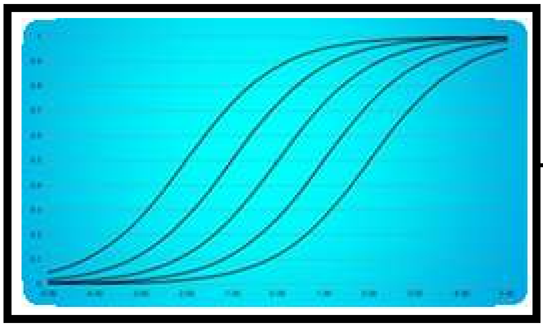


RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 28 No. 2 Autumn 2014 ISSN 1051-0796

The Sin of False Precision: Too Many Rating-Scale Categories

Kaplan (1996, p. 23) identifies "The Sin of False Precision." False precision is the assignment of a numeric value to a situation that has higher precision than the situation actually supports. Kaplan gives the following example: there are two urns. One contains 50 black balls and 50 white balls. The other contains an unknown number of balls, and the proportion of black to white is unknown. A ball is drawn from each urn. What is the probability that each ball is a black ball? From the first urn, the probability of a black ball is 50/100 = .5, in fact .50. What about the second urn? Since we have no evidence to expect one color more than the other, the probability of a black ball is also .5, but to say to ourselves that it is .50, or .500000, is to assign a false precision and to mislead ourselves about our ignorance.

Consider a typical 10 cm. visual-analog scale. We could measure the response in cm., on a rating scale of 0-10, or in mm. on a rating scale of 0-100. Is a rating of 91 mm. a stronger statement than a rating of 9 cm.? It would seem so. Then what about a rating of 91036 micromillimeters (microns)? Surely 91036 microns is in the same class of numbers as the commonly-encountered "22.6 grams of fat per serving" - a number labeled as "much too precise", in fact of "unwarranted" and "meaningless precision" (Paulos, 1995).

"A sharpening of the meaning of an indefinite statement results, as a rule, in strengthening the statement, i.e., in replacing the original statement by a stronger one. Now, are we entitled to do this? Does the evidence available to us justify the statement so modified? The answer in typical cases seems to be: no." (Przelecki, p. 215) In fact, Przelecki points out that a major source of imprecision is that many statements are metaphorical (p. 217). Applying this idea to measurement in the social and health sciences, our test items are metaphorical expressions of the latent variable. "Thus ... we are doomed to some kind of indefiniteness in our philosophical thinking." (p. 218).

"The postulate of precision requires that we strive for precision in so far as it is a necessary condition of the decidability of the problems under discussion. But, there are some limits to such an endeavor, because there are some limits of precision - of a linguistic and epistemic nature. Absolute precision is unattainable. We must learn to live with imprecision - to think and to speak rationally in spite of its all-pervasive nature." (p. 218).

In designing and analyzing rating scales, it is easy to construct response devices which exhibit false precision. "On a scale of 1-10, rate the cleanliness of your hotel room." Can anyone distinguish 10 levels of cleanliness? Surely “bad”, “good” and “excellent” are enough for practical purposes.

From the Rasch perspective, the data must not only be ordinal, but the empirical rating-scale categories must also express substantive qualitative advances along the latent variable. A major activity in the Rasch analysis of polytomous qualitative data is the identification of productive categorizations. This contrasts with the "graded response" model in which the empirical categorization is regarded as inconsequential: "it is nearly always appropriate to consider models [such as the Graded Response Model] that are invariant under the grouping of adjacent response categories" (McCullagh 1985 p.39).

John Michael Linacre

References

Mark Kaplan (1996) *Decision Theory as Philosophy*. Cambridge: Cambridge University Press.

Peter McCullagh (1985) Statistical and scientific aspects

Table of Contents

False Precision: Too Many Rating Scale Categories (Linacre) 1463

The Central Theoretical Problem of the Social Sciences (Fisher) 1464

Social Network Analysis and Rasch as Complementary Methods (Royal & Akers) 1466

Games Psychometricians Play (Marosszeky) 1467

Statistical Significance vs. Rasch (Linacre) 1470

C. H. Townes: A Mentor Among Mentors (Fisher) 1472

of models for qualitative data. p. 39-49, in P. Nijkamp et al. (Eds), *Measuring the Unmeasurable*, Dordrecht, The Netherlands: Martinus Nijhoff.

John Allen Paulos (1995) *A Mathematician Reads the Newspaper*. New York: Basic Books.

Marian Przelecki (1998) The Postulate of Precision. pp. 209-219 in K. Kijania-Placek and J. Wolensky (Eds.), *The Lvov-Warsaw School and Contemporary Philosophy*. Dordrecht: The Netherlands.

The Central Theoretical Problem of the Social Sciences

In a work first published in 1945, almost 70 years ago, Hayek (1948) describes the main theoretical problem of all social science as distinguishing between a true, social individual and a false, Cartesian, rationalist individual. In direct opposition to what he (Hayek, 1988) considered Descartes' (1999) "fatal conceit"—the claim that an informed individual could grasp the knowledge needed to organize society—Hayek (1948) says that "the central question of all social sciences" is this:

"How can the combination of fragments of knowledge existing in different minds bring about results which, if they were to be brought about deliberately, would require a knowledge on the part of the directing mind which no single person can possess? To show that in this sense the spontaneous actions of individuals will, under conditions which we can define, bring about a distribution of resources which can be understood as if it were made according to a single plan, although nobody has planned it, seems to me indeed an answer to the problem which has sometimes been metaphorically described as that of the 'social mind'" (p. 54).

Hayek's concern is with the broad social coordination of individual behaviors and decisions, especially in the economic domain. Writing just at the end of World War II and at the start of the Cold War, Hayek is concerned with how social and economic affairs can be managed effectively along a path that rejects socialized central planning as well as any form of capitalism requiring omniscient and fully rational individuals. The latter extreme of unbounded rationality has since been recognized as one of the major theoretical shortcomings of capitalist economics (Frantz and Leeson, 2013; Kahneman, 2003; Simon, 1982; Williamson, 1985). In the larger debates, Hayek's distinction between true and false individualisms was frequently lost—and less frequently appreciated—on both ends of the political spectrum.

Hayek had a strong interest in the philosophy of science, but he found little in the way of methods transferable from the natural sciences to economics or psychology (O'Brien, 1994, p. 354). His early training in psychology inclined him toward introspective, experiential methods and skepticism as to the roles of measurement and testing in economics. The close examination of networks of associations embodied by standards groups, metrologically traceable instrumentation, professional societies, educational cohorts, laboratory collaborations, etc. in recent studies of science and scientists (Callon, 2002; Latour, 1987, 2005), and economists, such as Tarde (Latour, 2002, 2010; Latour and Lepinay, 2009) has, however, redefined individuals and social relations in terms very similar to Hayek's. In this context, the philosophy of science comes to share Hayek's stress on the problem of how information is coordinated across individuals without the intervention of an omniscient rationality. Previous research has extended the concept of social individuality into considerations of how to re-invent social measurement in the social sciences (Fisher, 2000, 2005, 2009), but not in relation to Hayek's contributions.

The problem is one of understanding how to coordinate local behaviors and decisions over a variety of different kinds of decisions across wider swaths of society. Hayek positively quotes Whitehead's (1911, p. 61) observation that "Civilization advances by extending the number of important operations which we can perform without thinking about them." Indeed, everyday tools like telephones, computers, and automobiles are now so complex that individual engineering experts do not have the range of knowledge needed to master all of the component parts in a single device. Most people have little more than the most elementary grasp of how their homes, furniture, clothing, or food are produced, and have even less of an inkling when it comes to their medications or their electronic communications, computing, and entertainment systems.

Coordinating behaviors and decisions in the absence of full information requires trust and a basis in some kind of evidence that things are as others say they are. Entrepreneurial innovation, especially, depends on assurances that technologies and markets will be in place at specific points in time. The ability to plan positive outcomes for long term investments in education or manufacturing, for instance, requires foresight into what the future may hold concerning predictable patterns of employment, supply availability, and consumer preferences. One way in which trust develops even in the absence of full understanding is in terms of repeated patterns of investment documenting steady growth in a particular direction or area of the economy. For example, in the U.S. from the 1920s on, state and federal highway construction budgets assured automobile manufacturers that their customers would have roads on which to drive their vehicles. Similarly, rural electrification programs opened up markets for appliance manufacturers, who could be confident that consumers everywhere would have the voltage needed to power their products. The most famous example of an industry-wide shared understanding of coordinated and aligned investments is, however, Moore's Law, which projects a doubling of microprocessor speeds every two years and has served as a business model coordinating investment decisions in consumer electronics for over 50 years (Miller and O'Leary, 2007). The fact that a similar law based in Rasch measurement has been operating behind the scenes in education over the same time frame (Fisher and Stenner, 2011) seems to have gone unnoted, along with perhaps quite significant opportunities for improved outcomes in reading comprehension.

The pertinent questions for psychology and the social sciences following from these observations concern what technical developments might serve to provide a basis for growing similar networks of verifiable and contractually enforceable trust in education, health care, environmental management, social services, human resources, etc. For instance, if the claims made by Rasch measurement practitioners concerning objectivity and success in item banking, adaptive instrument administration, scale equating, construct mapping, causal modeling, quality assessment, and predictive control over items really are practical accomplishments, should not they provide theory, tools, and evidence relevant to forming the alliances of trust needed for countering the unavoidable imperfection of any one individual's knowledge? Would not a system of validated construct theories, interconnected instrumentation, efficiently shared experimental results, and critical comparability dramatically improve communication and mutual understanding?

Against the mainstream of widespread disregard for the problem, disregard that has persisted from Hayek (1948, p. 91) to the present, Rasch results produced to date suggest the unavoidable imperfection of knowledge and the need for processes by which knowledge can be constantly communicated and acquired are not insurmountable barriers to human progress. Identifying and formulating the problem, as is so often the case, is far more complex than actually solving it.

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

References

Callon, M. (2002). From science as an economic activity to socioeconomics of scientific research: The dynamics of emergent and consolidated techno-economic networks. In P. Mirowski & E.-M. Sent (Eds.), *Science bought and sold: Essays in the economics of science*. Chicago: University of Chicago Press.

Descartes, R. (1999). *Discourse on method and meditations on first philosophy*. Cambridge, MA: Hackett.

Fisher, W. P., Jr. (2000). Objectivity in psychosocial measurement: What, why, how. *Journal of Outcome Measurement, 4*(2), 527-563.

Fisher, W. P., Jr. (2005). Daredevil barnstorming to the tipping point: New aspirations for the human sciences. *Journal of Applied Measurement, 6*(3), 173-179.

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., & Stenner, A. J. (2011). A technology roadmap for intangible assets metrology. In Fundamentals of measurement science. International Measurement Confederation (IMEKO) TC1-TC7-TC13 Joint Symposium, http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24493/ilm1-2011imeko-018.pdf, Jena, Germany, August 31 to September 2.

Hayek, F. A. (1948). *Individualism and economic order*. Chicago: University of Chicago Press.

Hayek, F. A. (1988). *The fatal conceit: The errors of socialism* (W. W. Bartley, III, Ed.) (Vol. I). The Collected Works of F. A. Hayek. Chicago: University of Chicago Press.

Hyde, L. (2010). *As common as air: Revolution, art, and ownership*. New York: Farrar, Straus and Giroux.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review, 93*(5), 1449-1475.

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

Latour, B. (2002). Gabriel Tarde and the end of the social. In P. Joyce (Ed.), *The social in question: New bearings* (pp. 117-132). London: Routledge.

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

Latour, B. (2010). Tarde's idea of quantification. In M. Candea (Ed.), *The social after Gabriel Tarde: Debates and assessments* (pp. 145-162). London: Routledge.

Rasch Measurement Transactions
www.rasch.org/rmt
Editor: Kenneth Royal
Email Submissions to: Editor \at/ Rasch.org
Copyright © 2014 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

Latour, B., & Lepinay, V. A. (2009). *The science of passionate interests: An introduction to Gabriel Tarde's economic anthropology*. Chicago: Prickly Paradigm Press.

Miller, P., & O'Leary, T. (2007). Mediating instruments and making markets: Capital budgeting, science and the economy. *Accounting, Organizations, and Society, 32*(7-8), 701-734.

O'Brien, D. P. (1994). Hayek as an intellectual historian. In J. Birner & R. van Zijp (Eds.), *Hayek, coordination and evolution* (pp. 343-374). New York: Routledge.

Simon, H. A. (1982). *Models of bounded rationality: Empirically grounded economic reason* (Vol. 3). MIT Press.

Whitehead, A. N. (1911). *An introduction to mathematics*. New York: Henry Holt and Co.

Williamson, O. E. (1985). *The economic institutions of capitalism: Firms, markets, relational contracting*. New York: The Free Press.

Social Network Analysis and Rasch Measurement as Complementary Methods

The use of Social Network Analysis (SNA) has steadily gained in popularity in recent years. Regular applications of SNA occur in the fields of psychology, anthropology, sociology, epidemiology, and business. SNA is primarily used to examine connections and relationships between individuals and groups, rather than documenting the output of an isolated individual. Analyzing social networks at this level gives researchers the ability to view the structure of the network and understand how the connections among individuals within the network affect substantive outcomes.

Perhaps the most notable strength of SNA is its ability to view a social network in all its complexity. A thorough inspection of a single graphic can be very informative and useful for persons with an aversion to quantitative output (see Figure 1). Additionally, individual (ego) maps can be created that focus on an individual or even individuals with specific characteristics. Maps can be used to address questions such as: does a network exist? Do relationships exist among a network's members? Are there enough relationships to merit further investigation? Individuals, often called nodes, can be arranged by certain identifiable characteristics, such as their connections (as seen in Figure 1), or randomly. Additionally, the size and shape of the nodes and their connections, often called ties, can also be altered to offer a clearer picture of the network.

Consumers of SNA methodologies typically refer to the resulting output as "measures", but this conceptualization of a measure differs from that of a proponent for Rasch models. Most SNA analyses produce a variety of centrality measures, which indicate the key individuals within a data set. Centrality measures vary from simplistic (who has the most connections?) to much more complicated (who has the potential to have the most connections based on the current pattern of connections?). Centrality measures can be used to find persons that are considered "in the know". Matrix algebra provides the mathematical underpinnings for SNA techniques, as these matrices are a way of collecting data and serve as the foundation for quantitative data analysis. Formulae such as Freeman Degree centrality illustrate the types of mathematical computations occurring within the SNA methodology. Freeman degree centrality states for a given binary network with vertices v1....vn and maximum degree centrality cmax, the network degree centralization measure is S(cmax - c(vi)) divided by the maximum value possible, where c(vi) is the degree centrality of vertex vi. Most SNA software packages, such as UCINET, Gelphi and Pajek also include the ability to visualize the network. NetDraw is the freeware network visualization tool included with UCINET. It was also used to create the diagram in Figure 1.

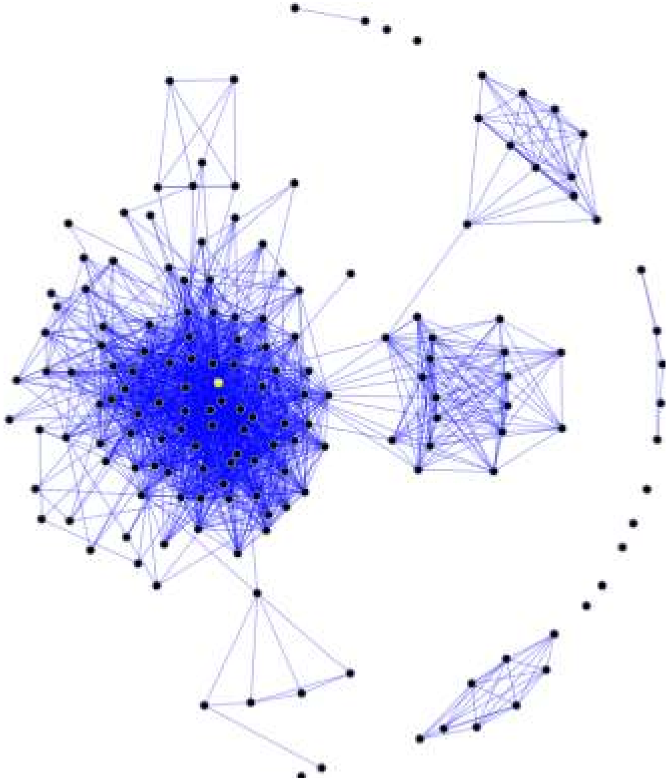


Figure 1. A sample social network map.

The brief example presented above illustrates a very basic application of SNA, and similar to Rasch modeling, analyses can become much more complex to answer more intricate questions. The form of SNA presented here is primarily mathematical, but other SNA models, such as stochastic network analysis models, often have a more statistical feel. Many of these models possess the limitations routinely discussed in the psychometrics literature regarding objective measurement. For this reason, it is helpful to consider the use of Rasch models as a complementary approach for obtaining more robust measures.

While many think of Rasch and SNA analyses as entirely distinct approaches (and they are), the fact remains that a wide variety of datasets collected for SNA purposes may also accord to Rasch measurement analyses. Although the methods typically are intended for different purposes, one can easily leverage the strengths of each methodology for maximum inferential benefit. More specifically, Rasch analyses may provide more robust and objective measurements of a network member's "social connectedness", "strength", or other construct of interest, while SNA analyses provide meaningful visuals of the network for easy interpretation. Table 1 provides an example of what a network's membership might look like in terms of demographic data and Rasch measures. Inferential statistical techniques can then be applied for additional analyses (e.g., Are males more likely to have more social connections within a particular network than females? Are females more likely to have stronger social relationships than males within a particular network? What role, if any, does race play in understanding a given network's social relationships? Etc.).

We encourage others to continually seek ways to use Rasch models as complementary approaches, where appropriate, to further evidence the utility of these models in a variety of applications and settings.

Table 1.

Row	Person	Gender	Race	Measure	SE
1	John	M	1	2.13	.28
2	Cheryl	F	3	1.49	.26
3	Mike	M	1	1.85	.25
4	Anna	F	2	2.02	.25
…	…	…	…	…	…
100	Jessica	F	1	1.76	.28

Kenneth D. Royal, North Carolina State University &
Kathryn S. Akers, Kentucky Center for Education and Workforce Statistics

References

Borgatti, S. P. (2002). NetDraw: Graph Visualization Software. Harvard, MA: Analytic Technologies.

Borgatti, S. P., Everett, M. G., & Freeman, L.C. (2002). Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Games Psychometricians Play

David Andrich is right! A Kuhnian paradigm shift is required in the data-model relationship in order to support Rasch measurement revolutionary science. Meanwhile... normal science continues in psychometric practice, as evidenced by the following publication by GlassLab (www.glasslabgames.org) entitled "Psychometric Considerations in Game-Based Assessment". It must be remembered that this publication occurs 60 years after Lord's football numbers controversy (Lord, 1953, 1954a, 1954b; Scholten & Borsboom, 2009). Here is just one quote from the GlassLab publication:

"The word psychometrics means 'mental measurement'. It originated more than a century ago with an aim of measuring traits - but the models don't know this. In educational assessment, psychometricians and statisticians have developed a toolkit to support reasoning from noisy data in real-world problems, to help monitor and guide learning. There are concepts and techniques for gathering information about what people know and can do, and methods for characterizing the amount and quality of evidence for given purpose. Our concern lies in this reasoning-about-evidence aspect of psychometrics. Quantitative methods for reasoning about evidence do not require any presumption of quantitative traits 'inside people's heads'." (p. 9)

As Borsboom (2006) states "... contemporary test analysis bears an uncanny resemblance to the psychometric state of the art as it existed in the 1950s" (p. 425). Further examples of psychometric normal science are highlighted by a number of other quotes from the GlassLab publication. These selected quotes are presented in the tables at the end of this article.

This normal science approach to psychometric practice is based on the epistemologically problematic concept of a-theoretical data modeling and this leads to a number of consequences. These include:

1) the tendency, need or desire for more data to model;
2) the tendency, need or desire for more and more complex models, in order to describe and deal with more and more data;
3) the tendency, need or desire to play with numbers.

Philosophical Justification

This section elaborates on the philosophical justification for the above comments, citing the originators or sources of these ideas in the measurement literature. Andrich (2002, 2004 and 2013) nicely sums up the standard operating procedures of the

Notable Quote

Commenting on GlassLabs (2014) advocacy for the excessive use of data modeling, Nick Marosszeky writes "*There are more models here than there are at a Paris Fashion Show!"*

a-theoretical data modelers / data describers of the normal science paradigm:

"... the traditional view of the relationship between model and data - the model should describe the data and if it does so, then that is sufficient to understand the data." (Andrich, 2002, p. 348)

"In the traditional paradigm, data are more sacrosanct from the point of view of being modeled, and the data are less likely to be abandoned if a better fitting model can be found." (Andrich, 2004, pp. 1-14)

"... in the traditional paradigm, the case for choosing one model over another is that it accounts better for the data. The data are given. In general, the model with a greater number of parameters accounts for the data. If, according to available statistical checks, it does not, the model with fewer parameters is favored." (Andrich, 2004, pp. 1-8)

Then he highlights the core difference between the paradigms, namely the relationship between the data and the model:

"To consider that when there is a mismatch between the data and the model it might be a problem with the data rather than the model, is in itself a considerable perceptual shift from the traditional perspective on the data-model relationship." (Andrich, 2002, p. 351)

"The class of models within Rasch measurement theory is based on the requirement of invariant comparisons, a requirement specified independently of any data set." (Andrich, 2013, p. 7)

Andrich (2013) then provides a proto-typical example of revolutionary science practice, where improved measurement is the focus of psychometric effort, rather than on collecting more data or developing more detailed models (i.e. data describing or data modeling).

"Rasch then applied the SLM to responses to Raven's (1940) nonverbal test of reasoning and to a Danish military intelligence test. He concluded that the Raven's data fitted the SLM, but that the military test did not: He had one success and one failure with his new model. However, instead of modifying the model to account for the data (in particular by adding a discrimination parameter in the first instance) he studied the pattern of misfit and concluded that the test seemed to be composed of four different kinds of items. As a result, the head of the military psychologists had four tests constructed, one for each class of items, where each test would conform to Rasch's new model. Thus, it was the data from set of intelligence items that was seen to fail, not the model. More important, and distinctively, however, this failure led to further experimentation in test development leading to the improvement in the assessment of the intended proficiencies, rather than to the application of models with more parameters which would have better modeled the original data. The test was required to be substantively valid, but in addition, and contributing to the validity, it was required to conform to the model, that is, conform to invariant comparisons of items and persons." (Andrich, 2013, p. 11)

Michell explains the third tendency of playing with numbers or numerical coding - "constructing number generating operations" (Michell, 1986, page 405). Applying numerical operations to data based on numerical coding grows out of operational thinking about measurement. Here numerical codes are used to describe or classify outcomes and events. The resultant numbers produce an ordered structure which can then be manipulated arithmetically. A flexible approach to numerical coding allows one to play with the resultant numerical data to describe and present your results in any ordered sequence you want. Mathematics then does the rest.

Bell, Staines and Michell (2001) summarise this view in their textbook "Evaluating, doing and writing research in psychology" (2001). They state very clearly that:

"Numerical data are data expressed using numerals. There are three kinds: descriptions of frequencies; descriptions of magnitudes; and descriptions employing numerical coding." (p. 251)

"All data are descriptions. Not all data are numerical. Not all numerical data are quantitative." (p. 251)

"Numerical data in psychology are obtained by counting, measuring, estimating and coding. All measureable attributes are quantitative (additive in structure). The distinction between quantitative and non-quantitative attributes is empirical." (p. 251)

"When non-quantitative data is coded numerically, conclusions drawn using standard arithmetic methods may not be validly entailed by the original non-quantitative data." (p. 251)

Bell, Staines and Michell (2001) then elaborate as follows:

"When the attributes that psychologists study are not quantitative, it is still always possible to obtain numerical data via numerical coding. The most common example is when the data are simply ordinal. A set of objects are ordered when they fall along a line, each object before or after each other object such that if any object, X, comes before any other, Y, and Y comes before another, Z, then X must also come before Z. Since the numbers are also an ordered sequence in this sense, any order can always be coded numerically by assigning numerals to the objects they are assigned to. However, the numerals used to code an order are never unique: a different set of ordered numerals could always have been used instead." (p. 239)

"Special problems arise when numerical data are produced by coding non-quantitative information. As noted, psychologists often mistakenly treat data as simply a set of numbers, and there is a widely shared view that 'the numbers do not know where they came from' (for example Gaito 1980, p 566). Believing this, there is a strong temptation to treat

numerical codings arithmetically, just the same as genuinely quantitative data. This may lead unsuspecting researchers to draw conclusions from numerical codings that depend only upon arbitrary features of the code used, rather than upon features representing the empirical information coded." (pp. 244-245)

"Where a conclusion based upon numerical data does not remain true when those data are transformed admissibly, then it is a sign the researcher is skating on thin ice. If, in such circumstances, a researcher wants to draw conclusions, not about the numbers used, but about the attributes coded, then before such conclusions may be treated as valid, the researcher needs to show that contradictory conclusions could not be drawn if the numerical data are transformed in admissible ways. Of course, passing this test does not ensure that the conclusion obtained follows validly from the data, but failing it does mean that the inference is invalid. It is necessary, but not sufficient, condition for validity. In numerical coding, the facts represented are not intrinsically numerical and, so, coding them numerically can mislead." (p. 248)

Concluding Comment

Finally, the GlassLab publication argues that psychometric involvement in game based assessment advances psychometric practice: "The field of psychometrics is challenged to extend insights it has developed over the past century for reasoning from simpler forms of evidence, to now support reasoning in 'the digital ocean'." (GlassLab, 2014, p. 12). However, from a Rasch measurement revolutionary science perspective, the selected quotes do not support effective scientific reasoning or psychometric practice. I would argue that when you go sailing out into an ocean it is always helpful to have a map and a compass (see Stone, Wright and Stenner, 1999).

Nick Marosszeky, Macquarie University, Australia

References

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

Andrich, D. (2004). Controversy and the Rasch Model: A characteristic of incompatible paradigms? *Medical Care, 42,* 7-16.

Andrich, D. (2013). The legacies of R.A. Fisher and K. Pearson in the application of the Polytomous Rasch Model for assessing the empirical ordering of categories. *Educational and Psychological Measurement, 20*, 1-28.

Bell, P., Staines, P., & Michell, J. (2001). Measurement and numerical reasoning (Chapter). In P. Bell, P. Staines & J. Michell (Eds.), *Evaluating, doing and writing research in psychology* (pp. p. 234-254). London: SAGE Publications Ltd.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*, 425-440.

GlassLab. (2014). Psychometric considerations in Game-Based Assessment (White paper). www.glasslabgames.org, Redwood City, CA.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8*, 750-751.

Lord, F. M. (1954a). Further comment on "Football Numbers" (Reply to Letters to the Editor). *American Psychologist, 9*, 264-265.

Lord, F. M. (1954b). Scaling. *Review of Educational Research, 24*, 375-392.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100*, 398-407.

Stone, M. H., Wright, B. D., & Stenner, A. J. (1999). Mapping Variables. *Journal of Outcome Measurement, 3*, 308-322.

Scholten, A. Z., & Borsboom, D. (2009). A reanalysis of Lord's statistical treatment of football numbers. *Journal of Mathematical Psychology, 53*, 69-75.

Appendix

GlassLab (2014), p. 105: This simple example illustrated a number of key ideas: Modeling salient aspects of students' proficiencies in terms of student-model variables. Modeling salient aspects of performance in terms of observable variables. Modeling distributions of observable variables in terms of conditional probabilities, given SMVs. Building and parameterizing the models in terms of theory, experience, and designed-in expectations. Using a Bayesian modeling framework so we can make coherent inferences about players, update the models as data become available, and assemble model fragments to suit evolving game situations. These same ideas obtain in exactly the same way conceptually with MIRT models and diagnostic classification models, even though the forms of the models and the details of calculation differ accordingly.

NM: *Consequence*: Lots of models are available to be used;
Comment: There are more models here than there are at a Paris Fashion Show!

GlassLab (2014), p. 85: Much of the excitement about game-based assessment is about being able to capture fine-grained data about player activity. The promise is that this data will help us understand the processes that players use to solve problems, not just their final products. It is argued that there is great potential for generating new insights regarding complex knowledge, skills and attributes.

However, the potential of games as assessment tools can be met only if methods for making sense of stream or trace data (in familiar terms, "scoring it" [6]) can be developed in evidentiarily sound and computationally feasible ways. Traditional psychometric models have commonly been focused on point-in-time models that overlook variation in activity over time (especially at the micro level). New interactive digital experiences such as on-line learning environments and games, however, elevate both the availability and importance of understanding student temporal micro-patterns, which can reflect variation in strategy or evolving psychological states. While the richness of the data holds promise for making important inferences, few standard methods for scoring and analysis exist.

A primary challenge in fulfilling the potential of log files for making inferences about students thus lies in evidence identification. We have tasks that are often open and multifaceted, in which learners can interact with the digital environment in a number of ways, choosing various paths through the game environment. What are the important features of a work product and how do we apply scoring rules? Log files present many types of data, including sequences, frequencies, and duration of actions. Potential evidence for each construct must be gleaned from the masses of potential data available. We must determine how to turn this evidence into values of observable variables. In traditional multiple choice tests, scoring is quickly accomplished by evaluating each response as correct or incorrect. When assessing new constructs with new forms of data, the simple notion of "correctness" may no longer be good enough.

[6] We sometimes use the term "scoring" at times because it is familiar, but familiarity is a disadvantage when it constrains thinking about what to look for, how to characterize it, and how to use it (Behrens et al., 2012). The less familiar terminology of ECD is more useful because it situates thinking in the realm of evidentiary argument more broadly, and allows us to talk in ways that apply to familiar assessment but also, in a rigorous way, to the more complicated challenges that arise in unfamiliar forms of assessment such as GBAs.

NM: *Consequence*: Playing with numbers; *Comment*: This extract gives new meaning to the phrase "scoring with models".

GlassLab (2014), p. 43: [2] We do not expect a description of a level to characterize a given student universally across systems and contexts. Evidence suggests that peoples' understanding of systems can vary substantially from one system to another; that increasing understanding need not follow well-defined levels; and different situations can evoke thinking at different levels even within the same person (Sikorski & Hammer, 2010). Rather, we use the learning progression to manage situations and demands in the game, and to organize a probabilistic summary of patterns of "noisy" performance of students as they work through challenges with increasingly complex aspects of systems. We can use the learning progression to help design situations and manage evidence, without having to take it as a "faithful" model of students' capabilities.

NM: *Consequence*: Context dependent modeling; *Comment*: Anything goes.

Statistical Significance vs. Rasch Measurement

Which provides more interpretable results? Maps of measures or lists of significance test results? A test case is Beh E.J. & Davy P.J. (1998) Partitioning Pearson's chi-squared statistic for a completely ordered three-way contingency table. *Australian & New Zealand Journal of Statistics 40(4),* 465-477. The Table shows their dataset:

Cross-classification of 1517 people according to happiness, schooling and number of siblings

Years of schooling	Number of siblings				
	0-1	2-3	4-5	6-7	8+
	Not too happy				
<12	15	34	36	22	61
12	31	60	46	25	26
13-16	35	45	30	13	8
17+	18	14	3	3	4
	Pretty happy				
<12	17	53	70	67	79
12	60	96	45	40	31
13-16	63	74	39	24	7
17+	15	15	9	2	1
	Very happy				
<12	7	20	23	16	36
12	5	12	11	12	7
13-16	5	10	4	4	3
17+	2	1	2	0	1

And here is the first quarter of their chi-squared table:

$\chi^2_{\text{Siblings-Years}}$	Value	d.f.	P Value
Siblings components			
Location	222.2234	3	0
Dispersion	7.7034	3	0.0528
Error	5.3720	3	0.5035
Years of Schooling components			
Location	209.9878	4	0
Dispersion	24.2943	4	0.0001
Error	1.0168	4	0.9156
Total	235.2988	12	0

A conclusion is that "the difference in number of siblings is significantly related to happiness as is the difference in years of schooling." But do happier people have more siblings or less? More years of schooling or less? These questions about happiness are not answered, but here is a more basic question: "How does number of siblings relate to years of schooling?"

Let's model "Number of Siblings" as a 5-category Rasch rating scale with "Years of Schooling" as 4 objects of measurement. Then the expected "Number of Siblings" for the different "Years of Schooling" is:

Logit Measure	Years of Schooling	Number of Siblings (Expected Value)
2		(0-1)

1		
	17+	2-3
	13-16	
	12	---
0		4-5
	<12	---
		6-7
-1		

-2		(8+)

Now let's model "Years of Schooling" as a 4-category Rasch rating scale with the "Number of Siblings" as 5 objects of measurement. Here are the expected "Years of Schooling" for the different "Number of Siblings":

Logit Measure	Number of Siblings	Years of Schooling (Expected Value)
2		(17+)

1		
		13-16
0	0-1	---
	2-3	
		12
	4-5	
-1	6-7	

	8+	
-2		(<12)

These Rasch maps tell us that more schooling goes with less siblings, but not symmetrically. With 17+ years of schooling we expect, on average, 2-3 siblings, but with 2-3 siblings we expect, on average, 12 years of schooling. Two pictures of Rasch measures tell us much more than a chi-squared table.

Rasch analysis was performed with *Facets.*

John Michael Linacre

IOMC 2015 - International Outcomes Measurement Conference

The *Journal of Applied Measurement* and JAM Press have organized a health outcomes measurement conference to be held at the Crowne Plaza Hotel in Chicago on Tuesday, April 21 and Wednesday, April 22, 2015. The hotel is located just off of Michigan Avenue at 160 E. Huron Avenue, two blocks south of Water Tower Place.

The conference schedule will allow two full days of presentations devoted to the application of Rasch measurement to health outcomes. The selection process for presentations will be completed by the program committee. The cost of registration has not yet been determined, but will be in the $50 to $100 range.

The deadline for proposals for paper presentations is October 3, 2014. Presenters will be notified of the acceptance of proposals by November 14, 2014. Proposals should contain all author information (name, mailing address, e-mail, and phone for all authors) and are limited to 1000 words. The presenting author and first author for publication should also be indicated. Proposals (in a pdf format) can be e-mailed to the organizing committee at IOMC2015@JAMPress.org.

The papers presented at this conference will be published in the *Journal of Applied Measurement* beginning in 2016. Authors will be required to provide a peer review copy of the **manuscript** that will be presented by March 20, 2015 to expedite publication. The *Journal of Applied Measurement* is indexed in Pub Med and *IndexMedicus*, which gives medical researchers easy access to the abstracts of the papers.

The conference will feature both plenary and concurrent presentations. Plenary presentations will be 35 minutes in length with each set of two papers followed by a 20 minute discussion. Concurrent presentations will be 20 minutes in length with each set of three papers followed by a 20 minute discussion. Similarly themed papers will be grouped for presentation. Authors are also invited to submit grouped papers. We are planning for between 44 and 52 slots on the program to be filled through the proposal selection process. Attendance at the meeting will be capped at 75, due to space limitations in the hotel.

The setting in Chicago provides an opportunity to celebrate the roots of the application of Rasch measurement to developing and improving health outcome scales. The work of Ben Wright in the 1970's and 1980's, with his collaboration with Ross Lambert at the Hines VA Blind Center, the redevelopment of the PECS system with Richard Harvey at Marianjoy Hospital, and the work with Carl Granger on the FIM, is central to the development of health outcome scales that allow parametric analyses.

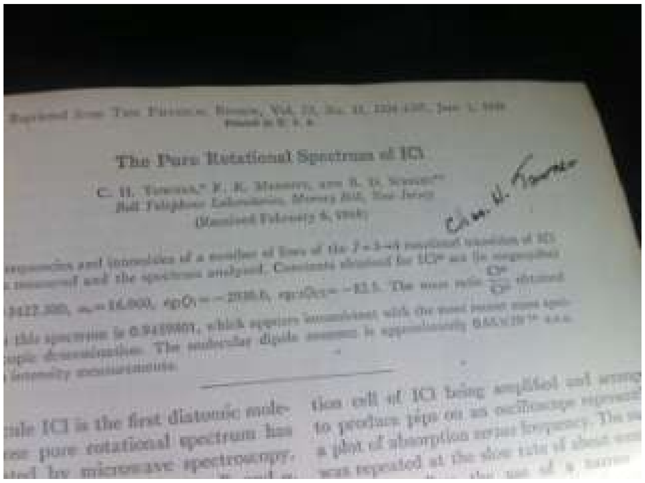
For further information, please contact us at
IOMC2015@jampress.org.

C. H. Townes: A Mentor Among Mentors

Charles Hard Townes ("C. H. Townes"), one of the world's most respected and influential physicists, was recently honored on his 99th birthday at the University of California-Berkeley. Townes is an American Nobel Prize-winning physicist and educator known for his work on the theory and application of the maser, on which he got the fundamental patent, and other work in quantum electronics connected with both maser and laser devices. Rasch measurement pioneer Ben Wright was an intern for Townes in 1947 at Bell Telephone Laboratories in NJ. The work in Townes' lab resulted in Ben Wright's first scientific publication, which he completed before entering graduate school. William Fisher was on hand for C. H. Townes birthday celebration and captured Townes' autograph on a copy of the manuscript published by Townes, Merritt, and Wright in 1948.

Pictured here: C. H. Townes' 99th birthday celebration.

Pictured here: Townes with wife Frances



Reprinted from The Physical Review, Vol. [illegible]

The Pure Rotational Spectrum of ICl

C. H. Townes,* F. R. Merritt, and B. D. Wright**
Bell Telephone Laboratories, Murray Hill, New Jersey
(Received February [illegible], 1948)

Chas. H. Townes

[illegible]

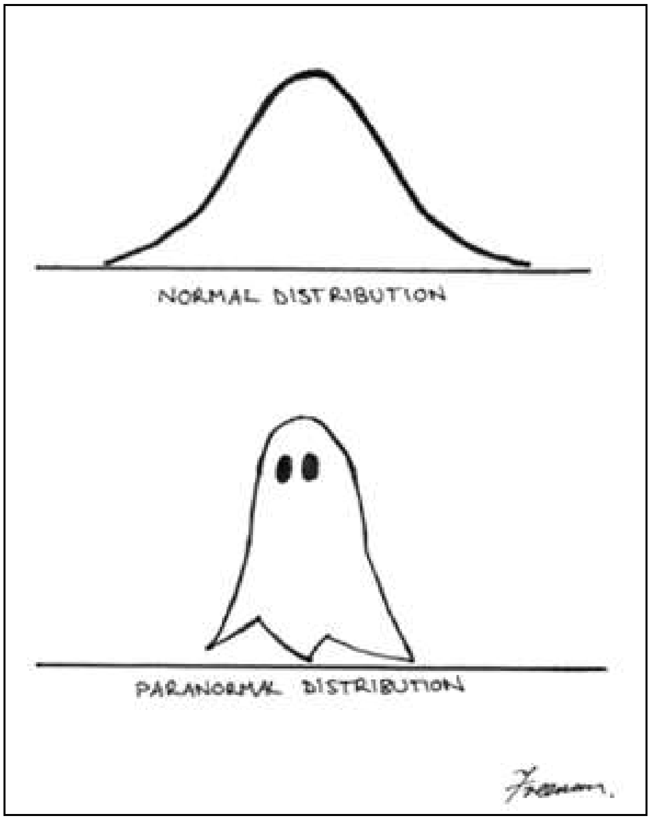
...cule ICl is the first diatomic mole-
...ose pure rotational spectrum has
...ted by microwave spectroscopy.

tion cell of ICl being amplified and ...
to produce pips on an oscilloscope ...
a plot of absorption versus frequency. The ...
was repeated at the slow rate of about ...

Pictured here: Signed manuscript of Townes work co-authored with Ben Wright

The whole story, with a video of Townes remarking on his career, is at: http://www.universityofcalifornia.edu/news/birthday-bash-celebrate-laser-inventor-charles-townes%E2%80%99-99th.

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



From: Matthew Freeman. (2006). A visual comparison of normal and paranormal distributions, *Journal of Epidemiology and Community Health, 60*, 1, 6. jech.bmj.com/content/60/1/6.full.pdf+html

New Rasch Book

Rasch Measurement in the Social Sciences and Quality of Life Research

Rasch measurement is a mathematical formulation of ideal measurement, if this were achievable. In the real world however, data are expected to deviate from ideal measurement due to measurement error, problematic test or scale design or inadequate construct definitions. Rasch analyses, apart from generating person estimates and item difficulties, investigate the discrepancies between the data observed and that which is expected by the model in an attempt to improve the instrument, the understanding of the construct and the overall measurement process. This book provides a brief introduction to Rasch measurement, explains its strengths and illustrates through examples how it can be used productively in the social sciences and quality of life research. The ultimate aim is to encourage researchers in the social sciences to appreciate the strengths of the models and to use the Rasch approach in their own work.

Panayiotis Panayides (2014). *Rasch Measurement in the Social Sciences and Quality of Life Research.* LAP LAMBERT Academic Publishing 92 pages. ISBN: 978-3659576072

Spring 2015 Workshop in Cincinnati, OH

Introduction to Rasch Measurement with Winsteps
March 26th & March 27th, 2015

A two day workshop will be led by Professor William Boone (boonewjd(at)gmail(dot)com), lead author of *Rasch Analysis in the Human Sciences*

The workshop will provide a hands-on introduction to Rasch analysis and Winsteps Rasch software. Attendees will learn how to conduct a Rasch analysis of rating scale data, multiple choice test data , and partial credit test data. At the end of the workshop attendees will-- 1) know how to apply Rasch to the design of instruments, 2) be able to create/interpret a Wright Map, 3) understand the basics of person measures/item measures/fit statistics, 4) have skills in reading Winsteps output, 5) be able to write basic text explaining Rasch (and Rasch results) for grants, talk proposals, and papers, 6) leave with ready to go Winsteps files for their own analysis, and 7) leave with easy to use (and interpret) handouts.

Mike Linacre, the author of Winsteps, has graciously agreed to provide a time limited copy of Winsteps for the workshop.

Visit raschmeasurementanalysis.com to register.

Rasch-related Coming Events

Sept. 30, 2014, Tues. Submission deadline: 6th Rasch Conference, Cape Town, South Africa, www.rasch.co.za/conference.php

Oct. 3, 2014, Fri. Submission deadline: IOMC 2015: International Outcomes Measurement Conference, Chicago IL www.jampress.org,

Oct. 8-10, 2014, Wed.-Fri. IACAT Conference: International Association of Computerized Adaptive Testing, Princeton, NJ, iacat.org/conference,

Nov. 14, 2014, Fri. In-person workshop: IX Workshop on Rasch Models in Business Administration, Tenerife, Canary Islands, Spain, www.institutos.ull.es/viewcontent/institutos/iude/46416/es

Journal of Applied Measurement
Vol. 15, No. 3, 2014

A Comparison of Stopping Rules for Computerized Adaptive Screening Measures Using the Rating Scale Model, *Audrey J. Leroux and Barbara G. Dodd*

Creating the Individual Scope of Practice (I-SOP) Scale, *Thomas O'Neill, Michael R. Peabody, Brenna E. Blackburn, and Lars E. Peterson*

Measuring Teacher Dispositions using the DAATS Battery: A Multifaceted Rasch Analysis of Rater Effect, *W. Steve Lang, Judy R. Wilkerson, Dorothy C. Rea, David Quinn, Heather L. Batchelder, Deirdre S. Englehart, and Kelly J. Jennings*

On Robustness and Power of the Likelihood-ratio Test as a Model Test of the Linear Logistic Test Model, *Christine Hohensinn, Klaus D. Kubinger, Manuel Reif*

Performance of the Likelihood Ratio Difference (G2 Diff) Test for Detecting Unidimensionality in Applications of the Multidimensional Rasch Model, *Leigh Harrell-Williams and Edward W. Wolfe*

Applying the Rasch Sampler to Identify Aberrant Responding through Person Fit Statistics under Fixed Nominal Alpha-level, *Christian Spoden, Jens Fleischer, and Detlev Leutner*

Power Analysis on the Time Effect for the Longitudinal Rasch Model, *M. L Feddag, M. Blanchin, J. B. Hardouin, and V. Sebille*

Application of Rasch Analysis to Turkish Version of ECOS-16 Questionnaire, *Pinar Gunel Karadeniz, Nural Bekiroglu, Ilker Ercan, and Lale Altan*

Erratum to Snijders's Correction of Infit and Outfit Indexes with Estimated Ability Level: An Analysis with the Rasch Model, *David Magis, Sébastien Béland, and Gilles Raîche*

Richard M. Smith, Editor, www.jampress.org