RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 29 No. 2 Autumn 2015 ISSN 1051-0796

If You Want to Get Ahead, Get a Theory

I jotted down the skeleton of the following as part of an email discussion between Andrew Stephanou (ACER), Mike Linacre and me about the role of substantive theories (especially those where *order* is paramount) in Rasch research. Given that such ‘history’ might be lost if not committed to publication, I have succumbed to the suggestion to share this with Rasch colleagues via *RMT*.

In the mid eighties, Geoff Masters and Mark Wilson introduced me to Rasch analysis at the annual meeting of the Australian Association for Research in Education (*c.f.* AERA in the US) in Hobart. I had been trying to use analyses based on Ordering Theory by William Bart and Peter Araisian (yes, that same PA involved in early development of Rasch software) to reveal the *order* of the acquisition of Piaget's formal operations during adolescence using data from my honours research thesis. (In retrospect, I reckoned that Mark and Geoff were trying to drum up a few participants for their Rasch measurement workshop held later in the week.)

The Australian Council of Educational Research, using ideas gleaned from Molly De Lemos, had earlier emphasized that using Rasch with my BLOT (Bond’s Logical Operations Test) would be a necessary, but not sufficient, precursor to its publishing the BLOT for use in Australian schools. Geoff Masters took me under his wing and during my visit to ACER where he was working, guided my analyses and interpretation of those data, especially in terms of what constitutes *ordering* from a Rasch analysis perspective.

In 1987, I went to King’s College in London to work with Michael Shayer on a sabbatical visit and to start my PhD research. Geoff sent over a copy of the DICOT software (developed with Wilson) to run our data on the tiny BBC personal computer in use in the Cognitive Acceleration through Science Education (CASE II) research project at King’s. Order via stages was everything for committed Piagetians such as Shayer and me! Quite serendipitously, Masters had picked up a copy of *Towards a Science of Science Teaching* by Shayer and Adey while browsing at the University of Chicago bookshop while on one of his frequent visits to Ben Wright at Judd Hall! He was impressed by the commitments to theory-based practice and to order in the acquisition of high school science concepts.

After my PhD research was successfully completed, I introduced Rasch analysis to fellow researchers at the Jean Piaget Society meetings in the US. Concurrently, I introduced Piagetian theory to Ben Wright via my presentations at the Midwest Objective Measurement Symposia in Chicago. That, of course, is where I met Mike Linacre, as well.

The watershed meeting came at Judd Hall between Terry Brown (Piagetian and psychiatrist), Ben Wright and me. Terry and I were French-speaking Piagetians; we had met at a Piaget annual conference in Geneva. Ben knew Terry via professional contact in Chicago, a few years earlier. I could understand Terry’s Piaget and Ben’s Rasch each well enough to bridge between their interpretations of my work. And, quite interestingly, we all had enough shared knowledge of Freudian theory to serve as a common conceptual ground, as well.

My chief fear was that fit statistics were, in practice,

Table of Contents

useless, because each of our Piagetian data analyses worked pretty well from the get-go. Ben Wright demurred, claiming that many worked for years (even ultimately successfully) trying to get fit statistics as good as we had first time round. After our delineating the extent of Piaget's theoretical and empirical *oeuvre* (over 70 books and more than 600 published articles), Ben expressed amazement by the depth and breadth of Piaget's theorizing, and simply quipped, "That's why you get good measures so easily, you have good theory."

A plan was hatched for me to run a series of Rasch workshops for interested JPS members in conjunction with the annual meetings, using Ben and Mark Wilson (by then at Berkeley) as key Rasch resource people. Kurt Fischer of Harvard was very enthusiastic. However, Ben, enthusiastic as ever, could not wait to get started and insisted on running the first JPS workshop in my absence. With his inimitable style and irrepressible manner, Ben managed to deeply offend Fischer and other Profs several times each day. Kaboom, end of series! I did manage to run the next as a rather low key version at Berkeley, thereby introducing Theo Dawson to Rasch. Fischer politely invited me to present at Harvard on my way to research in Geneva, and that's how Michael Commons caught the bug. Central to all this was the empirical revelation of developmentally sequential theoretical concepts by using Rasch analysis; the JPS website actually sponsored a section devoted to Rasch measurement.

While Mark Wilson was supportive of these initiatives, and claimed particular interest in development (remember his Saltus model), his view of developmentalism was too heavily informed by his US psychology colleagues for him to have the nuanced approach to stage development common to those who read Piaget - rather than those who read about Piaget as written by US authors.

Judy Amsel, wife of a prominent JPS member, was an editor for publishers Lawrence Erlbaum and Associates, and attended JPS meetings as the LEA rep. Larry Erlbaum supported the JPS - in his style as benefactor of academy - publishing annual books of collections of papers and the like. Usually, interested JPS members wanted to know, "How can I learn/understand Rasch in 10 minutes?" In response to my repeated frustration at the question, Judy, the LEA editor said, "When you are ready to write your answer as a book, please see me first. I know there is a space in the Erlbaum catalogue for such a book, and after watching your JPS presentations, I know you are someone who could write it."

Meeting co-author, Christine Fox, was yet another collusion of fortuitous circumstances. JPS stalwart Bill Gray and I were in ongoing email communication about our different attempts at quantifying the development of formal operations: Bill was very committed to factor analysis, and obviously, I was not. When I visited him at the University of Toledo, in conjunction with a Chicago-based conference, he introduced me to the Faculty's new hire, Christine Fox, fresh out of Duke. Although we have managed to write three editions of *Applying the Rasch Model,* neither separately, nor together, have we managed to coerce Bill to publish the results of our Rasch analyses of his data sets.

Trevor G. Bond
James Cook University, Australia

Note: The third edition of *Applying the Rasch model: Fundamental measurement in the human sciences*, published by Routledge, hit the bookstores (metaphorically speaking) on June 16, 2015.

See also: Bond, T. G. (2005). Past, present and future: an idiosyncratic view of Rasch measurement. In S. Alagumalai, D. Curtis & N. Hungi (Eds) *Applied Rasch Measurement: A Book of Exemplars. Papers in honour Of John P. Keeves.* Kluwer Academic Publishers. (pp. 329-341)

Illustrating the Psychometric Construct of Family Medicine on the American Board of Family Medicine's Examinations

At AERA 2015, we presented a paper entitled "Illustrating the Psychometric Construct of Family Medicine on the American Board of Family Medicine's Examinations" which may have been more appropriately entitled "How we stole Jack Stenner's Lexile Map...and how you can too". The objective of this paper was to: (1) to create a visual representation of the construct of the examination that is easy to understand for physicians who may not have the appropriate psychometric and measurement background necessary to fully grasp an abstract concept such as an exam construct, and (2) to illustrate the method used to create the visual representation so that others may utilize this resource for their own exams.

Rasch Measurement Transactions
www.rasch.org/rmt
Editor: Kenneth Royal
Email Submissions to: Editor \at/ Rasch.org

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

The Rasch model (Rasch, 1960) is particularly well suited for this because the items and people are mapped onto a single continuum, such that the probability of any examinee *with a known ability estimate* correctly answering any question *with a known difficulty calibration* can be computed.

Winsteps Rasch measurement software (Linacre, 2014) produces a DISFILE, which contains the counts for each response option for every item. From the DISFILE, items were identified and selected for the ruler if the sum of the responses for the correct answer and the most popular distractor accounted for 90% or more of the responses. These items can be described as being essentially about knowing the difference between A (the correct answer) and B (the popular distractor) in the context of C (the item stem). Content distillation was further refined by a practicing family physician to ensure that the kernel of the item was accurately captured in the abbreviated description.

The difficulty of each item was converted into a scaled score, so that examinee test scores could be directly compared to the content the examinee had either mastered (80% chance of answering correctly) or found challenging (50% chance of answering correctly). Items were plotted on a vertical ruler indicating their difficulty. A two-column ruler format permitted the user to identify any examinee score which content was likely mastered and which content was likely to be challenging. Thus, rather than looking at all candidates and all items, as in a typical Item Map, the MC-FP Content Ruler (Figure 1) shows distilled versions of items along the score continuum that can be referenced by an individual.

Item-level responses and distractor data from four years of exam administrations were used to populate the ruler. There were a total of 960 items administered on the 2010, 2011, 2012, and 2013 exams, of which 387 (40.3%) fulfilled the requirements for inclusion. Although a large number of items were available, the exam is highly targeted to the passing standard and so many of the items fell into the easy to moderate range. There were not as many difficult items to populate the high end of the scale or easy items to populate the low end of the scale. The ruler provides a hierarchical item structure that can serve as a way to translate a quantitative scaled score into a qualitative meaning implied by the score. It is hoped that the hierarchical structure clearly illustrates the relationship between scores and knowledge of family medicine.

The importance of this ruler lies in its utility to convey what is implied in the construct of family medicine. Hopefully, it will help examinees understand what is meant by family medicine on our examination and give them some sense of the ability level required to pass. This may reduce the anxiety level for those examinees who are likely to get moderate to high scores on the exam. With more content, it might be feasible to create several rulers with each one being specific to a particular content area.

Michael R. Peabody, Thomas R. O'Neill, and Lars E. Peterson
American Board of Family Medicine

References

Linacre, J. M. (2014). Winsteps Rasch Measurement computer program (Version 3.81.0). Beaverton, OR.: Winsteps.com.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Best Test Design Revisited: Validity Generalization in the Rasch Context

Because of the ways studies are designed and the ways data are analyzed, research results in psychology and the social sciences often appear to be nonlinear, sample- and instrument-dependent, and incommensurable, even when they are not. In contrast with what are common assumptions about the nature of the constructs involved, invariant relations may be more obscured than clarified by typically employed research designs and statistical methods.

To take a particularly salient example, the number of small factors with Eigenvalues greater than 1.0 identified via factor analysis increases as the number of modes in a multi-modal distribution also increases, and the interpretation of results is further complicated by the fact that the number of factors identified decreases as sample size increases (Smith, 1996).

Similarly, variation in employment test validity across settings was established as a basic assumption by the 1970s, after 50 years of studies observing the situational specificity of results. But then Schmidt and Hunter (1977) identified sampling error, measurement error, and range restriction as major sources of what was only the appearance of incommensurable variation in employment test validity. In other words, for most of the 20th century, the identification of constructs and comparisons of results across studies were pointlessly confused by mixed populations, uncontrolled variation in reliability, and unnoted floor and/or ceiling effects. Though they do nothing to establish information systems deploying common languages structured by standard units of measurement (Feinstein, 1995), meta-analysis techniques are a step forward in equating effect sizes (Hunter & Schmidt, 2004).

Wright and Stone's (1979) *Best Test Design*, in contrast, takes up each of these problems in an explicit way. Sampling error is addressed in that both the sample's and the items' representations of the same populations of persons and expressions of a construct are evaluated. The evaluation of reliability is foregrounded and clarified by taking advantage of the availability of individualized measurement uncertainty (error) estimates (following Andrich, 1982, presented at AERA in 1977). And range restriction becomes manageable in terms of equating and linking instruments measuring in different ranges of the same construct. As was demonstrated by Duncan (1985; Allerup, Bech, Loldrup, et al., 1994; Andrich & Styles, 1998), for instance, the restricted ranges of various studies assessing relationships between measures of attitudes and behaviors led to the mistaken conclusion that these were separate constructs. When the entire range of variation was explicitly modeled and studied, a consistent relationship was found.

Statistical and correlational methods have long histories of preventing the discovery, assessment, and practical application of invariant relations because they fail to test for invariant units of measurement, do not define standard metrics, never calibrate all instruments measuring the same thing in common units, and have no concept of formal measurement systems of interconnected instruments. Wider appreciation of the distinction between statistics and measurement (Duncan & Stenbeck, 1988; Fisher, 2010; Wilson, 2013a), and of the potential for metrological traceability we have within our reach (Fisher, 2009, 2012; Fisher & Stenner, 2013; Mari & Wilson, 2013; Pendrill, 2014; Pendrill & Fisher, 2015; Wilson, 2013b; Wilson, Mari, Maul, & Torres Irribarra, 2015), are demonstrably fundamental to the advancement of a wide range of fields.

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

References

Allerup, P., Bech, P., Loldrup, D., Alvarez, P., Banegil, T., Styles, I., & Tenenbaum, G. (1994). Psychiatric, business, and psychological applications of fundamental measurement models. International Journal of Educational Research, 21(6), 611-622.

Andrich, D. (1982). An index of person separation in Latent Trait Theory, the traditional KR-20 index, and the Guttman scale response pattern. Education Research and Perspectives, 9(1), 95-104 [http://www.rasch.org/erp7.htm].

Andrich, D., & Styles, I. M. (1998). The structural relationship between attitude and behavior statements from the unfolding perspective. Psychological Methods, 3(4), 454-469.

Duncan, O. D. (1985). Probability, disposition and the inconsistency of attitudes and behaviour. Synthese, 42, 21-34.

Duncan, O. D., & Stenbeck, M. (1988). Panels and cohorts: Design and model in the study of voting turnout. In C. C. Clogg (Ed.), Sociological Methodology 1988 (pp. 1-35). Washington, DC: American Sociological Association.

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

Fisher, W. P., Jr. (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. Measurement, 42(9), 1278-1287.

Fisher, W. P., Jr. (2010). Statistics and measurement: Clarifying the differences. Rasch Measurement Transactions, 23(4), 1229-1230.

Fisher, W. P., Jr. (2012, May/June). What the world needs now: A bold plan for new standards [Third place, 2011 NIST/SES World Standards Day paper competition]. Standards Engineering, 64(3), 1 & 3-5.

Fisher, W. P., Jr., & Stenner, A. J. (2013). Overcoming the invisibility of metrology: A reading measurement network for education and the social sciences. Journal of Physics: Conference Series, 459(012024), http://iopscience.iop.org/1742-6596/459/1/012024.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62(5), 529-540.

Mari, L., & Wilson, M. (2013). A gentle introduction to Rasch measurement models for metrologists. Journal of Physics Conference Series, 459(1), http://iopscience.iop.org/1742-6596/459/1/012002/pdf/1742-6596_459_1_012002.pdf.

Pendrill, L. (2014). Man as a measurement instrument [Special Feature]. NCSLi Measure: The Journal of Measurement Science, 9(4), 22-33.

Pendrill, L., & Fisher, W. P., Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. Measurement, 71, 46-55. doi: http://dx.doi.org/10.1016/j.measurement.2015.04.010

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. Structural Equation Modeling, 3(1), 25-40.

Wilson, M. R. (2013a). Seeking a balance between the statistical and scientific elements in psychometrics. Psychometrika, 78(2), 211-236.

Wilson, M. R. (2013b). Using the concept of a measurement system to characterize measurement models used in psychometrics. Measurement, 46, 3766-3774.

Wilson, M., Mari, L., Maul, A., & Torres Irribarra, D. (2015). A comparison of measurement concepts across physical science and social science domains: Instrument design, calibration, and measurement. Journal of Physics: Conference Series, 588(012034), http://iopscience.iop.org/1742-6596/588/1/012034.

Wright, B. D., & Stone, M. H. (1979). Best test design: Rasch measurement. Chicago, Illinois: MESA Press.

Mobile Computer-Adaptive Tests (CAT): Skin Cancer Risk Scale and Standard Errors

Computer Adaptive Testing (CAT) compared to traditional fixed form testing has advantages in term of precision and efficiency (Chien, 2009/2011; Ma, 2014). Most studies to date have used simulation studies instead of applying a test CAT to an actual existing dataset (Ma,2014; Chien,2011), and most have only tested the dichotomous model (Raîche,2006; Linacre,2006). In contrast, few studies are currently available that applied Rasch Partial Credit Model (PCM) to its adaptive testing in practice. We developed a mobile CAT survey procedure (see QR-code in Figure 1) to practically demonstrate the newly developed CAT application in action.

Figure 1. QR-code for conducting CAT

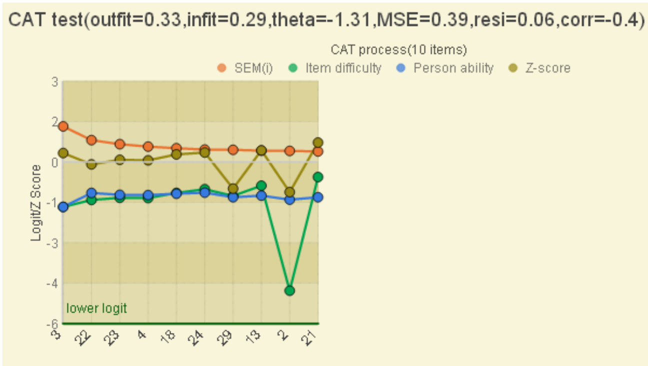


Figure 2. A graphical CAT report shown after each response

De-identified data from The QSkin Sun and Health study were used. This is a population-based cohort study of 43,794 men and women aged 40-69 years randomly sampled from the population of Queensland, Australia in 2011 (Olsen et al., 2012), of which two-thirds (29,314 cases) were randomly drawn. Based on the baseline questionnaire, we created a skin cancer risk item bank of 30 items. All NAT (Not Adaptive Testing) items were calibrated with the Rasch Partial Credit Model. A total of 1000 simulated respondents following a log normal distribution with a log mean of 0 and a log SD of 1 were simulated (Linacre, 2007) using 3 Rasch models with 3 respective fixed-item scenarios (see Table 1). We calculated the comparative efficiency and precision of CAT and Non Adaptive Test (NAT) and tested the difference using independent *t*-tests to count difference number ratios less than 5%. (Smith, 2002)

Table 1. Simulation data generated with 3 Rasch models

	DICHOTOMOUS		RSM		PCM	
Data sets	CRON	UNI	CRON	UNI	CRON	UNI
10 items	0.681	0.731	0.854	0.852	0.863	0.830
20 items	0.803	0.790	0.925	0.909	0.937	0.894
30 items	0.855	0.840	0.948	0.925	0.943	0.892

CRON=Cronbach's alpha; Uni=Dimension coefficient (Chien, 2011)

We ran a further developed VBA (Visual Basic for Applications) module in Microsoft Excel in compliance with CAT termination rules for a simulation study (see http://youtu.be/EUZUKMFCR9E). The results including theta, SE, Infit and Outfit are equivalent to Winsteps estimation. Cronbach alpha (=.80) were used to determine the CAT termination criterion using the standard error of measurement (SEM=SD × √1-reliability), whereas SEM = 1/sqrt(Σinformation(i)), i refers to the CAT item responded by a person (Linacre,2006).

We also set another rule that the minimum number of questions required for completion was 10. The first question was selected randomly from the 30 items. The provisional ability was estimated by a maximum log likelihood function using an iterative Newton-Raphson procedure after the 1st question was answered with its difficulty as the provisional ability. The next question selected was the one with the most information obtained from the remaining unanswered questions, interacting with the provisional person measures.

We found that CAT gains a smaller SE than NAT without compromising test precision and with a high efficiency by reducing response burdens: 48.20%, 66.70%, and 66.20% for dichotomous, RSM, and PCM models, respectively (see Table 2).

Table 2. Efficiency and precision of CAT and compared to using 10, 20 or 30 items in static NAT format

	DICHOTOMOUS		RSM		PCM	
Data sets	MEAN	SE	MEAN	SE	MEAN	SE
10 items	-0.007	0.829	0.030	0.414	-0.179	0.398
20 items	-0.008	0.555	0.020	0.289	-0.190	0.272
30 items	0.011	0.439	-0.080	0.235	-0.084	0.224
CAT	-0.194	0.432	0.029	0.154	-0.267	0.163
Precision	Diff(%)[a]	Corr.[b]	Diff(%)[a]	Corr.[b]	Diff(%)[a]	Corr.[b]
10 items	0.40%	0.863	0.30%	0.952	0.00%	0.931
20 items	0.00%	0.957	0.00%	0.988	0.00%	0.986
CAT	0.20%	0.925	4.00%	0.958	3.20%	0.946
Efficiency	CIL[c]	%[d]	CIL[c]	%[d]	CIL[c]	%[d]
CAT	15.55	48.20%	10.00	66.70%	10.13	66.20%

[a] Diff(%)=Different number ratio; [b] Corr = Correlation coefficient of person theta to NAT
[c] CIL = CAT item length; [d] Efficiency= 1-CIL/30

Tsair-Wei Chien[1], Ngadiman Djaja[2], Monika Janda[2], Catherine Olsen[3] and David Whiteman[4].
[1]Chi Mei Medical Center, Taiwan, [2]Queensland University of Technology, Australia. [3]National Health and Medical Research Council Centre for Research Excellence in Sun and Health (CRESH). [4]QIMR Berghofer Medical Research Institute, Australia

References

Chien , T. W., Wang, W. C., Huang, S. Y., Lai, W. P., Chow, J. C. (2011). A web-based computerized adaptive testing (CAT) to assess patient perception in hospitalization. *Journal of Medical Internet Research, 13*(3):e61.

Chien, T. W., Wu, H. M., Wang, W. C., Castillo, R. V., Chou, W. (2009). Reduction in patient burdens with graphical computerized adaptive testing on the ADL scale: tool development and simulation. *Health Quality Life Outcomes, 7*:39.

Chien, T. W. (2012). Cronbac's Alpha with the Dimension Coefficient to Jointly Assess a Scale's Quality. *Rasch Measurement Transactions, 26*(3): 1379.

Linacre, J. M. (2006). Computer-Adaptive Tests (CAT), Standard Errors and Stopping Rules. *Rasch Measurement Transactions, 20*(2), 1062.

Linacre, J. M. (2007) How to Simulate Rasch Data. *Rasch Measurement Transactions, 21*(3), 1125.

Ma, S. C.,Chien, T. W., Wang, H. H., Li, Y. C.,Yui, M. S.(2014). Applying Computerized Adaptive Testing to the Negative Acts Questionnaire-Revised: Rasch Analysis of Workplace Bullying. *Journal of Medical Internet Research, 16*(2):e50.

Olsen, C. M., Green, A. C., Neale, R. E., Webb, P. M., Cicero, R. A., Jackman, L. M. (2012). Cohort profile: The QSkin Sun and Health Study. *International Journal of Epidemiology, 41*(4).

Raîche,G., Blais, J.-G., Riopel, M. A. (2006). SAS Solution to Simulate a Rasch Computerized Adaptive Test. *Rasch Measurement Transactions, 20*(2),1061.

Smith, E. V. (2002). Detecting and evaluation the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*, 205-231.

Rasch Metrology: How to Expand Measurement Locally Everywhere

Though it undoubtedly sounds outrageously audacious to suggest, Rasch measurement will inevitably be seen as setting the stage for an amodern/unmodern (Dewey, 2012; Latour, 1990, 1993) resolution of the modern universalist vs. postmodern relativist culture wars. How? Rasch measurement can inform personalized instruction or health care in ways that respect individual uniqueness and local contexts while also facilitating rapid and efficient communication of comparable measures generally. This capacity for allowing the local and special to remain local and special while also facilitating its communication relative to standards is precisely what philosophers of science are pointing to as the way out of the modern-postmodern standoff.

Haraway (1996) for instance, points out that "embedded relationality is the prophylaxis for both relativism and transcendence" (pp. 439-440). Golinski (2012) similarly says, "Practices of translation, replication, and metrology have taken the place of the universality that used to be assumed as an attribute of singular science" (p. 35). Latour (2005) concurs, saying

Standards and metrology solve practically the question of relativity that seems to intimidate so many people: Can we obtain some sort of universal agreement? Of course we can! Provided you find a way to hook up your local instrument to one of the many metrological chains… …traceability is precisely what the whole of metrology is about! [Actor Network Theory] is the social theory that has taken metrology as the paramount example of what it is to expand locally everywhere, all while bypassing the local as well as the universal. The practical conditions for the expansion of universality have been opened to empirical inquiries. It's not by accident that so much work has been done by historians of science into the situated and material extension of universals. Given how much modernizers have invested into universality, this is no small feat.

As soon as you take the example of scientific metrology and standardization as your benchmark to follow the circulation of universals, you can do the same operation for other less traceable, less materialized circulations... (pp. 228-229).

Recent collaborations of metrologists and psychometricians (Pendrill & Fisher, 2013, 2015; Mari & Wilson, 2013; Wilson, Mari, Maul, & Torres Irribara, 2015) complement earlier work separately performed in the two fields exploring ways in which traceability to standards might be feasible for measured constructs not typically thought capable of supporting them (Finkelstein, 2003, 2009; Fisher, 2009, 2012; Mari, Lazzarotti, & Manzini, 2009; Mari & Sartori, 2007; Pendrill, 2014; Wilson, 2013a, 2013b). The less traceable, less materialized circulations of fields like education or patient-centered health care are made more stable, identifiable, and manageable in the context of Rasch-calibrated scales. The rigor and logic of this work is already informing a reconceptualization of the conceptual and operational relationships in educational measurement, for instance (Fisher & Stenner, 2013, 2015; Fisher & Wilson, 2015; Stenner & Fisher, 2013).

Though it is common to assume that quantification and measurement inherently assume a modern, universalist perspective antithetical to a postmodern, relativist perspective, close parallels between developments in mathematics and deconstructive thinking have been noted (Tasić, 2001), and reductionism in mathematics and physics has been thoroughly discredited (Chaitin, 1994; Garfinkel, 1991). Further, the assumption of measurement as modernist is contradicted even within the world of engineering standards itself, as Mari and Sartori (2007) note that

... measurement is so fundamental in epistemology and even ontology that the entities involved in it cannot be in their turn founded. Rather, such entities are connected in a network of mutual conceptual and operative relations, which prevents any reductionistic attempt to identify 'the definitive foundation' to our empirical knowledge (p. 241).

Traceability stochastically incorporates explanatory theory and empirical evidence even in laboratory sciences typically assumed to be built up from billiard-ball or clockwork-universe predictive models. For instance, Berg and Timmermans' (2000) study of the constitution of universalities in medicine found that

In order for a statistical logistics to enhance precise decision making, it has to incorporate imprecision; in order to be universal, it has to carefully select its locales. … Paradoxically, then, the increased stability and reach of this network was not due to more (precise) instructions: the protocol's logistics could thrive only by parasitically drawing upon its own disorder (pp. 55-56).

Galison (1997, pp. 883-884) notes a similar process at work in microphysics, saying that the disorder of science's separate instrument-, experiment-, and theory-focused communities is responsible for its overall strength and coherence.

Establishing metrological traceability for outcome and impact measures in education and other fields will likely succeed, then, only insofar as conceptual and operational relationships expand locally everywhere, in a kind of stochastic resonance (Fisher, 1992, 2011) or feminist diffraction pattern (Haraway, 1996). The stochastic basis of Rasch measurement allows local variations in student ability, item content, and response consistency to be contextualized relative to a given unit with a known uncertainty. When put into practice in communications systems incorporating theory-, instrument-, and data-based quality checks implemented by those who care about the relationships involved, such measures will offer new possibilities for realizing enhanced levels of human potential, social cohesion, and environmental quality.

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

References

Berg, M., & Timmermans, S. (2000). Order and their others: On the constitution of universalities in medical work. *Configurations, 8*(1), 31-61.

Chaitin, G. J. (1994). Randomness and complexity in pure mathematics. *International Journal of Bifurcation and Chaos, 4*(1), 3-15. http://www.worldscientific.com/doi/pdf/10.1142/S0218127494000022.

Dewey, J. (2012). *Unmodern philosophy and modern philosophy* (P. Deen, Ed.). Carbondale, Illinois: Southern Illinois University Press.

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement, 34*(1), 39-48.

Finkelstein, L. (2009). Widely-defined measurement—An analysis of challenges. *Measurement, 42*(9), 1270-1277.

Fisher, W. P., Jr. (1992). Stochastic resonance and Rasch measurement. *Rasch Measurement Transactions, 5*(4), 186-187 [http://www.rasch.org/rmt/rmt54k.htm].

Fisher, W. P., Jr. (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement, 42*(9), 1278-1287.

Fisher, W. P., Jr. (2011). Stochastic and historical resonances of the unit in physics and psychometrics. *Measurement: Interdisciplinary Research & Perspectives, 9*, 46-50.

Fisher, W. P., Jr. (2012). What the world needs now: A bold plan for new standards [Third place, 2011 NIST/SES World Standards Day paper competition]. *Standards Engineering, 64*(3), 1 & 3-5 http://ssrn.com/abstract=2083975

Fisher, W. P., Jr., & Stenner, A. J. (2013). Overcoming the invisibility of metrology: A reading measurement network for education and the social sciences. *Journal of Physics: Conference Series, 459*(012024), http://iopscience.iop.org/1742-6596/459/1/012024.

Fisher, W. P., Jr., & Stenner, A. J. (2015). The role of metrology in mobilizing and mediating the language and culture of scientific facts. *Journal of Physics Conference Series, 588*(012043).

Fisher, W. P., Jr., & Wilson, M. (2015). Building a productive trading zone in educational assessment research and practice. *Pensamiento Educativo*, in review.

Galison, P. (1997). *Image and logic: A material culture of microphysics*. Chicago: University of Chicago Press.

Garfinkel, A. (1991). Reductionism. In R. Boyd, P. Gasper & J. D. Trout (Eds.), *The philosophy of science* (pp. 443-459). Cambridge, Mass.: MIT Press.

Golinski, J. (2012). Is it time to forget science? Reflections on singular science and its history. *Osiris, 27*(1), 19-36.

Haraway, D. J. (1996). Modest witness: Feminist diffractions in science studies. In P. Galison & D. J. Stump (Eds.), *The disunity of science: Boundaries, contexts, and power* (pp. 428-441). Stanford, California: Stanford University Press.

Latour, B. (1990). Postmodern? No, simply amodern: Steps towards an anthropology of science. *Studies in History and Philosophy of Science, 21*(1), 145-71.

Latour, B. (1993). *We have never been modern*. Cambridge, Massachusetts: Harvard University Press.

Latour, B. (2005). *Reassembling the social: An introduction to Actor-Network-Theory*. (Clarendon Lectures in Management Studies). Oxford, England: Oxford University Press.

Mari, L., Lazzarotti, V., & Manzini, R. (2009). Measurement in soft systems: Epistemological framework and a case study. *Measurement, 42*(2), 241-53.

Mari, L., & Sartori, L. (2007). A relational theory of measurement: Traceability as a solution to the non-transitivity of measurement results. *Measurement, 40*, 233-242.

Mari, L., & Wilson, M. (2013). A gentle introduction to Rasch measurement models for metrologists. *Journal of Physics Conference Series, 459*(1), http://iopscience.iop.org/1742-6596/459/1/012002/pdf/1742-6596_459_1_012002.pdf.

Pendrill, L. (2014). Man as a measurement instrument [Special Feature]. *NCSLI Measure: The Journal of Measurement Science, 9*(4), 22-33.

Pendrill, L., & Fisher, W. P., Jr. (2013). Quantifying human response: Linking metrological and psychometric characterisations of man as a measurement instrument. *Journal of Physics: Conference Series, 459*, http://iopscience.iop.org/1742-6596/459/1/012057.

Pendrill, L., & Fisher, W. P., Jr. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. *Measurement*, http://dx.doi.org/10.1016/j.measurement.2015.04.010.

Stenner, A. J., & Fisher, W. P., Jr. (2013). Metrological traceability in the social sciences: A model from reading measurement. *Journal of Physics: Conference Series, 459*(012025), http://iopscience.iop.org/1742-6596/459/1/012025.

Tasić, V. (2001). *Mathematics and the roots of postmodern thought*. New York: Oxford University Press.

Wilson, M. R. (2013a). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika, 78*(2), 211-236.

Wilson, M. R. (2013b). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement, 46*, 3766-3774.

Wilson, M., Mari, L., Maul, A., & Torres Irribarra, D. (2015). A comparison of measurement concepts across physical science and social science domains: Instrument design, calibration, and measurement. *Journal of Physics: Conference Series, 588*(012034), http://iopscience.iop.org/1742-6596/588/1/012034.

Teaching Rasch Measurement: Teaching about Item and Category Difficulty Estimates

Within survey research is the tendency to simply report a mean score of a given item on a survey with other pertinent statistics (e.g., errors, standard deviations, and ranges, etc…). Often, accompanying these estimates will be weighting formulas to control for differences between the sub-strata of subjects and the population under investigation (for instance, Rea and Parker, 2005). This can be particularly useful for understanding a population parameter associated with item responses in measurement situations where the item represents the entire construct. One can think of any number of reasons why a person would want only population estimates and there is a great deal of excellent research addressing population parameter estimates (e.g., voting for a political candidate, serving on a jury where a verdict must be delivered, providing feedback on a customer satisfaction survey, etc). As always, it depends on the question a researcher is asking.

The downfall with the above approach, however, is that a social scientist cannot always accurately gauge how one item, and the categories that make up the item (as in the case of a likert scale), correspond to other items and categories on the same instrument. Thus, nuanced understandings of a social science construct can be lost *and* the ability to test a measurement hypothesis becomes practically impossible. A unidimensional model of measurement addresses this problem by placing all items and categories onto a single scale and solves the problem: how much latent trait does it take to endorse any given discreet category on a measure? Of course, to pull this off a few requirements must be in place: (A) a singular measurement construct should be defined and operationalized, (B) robust theory is the driver of all thinking (*both in terms of measurement/analysis and content*), (C) sound instrumentation must be employed and, (D) adequate sampling must be available to generate statistical power.

Table 1 provides an example of an Item Response Model *and* juxtaposes that to common measurements from the population estimates of the survey in question. Compare the inferences from the tables. By utilizing the Item Response Model, relative values of measurement can be estimated and the response categories themselves can then be dissected. *This cannot be accomplished in a more rudimentary data analysis approach.* Ideally, analysts want to look at all of the data to make determinations about the validity of a measurement construct and the instrument. And, the model gives us another set of data analysis tools to accomplish that goal.

Item	Item Response Model Statistics				Basic Descriptive Statistics		
	n	Item Location (standard error)	Andrich Thresholds 1→2; 2→3; 3→4	Overall Fit Statistics Infit (*Zstandard*) Outfit (*Zstandard*)	Mean	Error of mean	Standard Deviation
(1) Teacher unions should lead the way on educational policy issues such as NCLB, state-funding, class size reductions and issues of educational improvement.	736	-.60 (.06)	- 0.98 - 0.71 + 1.68	.92 *(-1.5)* .90 *(-1.8)*	3.16	0.026	0.713
(2) Teacher unions should only bargain for economic necessities such as salary and benefits.	736	1.62 (.06)	-1.79 +1.11 +0.68	1.28 *(4.3)* 1.27 *(4.4)*	1.79	0.025	0.683
(3) Bargaining is a tool to establish sound policies.	718	-.29 (.06)	-1.53 -0.75 2.28	1.05 *(.8)* 1.04 *(.7)*	2.95	0.025	0.662
(4) Teacher unions' first priorities should be related to learning needs.	724	-.08 (.05)	-2.18 +0.47 +1.71	.97 *(-.7)* .98 *(-.4)*	2.69	0.029	0.784
(5) Teacher unions' first priorities should be related to teacher needs.	730	-.44 (.05)	-1.41 -0.28 +1.7	1.28 *(5.0)* 1.35 *(6.1)*	3.03	0.028	0.756
(6) Teacher unions should be creating and advocating for systems of shared governance in K-12 schools.	692	-.41 (.06)	-1.83 -0.57 +2.4	.87 *(-2.2)* .87 *(-2.2)*	2.96	0.025	0.646
(7) Teacher unions should encourage building level control over working conditions.	716	.14 (.06)	-1.69 -0.44 +2.13	.87 *(-2.4)* .88 *(-2.4)*	2.74	0.027	0.735
(8) Teacher unions have not utilized their collective power to advocate for real issues related to teaching and learning.	714	.21 (.05)	- 1.68 +0.16 +1.53	1.03 *(.7)* 1.06 *(1.3)*	2.61	0.031	0.829
(9) Teacher unions should work with school boards and administrators to identify common goals, beliefs and values.	742	-1.50 (.07)	-0.91 -1.51 +2.42	.92 *(-1.7)* .91 *(-1.9)*	3.36	0.02	0.543
(10) Identification of common goals, beliefs and values should be the starting point for bargaining with school districts.	731	-.85 (.07)	-0.41 -1.9 +2.31	.88 *(-1.8)* .85 *(-2.3)*	3.22	0.02	0.554
(11) Teachers win when teacher unions adopt militant strategies.	712	1.95 (.06)	-1.87 +0.43 +1.44	1.19 *(3.5)* 1.21 *(3.9)*	1.74	0.025	0.679
(12) Teacher bargaining agreements should be flexible documents that respond to the needs of buildings.	722	.25 (.06)	-1.61 -0.69 +2.3	.87 *(-2.5)* .87 *(-2.5)*	2.72	0.027	0.717
(13) Teacher bargaining agreements should respond to the needs of districts.	724	.01 (.06)	-1.76 -0.8 +2.56	.91 *(-1.5)* .90 *(-1.6)*	2.83	0.024	0.648

Table 1. Item information.

With the data, all items had four Likert response options (1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree) and were administered via an online survey. Approximately 700 licensed educators responded to the survey. The item location can be thought of as the logit center of the item, the Andrich threshold is where an item-category probability becomes equal to the adjacent category, and the fit statistics are provided for overall evaluation.

At the end of the day, researchers need to utilize data models that are aligned to theory. One way to align to theory is to evaluate constructs (such as the agreement with bargaining reforms) through an Item Response approach. In the above data we see several examples of theoretical issues that could be discussed, simply by carefully analyzing the observed measurements. As an example, here are three questions and observations that analysts could ask when evaluating these items and their measurements:

(1) Why is item 11, related to union militancy, one of the most difficult items on the survey? How might that inform a theory related to agreement with labor reforms? Could there be some underlying theoretical reason that this became the most difficult item for people to endorse? Conversely, why is question 9 the easiest?

(2) Why did item 8, related to the power of the teaching union, create a perfectly ordered item? While, on the other hand, items two and ten, related to the purposes of bargaining, exhibit disordered thresholds?

(3) If an individual were to only look at the mean score of items, it might be assumed that items two and eleven are very close to one other on the scale of agreement (mean of 1.79 and mean of 1.74, respectively). Yet, when the measurement model is considered, these items appear .33 logits away from each other. *Almost one-half of an entire logit!* Why does this occur?

Big takeaways. As Figure 1 shows, there is a relationship between the location of the item on the measurement scale and the mean level of agreement: the easier the items, the stronger the level of average agreement by participants which is not surprising to most Rasch analysts. However, note that certain items sit above the fit of the line and certain items sit below the fit of the line. Now, in this case, this is not a terribly aberrant situation. But it is enough to wonder whether the mean is as useful as a logit placement for determining relative position. Why? This is because average agreement does not model the interaction between the location of the item on the scale and the location of the person on the scale. And, one category from one measure will not directly correspond to its same category from another measure. *So, why treat them the same?*

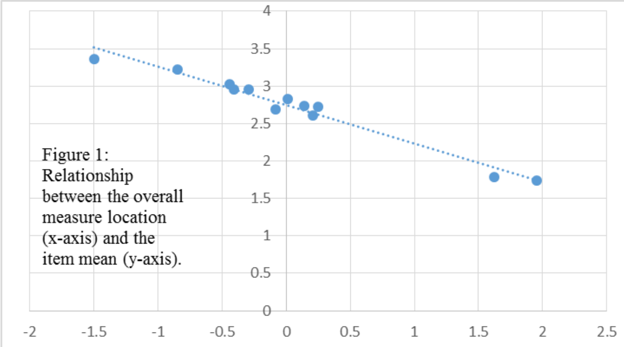


Figure 1.

The significant computational task of getting everything (respondents and measures and categories) onto a single scale was accomplished with the formula in the introduction. So, we are safe to consider all things related to the location of the item *and* the amount of the latent trait of the survey participants on the same scale, so long as the data fit the assumptions of the model.

Greg Sampson
Minot State University, Minot

References

Bond, T., Fox., C., (2001). *Applying the rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Erlbaum.

Kerchner, C., Koppich, J., & Werres, J. (1997). *United mind workers: Unions and teaching in the knowledge society.* San Francisco: Jossey-Bass.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach.* Mahwah, NJ: Erlbaum.
Rea, L,. & Parker, R., (2005). *Designing and conducting survey research: A comprehensive guide* (3rd ed.). San Francisco, CA: Jossey-Bass.

Sampson-Gruener, G., (2008). *An empirical analysis of educator beliefs related to post-industrial labor reforms in the State of Oregon.* (Unpublished doctoral dissertation). Oregon State University, Corvallis, OR.

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.

Introduction to Rasch Measurement Workshop

William Boone will be conducting a workshop on March 23-24, 2016 in Cincinnati, OH. Details about the workshop and registration information can be found at http://raschmeasurementanalysis.com/

Baseball Team Uses Computerized Video System to Call Balls and Strikes

The San Rafael Pacifics, an independent baseball team, recently relieved home plate umpires of their duty of calling balls and strikes after each pitch. Instead, the Pitch F/X system, the same system used to show viewers at home the location of each pitch, was used. Former MLB player Eric Byrnes served as the "strike zone umpire" and called pitches from behind a microphone. The technology used three cameras to judge the trajectory of each pitch and indicate a location of placement. The home plate umpire remained behind the plate and otherwise remained responsible for his normal duties. It is believed this experiment will reduce subjectivity in terms of umpire inconsistencies. If this technology proves useful over multiple experiments expect to see if in baseball parks everywhere in years to come.

Note: Thanks to Melanie Lybarger for sending this in. Readers can learn more about this story online at: http://sports.yahoo.com/blogs/mlb-big-league-stew/independent-team-to-use-computerized-video-system-to-call-balls-and-strikes-020858975.html

"Of course you still weigh 250 pounds! That's as high as the scale goes."

Journal of Applied Measurement
Vol. 16, No. 3, 2015

Comparison of Models and Indices for Detecting Rater Centrality, *Edward W. Wolfe and Tian Song*

Measuring Psychosocial Impact of CBRN Incidents by the Rasch Model, *Stef van Buuren and Diederik J. D. Wijnmalen*

Using the Partial Credit Model to Evaluate the Student Engagement in Mathematics Scale, *Micela Leis, Karen M. Schmidt, and Sara E. Rimm-Kaufman*

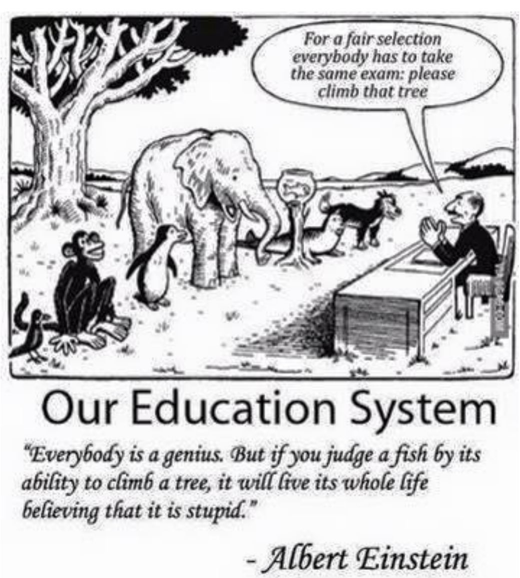
Estimation of Parameters of the Rasch Model and Comparison of Groups in Presence of Locally Dependent Items, *Mohand-Larbi Feddag, Myriam Blanchin, Véronique Sébille, and Jean-Benoit Hardouin*

Help Me Tell My Story: Development of an Oral Language Measurement Scale, *Patrick Charles, Michelle Belisle, Kevin Tonita, and Julie Smith*

A Dual-purpose Rasch Model with Joint Maximum Likelihood Estimation, *Xiao Luo and John T. Willse*

Using Rasch Analysis to Evaluate Accuracy of Individual Activities of Daily Living (ADL) and Instrumental Activities of Daily Living (IADL) for Disability Measurement, *Bruce Friedman and Yanen Li*

Richard Smith, Editor, www.jampress.org



Rasch-related Coming Events

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

Oct. 23-Nov. 20, 2015, Fri.-Fri. Online workshop: Rasch Applications, Part 2: Clinical Assessment, Survey Research, and Educational Measurement (W. Fisher), www.statistics.com

Nov. 13, 2015, Fri. In-person workshop: Rasch Models in Business Administration, in Homage to Professor Pedro Alvarez-Martinez, Tenerife, Canary Istlands, www.institutos.ull.es

Dec. 2-4, 2015, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

March 18, 2016, Fri., UK Rasch User Group Meeting, Durham, UK, www.rasch.org/uk

March 23-24, 2016, Wed.-Thurs., In-person workshop: Introduction to Rasch Measurement using Winsteps (W. Boone), Cincinnati, OH, http://raschmeasurementanalysis.com

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 Dec. 1, 2015, so please make your submission by Nov. 1, 2015 for full consideration. Please email Editor\at/Rasch.org with your submissions and/or ideas for future content.