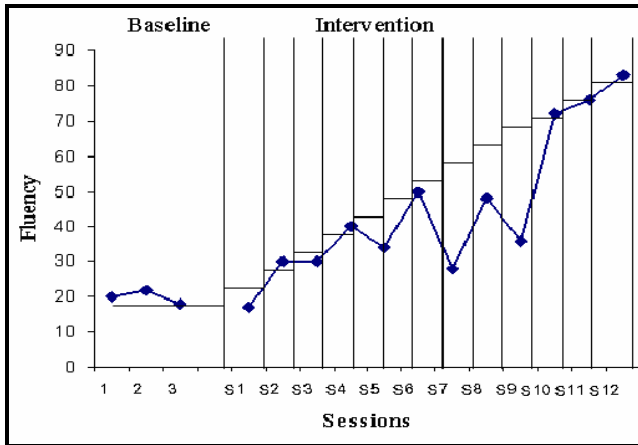# Noisy Responding and Intervention Effectiveness

The Rasch model is widely used for test development and item functioning. Noisy individuals are often a major concern because they distort item estimates, weakening inferences based on item difficulties. A typical practice is to omit and disregard individuals who are not constructive for item calibration purposes. But application of the Rasch model and information from misfitting participants can assist the evaluation of interventions in research on individuals. By examining misfitting participants useful information can be extracted regarding the strengths and weaknesses of an intervention or the specific behaviors of the individuals involved.

Figure 1. Tracy's fluency as a function of goal setting and reinforcement. Values above horizontal lines indicate Tracy achieved the goal for the session.

Figure 1 shows data from Tracy who was subjected to a "goal setting and reinforcement" intervention in order to improve her fluency. Using a changing criterion design, Tracy was asked to reach one goal after another and was reinforced contingently on her performance. The difficulty of the material was increased over sessions. As shown in Figure 1, Tracy reached some of the goals and missed some other, suggesting that the proposed intervention was not all that effective for her.

Using 5 students for comparative purposes I fit a Rasch model to the data of the 6 students by coding student's behavior during a session as 1 (if a student met the goal

for that session) or 0 (if the student missed the goal). Given that difficulty increased in every session we can hypothesize that student's behavior could fit the Rasch model (e.g., 1 1 1 1 1 0 1 0 1 0 0 0) for this particular research design. The presence of large fit statistics would suggest a pattern of responding that deviates from the Rasch model, or, in other words, that the intervention has not been as effective as hypothesized. An exception would be a one category responding (e.g., all 1s) which would still indicate that the student behaved contrary to expectations (as if the activity was too easy for the student).

Table 1 shows out-fit mean-square and point-measure correlations for all students, suggesting that Tracy's responding was not in accord to expectations, in comparison to her peers (see also Figure 2).

Raw Score	Measure	Outfit Mean-square	Point-measure correlation	Student
9	1.30	1.54	0.12	John
8	0.83	0.60	0.67	Lex
8	0.83	0.88	0.39	Bill
6	0.02	0.68	0.68	Mike
6	0.02	0.89	0.51	Steve
6	**0.02**	**1.58**	**-0.07**	**Tracy**

Table 1. Evaluation of students' response patterns.

Tracy's response pattern, 010101000111, is erratic, but provides useful information regarding her capabilities and her unique reaction to the intervention. She was able to reach some of the most challenging goals, but missed goals of medium and easy difficulty. She may fit the profile of an overachiever who got bored during some of the sessions or a student with attention problems, low motivation, etc. Alternatively, the intervention can be evaluated by examining, refining, and developing
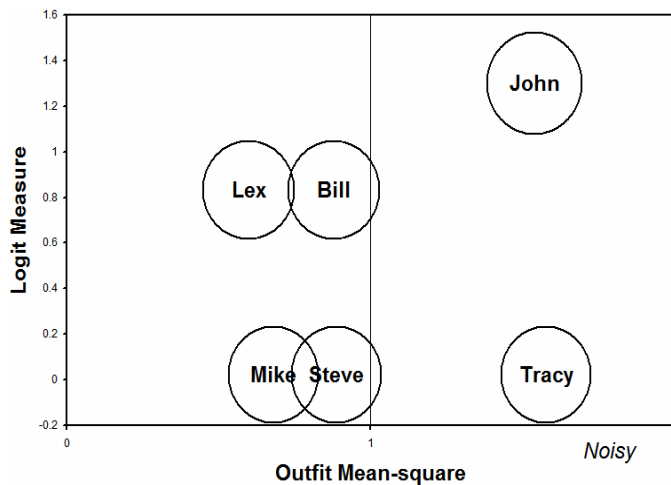
Figure 2. Erratic performance by Tracy and John.

the content of the sessions in which Tracy failed, if there is room to improve the content of the test.

Rasch Workshops

January 5-6, 2006, Thursday - Friday, Chicago IL
Introduction to *Facets*
conducted by Carol Myford and Lidia Dobria
www.winsteps.com/facwork.htm

January 16-20, 2006, Monday - Friday,
Canberra, Australia
Measurement in the Psychosocial Sciences:
from Raw Scores to Rasch Measures
conducted by Andrew Stephanou
www.acspri.org.au

February 25, 2006, Saturday, Taiwan
Rasch Analysis Workshop -
Introduction Course on Health care
conducted by Wen-Chung Wang
www.healthup.org.tw/rasch/950225.htm

March 20-21, 2006, Monday - Tuesday, Chicago IL
Introduction to *Winsteps*
conducted by Ken Conrad and Nick Bezruczko
www.winsteps.com/workshop.htm

April 6-7, 2006, Thursday-Friday, San Francisco CA
(pre-AERA)
Introduction to Rasch Measurement
conducted by Richard Smith and Everett Smith
www.jampress.org

June 25, 2006, Sunday, Hong Kong (pre-PROMS)
Introduction to *Winsteps*
June 26, 2006, Monday, Hong Kong (pre-PROMS)
Introduction to *Facets*
conducted by Mike Linacre
www.promshk.org

By focusing on a student's noisy performance, researchers in psychology and education may gain greater understanding of both student behaviors and intervention effectiveness. This knowledge base may assist them in developing effective educational interventions and in remediating learning difficulties.

Georgios D. Sideridis
Department of Psychology, University of Crete

An Introduction to Rasch Measurement:
Theory and Applications
April 6-7, 2006 (prior to AERA), San Francisco, CA

Directors: **Everett V. Smith Jr. and Richard M. Smith**

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.

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 breakfast each morning, assorted sandwiches and salads for lunch each afternoon, over 400 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. See www.jampress.org for more details on these publications.

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

For information and registration material, please visit www.jampress.org - *Rasch Measurement Workshops* or contact Everett Smith at evsmith -at- uic.edu

Registration Reminder

IOMW 2006

The 13th International Objective Measurement Workshop

April 5–7, 2006

University of California, Berkeley, CA, USA

Immediately preceding and across the San Francisco Bay from AERA 2006

Deadline for reduced rate at conference hotel: March 5, 2006

Hotel Durant, www.hoteldurant.com

The **International Objective Measurement Workshop** is the biennial conference of the objective measurement community. Researchers in this community are dedicated to the development and application of the Rasch family of item response psychometric models. All who are interested in the theory and practice of measurement in the human sciences, including education, psychology, sociology, and medicine, are invited to participate in IOMW 2006 as presenters and/or attendees.

Measurement as a constructive endeavor

This year's conference will offer participants an opportunity to reflect on **the centrality of the practice of measuring** within our work in objective measurement. A focus on the practice of measuring requires us to keep in mind the ultimate contributions of our work toward making meaningful, valid, and reliable characterizations of people and institutions that are often used as the basis for high-impact decisions about their future. In the service of measuring, the Rasch measurement model is a necessary but not sufficient condition for sustaining confidence in our work; we must consider other aspects of the design, development, and practical use of measures, including issues of construct theory, item design, interpretation, training, and implementation.

Conference location: International House, on the University of California, Berkeley campus. The Hotel Durant (one block from campus) is holding space for conference participants. Friday morning sessions and post-session workshops will be held at the Hotel.

For information on registration, hotel reservations, and transportation, please visit the conference website:

http://bearcenter.berkeley.edu/IOMW2006/

Proposals: Deadline for submissions: January 16, 2006 for individual papers; November 21 for proposals of symposia or workshops

In holding with the theme of the conference, we invite submitters to situate their work within the larger context of the practice of measuring. Papers focusing upon the relationship between the theoretical foundations of measurement and the practical use of measures in the field are particularly welcome. Proposals will be considered for symposia, individual papers, and post-conference workshops.

Hosted by:
Berkeley Evaluation & Assessment Research Center,
University of California, Berkeley.
Mark Wilson, Director.

Rasch Measurement in Health Sciences
Edited by Nikolaus Bezruczko

www.jampress.org

This compilation of Rasch health-science studies emerged out of presentations at the Third International Outcomes Measurement Conference at the University of Illinois at Chicago, June 2000. The book has the feel of sitting in Ben Wright's class held at University of Chicago. I, like Alan Heinemann (see Forward), had the pleasure of auditing Ben's class where graduate students and visiting researchers learned about the Rasch model by applying it to their own data. The dedication to Ben puts together many of the experiences he shared with us in class, presenting his amazing life and scientific journey. The chapters in this book, like the class, reflect the breath of studies and investigations in which the Rasch model can be applied in the health sciences.

Part I places the Rasch model in the context of measurement theory, highlighting the limitations of conventional psychometric approaches. **Part II** presents an overview of outcomes measurement in rehabilitation, presenting the Uniform Data System for Medical Rehabilitation, analysis of an instrument designed to measure a single dimension across impairment, activity and participation for brain injury, patient satisfaction showing that positively and negatively worded questions form different dimensions, methods to reduce differential item functioning across different languages for a quality of life measure, credentialing of health care professionals including computer adaptive testing, and the evaluation medical competence through standard patient-based testing. **Part III** addresses diagnosis and evaluation in healthcare, presenting the development of a shortened radiographic scale which maintains sensitivity to change, the relationship of blood chemistry to gout diagnosis, consistency of rater effects in inflammatory vascular disease in children, treatment effectiveness in post traumatic stress disorder, the monitoring functional status in patients with multiple sclerosis and measurement of air pollution exposure. **Part IV** presents special measurement topics including equating quality of life instrument for osteoporosis, anchoring measures to detect clinical change, comparing measures of depression, examining the dimensions of activities of daily living and instrumental activities of daily living, the measurement qualities of two quality of healthcare surveys and the development of a shortened quality of life survey that is statistically invariant across six cultures. **Finally,** commentary is presented, highlighting the contribution of the chapters in advancing measurement in healthcare noting the future challenges that must be addressed in order to advance fundamental measurement. While there is wide variability in the rigor of the studies presented, this limitation is offset by exposure to the almost infinite applications of the Rasch model in health care.

Craig A. Velozo, PhD, OTR/L
Research Health Scientist - Veterans Affairs Med. Ctr.
Dept. of Occupational Therapy, University of Florida

Frontiers in Educational Psychology
Editors: Russell F. Waugh
Edith Cowan University, 2005
www.novapublishers.com

On the Forefront of Educational Psychology
Editors: Russell F. Waugh
Edith Cowan Universiyt, 2004
www.novapublishers.com

Prologue; About the Authors

Meaningfulness, Measurement and Item Response Theory (IRT)

There is a basic principle of meaningfulness accepted by a wide cross-section of different philosophical viewpoints that justifies the use of fundamental measurement models like Rasch's to the exclusion of IRT models. That principle states that meaning is an abstract fictional ideal approximated to the extent that it is separable from the (metaphorical, geometrical, numerical, historical, etc.) figures representing it. According to the deconstructionist Jacques Derrida (1982, p. 229), for instance, "... the sense aimed at through these figures [of metaphor] is an essence rigorously independent of that which transports it." Similarly, the hermeneuticist Paul Ricoeur (1977, p. 293) concurs, saying "No philosophical discourse would be possible, not even a discourse of deconstruction, if we ceased to assume what Derrida justly holds to be 'the sole thesis of philosophy,' namely 'that the meaning aimed at through these figures is an essence rigorously independent of that which carries it over.'" Gadamer (1980, p. 100) also agrees, saying "It is clear to us that the figure which we draw to illustrate a mathematical relationship visually is not the mathematical relationship itself. ...in a manner of speaking one looks right through the drawn circle and keeps the pure thought of the circle in mind."

We find the same principle at work again in Mundy's (1986, p. 392) general theory of meaningful representation: "The hallmark of a meaningless proposition is that its truth-value depends on what scale or coordinate system is employed, whereas meaningful propositions have truth-value independent of the choice of representation, within certain limits. The formal analysis of this distinction leads, in all three areas [measurement theory, geometry, and relativity], to a rather involved technical apparatus focusing upon invariance under changes of scale or changes of coordinate system." The same focus on the independence of figure and meaning, or scale and proportion, emerges in a wide variety of other works on the creation of qualitative mathematical meaning (Heidegger, 1967; Luce, 1978; Narens, 1981, 2002; Roberts, 1985, 1994, 1999).

These issues of meaningfulness and measurement are explored at length by Fisher (2003a, 2003b, 2004). The basic point is that the content of tests and surveys ought to be used to illustrate mathematical relationships between abilities and difficulties visually and conceptually without confusing them with the mathematical relationships themselves. We need to look right through the sample of items used to illustrate the construct and keep the pure thought of the construct in mind, in the manner of the numerical and geometrical figures that are understood to paradigmatically define meaningful representation by, again, a wide range of diverse philosophers (Derrida, 1989, p. 66; Descartes, 1961, p. 8; Gadamer, 1989, pp. 412-3; Kant, 1970, p. 7; see Michell, 1990, pp. 6-8 for more).

Meaningful measurement requires what Rasch (1960) called parameter separation, what Ronald Fisher (1922) called statistical sufficiency, and what Luce and Tukey (1964) called conjoint additivity; all of these can be identified in Rasch models, but not in IRT models (Wright 1984, 1999). In purporting to produce meaningful results, IRT models assume, but do not test or establish, the separation of figure and meaning. Wood (1978, p. 31), accordingly, found himself "persuaded by Lumsden (Lumsden 1978) that two- and three-parameter models are not the answer - test scaling models are self-contradictory if they assert both unidimensionality and different slopes for the item characteristic curves." In David Andrich's (1988, p. 67) study of measurement, he notes that "[the 2-parameter IRT model] attempts to capture the differences in discriminations of the ICCs. The model destroys the possibility of explicit invariance of the estimates of the person and item parameters and will, therefore, not be pursued here."

Because of its internal contradictions, IRT discrimination and guessing parameter estimation requires much larger samples than are required for Rasch models (Lord, 1983). Even with very large samples, the estimation process may diverge instead of converge, a failure prevented by the authors of one popular IRT software program by eliminating the additional item parameters and reducing to a Rasch model on alternate iterations (Stocking, 1989). Given all of these issues, Lumsden (1978, p. 22) accordingly contends that "The two- and three-parameter logistic and normal ogive scaling models should be abandoned since, if the unidimensionality requirement is met, the Rasch (1960) one-parameter model will be realized."

It has lately been noted that some IRT advocates mistakenly think that unidimensionality is not tested in Rasch model applications, but is merely assumed. This has never been the case, but may be a perceptual by-product of a failure to recognize and accept the paradigmatic difference between the measurement perspective (data are fit to models specifying the relational structures needed for meaningfulness) and the statistical perspective (models are fit to data as a means of describing the data). IRT's item discrimination parameter captures and describes interactions between items and respondents in an essentially statistical exercise that compromises the requirements of measurement. This is, of course, quite a reasonable way to proceed in contexts where measurement has already been accomplished, and understanding of the relations between measures is at issue. But the need for information on these interactions does not require that they be estimated at the same time that item calibrations and respondent measures are estimated.

On the contrary, the internal contradictions and estimation problems introduced by the additional item parameter(s) are unnecessary complications that are easily overcome by removing them from the measurement environment and placing them in the statistical environment where they belong. Model fit statistics, for instance, usually correlate very highly (0.95 and up) with the discrimination parameter, do not confound the estimation process, have been available in Rasch software since the early 1970s, and are routinely employed in evaluating data quality. The fit statistics isolate specific scale dependencies that render measures and/or calibrations meaningless, and in doing so facilitate the deconstruction of the original research question in a critical search for a more fundamental articulation capable of supporting figure-meaning separation. As Derrida (in Wood and Bernasconi 1988, p. 88-9) put it, "I try to place myself at a certain point at which -- and this would be the very 'content' of what I would like to 'signify' -- the thing signified is no longer easily separable from the signifier." Model fit statistics flag failures of invariance in which the clarity of mathematical representation is compromised by content-dependencies that prevent the separation of the amount measured (the signified) from the number representing it (the signifier).

A popular Rasch measurement software package, Winsteps (Linacre, 2005), now makes it possible to estimate IRT item parameters without confounding the estimation of measures and calibrations (Linacre, 2004). The model fit statistics routinely produced in Rasch applications enable the study of individual residual differences between observed and expected responses, and are widely recognized for their diagnostic utility. The high correlation of the IRT discrimination parameter estimates with the fit statistics simply reproduces information already available and does not offer additional information beyond what is provided in the Rasch residuals. The recent inclusion of the IRT parameter estimates in Winsteps, then, provides an instructive point of contrast between the measurement and IRT perspectives, but does not enhance the existing substantive value of the information previously provided (Wright, 1992).

William P. Fisher, Jr.

Andrich, D. (1988). Rasch models for measurement. Beverly Hills, California: Sage Publications.

Derrida, J. (1982). Margins of philosophy. Chicago, Illinois: University of Chicago Press.

Derrida, J. (1989). Edmund Husserl's Origin of Geometry: An introduction. Lincoln: University of Nebraska Press.

Descartes, R. (1961). Rules for the direction of the mind (L. J. Lafleur, Trans.). The Library of Liberal Arts. Indianapolis: Bobbs-Merrill.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London, A, 222, 309-368.

Fisher, W. P., Jr. (2003, December). Mathematics, measurement, metaphor, metaphysics: Part I. Implications for method in postmodern science. Theory & Psychology, 13(6), 753-90.

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

Fisher, W. P., Jr. (2004, October). Meaning and method in the social sciences. Human Studies: A Journal for Philosophy and the Social Sciences, 27(4), 429-54.

Gadamer, H.-G. (1980). Dialogue and dialectic: Eight hermeneutical studies on Plato (P. C. Smith, Trans.). New Haven: Yale University Press.

Gadamer, H.-G. (1989). Truth and method (J. Weinsheimer & D. G. Marshall, Trans.) (Rev. ed.). New York: Crossroad (Original work published 1960).

Heidegger, M. (1967). What is a thing? (W. B. Barton, Jr. & V. Deutsch, Trans.). South Bend, Indiana: Regnery/Gateway.

Kant, I. (1786). (1970). Metaphysical foundations of natural science (J. Ellington, Trans.). Indianapolis, Indiana: Bobbs-Merrill.

Lord, F. M. (1983). Small N justifies Rasch model. In D. J. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing (pp. 51-61). New York: Academic Press, Inc.

Luce, R. D. (1978). Dimensionally invariant numerical laws correspond to meaningful qualitative relations. Philosophy of Science, 45, 1-16.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new kind of fundamental measurement. Journal of Mathematical Psychology, 1(1), 1-27.

Lumsden, J. (1978). Tests are perfectly reliable. British Journal of Mathematical and Statistical Psychology, 31, 19-26.

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

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

Narens, L. (1981). A general theory of ratio scalability with remarks about the measurement-theoretic concept of meaningfulness. Theory and Decision: An International Journal for Philosophy and Methodology of the Social Sciences, 13, 1-70.

Narens, L. (2002). Theories of meaningfulness (S. W. Link & J. T. Townsend, Eds.). Scientific Psychology Series. Mahwah, New Jersey: Lawrence Erlbaum Associates.

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

Roberts, F. S. (1985). Applications of the theory of meaningfulness to psychology. Journal of Mathematical Psychology, 29, 311-32.

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, The Netherlands: Elsevier.

Roberts, F. S. (1999). Meaningless statements. In R. Graham, J. Kratochvil, J. Nesetril & F. Roberts (Eds.), Contemporary trends in discrete mathematics, DIMACS Series, Volume 49 (pp. 257-274). Providence, RI: American Mathematical Society.

Stocking, M. L. (1989). Empirical estimation errors in item response theory as a function of test properties. Princeton, New Jersey: Educational Testing Service. ETS Research Reports.

Wood, R. (1978). Fitting the Rasch model: A heady tale. British Journal of Mathematical and Statistical Psychology, 31, 27-32.

Wood, D., & Bernasconi, R. (1988). Derrida and différance. Evanston, Illinois: Northwestern University Press.

Wright, B. D. (1984). Despair and hope for educational measurement. Contemporary Education Review, 3(1), 281-288.

Wright, B. D. (1992, Spring). IRT in the 1990's: Which models work best? Rasch Measurement Transactions, pp. 196-200 [http://www.rasch.org/rmt/rmt61.htm].

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), The new rules of measurement: What every educator and psychologist should know (pp. 65-104 [http://www.rasch.org/memo64.htm]). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Midwest Objective Measurement Seminar
May 20, 2005
School of Public Health
at the University of Illinois at Chicago

Accuracy of item difficulty estimates by subject matter experts. *Kirk Becker, Promissor, Inc.*

Impact of Altering Randomization Intervals on Precision of Measurement and Item Exposure. *Timothy Muckle and Betty A. Bergstrom , Promissor, Inc.*

Application of Variable Stop Rule and Pre-Calibration Procedures in a Computerized Adaptive Test for Substance Abuse Diagnosis. *Barth Riley, University of Illinois at Chicago , Kendon J. Conrad, University of Illinois at Chicago, Nikolaus Bezruczko, Measurement and Evaluation Consulting, Michael Dennis, Chestnut Health System*

Monitoring rater effects in communication skills assessment with Multi-Facet Rasch Measurement analysis. *Cherdsak Iramaneerat, Department of Educational Psychology, College of Education University of Illinois at Chicago*

Examining Unexpected Bias in Facets Analysis. *Ross Brown, Measurement Research Associates, Inc.*

Linking Subsets in Facets: An Issue not an Answer, *Lidia Martinez, Measurement Research Associates*

The psychometric properties of the Icelandic version of the Worker Role Interview (WRI-IS) using Facets. *Jessica Keller, University of Illinois at Chicago.*

What's so naïve about raw scores? *Nikolaus Bezruczko, Kendon J. Conrad, Michael Dennis, Chris Scott, & Rod Funk*

Using Concept Systems to develop a conceptual framework of expert's views of depression. *John W. Ridings, Institute for Clinical Social Work; Kendon J. Conrad, University of Illinois at Chicago*

Comparison of Resident and Faculty Perceived Preparedness of Residents for Surgical Cases. *Deborah Rooney; University of Illinois at Chicago, and Northwestern University, Northwestern Center for Advanced Surgical Education*

The development and testing of a scale to assess formal and informal strategies used by clinicians to encourage adherence to medication. *Beth Angell, School of Social Service Administration University of Chicago*

Special Poster Session: Rasch Analysis of a New Construct: Functional Caregiving for Adult Children with Intellectual Disabilities. *Shu-Pi C. Chen, School of Nursing St. Xavier University; Nikolaus Bezruczko, Measurement and Evaluation Consulting; Sheila Ryan-Henry, ACSW, Seguin Retarded Citizens Association.*

Thurstone on Social Science Measurement

"Although I have not attempted to gain competence in the general field of social psychology, our work in psychological measurement has naturally turned to the measurement of social values. This was largely due to our attempt to introduce some life and interest in psychophysics, which was dominated for a long time by the trivial problems of lifted weights and limen determinations. The extension of psychophysical methods to the measurement of social values was especially tempting when it turned out that the law of comparative judgment is entirely independent of the physical stimulus magnitudes. This circumstance enables one to use the law in the measurement of social and aesthetic values where physical stimulus measurement is entirely irrelevant.

"Our work on attitudes was started when I had some correspondence with Floyd Allport about the appraisal of political opinions, and there was discussion here at that time about the concept of social distance which was introduced by Bogardus. It was in such a setting that I speculated about the possible use of the new psychophysical toys. I wrote a paper entitled "Attitudes can be measured" (1928). Instead of gaining some approval for this effort, I found myself in a storm of criticism and controversy. **The critics assumed that the essence of social attitudes was by definition something unmeasurable.** There followed a number of other papers on the construction of particular attitude scales and on

methodology, including a little monograph on *The Measurement of Attitude* by Professor E. J. Chave and myself (1929). There was a good deal of interest in the subject and a lot of attitude scales were constructed for particular issues. .... Our best work in this field was a study, supported by the Payne Fund, on the effect of motion pictures on the social attitudes of high school children.

"There was heavy correspondence with people who were interested in attitude measurement, but they were concerned mostly with the selection of attitude scales on particular issues to be used on particular groups of people. **There seemed to be very little interest in developing the theory of the subject.** The construction of more and more attitude scales seemed to be unproductive, and I decided to stop any further work of this kind. Incomplete material for a dozen more attitude scales was thrown in the wastebasket and I discouraged any further work of that kind in my laboratory. I wanted to clear the place for work in developing multiple factor analysis."

"L. L. Thurstone." In Gardner Lindzey (ed.) A History of Psychology in Autobiography Vol VI.. Englewood Cliffs, NJ: Prentice Hall (1952): 294 - 321. [Emphasis RMT.]

Journal of Applied Measurement
Volume 6, Number 3. Autumn 2005

The Multilevel Measurement Model: Introduction to the Special Issue, *S. Natasha Beretvas and Akihito Kamata*

Demonstration of Software Programs for Estimating Multilevel Measurement Model Parameters, *J. Kyle Roberts and Rich Herrington*

Mixed Model Estimation Methods for the Rasch Model, *Frank Rijmen, Francis Tuerlinckx, Michael Meulders, Dirk J. M. Smits, and Katalin Balazs*

Some Links between Classical and Modern Test Theory via the Two-Level Hierarchical General Linear Model, *Yasuo Miyazaki*

Modeling Local Item Dependence with the Hierarchical Generalized Linear Model, *Hong Jaio, Shudong Wang, and Akihito Kamata*

The Cross-Classified Multilevel Measurement Model: An Explanation and Demonstration, *S. Natasha Beretvas, Jason L. Meyers, and Rolando A. Rodriguez*

Test Equating in the Presence of DIF Items, *Kwang-lee Chu and Akihito Kamata*

Richard M. Smith, Editor
JAM web site: www.jampress.org

Journal of Applied Measurement
Volume 6, Number 4. Winter 2005

Unique Properties of Rasch Model Item Information Functions, *Randall Penfield*

Evaluation of the Diabetes Self-Care Scale, *Nantawadee P. Lee and William P. Fisher, Jr.*

Rasch Analysis Examining Processing Mechanisms of the Object Location Memory, *Tracy L. Kline and Karen M. Schmidt*

Using the Rasch Model to Develop a Measure of Second Language Learners' Willingness to Communicate within a Language Classroom, *Christopher Weaver*

Knowledge and Understanding Asia: Using a Common Item Pool to Obtain National Estimates, *Patrick Griffin and Kerry Woods*

Measuring and Comparing Higher Education Quality between Countries Worldwide, *Anatoli A. Maslak, George Karabatsos, Tatijana S. Anisimova, and Sergi A. Osipov*

Understanding Rasch Measurement: The Relationship between the Rating Scale and Partial Credit Models and the Implication of Disordered Thresholds of the Rasch Models for Polychotomous Responses, *Guanzhong Luo*

Richard M. Smith, Editor
JAM web site: www.jampress.org

<div align="center">Demonstration of</div>

The Bootstrap Method of Rasch Conditional Item Difficulty Estimation

The bootstrap method was introduced by Efron (1979) as a general method for estimating standard errors and bias of parameters in statistical models. The bootstrap technique has become increasingly popular in educational testing because it provides researchers with an additional way of investigating research questions.

The application of the bootstrap method to the Rasch model seeks to provide a better understanding of what parametric Rasch models do. It is based on making no assumptions about the population distributions from which either items or persons are sampled. It also offers procedures for data with small numbers of items and persons. However, to the best of our knowledge, existing Rasch measurement software does not utilize the bootstrap method.

Two statistical software packages have been selected for illustration. One is the comprehensive public domain R statistical software, available for free (www.r-project.org). The other is the commercial Stata statistical package.

Conditional Item Difficulty Estimates in Rasch

Let X_{ij} be the binary or dichotomous (1,0) response by person i ($i = 1,..,N$) to item j ($j = 1,..,n$), where 1 denotes a correct response and 0 denotes an incorrect response. Then the probability of a correct response is $P_{ij} = P(X_{ij} = 1)$ and of an incorrect response is $Q_{ij} = 1 - P_{ij} = P(X_{ij} = 0)$. The simplest and the most widely quoted model for P_{ij} is the Rasch model (Rasch, 1960) and the Rasch conditional logistic model ($T_{i|j}$), for the difficulty of item j, β_i, is:

$$T_{i|j} = \ln\left[\frac{P_{i|j}(\theta)}{Q_{i|j}(\theta)}\right] = -\beta_j \quad (1)$$

where θ is the person ability parameter and β_j is the item parameter.

Data Structure

A typical dichotomous data structure for five individuals responding to three dichotomously-scored test items is shown in column 1-4 of Table 1. To estimate conditional Rasch item difficulty parameters, the data is reshaped lengthwise and the expression in (1) is used to create dummy variables, Th, which represent $-\beta_i$ in (1) (see column 5-10 in Table 1).

The dataset used in this demonstration consisted of 200 individual's responses to 10 dichotomously-scored examination questions. In both cases (R and Stata) each of the 10 items is replicated 1000 times to give the bootstrap estimate $(-\hat{\beta}_1^*,\ldots,-\hat{\beta}_9^*)$ while item 10 provides the local origin reference level when the conditional logistic Rasch method is applied.

R software demonstration and sample output

To apply the bootstrap method, we used Davison and Hinkley's (1997) "boot" function in the R-statistical software library. In the "boot" function, "statistic" is a function that returns the statistic to be bootstrapped. The first two argument of the function "boot.cond" specified the reshaped data set "combine" and the index vector gives the indices of the observations included in the bootstrap sample while "R" is the number of bootstrap replicates. The sample output is shown below. A complete step-by-step guide for users unfamiliar with the R package is available from the authors upon request.

Stata software demonstration and sample output

In the Stata example, we build upon the program written by Weesie in 1997. "bs" is the bootstrap sample while "reps" is the number of bootstrap replicates. In most bootstrap applications, an investigation using a B=1000 bootstrap sample will essentially be able to approximate the actual sampling distribution (Efron & Tibshirani, 1993). The sample output from the Stata statistical package is shown below.

In the programming output, the tenth item (item 10), with associated beta (1) = 0 is

---

Table 1. Data format for Rasch item difficulty estimation.

Response data Format				Estimation data format					
student_id	Ques1	Ques2	Ques3	student_id	item	Ques	Th1	Th2	Th3
1	1	0	0	1	1	1	-1	0	0
2	0	0	0	1	2	0	0	-1	0
3	0	0	1	1	3	0	0	0	-1
4	1	0	1	2	1	0	-1	0	0
5	0	1	0	2	2	0	0	-1	0
				2	3	0	0	0	-1
				3	1	0	-1	0	0
				3	2	0	0	-1	0
				3	3	1	0	0	-1
				4	1	1	-1	0	0
				4	2	0	0	-1	0
				4	3	1	0	0	-1
				5	1	0	-1	0	0
				5	2	1	0	-1	0
				5	3	0	0	0	-1

For R, Ques = resp, student_id = id and Th = i.

Stata example output

```
use "H:\Document\maths.dta"
reshape long Ques, i(student_id) j(item)
for num 1/10: gen ThX = -(item==X)
bs "clogit Ques Th1-Th10, group(student_id)"
   "_b[Th1] _b[Th2] _b[Th3] _b[Th4] _b[Th5]
   _b[Th6] _b[Th7] _b[Th8] _b[Th9]",reps(1000)
   cluster(student_id)

Variable Observed  Bias   Std. Err. [95% Conf.
   Interval]
_bs_1 0.708  0.018  0.303  0.113  1.303  (N)
               0.100  1.305  (P)
               0.037  1.260  (BC)
_bs_2 0.747  0.028  0.299  0.159  1.334  (N)
               0.199  1.405  (P)
               0.135  1.335  (BC)
_bs_3 0.670  0.011  0.293  0.094  1.247  (N)
               0.084  1.252  (P)
               0.038  1.230  (BC)
_bs_4 0.952  0.043  0.309  0.345  1.560  (N)
               0.385  1.582  (P)
               0.331  1.472  (BC)
_bs_5 0.597  0.009  0.296  0.015  1.178  (N)
               0.035  1.210  (P)
               0.031  1.174  (BC)
_bs_6 0.526  0.014  0.296  -0.056 1.108  (N)
               -4.51E-16 1.097  (P)
               -0.031 1.090  (BC)
_bs_7 0.491  0.007  0.290  -0.077 1.061  (N)
               -0.066 1.084  (P)
               -0.065 1.080  (BC)
_bs_8 0.561  0.016  0.293  -0.014 1.137  (N)
               1.15E-16  1.122  (P)
               -0.031 1.096  (BC)
_bs_9 0.561  0.011  0.295  -0.016 1.140  (N)
               -0.015 1.166  (P)
               -0.036 1.139  (BC)
 Note: N = normal, P = percentile and BC = bias-corrected.
```

the reference level. This is because the conditional logit function programming is not able to define a reference level if item 10 is added. The "Observed" values (in Stata) and the "original" values (in R) are the conditional logistic Rasch difficulty parameters.

Kingsley E. Agho, University of Newcastle, Australia
Kingsley.Agho -at - newcastle.edu.au

James A. Athanasou, University of Technology, Sydney

Baron, J. & Li, Y. (2004). Notes on the use of R for psychology experiments and questionnaires. www.psych.upenn.edu/~baron/rpsych/rpsych.html

Davidson AC, Hinckley DV (1997). Bootstrap Methods and Their Applications. Cambridge: Cambridge UP.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. Annals of Statistics, 7, 1-16.

Efron, B and Tibshirani, J.C (1993). An introduction to the bootstrap. New York: Chapman & Hall.

Weesie, J. (1999). The Rasch model in STATA. STATA statistical software 7.0 : STATA Corp.

R example output

```
library(boot) ## bootstrap library is needed
library(splines) ## need this library
library(survival)## need this library
####estimating the full sample
examboth                          <-
   read.table('H:\Document/maths.txt',
   header=T)
combine <- reshape(examboth, v.names="resp",
   idvar   =   "id",   timevar="item",
   varying=list(c("Ques1","Ques2","Ques3","Que
   s4","Ques5","Ques6","Ques7","Ques8","Ques9"
   ,"Ques10")), direction="long")
indices   <-      sample(length(combine[,1]),
   replace=T)
boot.cond   <-   function(combine,   indices,
   maxit=20){
full.i.dummy                           <-
   diag(nlevels(factor(combine$item)))[factor(
   combine$item),]
full.i.dummy <- 0 - full.i.dummy # turns (0,1)
   into (0, -1)
full.i.dummy   <-   data.frame(full.i.dummy,
   row.names=NULL)
dimnames(full.i.dummy) [[2]] <- paste("i",
   2:11, sep="")
attach(full.i.dummy)
examboth.clog <- clogit(resp ~ i2 + i3 + i4 +
   i5 + i6 + i7 + i8 + i9 + i10 + strata(id),
   data= combine[indices,])
coeff <-coefficients(examboth.clog)
return(coeff) }
Rasch.boot <- boot(combine, boot.cond, 1000,
   maxit=100)
Rasch.boot
ORDINARY NONPARAMETRIC BOOTSTRAP
Call:
boot(data = combine, statistic = boot.cond, R
   = 1000, maxit = 100)
Bootstrap Statistics :
    original     bias     std. error
t1* 0.7084904 -0.7024243  0.2957786
t2* 0.7470233 -0.7412647  0.2929619
t3* 0.6707143 -0.6701935  0.2994406
t4* 0.9527469 -0.9553256  0.2943158
t5* 0.5972723 -0.5869482  0.2862424
t6* 0.5264065 -0.5244834  0.3029250
t7* 0.4918603 -0.4915181  0.3067102
t8* 0.5615339 -0.5602604  0.2915319
t9* 0.5615339 -0.5531485  0.3065304
```

Well, That's a Relief ....

"Most obviously the favourite technique of this psychometric subset, a group that displays certain characteristics of a mafia[1], is something called the 'Rasch model', named after Danish mathematician George Rasch.

[1] I hasten to point out that I have absolutely no reason to believe that members of this group have any links to organised crime."

Harvey Goldstein (2004) The Education World Cup: international comparisons of student achievement. Plenary talk to Association for Educational Assessment - Europe, Budapest, Nov. 4-6, 2004.

MOMS
Midwest Objective Measurement Seminar
December 9, 2005, 9:00 – 4:00
Rehabilitation Institute of Chicago

The purpose of the MOMS is to allow students, teachers and practitioners to present research in progress or results of research studies, as well as, provide a forum for discussion of measurement issues. The seminar provides the opportunity for those interested in measurement to get together for serious discussion and social interaction.

Please send your proposals to me. The proposal should be one paragraph that explains the purpose of the study, data, and results summary. Be sure to include your name and affiliation. Email proposals to Mary Lunz at

mlunz-at-measurementresearch.com

[*www.measurementresearch.com*](www.measurementresearch.com)

Social Studies, Science and Invariance
"The excuse is often made that social phenomena are so complex that the relatively simple methods of the older sciences do not apply. This argument is probably false. The analytical study of social phenomena is probably not so difficult as is commonly believed. The principal difficulty is that the **experts in social studies are frequently hostile to science.** They try to describe the **totality** of a situation and their orientation is often to the market place or the election next week. They do not understand the thrill of discovering an **invariance** of some kind which **never** covers the totality of any situation. Social studies will not become science until students of social phenomena learn to appreciate this essential aspect of science."

"L. L. Thurstone." In Gardner Lindzey (ed.) A History of Psychology in Autobiography Vol VI.. Englewood Cliffs, NJ: Prentice Hall (1952): 294 - 321.

Pacific Rim Objective Measurement Symposium (PROMS) 2006

Hong Kong

June 27-29, 2005 (Tues.-Thur.)

Paper proposals and Registration are invited.

www.promshk.org

Trevor Bond
Conference Chair

Magdalena Mo Ching Mok
Programme Chair

June 25-26, 2006 - Sunday-Monday:
Pre-Conference Workshops on Winsteps and Facets conducted by Mike Linacre

---

Applied Rasch Measurement: A Book of Exemplars

Papers in Honour of John P. Keeves

www.springer.com

Series: Education in the Asia-Pacific Region: Issues, Concerns and Prospects, Vol. 4, 2005

Alagumalai, Sivakumar; Curtis, David D.; Hungi, Njora (Eds.)