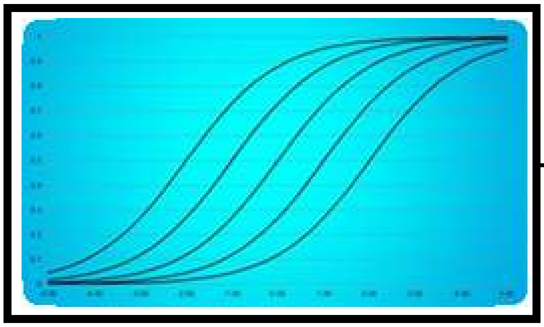


RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 29 No. 1 Summer 2015 ISSN 1051-0796

Coffee with Ben and IOMC 2015

Photographs presented here are from a recent meeting several participants had with Ben Wright, while *International Outcome Measurement Conference* (IOMC) was conducted in Chicago. This event was an unusual opportunity to congratulate Ben on celebrating his 90th birthday and reflect on the vision of scientific measurement he originally presented to health outcome measurement some years ago. It also forces recognition that advancing precision and objectivity in healthcare requires tenacity, as well as sensitive awareness of patient wellbeing.

IOMC was refreshing and reassuring, an enormous success. We are building the science that Ben dreamed about, implementing Thurstone's measurement principles with Rasch models to measure health outcomes. Participants from 12 countries attended 17 sessions across a broad range of outcome measurement issues. Rasch model integration with Medicare G, linear measurement of PROs, and computation of raw score mortality risk were among diverse topics. Those sessions are important achievements, as virtually all major federal funding for health outcome instrument development currently maintains and fortifies weaker, obsolete ordinal measurement methods.

Given the magnitude and intensity of those events over two conference days, I personally need to pause and reflect on the broader scene, especially after having attended National Council on Measurement (NCME) just couple of days before. A more somber contrast could not be made. I am driven to the conclusion that educational and psychological measurement practitioners actually like ordinal measures represented by true-score theory and item response theory (2, 3, or 4 item parameters). They really are not interested in linear measurement or scientific objectivity or the epistemology of scientific inquiry based on abstract scales. Contemporary psychometrics seems to be tripping over the stepping stones to scientific knowledge that Ben Wright laid out for outcome measurement. Feelings of despair and disappointment prevail.

Jack Stenner, Ben Wright, and Ong Kim Lee

Despite numerous examples of complacent implementation of raw scores and ordinal measures that can be harmful to patients, IOMC 2015 represents hope. The choice between ordinal and linear can never be arbitrary but must be judicious in recognition of patient effects. The wave of outcome measurement sweeping over healthcare must recognize measurement cannot be conducted without careful regard for *individual* patients. I am reminded of Luigi Tesio and Carl Granger's emphasis for several years and Richard Smith's concerns many years before them that outcome measurement must become person-metric. IOMC 2015 is a rejuvenation of that idea.

Nick Bezruczko

Table of Contents

Coffee with Ben

Starbucks at Oak and Rush
April 20, 2015

(From left to right): Jeremy Hobart, Filemon Cerda, Luigi Tesio, Richard Smith, Ben Wright, Jack Stenner, Ong Kim Lee, Craig Velozo, and William Fisher

(From left to right): Ong Kim Lee, Ben Wright, and William Fisher

Jack Stenner, Ben Wright and Ong Kim Lee

Ben Wright and Jeremy Hobart

(From left to right): Jeremy Hobart, Richard Smith, Matt Schulz, Jack Stenner and Luigi Tesio

Richard Smith and Matt Schulz

Craig Velozo and Luigi Tesio

William Fisher and Filemon Cerda

Benjamin Drake Wright (born 1926)

Update from the Rasch SIG Business Meeting

The annual Rasch SIG Business Meeting took place on Thursday, April 3 from 6:15 p.m. to 7:45 p.m. The meeting afforded a great opportunity to engage and network with colleagues. SIG Chair, Jim Tognolini, gave a presentation of the SIG's functioning over the past year and member of the leadership committee (also pictured) gave updates on the AERA conference program, finances and awards status. This year's keynote speaker was George Engelhard, Jr. who provided a thought-provoking presentation entitled "Invariant Measurement with Raters and Rating Scales". The meeting concluded with Stephanie Wind being presented the Georg William Rasch Early Career Publication Award.

Pictured: George Engelhard, Jr (2015 Rasch SIG Business Meeting Keynote speaker), Stephanie Wind (Recipient of the Georg William Rasch Early Career Publication Award) and Jim Tognolini (Rasch SIG Chair)

Pictured: Leigh Harrell-Williams (SIG Treasurer), Sara Hennings (Program Co-Chair), Jim Tognolini (SIG Chair), and Mikaela Raddatz (SIG Secretary)

Individual-Centered vs. Group-Centered Measures

The Preface to *Probabilistic Models for Some Intelligence and Attainment Tests* (Rasch, 1960) cites Skinner (1956) and Zubin (1955). In an argument whereby, "… individual-centered statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated" (Rasch 1960, p. xx). The Skinner reference is easily located. The mimeographed work by Zubin has not been found, but we did find another Zubin paper given at the 1955 ETS Invitational Conference on testing problems in which he writes, "An Example of the application of individual-centered techniques which keeps the sights of the experimenter focused on the individual instead of on the group…" (p. 116) may have helped Rasch situate his thinking. Rasch goes on to state "… present day statistical methods are entirely group-centered so that there is a real need for developing individual-centered statistics" (p. xx). What constitutes the differences in these statistics?

While it is individual persons and groups of persons that are the focus of discussion, we begin with an even more simple illustration because human behavior is complex, and a single mechanical-like variable is a better illustration to one that is complex. We choose temperature for this illustration because measuring mechanisms (Stenner, Stone & Burdick, 2004) for temperature are well established and all report out in a common metric or degree (disregarding wind-chill, etc.). A measuring mechanism consists of (1) guiding substantive theory, (2) successful instrument fabrication, and (3) demonstrable data by which the instrument has established utility in the course of its developmental history.

Consider six mercury-tube outdoor thermometers that are placed appropriately in a local environment, but near each other. They all register approximately the same degree of temperature, independently verified by consulting NOAA for the temperature at this location. One by one each thermometer is placed in a compartment able to increase/decrease the prevailing temperature by at least ten degrees. Upon verifying the artificially induced temperature change for each thermometer, it is returned to its original location and checked to see if it returns to its previous value and agrees with the other five.

If each of the six thermometers measured a similar and consistent degree of temperature before and after the induced environmental intervention/manipulation, this consistency of instrument recording validates a deep understanding of the attribute "temperature" and its measurement. Each thermometer initially recorded the same temperature, and following a change to and from the artificial environment returned to the base degree of temperature. Furthermore, all the measurements agree.

Interestingly, the experimentally induced change of environment also produced what may be called *causal validity*, not unlike constructive validity (Cronbach & Meehl, 1954) inasmuch as the temperature was manipulated, fabricated, engineered, etc. via construction and use of the artificial environment. When measuring mechanism(s) such as outdoor thermometers are properly manufactured this result is to be expected, and this experimental outcome and its replication would be predicted prior to environmental manipulation from all we know about temperature and thermometers. This outcome might further be termed *validity as theoretical equivalence* (Lumsden & Ross, 1973) because the replications produced by all six thermometer recordings might be considered "one" temperature. Our theoretical prediction is expected as a consequence of the causal process produced by the experiment, and reported by all the instruments. *Causal validity* is a consequence of the successful theoretical predictions realized in the experiment. Its essence is "prediction under intervention." The manipulable characteristics of our experiment involving the base environment, change made by way of an artificial environment, and the final change of recorded temperature are the consequence of a well-functioning construct theory and measuring mechanism. Each of the six individual thermometers records a similarly induced experimental deviation and a return to the base state. Each thermometer constitutes an individual unit, and the six thermometers constitute a group albeit without variation, which is exactly what would be predicted.

Now consider a transition to human behavior. Height is the new outcome measure and the determination of height at a point in time can be obtained from another well-established measuring mechanism – the ruler, which provides a point-estimate for one individual measured at a single point in time. When this process is continued for the same individual over successive time periods we produce a trajectory of height for the person over time (purely individual centered as no reference to any other person(s) is required). From these values one may determine growth over time intervals as well as any observed plateaus and spurts well-known to occur in individual development. The individual's trajectory rate may also vary because of illness and old age, so we could discover different rates over certain time periods as well as determine a curvilinear average to describe the person's total trajectory. Growth in height is a function of time, and the human characteristics entailed in a person's overall development result from genetic and environmental makeup. These statistics are intra-individually determined. Such statistical analyses produce the "individual-centered statistics" that Rasch spoke about.

Aggregating individual measurements of height into a group or groups is a common method for producing "group-centered statistics" often employing some frequency model such as the normal curve. This is most common when generalizing the characteristics of human growth in overall height based upon a large number of individuals. The difference between measuring a group of individuals compared to our first illustration using a group of thermometers is that while we expected no deviation among the thermometers, we do not expect all individuals to gain the same height over time, but rather to register individual differences. Hence, we resort to descriptive statistics to understand the central trend, and the amount of variation found in the group or groups. An obvious group-centered statistical analysis might aggregate by gender; comparing the typical height of females to males or provide norms tables

The measurement of height is straightforward and the measurement mechanism has been established over several thousand years. The same cannot be said for measuring mental attributes occurring in psychological, health, and educational investigations. Determining the relevant characteristics for their measurement is more difficult although the procedures for their determination should follow those already discussed. The major statistical hurdle is moving from the ordering of a variable's units to its "measurement application." The measurement models of Georg Rasch have been instrumental in driving this process forward.

Do we know enough about the measurement of reading that we can manipulate the comprehension rate experienced by a reader in a way that mimics the above temperature example? In the Lexile Framework for Reading (LFR) the *difference* between text complexity of an article and the reading ability of a person is *causal* on the success rate (i.e. count correct). It is true that short term manipulation of a person's reading ability is, at present, not possible, but manipulation of text complexity is possible because we can select a new article that possesses the desired text complexity such that any difference value can be realized. Concretely, when a 700L reader encounters a 700L article the forecasted comprehension rate is 75%. Selecting an article at 900L results in a decrease in forecasted comprehension rate to 50%. Selecting an article at 500L results in a forecasted comprehension rate of 90%. Thus we can increase/decrease comprehension rate by judicious manipulation of texts, i.e. we can experimentally induce a change in comprehension rate for any reader and then return the reader to the "base" rate of 75%. Furthermore, successful theoretical predictions following such interventions are invariant over a wide range of environmental conditions including the demographics of the reader (male, adolescent, etc.) and the characteristics of text (length, topic/genre, etc.).

Many applications of Rasch models to human science data are thin on substantive theory. Rarely proposed is an a priori specification of the item calibrations (i.e. constrained models). Causal Rasch Models (Stenner, Fisher, Stone & Burdick, 2013; Burdick, Stone, & Stenner, 2006; Stenner, Stone & Burdick, 2009; Stenner & Stone, 2010) prescribe (via engineering and manufacturing quality control) that item calibrations take the values imposed by a substantive theory. For data to be useful in making measures, those data must conform to the invariance requirements of both the Rasch model and the substantive theory. Thus, Causal Rasch Models are *doubly prescriptive*. When data meet both sets of requirements; the data are useful not just for making measures of some construct, but are useful for making measures of that precise construct specified by the equation that produced the theoretical item calibrations.

A Causal (doubly constrained) Rasch Model that fuses a substantive theory to a set of axioms for conjoint additive measurement affords a much richer context for the identification and interpretation of anomalies than does an unconstrained descriptive Rasch model. First, with the measurement model and the substantive theory fixed it is self-evident that anomalies are to be understood as problems with the data ideally leading to improved observation models that reduce unintended dependencies in the data (Andrich, 2002). Second, with both model and construct theory fixed it is obvious that our task is to produce measurement outcomes that fit the (aforementioned) dual invariance requirements. An unconstrained model cannot distinguish whether it is the model, data, or both that are suspect.

Over centuries, instrument engineering has steadily improved to the point that for most purposes "uncertainty of measurement," usually reported as the standard deviation of a distribution of imagined or actual replications taken on a single person, can be effectively ignored. The practical outcome of such successful engineering is that the "problem" of measurement error is virtually non-existent; consider most bathroom scale applications. The use of pounds and ounces also becomes arbitrary as is evident from the fact that most of the world has gone metric although other standards remain. What is decisive is that a unit is agreed to by the community and is slavishly maintained through substantive theory together with consistent implementation, instrument manufacture, and reporting. We specify these stages:

Theory ➡ Engineering ➡ Manufacturing ➡ Quality Control

The doubly prescriptive Rasch model embodies this process.

Different instruments qua experiences underlie every measuring mechanism; environmental temperature, human temperature, children's reported weight on a bathroom scale, reading ability. From these illustrations and many more like them we determine point estimates and individual trajectories and group aggregations. This outcome lies in well-developed construct theory, instrument engineering and manufacturing conventions that we designate *measuring mechanisms*.

Mark H. Stone and A. Jackson Stenner

References

Andrich, D. (2002). Understanding resistance to the data-model relationship in Rasch's paradigm: A reflection for the next generation. *J. Applied Measurement,* 3, 325-359.

Burdick, D. S., Stone, M. H., & Stenner, A. J. (2006). The combined gas law and a Rasch reading law, *Rasch Measurement Transactions, 20* (2), 1059-1060.

Cronbach, L & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Lumsden, J. & Ross, J. (1973). Validity as theoretical equivalence. *Australian Journal of Psychology, 25* (3), 191-197.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Skinner, B. F. (1956). A case history in scientific method. *The American Psychologist, 11*,
221-233.

Stenner, A. J., Burdick, D. S. & Stone, M. H. (2008). Formative and reflective models: Can a Rasch analysis tell the difference?, Rasch Measurement Transactions, 22 (1), 1152-1153.

Stenner, A. J., Stone, M. H., & Burdick, D. S. (2009). Indexing vs. measuring, Rasch Measurement Transactions, 22 (4), 1176-1177.

Stenner, A. J., Stone, M., Burdick, D. (2009). The concept of a measurement mechanism. Rasch Measurement Transactions, 23 (2), 1204-1206.

Stenner, A. J., Stone, M. H. (2010). Generally objective measurement of human temperature and reading ability: Some corollaries. *Journal of Applied Measurement, 11* (3), 244-252.

Stenner, A.J., Fisher, W.P., Stone, M.H. and Burdick, D.S. (2013). Causal Rasch Models. *Front. Psychol. 4:536.* doi: 10.3389/tpsyg.2013.00536

Zubin, J. (1955). *Experimental abnormal psychology* (mimeographed). New York: Columbia University Store.

Zubin, J. (1955). Clinical vs. actuarial prediction: a pseudo-problem. *In Proceedings of the 1955 Invitational Conference on Testing Problems*. Princeton: Educational Testing Service, 107-128

Individualized Medicine and Personalized Outcome Measures: Implications for Rasch Measurement

In recent years the notion of "individualized medicine" has become increasingly popular. Physicians have always known that everyone is different and that recovery for one person may look entirely different for another. The problem, however, has been that individualized medicine has not truly been possible (at least wide-scale) until more recently. Existing work in genetics and epigenetics coupled with everyday technologies are quickly paving the way for individualized medicine. In fact, on January 30, 2015, President Obama unveiled the "Precision Medicine Initiative" which specifically notes:

Most medical treatments have been designed for the 'average patient'. As a result of this 'one-size-fits-all-approach', treatments can be very successful for some patients but not for others. This is changing with the emergence of precision medicine, an innovative approach to disease prevention and treatment that takes into account individual differences in people's genes, environments, and lifestyles (The White House, 2015).

Eric Topol, a renowned cardiologist and geneticist, has discussed the notion of "individualized medicine" in a number of tangible ways, often illustrating what is possible when the human genome is mapped and how the use of personal data and technology are able to develop personalized treatments (Topol, 2014, 2015). Topol argues the future of medicine will involve using smartphones and similar devices to monitor our health and stay on top of treatments.

In another example Briggs (2015) discusses epigenomics and how the ability to turn genes on and off could help fight cancer. Briggs provides an excellent example answering the question "how is it that identical twins can share the exact same DNA but yet exhibit differences in growth, behavior and acquisition of illnesses?" He explains that cells read the genetic code embedded in DNA much like a script as opposed to a mold that replicates the same results. He suggest we should think of the genetic code much like a movie script in which the vision of the director (e.g., James Cameron vs. Woody Allen) could produce a significantly different film despite using the exact same script.

The aforementioned breakthroughs and initiatives show a great deal of promise for the future of medicine. However, many problems with measurement still remain that could potentially impede this progress. Most statistical approaches involve clustering, analyzing and reporting data at group levels. Certainly, these approaches have their purposes and are very useful when investigating relationships and trends. However, most "person-centered" analyses used in health outcomes measurement use statistical models that attempt to group individuals into homogenous subgroups and then determine predictors of subgroup membership (Muthén and Muthén, 2000). Since the 1950s, researchers (see Hayes, 1953; and Estes, 1956) have acknowledged that group summaries of data often obscure individual differences and make it difficult to draw appropriate inferences at the person level. It seems that in all our fervor to make sense of information and find solutions to problems many forget that all data sets consist of individual data points. Thus, while it often makes sense to group and norm some data, it is actually somewhat counter-intuitive to do this with outcomes measurement in the health, psychological and educational sciences. An example from sports medicine illustrates why.

Many familiar with baseball may be aware of a surgical procedure called the ulnar collateral ligament (UCL) reconstruction, also known as the "Tommy John Surgery" (TJS). The TJS procedure essentially consists of replacing the UCL with another tendon from elsewhere in the body. The procedure has been widely advertised to have success rates around 90-93% so many athletes, including those without UCL injuries, have opted to have the procedure as a means to enhance pitching performance (Ahmad, Grantham, & Greiwe, 2012).

However, such high success rates are based on group norms typically obtained from rehab scales that measure factors such as strength, mobility, comfort, sleep quality, etc. Of course, it is intuitive that what recovery looks like for one individual may not mirror that of another. A more accurate outcome measure would be to compare one's performance relative to a personal baseline measure. With baseball and the wide variety of data collected on pitchers (e.g., pitching velocity, accuracy, etc.), personal baseline measures have been increasingly investigated in recent years. Sports medicine physicians are now recognizing that when comparing a pitcher's post-UCL performance to his baseline statistics success rates typically drop to approximately 70-75% (Yurkiewicz, 2015). Thus, when outcomes are based on group norms they are, in many instances, inflated. The repercussions for such methodological carelessness is profound for persons considering these procedures, as an inflated success rate likely provides a false sense of security for potential patients and may impact ones decision to undergo an elective procedure. Further, assuming 25-30% of athletes are unable to match their personal baseline levels of performance, such procedures, especially when performed when medically unnecessary, could be devastating to one's career and future.

In the context of Rasch measurement, this difference between group- and individual-level comparisons takes on new significance, in genetics and elsewhere, as was noted by Markward and Fisher (2004, p. 131) in their study of 13 short tandem repeat marker loci from the FBI's CODIS database:

Rasch model parameters and fit statistics are estimated at the level of individual persons, marking an important methodological departure from classical statistical genetic, genetic epidemiological, and behavioral genetic approaches to measurement that rely exclusively on family- and group-level comparisons as the basis of inference and decision-making.

The need for this kind of important methodological innovation emerges also in comparisons of instrument sensitivities to change in health status, which sometimes produce clinically counter-intuitive results. Hobart, Cano, and Thompson (2010) show that two functional status instruments having the same sensitivity to change at the group level differed in their capacities to detect significant improvement by 50% to 31% at the individual level. Group level effect size indicators can be misleading, and Rasch measurement can provide meaningful and clinically interpretable quantitative comparisons at the individual level that are not otherwise available.

It is fair to say that if individualized medicine and personalized outcomes are indeed the future of medicine, then Rasch models should be expected to take on a more prominent role as these models are uniquely equipped to tackle many of these critical challenges.

Kenneth D. Royal, North Carolina State University
Melanie Lybarger, Independent Consultant
William P. Fisher, Jr., University of California - Berkeley

References

Ahmad, C. S., Grantham, W. J., & Greiwe, R. M. (2012). Public perceptions of Tommy John surgery. *The Physician and Sports Medicine*, *40*(2), 64-72.

Briggs, J. (2015). What is epigenomics and how is it helping in the fight against cancer? Mayo Clinic Individualized Medicine blog. Available at: http://individualizedmedicineblog.mayoclinic.org/discussion/what-is-epigenomics-and-how-is-it-helping-in-the-fight-against-cancer/

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*(2):134–40.

Hayes, K. J. (1953). The backward curve: A method for the study of learning. *Psychological Review, 60*, 269–75

Hobart, J. C., Cano, S. J., & Thompson, A. J. (2010). Effect sizes can be misleading: Is it time to change the way we measure change? *Journal of Neurology, Neurosurgery, & Psychiatry, 81*, 1044-1048.

Markward, N. J., & Fisher, W. P., Jr. (2004). Calibrating the genome. *Journal of Applied Measurement, 5*(2), 129-141.

Muthén, B. O., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24*(6), 882–91

The White House, Office of the Press Secretary (2015). Fact Sheet: President Obama's Precision Medicine Initiative [Press release]. Retrieved from: https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative

Topol, E. J. (2014). Individualized medicine from prewomb to tomb. *Cell, 157*(1), 241-253.

Topol, E. J. (2015). The future of medicine is in your smartphone. The Wall Street Journal. Available at: http://www.wsj.com/articles/the-future-of-medicine-is-in-your-smartphone-1420828632

Yurkiewicz, S. (2015, April 24). Did UCL Surgery Work? Baseball Stats May Have the Answer. *Medpage Today*. Available at:
http://www.medpagetoday.com/SportsMedicine/EliteSports/51184

Rasch-related Coming Events

July 3-31, 2015, Fri.-Fri. Online workshop: Practical Rasch Measurement – Further Topics (E. Smith, Winsteps), www.statistics.com

July 27-Nov. 20, 2015, Mon.-Fri. Introduction to Rasch Measurement (D. Andrich, I. Marais, RUMM), www.education.uwa.edu.au/ppl/courses

Aug. 14-Sept. 11, 2015, Fri-Fri. Online workshop: Many-Facet Rasch Measurement (E. Smith, Facets), www.statistics.com

Sept. 4-Oct. 16, 2015, Fri.-Fri. Online workshop: Rasch Applications, Part 1: How to Construct a Rasch Scale (W. Fisher), www.statistics.com

Sept. 9-11, 2015, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK,

Sept. 14-16, 2015, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), Leeds, UK,

Sept. 14-16, 2015, Mon.-Wed. IACAT Conference: International Association of Computerized Adaptive Testing, Cambridge, UK, www.iacat.org

Sept. 17-18, 2015, Thur.-Fri. In-person workshop: Advanced Rasch (A. Tennant, RUMM), Leeds, UK,

Oct. 16-Nov. 13, 2015, Thur.-Fri. Online workshop: Practical Rasch Measurement – Core Topics (E. Smith, Winsteps), www.statistics.com

Journal of Applied Measurement
Vol. 16, No. 1, 2015

A Mathematical Theory of Ability Measure Based on Partial Credit Item Responses, *Nan L. Kong*

Differential Item Functioning Analysis by Applying Multiple Comparison Procedures, *Paolo Eusebi and Svend Kreiner*

Visually Discriminating Upper Case Letters, Lower Case Letters and Numbers, *Janet Richmond, Russell F. Waugh, and Deslea Konza*

Testing the Multidimensionality of the Inventory of School Motivation in a Dutch Student Sample, *Hanke Korpershoek, Kun Xu, Magdalena Mo Ching Mok, Dennis M. McInerney, and Greetje van der Werf*

Measuring Teaching Assistants' Efficacy using the Rasch Model, *Zi Yan, Chun Wai Lum, Rick Tze Leung Lui, Steven Sing Wa Chu, and Ming Lui*

Detecting Measurement Disturbance Effects: The Graphical Display of Item Characteristics, *Randall E. Schumacker*

Criteria Weighting with Respect to Institution's Goals for Faculty Selection, *Sheu Hua Chen, Yen Ting Chen, and Hong Tau Lee*

Gendered Language Attitudes: Exploring Language as a Gendered Construct using Rasch Measurement Theory, *Kris A. Knisely and Stefanie A. Wind*

Richard Smith, Editor, www.jampress.org

Statistics Joke

A statistician goes hunting with two mathematicians. They spot a duck. The first mathematician levels his rifle, fires, and misses to the right. The second mathematician levels his rifle, fires, and misses to the left. The statistician turns to his friends and says "looks to me like we got him, boys" (p. 45).

Courtesy of Behar, R., Grima, P., & Marco-Almagro, L. (2013). Twenty-five analogies for explaining statistical concepts. *The American Statistician, 67*(1), 44-48.

Does Item Sequence Order Impact Local Dependence in Surveys?

Many survey researchers consider it a best practice to group related items together as it makes it easier for participants to complete the survey, gives the appearance of greater cohesiveness, and requires a lesser cognitive load from participants (Bradburn, Sudman, & Wansink, 2004; Dillman, 2000). However, a considerable number of studies have found that item order effects, often called "assimilation effects" or "carry-over effects", can result in biased participant responses (Heiman, 2002). A recent experience analyzing survey data provided an interesting case in which item ordering was hypothesized as the culprit for a number of statistically dependent item pairs. A simple experiment was conducted to test this hypothesis.

Background

Statistical dependency, also referred to as "local item dependence", refers to the extent to which a response to one item is directly influenced by a response to another item (Marais & Andrich, 2008). Survey researchers routinely investigate statistical dependency when evaluating the psychometric properties of an instrument. Typically, when statistically dependent items are discovered they are reviewed for content and a decision is made to either retain, revise, or discard one or more of the potentially dependent items.

Case Example

An academic misconduct survey was administered at North Carolina State University's College of Veterinary Medicine. The survey contained 23 items measuring the extent to which various actions and behaviors constitute academic misconduct. A 7-point semantic differential scale (1 = Not Misconduct to 7 = Severe Misconduct) was used to capture participants' perspectives. A total of 137 students completed the survey.

As part of the routine psychometric analysis, statistical dependence was investigated by reviewing residual item correlations. Items with residual correlations greater than 0.3 were considered statistically dependent (Smith, 2000). Results of the psychometric investigation indicated four pairs of items were statistically dependent (see Table 1).

Upon conclusion of the analysis, psychometric results (which included item pairs flagged as statistically dependent) were reviewed by a team of veterinary faculty. The faculty had difficulty understanding why each presumably dependent item pair was so highly correlated, as they perceived each item pair to represent substantively different questions. One faculty member noticed each pair of dependent items appeared in sequential order on the instrument and hypothesized that respondents may have perceived the items to be somewhat redundant based on their adjacent positioning. Curious if item sequence order may be the culprit for statistical dependency in this case, a simple experiment was conducted.

Table 1. Statistically dependent item pairs based on veterinary students' responses.

Item #	Item	Correlation
1	Copying from another student during a quiz or exam	.69
2	Using unauthorized cheat sheets or other materials during a quiz or exam	
12	Missing class or lab due to a false excuse	.37
13	Claiming to have attended class when you actually did not	
19	Failing to prepare adequately for a group assignment or laboratory	.67
20	Doing less than your fair share in a group project or a laboratory	
22	Presenting your clinical skills book for signing without actually completing the skill	.60
23	Listing false completions on your online clinical skills completion summary	

Experiment

Two questionnaires were created with one version containing the items in the same order as originally presented (control) and the other version containing items presented in random order (experiment). The survey was published on Amazon Mechanical Turk to a national panel of respondents. The first 100 participants completing each survey was awarded a small stipend for their time and effort. Individuals who participated in one survey were ineligible for participation in the other, ensuring 200 distinct individuals completed the surveys.

Quality Control

Upon data collection, a series of routine quality control checks were performed as part of the initial Rasch analysis. Data from both sets were evidenced to be mostly unidimensional, highly reproducible (reliability > .90), and fit the Rating Scale Model (Andrich, 1978) quite well. In order to obtain excellent fit, 10 misfitting persons were removed from the control group and 2 misfitting persons were removed from the experimental group.

Results

Item pairs that were previously flagged as potentially statistically dependent were investigated in both the control and experimental data sets (see Table 2).

Table 2. Investigation of Potentially Dependent Item Pairs

Item Pairs	Control Group Correlation	Experimental Group Correlation
1 & 2	.40	.32
12 & 13	.32	.08
19 & 20	.68	.32
22 & 23	.56	.28

Results from the control study indicated that when the items were presented in the same order as the veterinary student survey, each pair of items was once again flagged as being potentially statistically dependent. However, when the items were randomly presented to participants in the experimental group evidence of statistical dependency was greatly reduced. In fact, two pairs of items fell below the suggested threshold of .30 and the remaining two pairs fell to .32 (just slightly above the suggested threshold). Based on this evidence, it appears that item ordering may impact local item dependence, at least in some situations.

Implications and Recommendations

Many researchers and practitioners routinely revise or discard one or more survey items that are flagged as statistically dependent. The results of this experiment suggest one should use additional caution when considering revising or discarding items based on initial inspection of residual correlations. It is recommended that one pays particular attention to the order in which items were presented to respondents. If the flagged items were adjacent to one another it may be a false-positive detection due (presumably) to respondents' acquiescing to what they may perceive as a redundant question. Researchers using Rasch models to analyze survey data may be wise to randomize survey item order when possible to minimize threats to false-positive detections of statistical dependency and potentially other threats to item location estimates and stability.

Kenneth D. Royal, North Carolina State University

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-73.

Bradburn, N., Sudman, S., & Wansink, B. (2004). A*sking questions: the definitive guide to questionnaire design*. San Francisco: Jossey-Bass.

Dillman, Don. (2000). *Mail and internet surveys*. New York: John Wiley & Sons.

Heiman, G. W. (2002). *Research Methods in Psychology*. 3rd Edition. Boston & New York. Houghton Mifflin Company.

Marais, I., & Andrich, D. (2008). Effects of varying magnitude and patterns of response
dependence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(2),
105-124.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied
Measurement, 1*(2), 199-218.

A Probabilistic Model of the Law of Supply and Demand

Smith (1962) presents an experimental study of competitive market behavior notable (Cowen and Tabarrok, 2009, p. 39) for being a rare and early instance of a controlled study in economics. Chart 1 (Smith, 1962, p. 113; reproduced below) shows the convergence of prices, supply, and demand, as predicted by economic theory. The idea of a lawful relationship in which prices are predictable given the difference between supply and demand immediately suggests the applicability of a probabilistic model structured in the same form as a natural law, such as Rasch (1960, pp. 110-115; 1972/2010; Burdick and Stenner, 1996; Burdick, Stone, and Stenner, 2006) proposed, and as has been suggested repeatedly throughout the history of economics (Boumans, 1993, 2005; Fisher, 2010a, 2010c; Grattan-Guiness, 2010; Mirowski, 1991; Myers, 1983, pp. 65-75). However, the resemblance of Smith's Chart 1 to a Wright map (Wilson, 2005, 2013) is only superficial, since what appear to be stair-stepped vertical histograms are in fact crossing representations of the available supply and demand quantities.

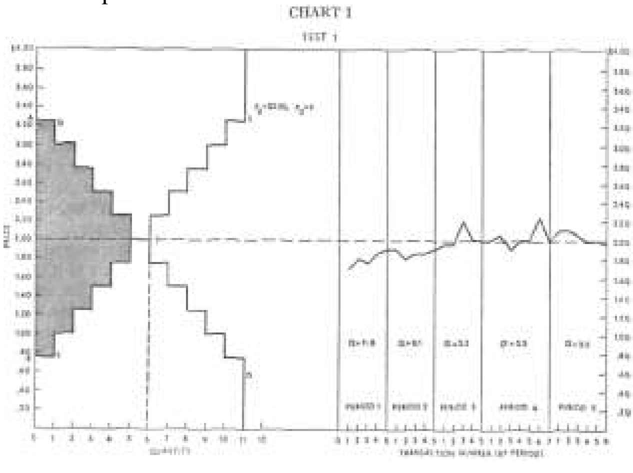


Figure 1. Supply (S), Demand (D), and Prices (P) (Smith, 1962, p. 113)

The usual approach to estimating the relationship of supply, demand, and prices is through a series of simultaneous equations. Roughly the same results can be obtained using algebraic approximations (Cohen, 1979).

A simple model for the law of supply and demand can be expressed as:

(1) $p = s / d$

where price p equals supply divided by demand. Applying the natural logarithm, the same result can be obtained from:

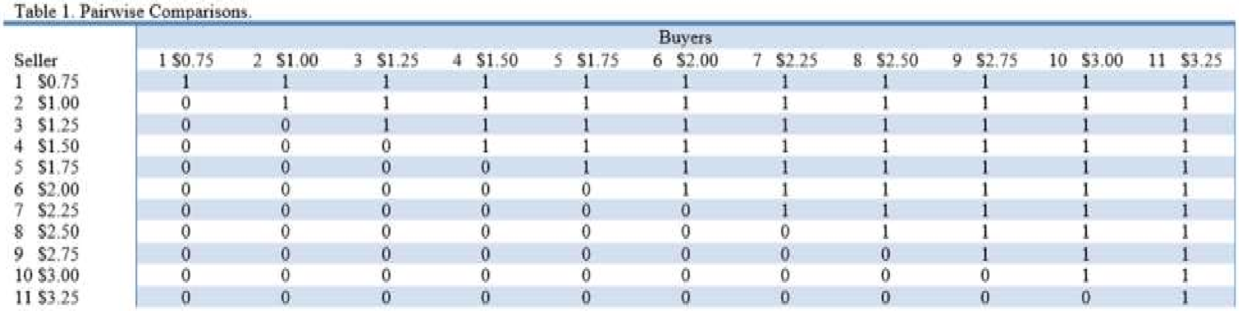
(2) $\ln(p / (1 - p)) = s - d$

where p now stands for the probability of a successful trade. Do Smith's (1962) data fit this model? If so, how can the equilibrium price be determined?

Each implied buyer-seller interaction could be evaluated in terms of pairwise comparisons, with profitable exchanges scored 1, and unprofitable exchanges, 0. Following this method, the seller with the lowest price would be able to sell to any buyer, and the buyer with the highest price could buy from any seller. The data matrix of all 121 potential trades in Smith's Chart 1 is shown in Table 1. Smith's Chart 1 counts can be recovered in Table 1, since, at price of $1.40, there is a supply of 3 sellers and demand from 8 buyers.

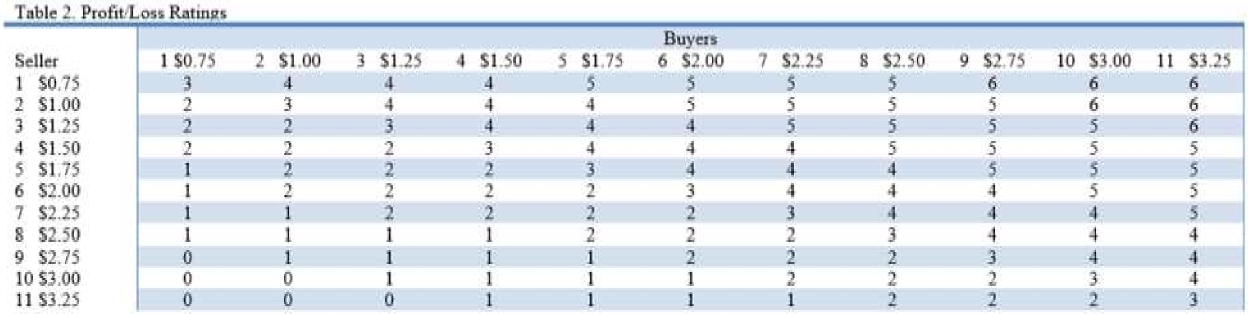
Alternatively, differences in the offering and asking prices could be rated as large or small profits or losses on a positive to negative continuum, where large positive differences of $2.00 or more might be categorized as 6, no difference as 3, and large negative differences of $2.00 or more, as 0. This scoring method spreads variation evenly throughout the matrix, which is advantageous for maximizing the ratio of the variance to measurement uncertainty (precision, or reliability). Other scoring methods might not achieve the desired spread in scores.

Table 1. Pairwise Comparisons.



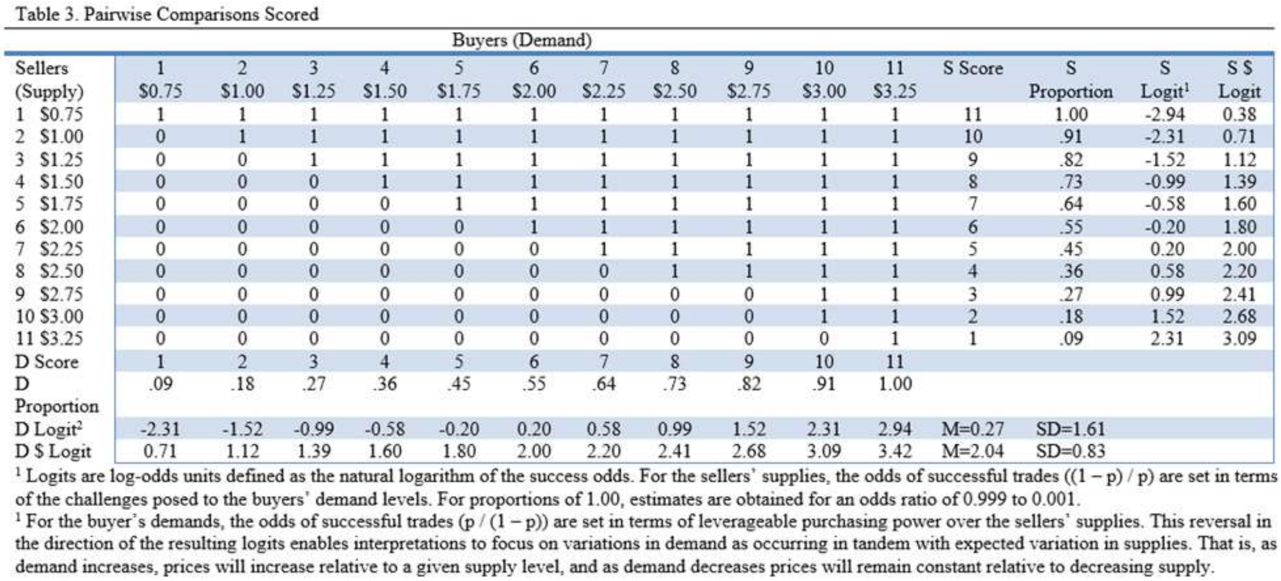
	Buyers										
Seller	1 $0.75	2 $1.00	3 $1.25	4 $1.50	5 $1.75	6 $2.00	7 $2.25	8 $2.50	9 $2.75	10 $3.00	11 $3.25
1 $0.75	1	1	1	1	1	1	1	1	1	1	1
2 $1.00	0	1	1	1	1	1	1	1	1	1	1
3 $1.25	0	0	1	1	1	1	1	1	1	1	1
4 $1.50	0	0	0	1	1	1	1	1	1	1	1
5 $1.75	0	0	0	0	1	1	1	1	1	1	1
6 $2.00	0	0	0	0	0	1	1	1	1	1	1
7 $2.25	0	0	0	0	0	0	1	1	1	1	1
8 $2.50	0	0	0	0	0	0	0	1	1	1	1
9 $2.75	0	0	0	0	0	0	0	0	1	1	1
10 $3.00	0	0	0	0	0	0	0	0	0	1	1
11 $3.25	0	0	0	0	0	0	0	0	0	0	1

Table 2. Profit/Loss Ratings



	Buyers										
Seller	1 $0.75	2 $1.00	3 $1.25	4 $1.50	5 $1.75	6 $2.00	7 $2.25	8 $2.50	9 $2.75	10 $3.00	11 $3.25
1 $0.75	3	4	4	4	5	5	5	5	6	6	6
2 $1.00	2	3	4	4	4	5	5	5	5	6	6
3 $1.25	2	2	3	4	4	4	5	5	5	5	6
4 $1.50	2	2	2	3	4	4	4	5	5	5	5
5 $1.75	1	2	2	2	3	4	4	4	5	5	5
6 $2.00	1	2	2	2	2	3	4	4	4	5	5
7 $2.25	1	1	2	2	2	2	3	4	4	4	5
8 $2.50	1	1	1	1	2	2	2	3	4	4	4
9 $2.75	0	1	1	1	1	2	2	2	3	4	4
10 $3.00	0	0	1	1	1	1	2	2	2	3	4
11 $3.25	0	0	0	1	1	1	1	2	2	2	3

Table 3. Pairwise Comparisons Scored



	Buyers (Demand)														
Sellers (Supply)	1 $0.75	2 $1.00	3 $1.25	4 $1.50	5 $1.75	6 $2.00	7 $2.25	8 $2.50	9 $2.75	10 $3.00	11 $3.25	S Score	S Proportion	S Logit[1]	S $ Logit
1 $0.75	1	1	1	1	1	1	1	1	1	1	1	11	1.00	-2.94	0.38
2 $1.00	0	1	1	1	1	1	1	1	1	1	1	10	.91	-2.31	0.71
3 $1.25	0	0	1	1	1	1	1	1	1	1	1	9	.82	-1.52	1.12
4 $1.50	0	0	0	1	1	1	1	1	1	1	1	8	.73	-0.99	1.39
5 $1.75	0	0	0	0	1	1	1	1	1	1	1	7	.64	-0.58	1.60
6 $2.00	0	0	0	0	0	1	1	1	1	1	1	6	.55	-0.20	1.80
7 $2.25	0	0	0	0	0	0	1	1	1	1	1	5	.45	0.20	2.00
8 $2.50	0	0	0	0	0	0	0	1	1	1	1	4	.36	0.58	2.20
9 $2.75	0	0	0	0	0	0	0	0	0	1	1	3	.27	0.99	2.41
10 $3.00	0	0	0	0	0	0	0	0	0	1	1	2	.18	1.52	2.68
11 $3.25	0	0	0	0	0	0	0	0	0	0	1	1	.09	2.31	3.09
D Score	1	2	3	4	5	6	7	8	9	10	11				
D Proportion	.09	.18	.27	.36	.45	.55	.64	.73	.82	.91	1.00				
D Logit[2]	-2.31	-1.52	-0.99	-0.58	-0.20	0.20	0.58	0.99	1.52	2.31	2.94	M=0.27	SD=1.61		
D $ Logit	0.71	1.12	1.39	1.60	1.80	2.00	2.20	2.41	2.68	3.09	3.42	M=2.04	SD=0.83		

[1] Logits are log-odds units defined as the natural logarithm of the success odds. For the sellers' supplies, the odds of successful trades ((1 – p) / p) are set in terms of the challenges posed to the buyers' demand levels. For proportions of 1.00, estimates are obtained for an odds ratio of 0.999 to 0.001.

[1] For the buyer's demands, the odds of successful trades (p / (1 – p)) are set in terms of leverageable purchasing power over the sellers' supplies. This reversal in the direction of the resulting logits enables interpretations to focus on variations in demand as occurring in tandem with expected variation in supplies. That is, as demand increases, prices will increase relative to a given supply level, and as demand decreases prices will remain constant relative to decreasing supply.

For instance, it might seem that only the most profitable trade should be scored 2, with all other profitable trades set to 1, and all ties and losses, 0, but this results in all of the 2s being assigned only to the highest offer ($3.25).

The kinds of ordered matrices shown in Tables 1 and 2 define the ideal patterns of observations typically identified and constructed in psychometric applications. In both cases, the probability of any buyer's success in buying depends only on their purchasing power and the seller's supply. For any given buyer, the probability of a successful trade increases monotonically with demand (purchasing power), no matter which seller is involved. Conversely, for any given seller, the probability of a successful trade increases monotonically with supply, no matter which buyer is involved.

In either case, the deterministic, staircase-shaped structure of Smith's supply and demand distributions would have to be modified for prices to be estimable in a probabilistic context using maximum likelihood methods. With no stochastic overlap across buyers for the sellers, and across sellers for the buyers, it is impossible to locate any buyer or seller relative to any other (Engelhard, 1993; Loevinger, 1954). Smith's (1962) data indicate that some small departures from the expected pattern occurred in his experiments, but because of the format of the data presented it is impossible to determine the precise exchanges in which those departures took place. For the purposes of illustration, a simple analysis will suffice, without compromising the principles involved.

Table 3 shows the counts of successful trades, the overall proportion of successes for each buyer and seller, and the associated logit (log-odds unit) values. An advantage of this form of model is that both axes of the matrix are expressed in the same unit. Comparisons of a buyer's level of demand (purchasing power) with any seller's offer immediately shows the likelihood of a successful trade. Moreover, the log-odds unit could be fixed at any arbitrarily defined convention, such as dollars, so that the equilibrium price in any given market would be the simple mean of the trades. In this way, all markets could potentially be equated to a common unit that has a built in expression of the relative purchasing power.

The bottom row of Table 3 shows the demand logit dollar equivalent with the mean set at the equilibrium value of $2.00 and the standard deviation set to the original dollar purchase price standard deviation (0.83). The equating was obtained by multiplying each demand logit by ratio of the standard deviations (0.516) and adding 1.73. The same process was applied to the supply logits to arrive at a supply logit dollar equivalent. Of course, the nonlinearity of the dollar unit denomination of the original purchasing prices becomes apparent relative to the linear logit.

Finally, the data are not symmetrically distributed, as there is one level at which a buyer has perfect purchasing power ($3.25) and one at which a seller can meet all demand ($0.75), but there are no corresponding opposite levels at which no buyer has any purchasing power (which could have been obtained, for instance, at $0.50) and at which a seller can meet no demand (which may have been at $3.50). This results in a shift of the sellers' supply logit distributions relative to the buyers' distributions. It was arbitrarily decided to center the scale at the buyers' equilibrium price of $2.00. Other situations may entail other centering decisions.

Rasch (1960, pp. 110-115) formed his models on the pattern of Maxwell's analysis of mass, force, and acceleration, approximating Maxwell's own method of analogy, and setting up what has been to date a largely unrealized potential for more meaningfully situating psychological and social measurement in the history of science (Fisher, 2010a, 2010c; Fisher and Stenner, 2013). It is of particular interest here that Rasch likely learned not only a great deal about probabilistic models, but also about Maxwell's method of analogy, through his documented association with the economists Koopmans and Frisch (Bjerkholt and Dupont-Kieffer, 2011; Andrich, 1997; Wright, 1980). Koopmans was an economist who had studied with Tinbergen, well known for his use of Maxwell's method (Boumans, 1993, 2001, 2005).
The implications of Rasch's models for economics remain largely unexplored, though they correspond with a number of developments in econometrics (Kærgård, 2003), as also do related models (Andrich, 1978) previously offered by Thurstone (McFadden, 2005). The question remains open, however, as to how further study of Rasch's measurement ideas relative to the extrapolations of natural laws into economics (Boumans, 1993; Mirowski, 1991) might lead to their improvement, extension, or abandonment.

William P. Fisher, Jr., University of California-Berkeley

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2,* 449-460.

Andrich, D. (1997). Georg Rasch in his own words [excerpt from a 1979 interview]. *Rasch Measurement Transactions, 11*(1), 542-3. [http://www.rasch.org/rmt/rmt111.htm#Georg].

Bjerkholt, O., & Dupont-Kieffer, A. (2011). Ragnar Frisch and the probability approach. *History of Political Economy, 43,* 109-139.

Boumans, M. (1993). Paul Ehrenfest and Jan Tinbergen: A case of limited physics transfer. In N. De Marchi (Ed.), *Non-natural social science: Reflecting on the enterprise* *of "More Heat than Light"* (pp. 131-156). Durham, NC: Duke University Press.

Boumans, M. (2001). Measure for measure: How economists model the world into numbers. *Social Research, 68*(2), 427-53.

Boumans, M. (2005). *How economists model the world into numbers.* New York: Routledge.

Burdick, H., & Stenner, A. J. (1996). Theoretical prediction of test items. *Rasch Measurement Transactions, 10*(1), 475 [http://www.rasch.org/rmt/rmt101b.htm].

Burdick, D. S., Stone, M. H., & Stenner, A. J. (2006). The Combined Gas Law and a Rasch Reading Law. *Rasch Measurement Transactions, 20*(2), 1059-60 [http://www.rasch.org/rmt/rmt202.pdf].

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology, 32,* 113-120.

Cowen, T., & Tabarrok, A. (2009). *Modern principles: Macroeconomics.* New York: Worth Publishers.

Engelhard, G., Jr. (1993). What is the attenuation paradox? *Rasch Measurement Transactions, 6*(4), 257 [http://www.rasch.org/rmt/rmt64.htm].

Fisher, W. P., Jr. (2010a, June 13-16). Rasch, Maxwell's method of analogy, and the Chicago tradition. In G. Cooper (Chair), *Https://conference.cbs.dk/index.php/rasch/Rasch2010/paper/view/824*. Probabilistic models for measurement in education, psychology, social science and health: Celebrating 50 years since the publication of Rasch's Probabilistic Models, University of Copenhagen School of Business, FUHU Conference Centre, Copenhagen, Denmark.

Fisher, W. P. J. (2010b). *Measurement, reduced transaction costs, and the ethics of efficient markets for human, social, and natural capital.*, Bridge to Business Postdoctoral Certification, Freeman School of Business, Tulane University (p. http://ssrn.com/abstract=2340674).

Fisher, W. P., Jr. (2010c). The standard model in the history of the natural sciences, econometrics, and the social sciences. *Journal of Physics: Conference Series, 238*(1), http://iopscience.iop.org/1742-6596/238/1/012016/pdf/1742-6596_238_1_012016.pdf.

Fisher, W. P., Jr., & Stenner, A. J. (2013). On the potential for improved measurement in the human and social sciences. In Q. Zhang & H. Yang (Eds.), *Pacific Rim Objective Measurement Symposium 2012 Conference Proceedings* (pp. 1-11). Berlin, Germany: Springer-Verlag.

Grattan-Guinness, I. (2010). How influential was mechanics in the development of neoclassical economics? A small example of a large question. *Journal of the History of Economic Thought, 32*(4), 531-581.

Kærgård, N. (2003, May 2-4). *Georg Rasch and modern econometrics.* Presented at the Seventh Scandinavian History of Economic Thought Meeting, Molde University College, Molde, Norway.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51,* 493-504.

McFadden, D. (2005, August 20). *The new science of pleasure: Consumer behavior and the measurement of well-being.* In *Frisch Lecture.* Econometric Society World Congress, London, England [http://emlab.berkeley.edu/wp/mcfadden0105/ScienceofPleasure.pdf].

Mirowski, P. (1991). *More heat than light: Economics as social physics, physics as nature's economics.* (Historical Perspectives on Modern Economics). New York: Cambridge University Press.

Myers, M. (1983). *The soul of modern economic man: Ideas of self-interest Thomas Hobbes to Adam Smith.* Chicago, Illinois: University of Chicago Press.

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

Rasch, G. (1972/2010). Retirement lecture of 9 March 1972: Objectivity in social sciences: A method problem (Cecilie Kreiner, Trans.). *Rasch Measurement Transactions, 24*(1), 1252-1272 [http://www.rasch.org/rmt/rmt241.pdf].

Smith, V. (1962). An experimental study of competitive market behavior. *Journal of Political Economy, 70*(2), 111-137.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wilson, M. R. (2011). Some notes on the term: "Wright Map." *Rasch Measurement Transactions, 25*(3), 1331 [http://www.rasch.org/rmt/rmt253.pdf].

Wright, B. D. (1980). Foreword, Afterword. In *Probabilistic models for some intelligence and attainment tests, by Georg Rasch* (pp. ix-xix, 185-199). http://www.rasch.org/memo63.htm) Chicago: University of Chicago Press.

R-Program Code for Calibration of Item Difficulties using Conditional Pairwise Estimation with the Application of Principal Components in Quality Audit of Large-Scale Public Examinations

In a large-scale public examination, the number of students sitting a compulsory subject could amount to some 100,000. The maximum mark of an item could be large, say, up to 50. In such a situation, direct application of the polytomous Rasch Model, Partial Credit Model (PCM), would simply lead to a crash during the computational process due to the need of estimating "too many" parameters. Therefore, in quality audit of exam for Hong Kong Diploma of Secondary Education (HKDSE), the following model, which is derived from Partial Credit Model using four principal components (Andrich & Luo, 2003), is adopted instead.

$$P_{xi} = \Pr\{X_{ni} = x\} = \frac{1}{\gamma_{ni}} \exp[x\beta_n + \sum_{l=1}^{4} f_{li}(x)\omega_{li}].$$

where $f_{li}(x)$ is an arithmetic expression based on x and maximum mark of the item i, m_i, (see R program code below)

ω_{li} is the parameter for item difficulty, called principal components,

γ_{ni} is the normalizing factor, and

β_n is the student ability

Unlike PCM, the number of parameters for an item could be restricted to at most four; even the maximum mark of the item is very large. To eliminate the need of estimating a large number of student abilities, the following conditional pairwise probabilities are considered instead: With respect to a student n, given that the total scores of two items (i and j) are r consider the probability that the student obtain x marks for item i (i.e., $p_{ij}(x|r)$). By simple manipulation, it can be shown that the student ability β_n involved would be cancelled off. Accordingly, a pseudo-log-likelihood function is derived.

$$\log L(X) = 2\sum_{i=1}^{I}\sum_{j\neq i}\sum_{r=0}^{m_i+m_j}\sum_{x=L}^{U} n_{ij}(x|r)\log p_{ij}(x|r).$$

where $n_{ij}(x|r)$ is the number of students obtaining x for an item i and his total score of item i and j is r.

It should be noted that such a log-likelihood function is derived by assuming pairwise responses are statistically independently which do not hold in general. Nevertheless, it is expected that the parameter estimates that maximize *log L(X)* are consistent. The item difficulty of each item could be computed by deriving 1st derivate and 2nd derivate of *log L(X)* respect to individual principal components, and applying Newton- Raphson Method. The following provides the corresponding detailed R-program code annotated with comments, where appropriate. We hope that with the provision of the R-program code, the method could be utilized by more users.

```
#---Compute nij(x|r)
count.n<-function(i,j,x,r)
{
  cnt=0
  for (s in 1:Ns) {
    score.i   =   raw.Data[s,   i];   score.j   =
raw.Data[s,j]
    tot.score.ij =score.i+score.j
    if (score.i == x && tot.score.ij == r)
      cnt =cnt+1
  }
  return(cnt)
}

#---Compute  Nij(r)--The  number  of  students  who
has a total score r
count.N<-function(i,j,r)
{
  cnt = 0
  lower=max(0,                 r-max.ItMark[j]);
upper=min(max.ItMark[i], r)
  for (x in lower:upper){
   cnt=cnt+count.n(i,j,x,r)
  }
  return(cnt)
}

#---Compute fli(x)
f.coeff<-function(n,i,x)
{
if(n==1) return(-x)
if(n==2) return(x * (max.ItMark[i] - x) )
if(n==3) return( x * (max.ItMark[i] - x) * (2*x
- max.ItMark[i]) )
if (n==4) return( x* (max.ItMark[i]- x)*(5*x^2 -
5*x*max.ItMark[i] + max.ItMark[i]^2 + 1) )
}

#---compute pij(x|r)
prb<-function(i,j,x,r){
 num=0
 for(l in 1:Np){
   num=num + (f.coeff(l, i, x)*it.para[l, i] +
f.coeff(l, j, r-x)*it.para[l, j] )
 }
 denom=0
 lower=max(0,                r-max.ItMark[j]);
upper=min(max.ItMark[i], r)
 for(k in lower:upper){
   num.1 = 0
        for(l in 1:Np)
       {num.1=num.1+(f.coeff(l,i,k)*it.para[l,i]
+f.coeff(l, j, r-k) * it.para[l, j] ) }
          denom=denom+exp(num.1)
 }
 return(exp(num)/denom)
}

#---Compute sufficient statistics Tli, useful for
deriving the 1st derivate
```

```
cal.T<-function(l,i)
{
 cnt=0
 for(j in 1:Nt){
       if (!(i==j)) {
         index.r = max.ItMark[i]+max.ItMark[j] -
1
         for (r in 1:index.r){
             lower=max(0,       r-max.ItMark[j]);
upper=min(max.ItMark[i], r)
        for (x in lower:upper) {
         cnt=
cnt+count.n(i,j,x,r)*f.coeff(l,i,x)
        } # 3rd for-loop
       } # 2nd for-loop
        } # if statement
 } # 1st for-loop
 return(cnt)
}

#---Store the sufficient statistics in a matrix
for latter use
T=matrix(0,Np, Nt)
for (l in 1:Np){
 for (i in 1:Nt){
   T[l,i] = cal.T(l,i)
 }
}

#---Derive the 1st derivate of log L(X) wrt a
principal component, l of an item, i
first.der<-function(l,i){
 res=0
  for (j in 1:Nt){
   if(!(j==i)){
          index.r  =  max.ItMark[i]+max.ItMark[j]
- 1
     for(r in 1:index.r){
              cnt=0;        lower=max(0,        r-
max.ItMark[j]); upper=min(max.ItMark[i], r)
         for(k in lower:upper){
        cnt=cnt+ f.coeff(l,i,k)*prb(i,j,k,r)
              }
              res=res+count.N(i,j, r)*cnt
          }
         }
   }
 return( T[l,i] - res)
}

#---Derive the 2nd derivate of log L(X) wrt a
principal component, l of an item, i
sec.der<-function(l,i)
{
 res=0
  for (j in 1:Nt){
   if(!(j==i)){
          index.r  =  max.ItMark[i]+max.ItMark[j]
- 1
     for(r in 1:index.r){
             term.1=0; term.2=0; lower=max(0, r-
max.ItMark[j]); upper=min(max.ItMark[i], r)
         for(k in lower:upper){
```

```
                 f.lik     =     f.coeff(l,i,k);
p.ijkr=prb(i,j,k,r)
           term.1=term.1 + (f.lik^2)*p.ijkr
                 term.2=term.2 + f.lik*p.ijkr
              }
             res=res+count.N(i,j,  r)*(term.1  -
(term.2)^2)
           }
         }
   }
 return(-res)
}

#---initial  values  for  first  component,  the
others are simply set to zeros
for(i in 1:Nt) {
  obs.TMark=sum(raw.Data[,i])
  eff.Num=obs.TMark/max.ItMark[i]
  it.para[1,i] = log(( Ns - eff.Num)/eff.Num)
}
it.para[1,]= it.para[1, ]-mean(it.para[1, ])

#------------Start the iterative looping
prev.para=matrix(0,Np,Nt)
new.para=matrix(0, Np,Nt)
prev.para=it.para; new.para=it.para
max.chg=seq(0,0, length.out=Np); max.change=10
while (max.change > 0.01){
  for(l  in  2:Np){  #---for  individual  principal
components, starting from second
    for(i in 1:Nt){#---for each item
   prev.para[l,i] = it.para[l,i]

   #---------Newton-Raphson Methiod
      chg=100
   while (chg>0.01){
        new.est      =      it.para[l,i]      -
(first.der(l,i)/sec.der(l,i))
           chg = abs(new.est - it.para[l,i])
           it.para[l,i] = new.est
        }
        new.para[l,i]=new.est;
it.para[l,i]=prev.para[l,i]
    }# for i
   max.chg[l]     =     max(abs(     it.para[l,]-
new.para[l,]))
   it.para[l,] = new.para[l,]
  }# for l
  l=1  #---cater the first principal component
  for(i in 1:Nt){
   prev.para[l,i] = it.para[l,i]

   #---------Newton-Raphson Methiod
    chg=100
    while (chg>0.01){
          new.est      =      it.para[l,i]      -
(first.der(l,i)/sec.der(l,i))
          chg = abs(new.est - it.para[l,i])
          it.para[l,i] = new.est
        }
    new.para[l,i]=new.est;
it.para[l,i]=prev.para[l,i]
    }# for i
    new.para[l,]=new.para[l,]              -
mean(new.para[l,])#the mean of item difficulties
set to 0
    max.chg[l]       =       max(abs(it.para[l,]-
new.para[l,]))
    it.para[l,] = new.para[l,]
    max.change=max(max.chg[])
}
```

Fung Tze Ho, Eric, Hong Kong Examinations and Assessment Authority

Reference

Andrich, D., & Luo, G. (2003). Conditional Pairwise Estimation in the Rasch Model for Ordered Response Categories using Principal Components. *Journal of Applied Measurement*, 4(3), 205-21.

Ohio River Valley Objective Measurement Seminar (ORVOMS)

The fifth annual Ohio River Valley Objective Measurement Seminar (ORVOMS) will be held on October 16, 2015 at the University of Kentucky's College of Education in Lexington, KY.

We are currently accepting submissions for presentation consideration. To be accepted, submissions should be about the Rasch Model or an application of the Rasch Model to a particular problem. All submissions should include an abstract of approximately 200 words summarizing your research question, why it's important, and results. **A paper is not required. Please submit your proposal by Wednesday, July 1, 2015** to Michael Peabody (mpeabody@theabfm.org).
Late submissions may be accepted if slots are available

There is no fee to attend!

Rasch Measurement Transactions
www.rasch.org/rmt
Editor: Kenneth Royal
Email Submissions to: Editor \at/ Rasch.org
Copyright © 2015 Rasch Measurement SIG, AERA
Permission to copy is granted.
RMT Editor Emeritus: John M. Linacre
Rasch SIG Chair: Jim Tognolini
Secretary: Mikaela Raddatz
Treasurer: Leigh Williams
Program Chairs: Jessica Cunningham& Sara Hennings
Rasch SIG website: www.raschsig.org

Rasch-related Paper Wins AARC Award

The Association for Assessment and Research in Counseling (AARC) Journal Editor's Research Award was award to Larry Ludlow, Christina Matz-Costa, Clair Johnson, Melissa Brown, Elyssa Besen, and Jacquelyn B. James.

According to the organization:

Each year, the MECD Editor, CORE Editor, and the AARC Member-at-Large for Awards have the responsibility for selecting the Journal Editor's Research Awards. These awards are given to the author or authors of the manuscripts deemed to have made the greatest contribution to the professional literature. Authors are eligible for the Association for Assessment and Research in Counseling/Measurement and Evaluation in Counseling and Development Patricia B. Elmore Award for Outstanding Research in Measurement and Evaluation if their manuscript was published in Measurement and Evaluation in Counseling and Development from July 1 to June 30 of the previous year.

The citation for the article is:

Ludlow LH, Matz-Costa C, Johnson C, Brown M, Besen E, & James JB (2014). Measuring engagement in later life activities: Rasch-based scenario scales for work, caregiving, informal helping, and volunteering. *Measurement and Evaluation in Counseling and Development, 47*(2), 127-149.

An abstract from the paper is presented below:

The development of Rasch-based "comparative engagement scenarios" based on Guttman's facet theory and sentence mapping procedures is described. The scenario scales measuring engagement in work, caregiving, informal helping, and volunteering illuminate the lived experiences of role involvement among older adults and offer multiple advantages over typical Likert-based scales.

Call for Submissions

Research notes, news, commentaries, tutorials and other submissions in line with *RMT*'s mission are welcome for publication consideration. All submissions need to be short and concise (approximately 400 words with a table, or 500 words without a table or graphic). The next issue of *RMT* is targeted for Sept. 1, 2015, so please make your submission by Aug. 1, 2015 for full consideration. Please email Editor\at/Rasch.org with your submissions and/or ideas for future content.