dweck, C. S. (1998). Stereotype formation and endorsement: The role of implicit theories. *Journal of Personality and Social Psychology*, 74(6), 1421-1436.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Rasch, G. 1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut. [Reprinted, 1980, Chicago, IL: University of Chicago Press.]

Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press.

## Rasch Online Courses

### Winsteps and Facets

by Mike Linacre

May 4 - June 1, 2007, Fri.-Fri. **Facets** online course (Mike Linacre) [www.statistics.com/courses/rasch](http://www.statistics.com/courses/rasch)

Aug. 3 - Aug. 31, 2007, Fri.-Fri. Practical Rasch Measurement with **Winsteps** online course (Mike Linacre) [www.statistics.com/courses/facets](http://www.statistics.com/courses/facets)

Feb. 16 - Mar. 16, 2007, Fri.-Fri. Practical Rasch Measurement with Winsteps online course (Mike Linacre) [www.statistics.com/courses/rasch](http://www.statistics.com/courses/rasch) (Sorry! Fully booked.)

Each Course consists of 4 weeks of detailed step-by-step downloadable tutorials on Rasch theory and software operation. There are Discussion Boards for Q-&A and group interaction. Free time-limited versions of the software are provided. These Courses are the next-best-thing to in-person Workshops. You work at your own pace in your own location on your own schedule. Give yourself at least 10 hours per week to fully benefit.

**ACSPRI Conference, Sydney, Australia, December 2006**  
**Australian Consortium for Social and Political Research Incorporated**  
**Objective Measurement in the Social Sciences Stream**  
**Coordinator: Andrew Stephanou**

The Polytomous Unidimensional Rasch Model:  
Understanding its Response Structure and Process, *David Andrich, School of Education, Murdoch University*

Psychometric Properties of the PsychoSomatic Problems  
Scale – an Examination Using the Rasch Model, *Curt Hagquist, Karlstad University, Karlstad, Sweden*

Predicting Fitness to Drive in People with Physical and/or  
Cognitive Impairment Using a Clinical Test, *Lynn Kay, Anita Bundy & Lindy M. Clemson, School of Occupation and Leisure Sciences, Faculty of Health Sciences, The University of Sydney*

Measuring Task Embedded Information Processing  
Capacity During Occupational Performance: an  
Application of Rasch Measurement, *Melissa Nott & C. Chapparo, School of Occupation and Leisure Sciences, Faculty of Health Sciences, The University of Sydney*

Does Attitudinal Ambivalence Necessitate the Bivariate  
Measurement of Attitudes? An Application of the Quasi-  
Rasch Hyperbolic Cosine Model, *Joshua McGrane, School of Psychology, The University of Sydney*

Developing a Diagnostic Assessment Instrument for  
Identifying Students' Understanding of Fraction  
Equivalence, *Monica Wong, David Evans & Judy Anderson, The University of Sydney*

The Motivation of Stereotypic and Repetitive Behaviour:  
Examination of Construct Validity of the Motivation  
Assessment Scale, *Annette V. Joosten and Anita C. Bundy, School of Occupation and Leisure Sciences, Faculty of Health Sciences, The University of Sydney*

Refining an Instrumental Activities Daily Living Measure  
by Determining Category Functioning, *Lindy Clemson, Anita Bundy, Lynnette Kay & Tim Lockett, Faculty of Health Sciences, The University of Sydney*

Maintaining a Common Unit in Social Measurement,  
*Steve Humphry, University of Western Australia*

PISA – The Programme for International Student  
Assessment – An Overview, *Ross Turner, Australian Council for Educational Research*

Using Differential Item Functioning to Enhance the  
Curriculum, *Juho Looveer, NSW Department of Education and Training*

Development of a Numeracy Achievement Scale to  
Assess Progress from Kindergarten Through Year 6, *Juho Looveer, Joanne Mulligan & Susan Busatto, NSW Department of Education and Training; Macquarie University*

Changes in Students' Mathematics Achievement in  
Australian Lower Secondary Schools Over Time, *Tilahun Afrassa, SA Department of Education and Children's Services*

The Impact of Moving Testing From August to May on  
Students' Achievement in Numeracy and Literacy: a  
Rasch Analysis, *Tilahun Afrassa, SA Department of Education and Children's Services*

Ameliorating Culturally Based Extreme Response  
Tendencies to Attitude Items: The Use of Item Response  
Models to Explore the Alternatives, *Maurice Walker, Australian Council for Educational Research Melbourne, Australia*

Using the Rasch Model in the Design of a New  
Curriculum Framework and to Moderate Teacher  
Assessments Within it, *Andrew Smith, Office for Educational Review, Department of Education (Tasmania)*

Norming the Progressive Achievement Tests in  
Mathematics with the Rasch Model, *Charles Darr & Andrew Stephanou New Zealand Council for Educational Research; Australian Council for Educational Research*

**Abstracts and Full Papers available via**

[www.acspri.org.au/conference2006/proceedings/](http://www.acspri.org.au/conference2006/proceedings/)

---

**Science ...**

Science consists of two general areas: there is the act of measurement, which is the empirical side of science, and there is the development of mechanisms, which is its theoretical side.

*Dr. Dean Radin on "Closer to Truth", PBS, 2000*

**Rasch-related Coming Events: 2008**

Jan. 7-11, 2008, Mon.-Fri. Introductory course on Rasch measurement (Andrich, RUMM), Australia  
[www.rasch.org/i2008.htm](http://www.rasch.org/i2008.htm)

Jan. 14-18, 2008, Mon.-Fri. Advanced course on Rasch measurement (Andrich, RUMM), Australia  
[www.rasch.org/i2008.htm](http://www.rasch.org/i2008.htm)

Jan. 21, 2008, Mon. One-day RUMM Workshop (Andrich, RUMM), Australia [www.rasch.org/i2008.htm](http://www.rasch.org/i2008.htm)

Jan. 22-24, 2008, Tues.-Thurs. 3rd International Conference on Measurement in Health, Education, Psychology and Marketing: Developments with Rasch models, Australia [www.rasch.org/i2008.htm](http://www.rasch.org/i2008.htm)

March 24-28, 2008, Mon.-Fri. AERA Annual Meeting New York [www.aera.net](http://www.aera.net)

# An Adjustment for Sample Size in DIF Analysis

A statistical test for differential item functioning (DIF, item bias) between two groups, 1 and 2, is:

$$t_i(12) = \frac{d_i(1) - d_i(2)}{\sqrt{se_i(1)^2 + se_i(2)^2}} > 1.96 \quad [1]$$

where  $d_i(1) - d_i(2)$  is the shift of the item  $i$  measures between groups 1 and 2;  $se_i(1)$  and  $se_i(2)$  are the standard errors of the item measures. A shift greater than 0.5 logits is considered evidence of DIF, and a  $t$  value of 1.96 or more is statistically significant ( $p < .05$ ) for large samples (Draba, 1977). The ETS DIF Classification is similar, but more elaborate. Draba's rule approximates the ETS Category B rule.

| ETS DIF Category                                                                                                                                                                                                            | DIF Size (Logits)              | DIF Significance                 |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------|----------------------------------|
| C = moderate to large                                                                                                                                                                                                       | $ DIF  \geq 1.5 / 2.35 = 0.64$ | $p( DIF  > 1/2.35 = 0.43) > .05$ |
| B = slight to moderate                                                                                                                                                                                                      | $ DIF  \geq 1/2.35 = 0.43$     | $p( DIF  > 0) > .05$             |
| A = negligible                                                                                                                                                                                                              | -                              | -                                |
| C-, B- = DIF against focal group<br>C+, B+ = DIF against reference group                                                                                                                                                    |                                |                                  |
| Zwick, R., Thayer, D.T., Lewis, C. (1999) An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis.<br><i>Journal of Educational Measurement</i> , 36, 1, 1-28<br>Note: ETS use Delta units. 1 logit = 2.35 Delta units. |                                |                                  |

Formula [1] has been employed in a 21 items test, answered by 21820 persons, and the purpose is to detect gender bias. Standard errors are deliberately shown with 4 decimal places. It appears that even for small shift values, 20 of the 21 items produce significant  $t_i(12)$  values, indicating that practically all the items show gender bias according to the significance rule alone. The same

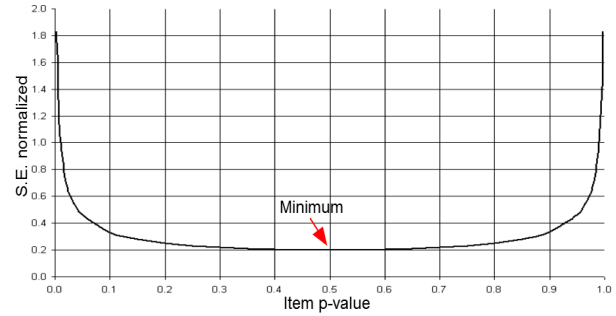


Figure 1. Values of S.E.normalized for different item p-values.

analysis was developed with sub-samples of 250 women and 250 men, now only 5 of the 21 items produce significant  $t$  values and so it is a minority of items that show gender bias. The main reason of the difference in results is that the standard errors of the item measures depend on the sizes of the focus and reference groups. But, according to Clauser & Hambleton (1994), "DIF analysis should be based on the largest sample available", so their guideline implies that even the smallest difference could be significant, nullifying the ETS Significance rule.

It can be seen that the conventional  $t$  values,  $t_i(12)$ , differ considerably across calculations. The measure estimates for the large sample are more precise than for the sub-sample, so the large sample produces smaller standard errors of measurement and higher  $t$  values. This suggests that a DIF Significance test is needed that is robust against sample size. Here is one based on computationally normalizing the empirical sample size of  $N$  to a standard sample size of 100:

$$S.E._{normalized} = \frac{S.E. \cdot \sqrt{N}}{\sqrt{100}} = S.E. \cdot \frac{\sqrt{N}}{10} \quad [2]$$

The reference value of 100 has been chosen because it is not only a simple number to remember, but also because

the mean  $S.E._{normalized}$  for item p-values between 0.001 and 0.999 is 0.965, i.e., close to 1, and, as can be seen in Fig. 1, the minimum possible value of  $S.E._{normalized}$  is 0.2. These values are of the same order of magnitude as the  $t$  values expected according to formula [1]. In addition,  $t$  values bigger than 1.96 may occur for shifts of 0.55 (Fig. 2), which closely corresponds to the half logit rule, so the conclusions of the DIF size and DIF significance rules are in accord. For a closer match with the ETS criteria for DIF Category B, instead of 100, normalize with  $100 \cdot (0.43/0.55)^2 \approx 60$ .

|      | Men=11320 |        | Women=10500 |        | Shift | $t_i(12)$ | Men             | Women | $t_i(12)_n$ |
|------|-----------|--------|-------------|--------|-------|-----------|-----------------|-------|-------------|
| Item | Measure   | S.E.   | Measure     | S.E.   |       |           | S.E. normalized |       |             |
| 1    | -1.4830   | 0.0229 | -1.5731     | 0.0233 | -0.09 | -2.76     | 0.24            | 0.26  | -0.254      |
| 5    | -0.8466   | 0.0207 | -1.4895     | 0.0209 | -0.64 | -21.86    | 0.22            | 0.23  | -2.017      |
| 9    | -0.6947   | 0.0223 | -0.0767     | 0.0214 | 0.62  | 20.00     | 0.24            | 0.24  | 1.847       |
| 13   | 0.3486    | 0.0210 | 0.1849      | 0.0197 | -0.16 | -5.69     | 0.22            | 0.22  | -0.525      |

|      | Men=250 |        | Women=250 |        | Shift        | t <sub>i</sub> (12) | Men             | Women | t <sub>i</sub> (12) <sub>n</sub> |
|------|---------|--------|-----------|--------|--------------|---------------------|-----------------|-------|----------------------------------|
| Item | Measure | S.E.   | Measure   | S.E.   |              |                     | S.E. normalized |       |                                  |
| 1    | -1.6269 | 0.1714 | -1.9528   | 0.1966 | -0.33        | -1.25               | 0.27            | 0.31  | -0.790                           |
| 5    | -0.6214 | 0.1436 | -1.3438   | 0.1652 | <b>-0.72</b> | <b>-3.30</b>        | 0.23            | 0.26  | <b>-2.087</b>                    |
| 9    | -0.6214 | 0.1572 | 0.0047    | 0.1479 | <b>0.63</b>  | <b>2.90</b>         | 0.25            | 0.23  | 1.835                            |
| 13   | 0.2751  | 0.1399 | 0.3601    | 0.1372 | 0.09         | 0.43                | 0.22            | 0.22  | 0.274                            |

Table 1 (left-hand columns) show some measures and S.E. of items for the complete sample and for a sub-sample.

Figure 2 shows that, independently of the S.E., shifts below 0.55 imply that no item bias will be present and that shifts above 1.3 imply item bias (both are close to the 0.5 logit rule and the 1.5 logit rule). Therefore bias depends on *S.E.normalized* (i.e., item-sample targeting) only for shifts between 0.55 and 1.3 logits.

The right-hand columns of Table 1 show the normalized standard errors and *t*-tests. It can be seen that the size of *S.E.normalized* is comparable between the complete sample and the sub-sample and that the *t* values,  $t_i(12)n$ , are similar. The DIF sizes and significances are now in closer accord. It can now be seen that it is meaningful to apply both the size and significance criteria in classifying items for DIF.

Agustín Tristán

Clauser B.E. & Hambleton, R.K. (1994) Review of *Differential Item Functioning*, P. W. Holland, H. Wainer.

### Rasch Workshop

#### An Introduction to Rasch Measurement: Theory and Applications

by Everett V. Smith Jr. & Richard M. Smith

April 7-8, 2007 - Saturday-Sunday  
immediately before AERA

University of Illinois - Chicago

This training session on the theory and applications of Rasch measurement 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.

The format will consist of eight units:

- 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
- Analysis of Participants' Data.

The material covered in these units is an overview of material that would normally be covered in approximately three graduate level measurement courses. Registration includes the full 2-day workshop, a continental breakfast each morning, over 550 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*.

For more details and registration: [www.jampress.org](http://www.jampress.org) under Rasch Measurement Workshops

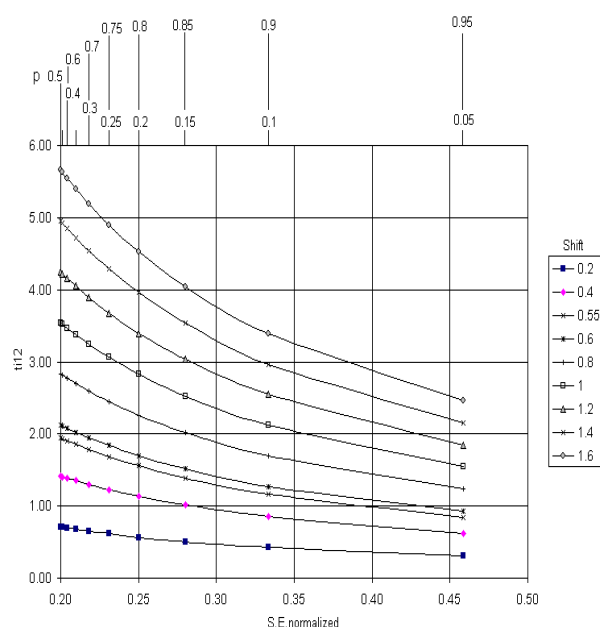


Figure 2. Theoretical *t* values as a function of shift and S.E. normalized

*Journal of Educational Measurement*, 31, 1, 88-92.

Draba R.E. (1977) *The identification and interpretation of item bias*. Educational Statistics Laboratory. Memo 25. University of Chicago. [www.rasch.org/memo25.htm](http://www.rasch.org/memo25.htm)

#### Statistical Significance vs. Substantive Boundaries

“The major scientific disadvantage of [significance testing and confidence intervals] is that their significance is merely an inference derived from principles of mathematical probability, not an evaluation of substantive importance for the big or small magnitude of the observed distinction. ... They offer no guidance for the basic quantitative scientific appraisals that depend on purely descriptive rather than inferential boundaries. ... The latter evaluation has not received adequate attention during the emphasis on probabilistic decisions; and careful principles have not been developed either for the substantive reasoning, or for setting appropriate boundaries, for big or small. After a century of significance inferred exclusively from probabilities, a basic scientific challenge is to develop methods for deciding what is substantively impressive or trivial.”

Feinstein, A. R. (1998). *P-values and confidence intervals: Two sides of the same unsatisfactory coin*. *Journal of Clinical Epidemiology*, 51(4), 355-60.

Courtesy of William P. Fisher

#### Psychological Measurement Impossible!

“Measurement can belong, therefore, only to that which is objective and spacial, and the psycho-physical quanta must stand for the physiological elements of our reactions, expressed in personal equations.”

George Herbert Mead. *"The Problem of Psychological Measurement"*, *Proc. of the American Psychological Association*, New York: MacMillan & Co. (1894): 22-23.

## Survey Design Recommendations

The field of survey-based measurement is maturing. Calibrated survey tools are proliferating rapidly in a number of fields. Much of this work stops with the calibration of an instrument that was not designed with the intention to produce invariant measures based in sufficient statistics. The description of data-based calibrations and measures is often the only goal of publications reporting this work. That is, little or no attention is typically accorded a theory of the construct measured. This is so, even though it has long since been recognized that “there is nothing so practical as a good theory” (Lewin, 1951, p. 169), and even though the practicality of theory has been made glaringly evident in the success of the Lexile Framework for Reading (Stenner, Burdick, Sanford, & Burdick, 2006).

Accordingly, even when instruments are precisely calibrated, survey content still tends to dominate the reporting of the results and the implicit definition of the construct. But is it not likely, in this scenario, that supposedly different constructs measured in supposedly different units might actually be one and the same? Is not the real proof of understanding a capacity to construct parallel instruments from theory in a way that results in equivalent measures across samples? Would not a focus on experimental tests like this work to spur consensus on what is measurable and on what works to change measures in desired directions?

To advance survey-based science in this direction, item writers and survey data analysts should follow nineteen basic rules of thumb to create surveys that

- are likely to provide data of a quality high enough to meet the requirements for measurement specified in a probabilistic conjoint measurement (PCM) model (Suppes, Krantz, Luce & Tversky, 1989),
- implement the results of the PCM tests of the quantitative hypothesis in survey and report layouts, making it possible to read interpretable quantities off the instrument at the point of use with no need for further computer analysis (Masters, Adams & Lokan, 1994); and
- are joined with other surveys measuring the same variable in a metrology network that ensures continued equating (Masters, 1985) with a single, reference standard metric (Fisher, 1997).

An initial experiment in this direction has been sponsored under the auspices of the *National Center for Special Education Accountability Monitoring (NCSEAM)*. With the November, 2004, reauthorization of the *Individuals with Disabilities Education Act (IDEA)*, states are required to report parents and families perceptions of the quality of the services received by them and their children. NCSEAM designed and piloted surveys in a research study intended to provide states with scientifically defensible measures. The research began with intensive qualitative research into the constructs to

be measured, involving literature reviews, focus sessions with stakeholders in several states, and Rasch analyses of survey data from several other states. This work paved the way for the intentional conceptualization of theories pertaining to several distinct constructs. A pilot study employing these surveys was itself designed so as to demonstrate as conclusively as possible the invariant comparability of the measures across independent samples of item. The success of this research has culminated in the practical application of the NCSEAM surveys to the new reporting requirement by a number of states, and with the emergence of a small community of special education and early intervention researchers, administrators, parents, and advocates who are learning how to use these tools to assess, compare, and improve the quality of programs.

For those wishing to emulate this program, the following recommendations are offered:

1. **Make sure all items are expressed in simple, straightforward language.** Use common words with meanings that are as unambiguous as possible.
2. **Restrict each item to one idea.** This means avoiding conjunctions (and, but, or), synonyms, and dependent clauses. A conjunction indicates the presence of at least two ideas in the item. Having two or more ideas in an item is unacceptable because there is no way to tell from the data which single idea or combination of ideas the respondent was dealing with. If two synonymous words really mean the same thing, only one of them is needed. If the separate words are both valuable enough to include, they need to be expressed in separate items. Dependent (if, then) clauses require the respondent to think conditionally or contingently, adding an additional and usually unrecoverable layer of interpretation behind the responses that may muddy the data.
3. **Avoid “Not Applicable” or “No Opinion”** response categories. It is far better to instruct respondents to skip irrelevant items than it is to offer them the opportunity in every item to seem to provide data, but without having to make a decision.
4. **Avoid odd numbers of response options.** Middle categories can attract disproportionate numbers of responses. Like “Not Applicable” options, middle categories allow respondents to appear to be providing data, but without making a decision. If someone really cannot decide which side of an issue they come down on, it is better to let them decide on their own to skip the question. If the data then show that two adjacent categories turn out to be incapable of sustaining a quantitative distinction, that evidence will be in hand and can inform future designs.
5. **Have enough response categories.** Not too few and not too many. Do not assume that respondents can make only one or two distinctions in their responses, and do not

simply default to the usual four response options (Strongly Agree, Agree, Disagree, Strongly Disagree, or Never, Sometimes, Often, and Always, for instance). The LSU HSI PFS, for example, employs a six-point rating scale and is intended for use in the Louisiana statewide public hospital system, which provides most of the indigent care in the state. About 75% of the respondents in the study reported have less than a high school education, but they provided consistent responses to the questions posed. Part of the research question raised in any measurement effort concerns determining the number of distinctions that the variable is actually capable of supporting, as well as determining the number of distinctions required for the needed comparisons. Starting with six (adding in Very Strongly Agree/Disagree categories to the ends of the continuum) or even eight (adding Absolutely Agree/Disagree extremes) response options gives added flexibility in survey design. If one or more categories blends with another and isn't much used, the categories can be combined. Research that starts with fewer categories, though, cannot work the other direction and create new distinctions. More categories have the added benefit of boosting measurement reliability, since, given the same number of items, an increase in the number of functioning (used) categories increases the number of distinctions made among those measured.

6. Write questions that will **provoke respondents to use all of the available rating options**. This will maximize variation, important for obtaining high reliability. This is a start at conceptualizing a theory. What kinds of questions will be most likely to consistently provoke agreeable responses, no matter how agreeable a respondent is? Conversely, what kinds of questions will be most likely to consistently provoke disagreeable responses, no matter how agreeable a respondent is? What is it that makes the variable evolve in this manner, along the hierarchy defined by the agreeability continuum of the questions? Articulating these questions in advance and writing survey items that put an explicit theory into play propels measurement into higher likelihoods of obtaining the desired invariance.

7. Write enough questions and have enough response categories to obtain an **average error of measurement low enough** to provide the needed measurement separation reliability, given sufficient variation. Reliability is a strict mathematical function of error and variation and ought to be more deliberately determined via survey design than it currently. For instance, if the survey is to be used to detect a very small treatment effect, measurement error will need to be very low relative to the variation, and discrimination will need to be focused at the point where the group differences are effected, if statistically significant and substantively meaningful results are to be obtained. On the other hand, a reliability of .70 will suffice to simply distinguish high from low measures. Given that there is as much error as variation when reliability is below .70, and it is thus not possible to distinguish two groups of measures in data this

unreliable, there would seem to be no need for instruments in that range.

8. Before administering the survey, **divide the items into three or four groups according to their expected ratings**. If any one group has significantly fewer items than the others, write more questions for it. If none of the questions are expected to garner very low or very high ratings, reconsider the importance of step 6 above.

9. Order the items according to their expected ratings and **consider what it is about some questions that make them easy** (or agreeable or important, etc.), and what it is about other questions that make them difficult (or disagreeable, unimportant, etc.). This exercise in theory development is important because it promotes understanding of the variable. After the first analysis of the data, compare the empirical item order with the theoretical item order. Do the respondents actually order the items in the expected way? If not, why not? If so, are there some individuals or groups who did not? Why?

10. Consider the intended population of respondents and **speculate on the average score** that might be expected from the survey. If the expected average score is near the minimum or the maximum possible, the instrument is off target. Targeting and reliability can be improved by adding items that provoke responses at the unused end of the rating scale. Measurement error is lowest in the middle of the measurement continuum, and increases as measures approach the extremes. Given a particular amount of variation in the measures, more error reduces reliability and less error increases it. Well-targeted instruments enhance measurement efficiency by providing lower error, increased reliability, and more statistically significant distinctions among the measures for the same number of questions asked and rating options offered.

11. If it is possible to write enough questions to calibrate a bank of more items than any one respondent need ever see, **design the initial calibration study to have two forms** that each have enough items to produce the desired measurement reliability. Use the theory to divide the items into three equal groups, with equal numbers of items in each group drawn from each theoretical calibration range. Make sure that each form is administered to samples of respondents from the same population who vary with respect to the construct measured, and who number at least 200. Convincing demonstrations of metric invariance and theoretical control of the construct become possible when the separate-sample calibrations of the items common to the two forms plot linearly and correlate highly, and when the common- and separate-form items each produce measures of their respective samples that also plot linearly and correlate highly.

12. Be sure to obtain enough **demographic information** from respondents to be able to test hypotheses concerning the sources of failures of invariance. It can be frustrating to see significant differences in calibration values and be

unable to know if they are due to sex, age, ethnic, educational, income or other identifiable differences.

13. **As soon as data from 30-50 respondents are obtained, and before more forms are printed and distributed, analyze the data** and examine the rating scale structure and the model fit using a measurement analysis that evaluates each item's rating scale independently. Make sure the analysis was done correctly by checking responses in the Guttman scalogram against a couple of respondents' surveys, and by examining the item and person orders for the expected variable. Identify items with poorly populated response options and consider combining categories or changing the category labels. Study the calibration order of the category transitions and make sure that a higher category always represents more of the variable; consider combining categories or changing the category labels for items with jumbled or reversed structures. Test out recodes in another analysis; check their functioning, and then examine the item order and fit statistics, starting with the fit means and standard deviations. If some items appear to be addressing a different construct, ask if this separate variable is relevant to the measurement goals. If not, discard or modify the items. If so, use these items as a start at constructing another instrument.

14. When the full calibration sample is obtained, **maximize measurement reliability and data consistency**. First identify items with poor model fit. If an item is wildly inconsistent, with a mean square or standardized fit statistic markedly different from all others, examine the item itself for reasons why its responses should be so variable. Does it perhaps pertain to a different variable? Does the item ask two or more very different questions at once? It may also be relevant to find out which respondents are producing the inconsistencies, as their identities may suggest reasons for their answers. If the item itself seems to be the source of the problem, it may be set aside for inclusion in another scale, or for revision and later re-incorporation. If the item is functioning in different ways for different groups of respondents, then the data for the two groups on this item ought to be separated into different columns in the analysis, making the single item into two. Finally, if the item is malfunctioning for no apparent reason and for only a very few otherwise credible respondents, it may be necessary to omit temporarily only specific, especially inconsistent responses from the calibration. Then, after the highest reliability and maximum data consistency are achieved, another analysis should be done, one in which the inconsistent responses are replaced in the data. The two sets of measures should then be compared in plots to determine how much the inconsistencies affect the results.

15. The instrument calibration should be **compared with calibrations of other similar instruments** used to measure other samples from the same population. Do similar items calibrate at similar positions on the measurement continuum? If not, why not? If so, how well

do the pseudo-common items correlate and how near the identity line do they fall in a plot? If the rating scale category structures are different, are the transition calibrations meaningfully spaced relative to each other?

16. Calibration results should be fed back onto the instrument itself. When the variable is found to be quantitative and item positions on the metric are stable, that information should be used to **reformat the survey into a self-scoring report**. This kind of worksheet builds the results of the instrument calibration experiment into the way information is organized on a piece of paper, providing quantitative results (measure, error, percentile, qualitative consistency evaluation, interpretive guidelines) at the point of use. No survey should be considered a finished product until this step is taken.

17. Data should be **routinely sampled and recalibrated** to check for changes in the respondent population that may be associated with changes in item difficulty.

18. For maximum utility, the instrument should be **equated with other instruments** intended to measure the same variable, creating a reference standard metric.

19. Everyone interested in measuring the variable should set up a **metrology system**, a way of maintaining the reference standard metric via comparisons of results across users and brands of instruments. To ensure repeatability, metrology studies typically compare measures made from a single homogeneous sample circulated to all users. This is an unrealistic strategy for most survey research, so a workable alternative would be to occasionally employ two or more previously equated instruments in measuring a common sample. Comparisons of these results should help determine whether there are any needs for further user education, instrument modification, or changes to the sampling design.

*William P. Fisher, Jr.*

Fisher, W. P., Jr. (1997). What scale-free measurement means to health outcomes research. *Physical Medicine & Rehabilitation State of the Art Reviews*, 11(2), 357-373.

Lewin, K. (1951). *Field theory in social science: Selected theoretical papers* (D. Cartwright, Ed.). New York: Harper & Row.

Masters, G. N. (1985, March). Common-person equating with the Rasch model. *Applied Psychological Measurement*, 9(1), 73-82.

Masters, G. N., Adams, R. J., & Loken, J. (1994). Mapping student achievement. *International Journal of Educational Research*, 21(6), 595-610.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement*, 7(3), 307-22.

Suppes, P., Krantz, D., Luce, R., & Tversky, A. (1989). *Foundations of Measurement, Volume II: Geometric and Probabilistic Representations*. NY: Academic Press.