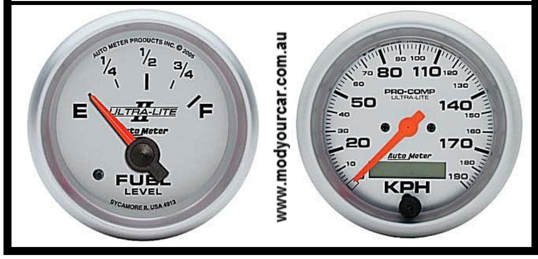
Teaching Rasch



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 24 No. 3 Winter 2010 ISSN 1051-0796

Fred Lord and Ben Wright discuss Rasch and IRT Models

These letters illustrate how Lord's and Wright's explorations intersected and then diverged, with Lord basing his thinking on the normal ogive and the frailties of empirical data, but Wright basing his on objective measurement and the demands of an ideal model.

Letter from Frederic M. Lord to Benjamin D. Wright, November 18, 1965: "Rasch's model for unspeeded tests [the Rasch dichotomous model] can be considered as a special case of the normal-ogive model, as Rasch himself points out extremely briefly at the end of Section 8 of his Chapter VII. The usual normal-ogive model has two parameters for each item, whereas Rasch uses only one of these. Rasch's model is thus a somewhat special case. Birnbaum's [2-PL] logistic model seems to provide a very satisfactory approximation to the normal-ogive model with two parameters per item. Altogether, we are devoting six chapters to the normal-ogive model and to Birnbaum's logistic model in our book ["Statistical Theories of Mental Test Scores"]."

Ben Wright to Fred Lord, November 23, 1965: "About Rasch's item analysis model as described in the latter part of his book, I think he would be horrified to learn that you regard his model as a special case of the normal-ogive model. The special feature of his model is that it allows for separating parameters of objects and agents, that is of children and test items. This is not possible with the normal-ogive model, and, in fact, if one sets down a few reasonable characteristics of objectivity, it can be proven that in the special case where observations are limited to ones and zeros, that the Rasch item analysis model is the only model which retains parameter separability. From Rasch's point of view this separability is a *sine qua non* for objective measurement."

Fred Lord of Ben Wright, November 26, 1965: "I am aware of the virtue of Rasch's model, which he elucidates very well in his chapter in the *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. On the other hand, it is quite clear that his model cannot really apply to the types of test items usually used in our tests. We all know that test items can have the same difficulty level and still differ very much in discriminating power -- some items have high discriminating power and some have none at all. This means that the item characteristic curves of typical test items frequently cross each other. In Rasch's model, it is impossible for the characteristic curves to intersect (except, of course, at the extreme ends where all curves meet in the same points).

"This leaves us with a dilemma. Shall we have *objective measurement*, which does not really hold for the test items we use? Or shall we allow the term *measurement* to include what we get from actual test items? I suppose one possible solution would be to discard all of those items that violate Rasch's assumptions. This possibility would certainly be an interesting one to explore."

Melvin R. Novick to Ben Wright, November 30, 1965: "I enjoyed reading your comment [in the letter to Lord] on Rasch's work as I too share a certain enthusiasm for it despite certain reservations and qualifications. ... More to the point, however, is Birnbaum's demonstration (see page 15 of part V of our text that the third Rasch model is a special case of his more general logistic model which obtains when all items have the same discriminating power. Since *few* tests are composed of items all having the same discriminating power, the practical utility of the third Rasch model would seem to be limited."

Ben Wright to Fred Lord, December 3, 1965: "If you write out Rasch's model for the binary case, that is where the alternative answers are right and wrong, and introduce a second item parameter, ... you can then take account of the variation and discriminating power of the items. This puts the model into the situation of there being one person parameter and two item parameters. The situation has a slightly unfortunate consequence as far as the estimation of item parameters are concerned. At least at present it

Table of Contents

seems to me that they now cannot be estimated entirely independently of the standardizing population.

"The other line, of course, is the one that you end up with and that is to only to accept items which conform to the simpler model, that is where the second parameter ... are all the same, let us say all one. Rasch believes that this is the only case where full objectivity can be reached. He has developed a proof which shows that only models of his kind, or models which reduce in a trivial way to his kind, allow for the specific objectivity in which he is interested.

"Should this proof stand the test of other people's scrutiny, well then I think the solution to discard all items that violate the Rasch assumptions may be the most attractive one and may even come to define the domain in which objective measurement is possible."

Ben Wright to Fred Lord, June 12, 1967: "Is there any reason for working for a normal ogive rather than a logistic ogive, or to put it in another way, is there a reason worth the added computing difficulty of working with the normal ogive?"

Fred Lord to Ben Wright, June 20, 1967: "You asked about the relative merits of the normal-ogive and logistic models. It is true that there is better *a priori* reason to use the normal ogive than the logistic; on the other hand, the difference between the two is so small that it would be very difficult to prove that one model was better than the other. The real answer to the dilemma is surely both models are wrong. Since they are so much alike, it seems futile to wonder whether one is slightly more wrong than the other. For this reason, I would use whichever is most convenient, until such time as we know a better model to use."

Ohio, Kentucky, & Indiana

On Friday April 9th, 2010 the initial meeting of the **Rasch Outcome Measures Research Group for Ohio, Kentucky, & Indiana** was held. The kick off talk and venue was supported by the Xavier University (Cincinnati, Ohio) Department of Nursing. The meeting was organized by Dr. Cynthia Kelly (Professor of Nursing, Xavier University) and Dr. Bill Boone (Professor of Educational Psychology, Miami University, Oxford, Ohio). In attendance were Xavier professors of nursing (Linda Moore, Lisa Niehaus, Marie Reynolds and Cathy Leahy), Miami professor of speech pathology and audiology (Geralyn Timler), as well as Dr. Tom O'Neill from the American Board of Family Medicine in Lexington, Kentucky.

ROM ReGroup plans to hold meetings every 2-3 months. If you are interested in being added to our mailing list, which will announce upcoming meetings, talks, and workshops, please contact:
Cynthia Kelly *kellyc3~at~xavier.edu*
or Bill Boone *boonewj~at~muohio.edu*

Rasch-related Coming Events

Jan. 7 - Feb. 4, 2011, Fri.-Fri. Online course: Rasch - Core Topics (Winsteps, introductory) (M. Linacre, Winsteps), www.statistics.com

Jan. 26, 2011, Wed. 5th UK Rasch User Group meeting, Warwick, UK www.rasch.org.uk

Feb. 28 - June 24, 2011, Mon.-Fri. Online course: Advanced course in Rasch Measurement of Modern Test Theory (Andrich, Marais, RUMM2030), www.education.uwa.edu.au

March 4 - April 1, 2011, Fri.-Fri. Online course: Many-Facets Rasch Measurement (Facets, intermediate) (M. Linacre, Winsteps), www.statistics.com

March 23-25, 2011, Mon.-Wed. In-person workshop: Introductory Rasch (A. Tennant, RUMM), UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Apr. 8-12, 2011, Fri.-Tues. AERA Annual Meeting, New Orleans, LA, www.aera.net

April 29 - May 27, 2011, Fri.-Fri. Online course: Rasch (Winsteps, introductory) online course (M. Linacre, Winsteps), www.statistics.com

May 4-6, 2011, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), UK,
May 9-11, 2011, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

June 23-25, 2011, Thurs.-Sat. 33rd Language Testing Research Colloquium LTRC, Ann Arbor, MI, USA, www.lsa.umich.edu/eli/LTRC2011

July 4-5, 2011, Mon.-Tues. International Workshop on Patient Reported Outcomes and Quality of Life, France, www.lsta.upmc.fr/mesbah/PROQOL/

Aug. 31 - Sept. 2, 2011, Wed.-Fri. IMEKO Conference, Jena, Germany, www.tu-ilmenau.de

Sept. 14-16, 2011, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), UK,
Sept. 19-21, 2011, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), UK,
Sept. 22-23, 2011, Wed.-Fri. In-person workshop: Advanced Rasch (A. Tennant, RUMM), UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Jan. 9-15, 2012, Mon.-Wed. In-person workshop: Introductory Rasch course (Andrich, RUMM2030),
Jan. 16-20, 2012, Mon.-Wed. In-person workshop: Advanced Rasch course (Andrich, RUMM2030), Perth, Australia, www.education.uwa.edu.au

Jan. 23-25, 2012, Mon.-Wed. Fifth International Conference on Probabilistic Models for Measurement in Education, Psychology, Social Science and Health, Perth, Australia, www.education.uwa.edu.au

How Skeptical are Magicians?

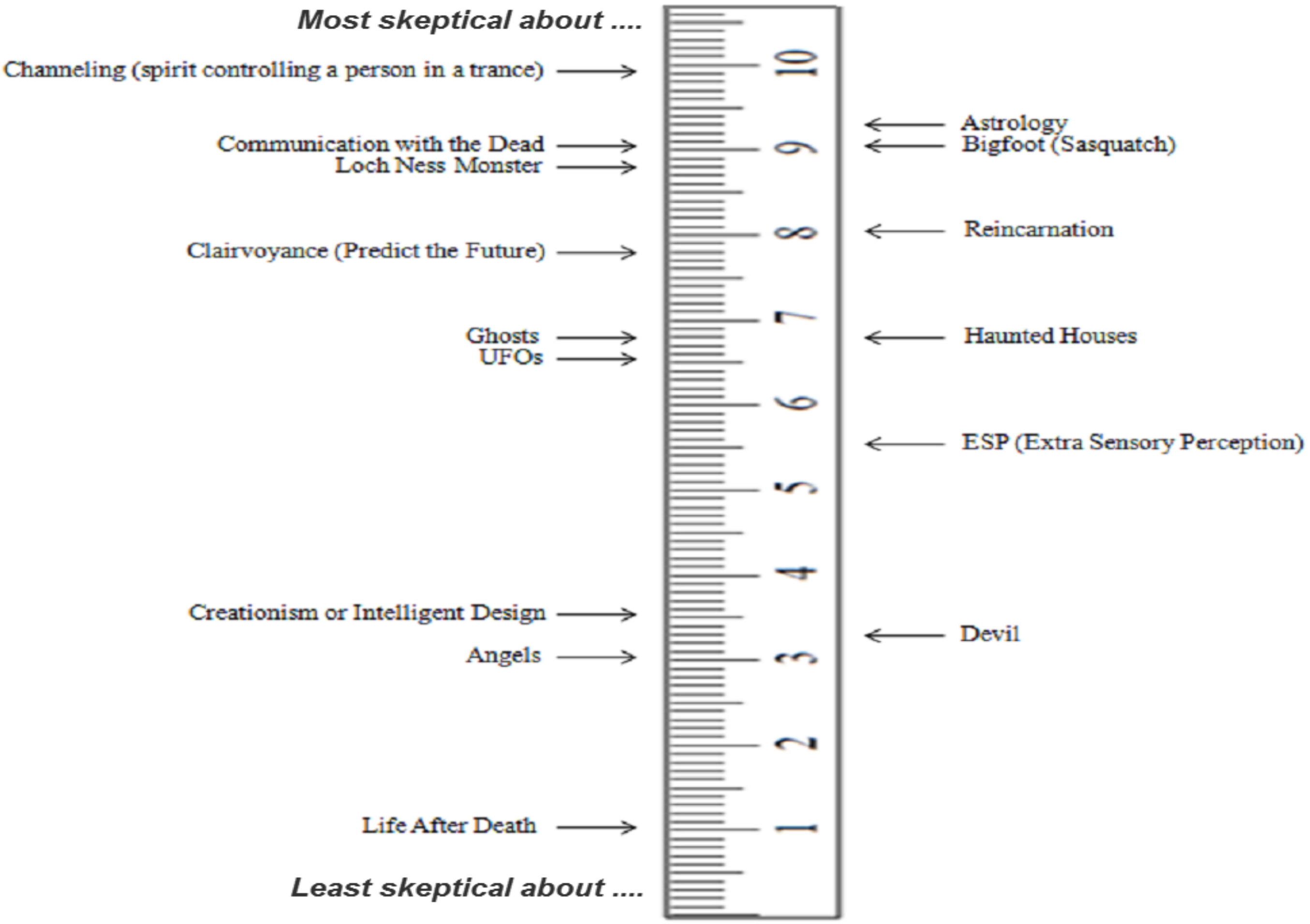
In 2010, renowned sociologist Dr. Peter Nardi published a study of magicians' beliefs about the paranormal. He was particularly interested in learning to what extent magicians believed various paranormal phenomena were possible. Nardi hypothesized that magicians would make a very interesting research sample because they are either true believers of paranormal phenomena, or because they are essentially "in on the secrets", the biggest skeptics of all.

Nardi administered a web-based survey in various magician Websites, discussion boards, and Internet chat rooms and was able to obtain a sample of 227 responses. I contacted Dr. Nardi and requested his data. I used the Rating Scale Model to analyze survey responses and rescaled the item logit values to fit a continuum from 1 to 10. Items located at the bottom of the scale (1) are the easiest for magicians to endorse (i.e., *Life after Death*). Items located at the top of the scale (10) are the most difficult for magicians to endorse (i.e., *Channeling*). As one might expect, items pertaining to religious notions are not very difficult to endorse, as magicians are a cross-section of the general public. However, what is especially interesting is that magicians believe UFOs, the Loch Ness Monster and Bigfoot are more plausible than astrology or channeling spirits. Perhaps the lesson here is to beware of card readers and psychics!

Kenneth D. Royal

Nardi, P. M. (2010). Magic, skepticism, and belief: An empirical study of what magicians believe about the paranormal. *Skeptic Magazine, 15(3),* 58-64.

Items	Measure	S.E.
(Most skeptical about ....)		
Channeling (spirit controlling a person in a trance)	10.00	.45
Astrology	9.30	.40
Communication with the Dead	9.08	.40
Bigfoot (Sasquatch)	9.05	.41
Loch Ness Monster	8.87	.42
Reincarnation	8.05	.37
Clairvoyance (Predict the Future)	7.81	.36
Ghosts	6.84	.34
Haunted Houses	6.83	.33
UFOs	6.56	.34
ESP (Extra Sensory Perception)	5.52	.32
Creationism or Intelligent Design	3.63	.32
Devil	3.25	.32
Angels	3.01	.32
Life After Death	1.01	.36
(Least skeptical about ....)		



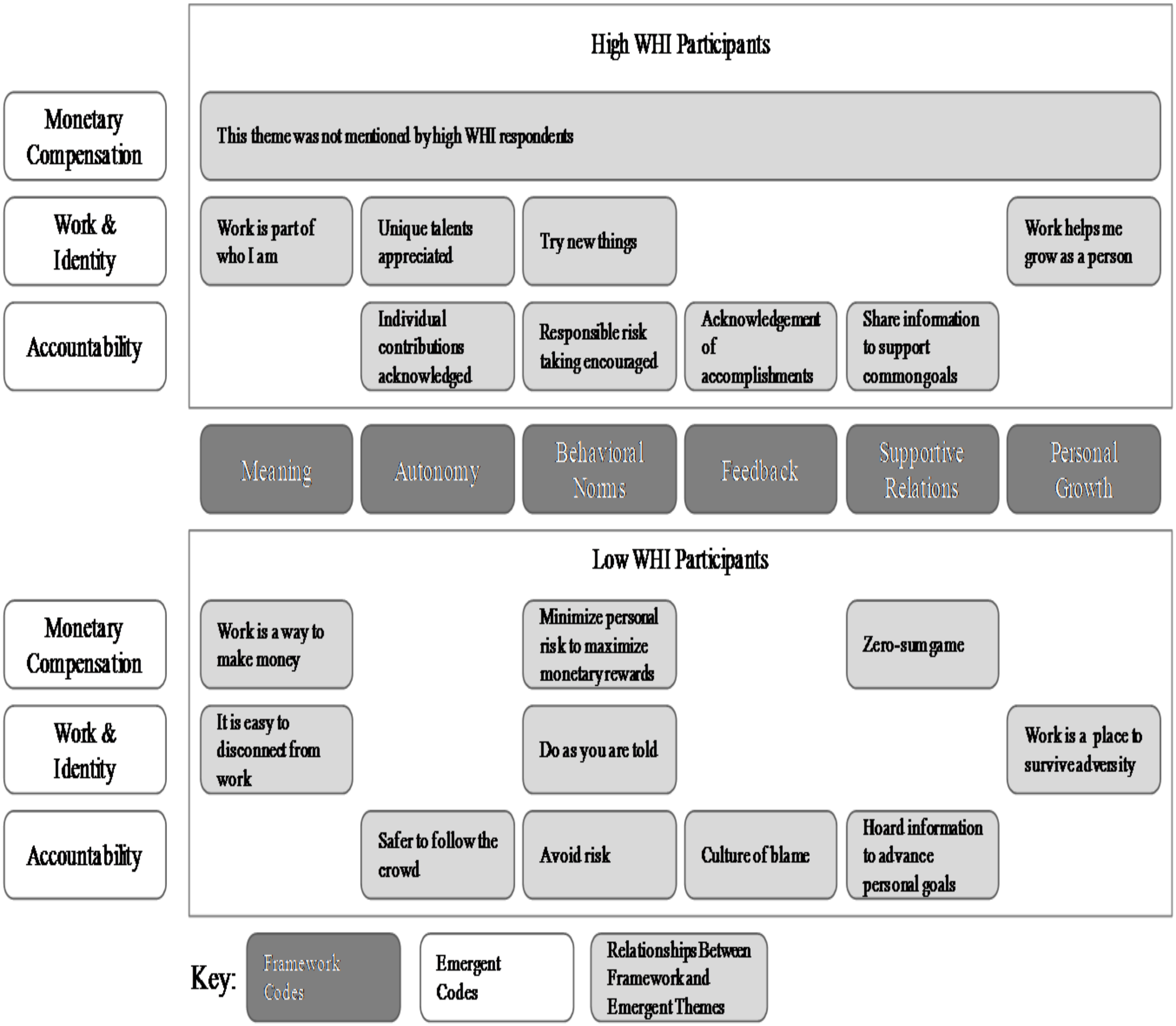
Lessons Learned While Developing the *Workplace Happiness Index*

This project was motivated by the observation that the majority of extant workplace measures focus on the alignment of individual employees to the goals and objectives of the organization. The goal of this project was to develop a measure that indicates an individual's level of satisfaction with the experience of work on a personal, psychological level. The project methodology was composed of three major methods: (1) develop a theoretical framework for the Workplace Happiness Index (WHI), (2) create the Rasch-based WHI measure, and (3) conduct semi-structured interviews to collect data to validate the WHI with respect to the theoretical framework (for full details see, Albano, 2010).

Completing this project yielded some valuable lessons for developing measures using this methodology.

1. Develop a solid theoretical footing

The proposed stems for the WHI were developed by a panel of expert practitioners in fields including organizational psychology, management consulting, and human resources management. Because the colloquial use of the term "happiness" is so varied, it was important to develop a precise definition of the phenomenon the WHI was intended to measure. Basing this definition on a thorough review of the literature accomplished task. The intent of the WHI is to measure happiness in a civic context. Aristotle's (2001) notion of eudemonic happiness provided both a civic anchor and historical context for the measure of happiness. Eudemonic happiness also provided a conceptual thread that lead to the inclusion of identity formation (Waterman, 2004) and psychological well-being (Ryff, & Keyes, 1995) as important pillars in the theoretical foundation upon which the WHI is based. This rich framework provided good guidance for focusing the efforts of the expert panel and providing a conceptual anchor for researchers using the WHI.


High WHI Participants
Monetary Compensation
This theme was not mentioned by high WHI respondents
Work & Identity
Work is part of who I am
Unique talents appreciated
Try new things
Work helps me grow as a person
Accountability
Individual contributions acknowledged
Responsible risk taking encouraged
Acknowledgement of accomplishments
Share information to support commongoals
Meaning
Autonomy
Behavioral Norms
Feedback
Supportive Relations
Personal Growth
Low WHI Participants
Monetary Compensation
Work is a way to make money
Minimize personal risk to maximize monetary rewards
Zero-sum game
Work & Identity
It is easy to disconnect from work
Do as you are told
Work is a place to survive adversity
Accountability
Safer to follow the crowd
Avoid risk
Culture of blame
Hoard information to advance personal goals
Key:
Framework Codes
Emergent Codes
Relationships Between Framework and Emergent Themes

2. Use Rasch statistics to examine fidelity to the theoretical model

The theoretical model for happiness used in the WHI identifies six elements that are indicative of an experience of happiness. The expert panel developed stems based on each of these elements and the resulting instrument was tested (*N* = 86) using the Rasch rating scale model. During this testing, misfitting stems were identified and examined for possible exclusion from the final version of the instrument. One stem–" My work is stressful"—is indicative of the importance of this analysis and its relationship to the theoretical model. In testing, this stem showed poor fit characteristics (IN.MSQ = 2.12, IN.ZSTD = 5.78). Examination of the stem showed that it was developed to test the theoretical element "A sense of meaningfulness in one's work". Upon further examination, the panel speculated that the stem was not indicative of the underlying construct—as an example, and emergency room doctor might find work both stressful and meaningful—and dropped the stem because of its lack of fidelity to the underlying model indicating the importance of using both Rasch statistics and an understanding of the underlying model to decide when to remove stems and when to attempt to rewrite them.

3. Interview data can provide rich evidence of validity

After administering the final version of the WHI to a second respondent pool (*N* = 67), I selected a group of high-(N=4) and low-scoring (N=4) respondents to participate in a follow-up semi-structured telephone interview. Interview data were examined and coded first with respect to the six themes developed in the theoretical model and then with respect to emergent themes (Bazeley, 2007). This analysis provides evidence of the validity of the instrument for separating respondents based on their experience of each of the theoretical themes and suggests additional themes for further investigation of workplace happiness.

Joseph F. Albano, Jr.

Albano, J. F., Jr. (2010). *Developing a measure and an understanding of the individual experience of happiness at work*. Retrieved from ProQuest Dissertations and Theses database. (AAT 3371929)

Aristotle. (2001). Ethica Nicomachea [The Nicomachean Ethics] (W. D. Ross, Trans.). In R. McKeon (Ed.), *The basic works of Aristotle* (pp. 927-1112). New York: Random House.

Bazeley, P. (2007). *Qualitative data analysis with NVivo*. London: Sage.

Ryff, C. D. & Keyes, C. L. M. (1995, October). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology, 69*(4), 719-727.

Waterman, A. S. (2004, July). Finding someone to be: Studies on the role of intrinsic motivation in identity formation. *Identity, 4*(3), 209-228.

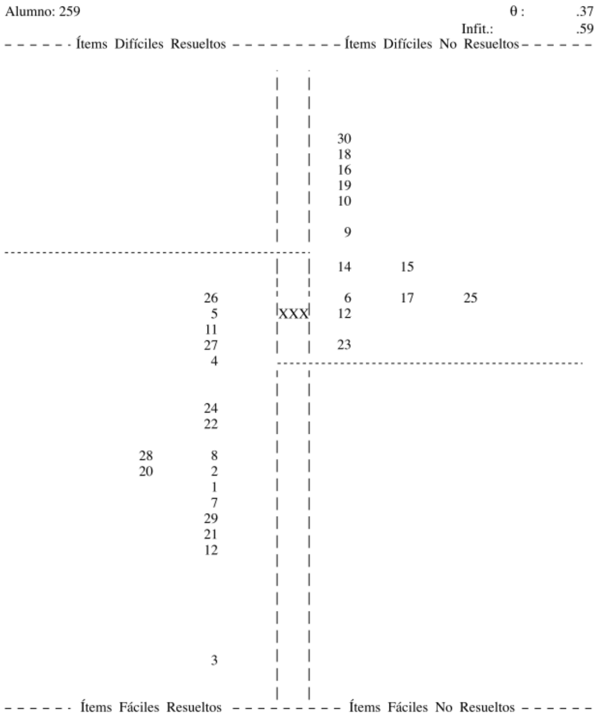


Figure 2. Map of the performance of a student. In G. Prieto and A. R. Delgado (2003). Análisis de un test mediante el modelo de Rasch. Psicothema, 15:1, 94-100

CAT Requires a Change in Thinking

"The condition-specific CATs (Computer-Adaptive Tests) were received well by clinicians and patients with respect to limited time to complete, but the [we] continued to receive calls related to *changing of items* from one CAT administration to another over the time of rehabilitation. Researchers considered these comments supportive of the *CAT functioning properly*, administering more difficult items to higher functioning patients as they improve during rehabilitation, but *clinicians wanted to serially track the patient's improvement per question*, which the CATs did not permit." (p. 296)

Hart D. et al. (2010). Implementing Computerized Adaptive Tests in Routine Clinical Practice: Experience Implementing CATs. *Journal of Applied Measurement, 11:3,* 288-303.

CAT requires a change of perspective from the details of the protocol to the meaning of the measures on the latent variable.

Online Educational Research Journal

www.oerj.org/View?action=frontpage

"OERJ is an entirely internet-based educational research journal. It is available to anyone who can access the web and all articles can be read and downloaded online. Anybody can submit articles as well as comment on and rate articles. Submissions are published immediately provided certain rules are followed. The language of the journal is English."

The Rasch Model in Europe: A History

Dr. Gerhard Fischer
Vienna, Austria

25. October 2010

Dear Mr. Purya Baghaei,

Let me answer your request for the history of the Rasch model in Europe.

Erling Andersen was a statistics student of Georg Rasch. He wrote a master's thesis on discrete measurement models with applications to data of social psychology (1966, in Danish), spent a year or so in the USA, then became a coworker of Rasch (as an assistant professor) and, after his doctoral thesis (*Conditional Inference for Multiple Choice Questionnaires,* 1973, in English) attained a position as senior lecturer (or similar) at the University of Economics in Copenhagen. When Georg Rasch retired, he applied for the vacant chair and was appointed Rasch's successor. Erling Andersen had a very important influence on the spread of knowledge about, and the further development of, the theory of the Rasch Model (RM) owing to his many original articles in international journals and the books he published (for some references to Erling's work, see, e.g., Fischer & Molenaar, 1995, *Rasch Models*, New York: Springer.) Unfortunately, Erling deceased in 2004.

I came to meet Georg Rasch in 1966 when he was one of the five lecturers at the NUFFIC (Netherland's Universities' Foundation For International Cooperation, sponsored by NATO) summer seminar on Psychological Measurement Theory at The Hague, Netherlands. At that time I had just become assistant professor at the Department of Psychology of the University of Vienna and had no previous knowledge of the RM. I was greatly taken by Georg's new approach to measurement in psychology because it appeared to solve some old fundamental problems of psychology. Apparently Georg noticed my sincere interest, for he invited me to come to Copenhagen for a deeper study of his approach, and he even raised a little money for employing me as his personal *amanuensis* for a period of two months in 1967. I did not meet Erling at that time in Copenhagen because he was staying in the USA; mostly I had contacts with Georg and with Peter Allerup, another young assistant of Georg, partly also with Jon Stene, a more senior lecturer at the department. (Please note that I am not quite sure about the exact positions of these persons, all the more so as the university system and academic degrees there were different from those in Austria.)

During my stay in Copenhagen, I wrote a computer program for conditional maximum likelihood (CML) estimation of the parameters of the RM, and Peter Allerup helped me to test it at the Nordisc Computer Centre (or similar name; it was a jointly Danish and Swedish closed shop facility). At that time no operational program of that kind was available and the algorithmic/numerical problems seemed all but trivial. One or maybe two years before, Ben Wright had undertaken to write such a program, but his approach to the computation of the elementary symmetric functions was too simplistic, and thus, unsuccessful. As a consequence, the Chicago group around Ben Wright as well as the researchers at ETS claimed that the CML method was "impractical". My algorithm and program were published by myself and Peter Allerup in the book *Psychologische Testtheorie* (G.H. Fischer, Ed., 1968. Berne: Verlag Huber) which was a proceedings volume of a symposium on test theory I had organized at a conference 1967 in Dusseldorf, Germany. The editing company wanted me to write an introduction to test theory in order to give the book a broader readership, so I wrote a brief exposition of classical test theory, a chapter on a fundamental critique of the classical theory, including of factor analysis, and an introduction to IRT, comprising a derivation of the RM and of the 2PL from their respective sufficient statistics (or raw scores) for the ability parameter. (A minor technical error in the derivation was corrected later in my 1974 book *Einführung in die Theorie Psychologischer Tests*; Berne: Verlag Huber; in German).

The 1968 proceedings volume had an unexpected effect; obviously many readers had also been subconsciously longing for a better theoretical foundation of quantification in psychology. The computer program was also used a great deal in the following years in the German language countries. An improved and extended set of programs was published in my 1974 book (see above). Both the greatly extended text (606 pages) and the programs were much read and used both in Germany and the Netherlands. Besides myself, my early assistants Hartmann Scheiblechner, who 1972 became professor at the University of Marburg, Germany; Hans Spada, who in 1972 was appointed to a research position at the IPN (an Institute for Science Education associated with the University of Kiel, Germany) and later became professor at the University of Freiburg, Germany; and my early student Wilhelm Kempf, who also became researcher at the IPN and later professor at the University of Konstanz, Germany, contributed a great deal to the understanding and further development of Rasch's approach to measurement in psychology. Our individual and/or joint publications and activities secured for the RM a fixed role both in teaching and research of psychology in Austria and Germany. When Scheiblechner and Spada in 1972 had left Vienna, Klaus Kubinger, now professor at the University of Vienna, and Anton Formann, who eventually became my successor as professor in psychological methodology, were appointed assistants at my department and made their academic career there. Both of them also have contributed much to the development and applications of probabilistic measurement models; Formann particularly specialized in latent class analysis as a generalization of IRT and has published a large number of papers on that topic in many international journals. Tragically, he suddenly perished in summer 2010. (Please note that all these remarks are made from memory and thus may be imprecise in some details or in their dates. The selection of what I mention here is also subjective, of course.) The German, Jürgen Rost, a well-known author in IRT in Germany, was a younger colleague of Spada and Kempf at the IPN at Kiel.

Another person with influence on the spread of the RM in Germany was Hans Christoph Micko, one more Austrian, eventually to become professor at the Technical University of Braunschweig, Germany, who published the multifactorial (or multifacet) generalization of the dichotomous RM already in 1969 and 1970, both in English and German languages (references in Fischer & Molenaar, 1995, see above). This was long before US researchers took interest in that topic. Maybe the first German who delved into the theoretical foundation of the RM was Hans Colonius; he published a very fundamental paper on the scale or measurement issue in the RM in 1979 in German. I think, however, that his primary interest was in mathematical psychology rather than psychometrics.

To me it seems that the first people who became interested in the RM in the Netherlands were (i) Leo van der Kamp, one of the two editors of the NUFFIC seminar proceedings (Leyden University, 1967, mimeographed) and co-editor of the volumes by Dato de Gruijter and Leo van der Kamp (Eds., 1976), *Advances in Psychological and Educational Measurement* (Proceedings volume of the *Second International Symposium on Educational Testing,* held in Montreux, Switzerland, 1975; London: J. Wiley) and by Leo van der Kamp, Willem Langerak, and Dato de Gruijter (Eds., 1980), *Psychometrics for Educational Debates* (Proceedings volume of the *Third International Symposium on Educational Testing,* held in Leyden, Netherlands, 1977; London: J. Wiley); and (ii) Dato de Gruijter, who also published several journal articles about Rasch modeling, partly in Dutch. Of course there will have been others of whose activities I was just not aware to the same extent.

In the mid seventies (maybe it was 1975) Leo van der Kamp with a group of some 20 younger staff members or graduate students traveled to Vienna to meet with my little psychometric group, and we presented a one-week intensive seminar on our IRT work to them. This apparently triggered interest and research on IRT in the Netherlands. In 1977 and 1978 I was invited to teach similar seminars at the Universities of Nijmegen, Groningen, and Twente (Enschede), all in the Netherlands. Again, the audiences were staff members and graduate students. At that time, the interest in IRT was growing fast in the Netherlands, and they were very interested in knowing about our research in Vienna. In subsequent years it was mainly at CITO (Arnhem; Norman Verhelst and Cees Glas), at the University of Twente (Enschede; Wim van der Linden), at the University of Groningen (Ivo Molenaar), at the University of Utrecht (Gideon Mellenbergh), and in part at the University of Nijmegen (Eddie Roskam) where the RM and related models were studied and further developed. (Again, this is just a personal recollection and may be quite incomplete.) Anyway, our colleagues and friends in the Netherlands were then catching up very fast and soon took the lead in IRT in Europe. As I have often experienced, their resources were far superior to ours in Vienna, so that it was impossible for us to keep pace with them.

If you want to know more about references to early work on the RM and/or on my personal views on the RM-related models, see the 1995 (2nd revised printing 1998) volume by Fischer & Molenaar, *Rasch Models;* New York: Springer-Verlag. More recently I have given a concise account of my views, including some of my later research, in the chapter *"Rasch Models"* in the volume of Rao & Sinharay (Eds., 2007), *Psychometrics. Handbook of Statistics, Vol .26;* Amsterdam: Elsevier.

I hope these remarks are useful for you, but please do not hold me responsible for the correctness of all details and for completeness. When I quit the Department of Psychology upon retirement in 1999, I sat on a huge amount of material after 39 years of academic work. I was neither able nor willing to take all this home. So I dumped most of it, including the documents concerning most seminars, conferences, presentations, travels, etc. Therefore, I cannot easily reconstruct exactly in what years I have been here or there. Moreover, I apologize if I am overlooking or just not mentioning persons of equal importance in the early European IRT work.

Best regards,

Gerhard Fischer

Is the Partial Credit Model a Rasch Model?

Question: How is it possible for the Andrich Rating-Scale version of the Rasch polytomous model to satisfy the requirements of statistical sufficiency of person and item raw scores and separability of model parameters, but not the Masters Partial-Credit version?

Answer: It isn't. Every parameter of a Rasch model (including the Partial Credit Model) has a matching raw score, its "sufficient statistic". This leads to parameter separability. Of course, those raw scores (and the observations summed to make them) are not independent, because every observation manifests more than one parameter. In the dichotomous model, every observation contributes to one person and one item. In most Rasch polytomous models, every observation contributes to one person, one item and one ordinal category.

Gerhard Fischer has a chapter in the book *"Rasch Models"* in which he derives Rasch polytomous models from sufficiency. His example is a multidimensional form of the Partial Credit Model. He also derives the Partial Credit Model from other Rasch-related criteria.

Rasch Models *(Gerhard Fischer, Handbook of Statistics, Vol. 26, 2007)*

Gerhard Fischer, a leading Rasch theoretician since the 1960s, and a pillar of the Rasch community, has retired. His article in the *Handbook of Statistics* summarizes his legacy and his perspective on Rasch measurement. The article is 71 pages long. It is a thorough algebraic exposition of major aspects of the Rasch dichotomous model (RM), supported by numerical examples. Here are the section headings:

Rasch Models

1. Some history of the Rasch Model (2 pages). This recounts the work of Georg Rasch, focusing on the Poisson model, and ending with the RM.

2. Some basic concepts and properties of the RM (6 pages). Local independence, likelihood functions, the raw score as a sufficient statistic.

3. Characterizations and scale properties of the RM (10 pages). Mathematical comparisons with other IRT models based on item response functions. Specific objectivity.

4. Item parameter estimation (15 pages), including
   4.1 Joint maximum likelihood estimation (2 pages)
   4.2 Conditional maximum likelihood estimation (4 pages)
   4.3 Marginal maximum likelihood estimation (5)
   4.4 An approximate estimation method (1 page). This minimizes a variance-weighted sum of squares.
   [Absent from this list is the pairwise estimation method (Rasch, 1980, pp. 171-2, Choppin, RUMM2020, etc.)]

5. Person parameter estimation. (2 pages). With known item difficulties, Maximum Likelihood Estimation and Warm Likelihood estimation.

6. Testing of fit (15 pages)
   6.1 Conditional likelihood ratio tests for comparing person groups. (3 pages). Item response functions (IRFs), model and empirical.
   6.2 Pearson-type tests. (3 pages). Glas-Verhelst tests of deviations between observed and expected frequencies.
   [Absent: mention of Wright's INFIT and OUTFIT statistics.]
   6.3 Wald-type tests. (1 page)
   6.4 Lagrange multiplier tests (1 page)
   6.5 Exact tests and approximate Monte Carlo tests (5 pages). Fit tests for person response strings when the item difficulties are known.

7. The linear logistic test model (7 pages)
   7.1 Testing the fit of an LLTM (1 page)
   7.2 Differential item functioning (DIF) [using LLTM] (3 pages)

8. Longitudinal linear logistic models (LLTM). (6 pages). Using LLTM across time-points.
   8.1 a unidimensional LLTM of change (1 page)
   8.2 A multidimensional LLTM of change (3 pages). Multiple dimensions modeled as parallel unidimensional LLTMs.
   8.3 The special case of two time points: The LLRA (1 page). Linear Logistic Test Model With Relaxed Assumptions

9. Some remarks on applications and extension of the RM. (2 pages)
   9.1 Dichotomous generalizations (1 Page). Mentioned are multifactorial RM, FACETS model, Mixed RM, One Parameter Logistic Model (OPLM), dynamic RMs.
   9.2 Polytomous generalizations (1 Page). Mentioned are Rating Scale Model (RSM). Partial Credit Model (PCM), IRT models and multidimensional IRT models, and a Rasch model for continuous data.

References (7 pages). Gerhard Fischer has 20 references as first author, Erling Andersen 9, Georg Rasch 7, no one else more than 4 references.

Gerhard Fischer's Insights:

"G. Rasch generally showed a preference for heuristic graphical methods over significance tests." p. 549

"Given these results, an applied research worker might be inclined to conclude that "the RM fits the data". Statisticians are usually more reserved: they know that models never fit; models can at best fit to some degree of approximation. If the sample is large and the test powerful enough, the model will always be refuted." p. 552.

Comment: other Rasch philosophers would have worded this: "the data fits the RM", and "If the sample is large enough and the (statistical) test powerful enough, empirical data will always be shown to be defective."

But most revealing of Fischer's measurement philosophy is this critique of the modern use of the Rasch model from section 9:

"Applying the RM has recently become quite popular not only in psychology and education, but also in many other scientific domains. It is tempting to use the RM whenever a 'scale' consists of dichotomous observations ('items') and the raw score suggests itself as a useful data reduction or 'measure'. More often than not, such enterprises are futile, however, because the strict limitations of the RM are violated: unidimensionality, no guessing, parallel IRFs (or SO), no DIF with respect to gender, age, education, etc. Unidimensionality of the items requires, on the substantive level, that the items are of very homogeneous content; this often conflicts with psychologists' diagnostic aims. The requirement of no guessing strictly speaking excludes the popular multiple choice item format; in particular, it is hopeless to fit a RM to personality or attitude questionnaires with two (`yes' vs. `no') answer categories, because answers determined by the latent trait to be measured are indistinguishable from random responses. Absence of DIF is also a very hard criterion:

experience with many intelligence and achievement tests shows that all verbal items — or items having a substantial verbal component — are prone to DIF with respect to the subpopulations mentioned above. Even elementary arithmetic or science problems often show considerable DIF with respect to gender, depending on their content. Therefore, the RM can by no means be considered as an omnibus method for the analysis and scoring of all sorts of tests. Rather, it should be viewed as a guideline for the construction or improvement of tests, as an ideal to which a test should be gradually approximated, so that measurement can profit from the unique properties of the RM."

Comment: Fischer's conclusion that the RM is "an ideal to which a test should be gradually approximated" is one with which surely all Rasch practitioners agree. However, Fischer's intermediate finding, "More often than not, such enterprises are futile", is strongly contradicted by 40 years of the practical application of Rasch methodology by numerous analysts to messy ordinal data originating from many different sources. Rasch measures, and the insights obtained from Rasch analysis of the data, have generally proved to be informative.

Journal of Applied Measurement
Vol. 11, No. 2 Summer 2010

Using Item Response Modeling Methods to Test Theory Related to Human Performance. *Diane D. Allen, 99-111*

From Model to Measurement with Dichotomous Items. *Don Burdick, A. Jackson Stenner, and Andrew Kyngdon, 112-121*

Using Guttman's Mapping Sentences and Many Facet Rasch Measurement Theory to Develop an Instrument that Examines the Grading Philosophies of Teachers. *Jennifer Randall and George Engelhard, Jr., 122-141*

Development of Scales Relating to Professional Development of Community College Administrators. *Edward W. Wolfe and Kim E. Van Der Linden, 142-157*

Comparing Décalage and Development with Cognitive Developmental Tests. *Trevor Bond, 158-171*

Reliability of Performance Examinations: Revisited. *Mary E. Lunz and John M. Linacre, 172-181*

Understanding Rasch Measurement: Equating Designs and Procedures Used in Rasch Scaling. *Gary Skaggs and Edward W. Wolfe, 182-195*

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

"No problem can be solved until it is reduced to some simple form. The changing of a vague difficulty into a specific, concrete form is a very essential element in thinking." - J. P. Morgan

Journal of Applied Measurement
Vol. 11, No. 3 Fall 2010

Special Issue:
The Improving Efficiency in Outcome Measurement

Preface, Improving Efficiency in Outcome Measurement. *Allen W. Heinemann, 196*

Foreword, Emergence of Efficiency in Health Outcome Measurement. *Nikolaus Bezruczko, Guest Editor, 197-213*

Measuring One Variable at a Time: The Wright Way. *Ed Bouchard, 214-229*

Rasch-Derived Latent Trait Measurement of Outcomes: Insightful Use Leads to Precision Case Management and Evidence-Based Practices in Functional Healthcare. *Carl V. Granger, Marsha Carlin, John M. Linacre, Ronald Mead, Paulette Niewczyk, A. Jackson Stenner, and Luigi Tesio, 230-243*

Generally Objective Measurement of Human Temperature and Reading Ability: Some Corollaries. *A. Jackson Stenner and Mark Stone, 244-252*

A Clinically Meaningful Theory of Outcome Measures in Rehabilitation Medicine. *Robert W. Massof, 253-270*

Embedding Measurement within Existing Computerized Data Systems: Scaling Clinical Laboratory and Medical Records Heart Failure Data to Predict ICU Admission. *William P. Fisher, Jr. and Elizabeth C. Burton, 271-287*

Implementing Computerized Adaptive Tests in Routine Clinical Practice: Experience Implementing CATs. *Dennis L. Hart, Daniel Deutscher, Mark W. Werneke, Judy Holder, and Ying-Chih Wang, 288-303*

The Use of PROMIS and Assessment Center to Deliver Patient-Reported Outcome Measures in Clinical Research. *Richard C. Gershon, Nan Rothrock, Rachel Hanrahan, Michael Bass, and David Cella, 304-314*

Opportunities for CAT Applications in Medical Rehabilitation: Development of Targeted Item Banks. *Pamela A. Kisala and David S. Tulsky, 315-330*

Postscript, Emergence of Efficiency in Health Outcomes Measurement. *Karon F. Cook, 331-336*

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

More Objections to the Rasch Model

More objections have been raised to the application of the Rasch model to empirical data.

1. “The purpose of the Rasch model is to describe the data, so a poor fit of the Rasch model to the data invalidates the use of the Rasch model.”

Describing the data is the purpose of many statistical models, such as regression models, but it is not the purpose for using the Rasch model. The purpose of the Rasch model is to use the data to construct additive measures on a latent variable. These measures may or may not be a good description of the data. For instance, if the data contain lucky guesses, the data will be intentionally badly described by a Rasch model. The lucky guesses will contradict the Rasch measures and be detected with misfit statistics. For more, see “Rasch model as Additive Conjoint Measurement” www.rasch.org/memo24.htm

2. “The Rasch-Andrich Rating-Scale model and the Rasch-Masters Partial Credit model assume that the respondent is making a series of consecutive choices between neighboring categories.”

Those polytomous models specify that the respondent is making a choice from all categories simultaneously. Consecutive choices are specified in other models such as the Glas-Verhelst “Steps” (“Success”) Model or the “Failure” model, see *RMT 5:2, 155* www.rasch.org/rmt/52j.htm. However, experience indicates that even in situations where consecutive decisions are made, the Andrich or Masters models are often a better basis for measurement than consecutive-choice models. This may be because the respondent is aware of the other choices, even if they are not currently available for selection.

3. “Empirical items never measure in the same scale units. Real items have different discriminations. Consequently the Rasch model cannot be used.”

This is true about real items, but not about the Rasch model. We do not need exact concordance between items, we need useable concordance. Then we need to be alerted to where the lack of concordance has become a threat to useful measurement. Rasch analysis constructs as-concordant-as-possible additive measures based on items with different scale units (discriminations). Rasch analysis then reports the degree of non-concordance of each item using misfit statistics. Items with exceedingly high or exceedingly low discrimination are usually defective items for other reasons, see *RMT 7:2, 289* www.rasch.org/rmt/rmt72f.htm

4. “The responses by each respondent to each item must be independent for Rasch analysis to be successful.”

The Rasch ideal is local independence. Each item has a difficulty, a location on the latent variable. Each respondent has an “ability”, also a location on the same latent variable. A Rasch model predicts the expected response for each respondent to each item based on those locations. When the expected responses are subtracted from the observed responses, the resulting residuals are modeled to be independent. Of course, they never are! Again, misfit analysis comes to our rescue. Is the lack of local independence in the data sufficiently large and sufficiently pervasive to be a threat to the meaning of the additive measures? Experience indicates that thoughtfully-constructed instruments produce observations that are locally independent enough for the additive Rasch measures to be useful for inference.

5. “Rasch analysis can cause unidimensional data to appear multidimensional.”

No empirical data are strictly unidimensional. Imagine a perfectly constructed test. Each item implements the intended unidimensional latent variable. But each item also differs from every other item. The ways in which two items differ from each other must be independent of any other item, otherwise they will be locally dependent. Thus each item must implement the intended dimension and also its own “difference” dimension, unique to the item, and uncorrelated with the “difference” dimension of any other item. Of course, empirical items fall short in both regards. They do not exactly implement the intended variable, and their “difference” dimensions are somewhat correlated with the “difference” dimensions of other items.

The choice of variant of the Rasch model, and other decisions made by the analyst, can alter the impact of the inherent multidimensionality of the items. For instance, if polytomous items are rescored as dichotomies, the choice of cut-point in the rating-scale may exacerbate or ameliorate the unwanted correlations in the data. Accordingly, the analyst must be aware of this and may adjust the scoring accordingly. See for instance “Communication validity and rating scales”, *RMT 10:1, 482* www.rasch.org/rmt/rmt101k.htm

6. “Factor Analysis of the original responses is more accurate for investigating possible multidimensionality than unidimensional Rasch analysis.”

Factor analysis (FA) can report too many factors, *RMT 8:1, 347,* www.rasch.org/rmt/rmt81p.htm. But let us consider a practical situation, suppose that FA reports one substantial factor in the inter-item correlation matrix (according to Kaiser’s rule or whatever), but the Rasch analysis (PCA of residuals) reports that there is a sizable secondary dimension in the inter-item correlation matrix of the Rasch residuals (or *vice-versa*). Which is correct?

An obvious solution is to split the set of items into two subsets based on their dimensionality in the analysis which reports two possible dimensions. Then cross-plot the person raw scores or Rasch measures on the two subsets. If the correlation is close to 1.0 (especially when disattenuated for measurement error - *RMT 10:1, 479* www.rasch.org/rmt/rmt101g.htm) then we have falsified the empirical two-dimensional finding for this sample.

If the correlation between the two subsets is close to 0.0, then clearly there are two dimensions. Two different dimensions have been combined into one instrument. Inferences based on either dimension are weakened by the other. Suppose that the correlation of person scores or measures is not close to 1.0, but is, say, 0.8. Then is this one dimension or two? For instance, suppose the dimensions are reading and arithmetic for grade-school children. We see immediately that, for the purposes of instruction, they are different dimensions, but for the purposes of school administration, such as advancing the child to the next grade, they are different strands within the same "educational achievement" variable.

Consequently, from the Rasch perspective, the more accurate method for investigating multidimensionality is the method which provides the best guidance about the threat to the validity of the additive measures. FA may identify (or fail to identify) dimensions, but it provides uncertain information on which to base decisions about the threat to additive measurement.

John Michael Linacre

Measurement In Marketing Research

An Alternative Framework

Thomas Salzberger analyses current measurement approaches in terms of their compliance with the scientific requirements of measurement. He reaches the conclusion that the predominantly applied practices, to a varying extent, suffer from substantial shortcomings, and suggests an alternative framework of measurement based on the philosophy of Rasch modeling. In the Rasch model great importance is attached to the mathematical principles of measurements, which take precedence over 'flexibility' in terms of accommodating idiosyncrasies of the data. The Rasch model promises to narrow the gap between the quality of measurement in the natural sciences and in the social sciences.

The future of measurement in marketing is about to be set. This book aims to raise researchers' awareness of measurement issues and to contribute to a transfer of knowledge from psychometrics into marketing research.

Book published by Edward Elgar Publishing, 2009.

Benjamin D. Wright in Wikipedia

Please contribute your knowledge of Ben Wright to:
http://en.wikipedia.org/wiki/Benjamin_Drake_Wright
Courtesy of William P. Fisher, Jr. and Edward Bouchard.

Teaching Rasch

As a former high school and college physics teacher, I often had to dream up ways of showing students topics that were too fast (the speed of light) or too slow (the movement of tectonic plates) or too big (the galaxy) to pull off in class. I have also found that a few basic demonstrations in Rasch Measurement courses have helped my students better apply and remember Rasch theory and techniques.

In addition to using a ruler as Ben Wright did in Chicago, I have some added props. I went to a junkyard and purchased a speedometer and a car's fuel gauge. When I introduce measurement to my students I can pull out these props and ask participants to discuss what each device is measuring. "Is one device better than the other"?

Invariably the students comment that a speedometer works pretty well, and that as far as they know the difference between 30 mph and 35 mph is the same as the difference between 65 mph and 70 mph. And they always comment that the fuel gauge is kind of odd, in that once you fill up the car, the gauge does not seem to move very much as you drive. But when you have about a half a tank of gas, the needle seems to move more quickly with each mile driven. The speedometer is linear, but the gas gauge is non-linear!

They seem to really get it when I then float the idea that some of our measurement devices in education, medicine and other fields might sometimes act like the gas gauge. Usually I pull out a rating scale survey and place it next to the speedometer and the gas gauge. Which is it more like?

I finish up by stressing that if we want to improve what we do, we need to make sure we have gauges that mean what they say in all situations.

Recently some of my students have remarked that some cars have digital gauges that tell you how many miles you have left in your tank and that this might be an improvement over the old gauge design. Interestingly some students also mention that even if they have an accurate digital device that might say "105 miles until empty", they probably would still find themselves looking at the gauge with a needle, for it is easier to read with a quick glance. This of course ties into later topics that involve the use of devices such as Wright Maps to quickly communicate a picture of what is taking place!

William Boone. Miami University (Oxford, Ohio, USA)

Rasch Measurement Transactions
www.rasch.org/rmt
Editor: John Michael Linacre
Copyright © 2010 Rasch Measurement SIG, AERA
Permission to copy is granted.
SIG Chair: Michael Young
Secretary: Kenneth Royal
Program Chairs: Leigh Harrell & Stephen Jirka
SIG website: www.raschsig.org

Should Hypothesis Tests be Rejected?

"Rigorous mathematical methods have secured science's fidelity to fact and conferred a timeless reliability to its findings. ... But in practice, widespread misuse of statistical methods makes science more like a crapshoot. It's science's dirtiest secret: The "scientific method" of testing hypotheses by statistical analysis stands on a flimsy foundation. Statistical tests are supposed to guide scientists in judging whether an experimental result reflects some real effect or is merely a random fluke, but the standard methods mix mutually inconsistent philosophies and offer no meaningful basis for making such decisions. Even when performed correctly, statistical tests are widely misunderstood and frequently misinterpreted. As a result, countless conclusions in the scientific literature are erroneous, and tests of medical dangers or treatments are often contradictory and confusing."

Tom Siegfreid, Science News, 3/27/2010, 177:7, 26.
www.sciencenews.org/view/feature/id/57091/

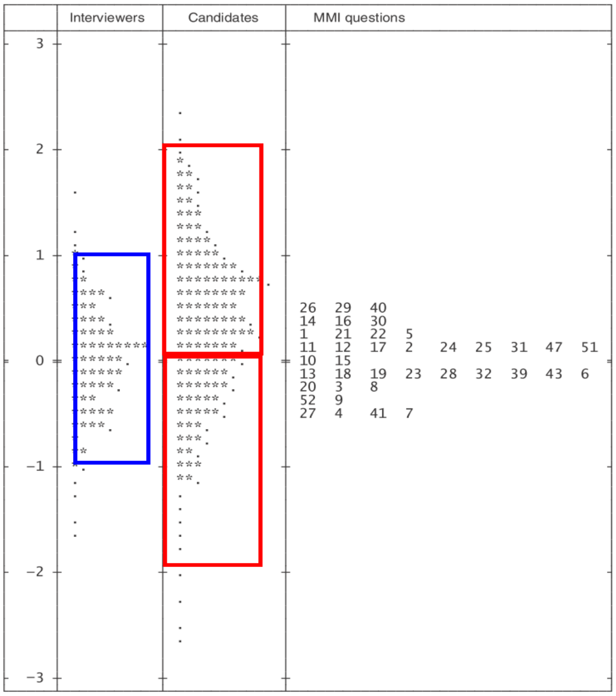


Figure 2 in "**Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview**?" by Chris Roberts, Imogene Rothnie, Nathan Zoanetti & Jim Crossley, *Medical Education 2010: 44:* 690–698

Notice that the practical logit range of the interviewers (raters, my blue box) is about half that of the candidates (my red boxes). So that:

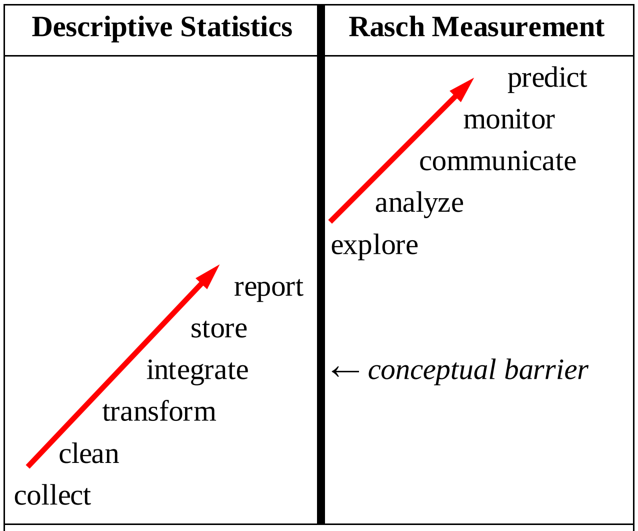
smartest candidate + most severe rater ≈
least smart candidate + most lenient rater

This was first noticed by Francis Ysidro Edgeworth around 1890, but 120 years later, Examination Boards continue to rely on the "luck of the draw" (as stated in Shavelson & Webb, *Generalizability Theory*, 1991, p.8).

Quality Control and Waste

Does Rasch quality-control fit-analysis waste items? Quality-control always generates "waste" along the road to creating good products, but there are not many car-drivers who would consider that binning (rejecting) misfitting tires is a waste. Nor many soldiers who would consider that binning misfitting ammunition is a waste.

W.E. Deming used to point out that proper quality-control reduces waste overall because the manufacturing process is improved. Using 2-PL and 3-PL facilitates sloppy item-writing, poor test-administration procedures and inadequate conceptualization of the latent variable. Perhaps this attitude toward psychometric "waste" is yet another reason why advances in the social sciences lag so far behind those in the physical sciences.



Descriptive Statistics

report
store
integrate
transform
clean
collect

Rasch Measurement

predict
monitor
communicate
analyze
explore

← *conceptual barrier*

After Stephen Few, www.perceptualedge.com

Rasch Mixture Models

Rasch Mixture (or Mixed) models (Rost, 1990) combine Latent Class Analysis (LCA) with Rasch analysis. LCA is based on Lazarsfeld PF, Henry NW. *Latent structure analysis*. Boston: Houghton Mifflin, 1968. This technique identifies classes or types within a sample, and then estimates each sample-member's probability of belonging to each class. Rasch models estimate each sample-member's ability within each class, and each item's difficulty for each class.

In order for the probabilities, abilities and difficulties to be uniquely estimable, the Mixture model analysis must be constrained. Typical constraints include: each person's ability (or each item's difficulty) is the same across classes, also the person abilities (and/or the item difficulties) are distributed normally within classes.

Software for estimating Rasch Mixture models includes WINMIRA and Latent GOLD.

Rost, Jürgen. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. Applied Psychological Measurement, 14, 271-282.