XXI IMEKO World Congress “Measurement in Research and Industry”
August 30 – September 4, 2015, Prague, Czech Republic

HISTORICIZING SCALE DEVELOPMENT: SHIFTING EPISTEMIC PRACTICES IN RASCH MEASUREMENT

Jacob Pearce[1]

[1] Australian Council for Educational Research, Assessment and Psychometric Research Division, Camberwell, Australia, jacob.pearce@acer.edu.au

Abstract – This paper traces the history of conceptual pre-conditions and shifts in epistemic practice, which laid foundations for the proliferation of Rasch analysis in scale development. New methods (1960s) saw shifts in educational measurement. With new computing techniques (1990s), Rasch measurement transformed scale development practice in social and clinical sciences. It is illuminating to uncover latent historical conditions that have contributed to Rasch measurement’s current realm of application.

Keywords: Rasch, Scale Development, History, Epistemic practice, Philosophy

1. THE APPROACH: THE SPIRIT OF HISTORICAL EPISTEMOLOGY

This paper is written in the spirit of an intellectual history of measurement, in the tradition of historical epistemology. It follows the terms set out by Hans-Jörg Rheinberger—as an investigation into “the historical conditions *under* which, and the means *with* which, things are made into objects of knowledge” [1]. From this perspective, science(s) becomes fragmented over time due to historical developments and contingencies. Regardless of any claims to truth, the norms of scientific practice and the content of scientific knowledge are products of these long and convoluted histories. The fundamental tenet is that in order to understand what science is philosophically, we have to study its history. The phrase that I have used in the title of this paper—‘historicizing scale development’—can be understood in such terms.

Rather than presenting a Whig history of certain measurement theories, I trace the conceptual and practical conditions under which Rasch analysis, as we know it today, become possible. I take cue from Arnold Davision, that “Unless we examine the historical conditions under which our concepts emerge, we are liable to find ourselves surrounded by philosophical perplexity” [2].

Following Nicolas Rasmussen’s reading of Nicholas Jardine’s *Scenes of Inquiry*, my approach is to focus on the history of scale development through the lens of the shifting statements, questions, problems, solutions, practices and presuppositions among the community of practitioners [3]. Thus, I look for shifting practices and changes in conceptual pre-conditions in the initial emergence and later consolidation of Rasch measurement as a powerful and ubiquitous tool in scale development.

In the history of science, certain theories gain traction when the inquirers are able to tackle questions that are intelligible to them. Rasmussan places great emphasis on the way scenes of inquiry are transformed. The development of new methods and techniques (new forms of scientific practice) influence the sets of questions which are deemed real, pressing, or even answerable. However, this process is not one-directional. There is a dynamic dialectic between questions and techniques—one can motivate the other. The trajectory taken by Rasch analysis in the context of scale development highlights this dialectic.

2. BRIEF HISTORY OF SCALE DEVELOPMENT

A brief history of scale development will be presented. In the history of measurement, the scaling tradition (based on Item-Response Theory (IRT)) is distinct from the more traditional test-score tradition (also known as classical test theory (CTT)). Scaling involves the calibration of individual items and the measurement of individual persons against an assumed shared latent trait.

In educational measurement there is a long history of using instruments to discriminate candidates, ranking their achievement against a construct (or multiple constructs) [4]. In high-stakes aptitude tests, for instance, ranking candidates by locating their ability on a scale is standard practice.

In many other settings, scales are used to measure a range of latent variables; be they objective measures of physical aspects, subjective symptoms, functional performance, and feelings of satisfaction in surveys. Measures of opinions, preferences and behaviours are often operationalized through measurable variables, which relate to conceptual variables. There are multifarious instruments of varying psychometrical soundness which aim to measure psychological constructs. These are often deployed in social research or as clinical tools.

In such measurement undertakings, construct and content validity, along with reliability, remain important considerations in designing and deploying instruments to locate individuals on a pre-defined scale.

Measurement instruments are comprised of a collection of items which, when collated, can be combined into a composite score. This score when reviewed against a scale is supposed to reveal levels of variables which are not readily observable by direct means [5]. For instance, psychologists postulate constructed notions of anxiety and depression to explain observed behaviour. Measuring anxiety or depression, however, requires operationalization. An example of such a measure is the Hospital Anxiety and Dpression Scale (HADS) developed in 1983 by Zigmond and Snaith [6]. This scale contains 14 polytomous items scored as two 7-item subscales (for depression and anxiety).

Against the backdrop of a proliferation of scale use in research, in 1995 Clark and Watson [7] summarized and discussed the basic issues in developing scales as objective measures. These include theoretical issues, notions of validity and reliability, the creation of an item pool, the basic principles of item writing, test construction methods, principles of analysis, sample considerations, psychometric evaluations, and the creation of sub-scales.

3. RASCH ANALYSIS

Measurement models developed by Georg Rasch in 1960 [8] have become widely used in educational measurement (achievement tests), becoming the accepted minimum standard in many international high-stakes assessment contexts. However, Rasch analysis is becoming far more common as a technique in the health sciences, clinical medicine, psychology, social science, marketing and business contexts.

The Rasch model is a powerful statistical technique for analyzing response options on collections of items, by linking candidate ability with item difficulty. Using the model, individuals can be compared without requiring the same items to be administered, as it is assumed that individuals and items can be located against a unidimensional latent trait. Rasch measures are thus generalizable across various samples and test items.

For dichotomous data, the Rasch model takes the form:

$$Pr\{X_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1+e^{\beta_n - \delta_i}} \qquad (1)$$

where β_n is the ability of person *n*, and δ_i is the difficulty of item *i*.

For polytomous data, the ‘partial credit’ Rasch model which deals with ordered categories used in ratings scales [9, 10] takes the generalized form:

$$Pr\{X_{ni} = x\} = \frac{e^{-\tau_{1i}-\tau_{2i}\cdots-\tau_{xi}+x(\beta_n-\delta_i)}}{\sum_{x'=0}^{m_i} e^{-\tau_{1i}-\tau_{2i}\cdots-\tau_{x'i}+x'(\beta_n-\delta_i)}} \qquad (2)$$

where β_n is the ability of person *n*, and δ_i is the difficulty of item *i*, and τ_{xi} ; x = 1, 2, ... m_i are thresholds which separate the latent trait of item *i* into m_i + 1 distinct ordered categories.

4. SHIFTING EPISTEMIC PRACTICES

4.1. New methods: shifts in educational measurement

Rasch analysis was originally only applied in educational measurement contexts, used for achievement testing in the USA. This paper looks at the emergence of these methods, and the shifts which occur in the following decades in the field of educational measurement.

By 1920s, the field of educational measurement was keen on the idea of objective testing. Several issues of the *Journal of Educational Measurement* were devoted to a 1921 symposium on the scientific measurement of intelligence. A great deal of work on objective measurement in educational contexts was done in the following decades. In 1951, Frederic Lord published his doctorial dissertation, which explored the application of latent trait theory to test theory [11]. However, he believed that the theory would be difficult to use in practice due to the large data sets required.

Wendt, Bos and Goy offer a detailed historical review on the applications of Rasch models in international comparative large-scale assessments [12]. Today, using large data sets is commonplace. IRT is deployed in countless large-scale educational assessment contexts. International comparative studies such as the Programme for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS) and Trends in International Mathematics and Science Study (TIMSS) all use Rasch analyses.

Since emerging, Rasch measurement was constrained by the tools available to the community of practitioners. In order to effectively use Rasch analysis, highly competent mathematicians were required in order to manually calculate the particular Rasch estimates for items and persons. Plotting this data was also a slow and painstaking exercise. Consider the example of an item characteristic curves for items that fit to the Rasch model by Chopping in Fig. 1 [13]. Prior to the development of software, competent statistical mathematicians were required to calculate the probability for each location on the ability scale and for each item.

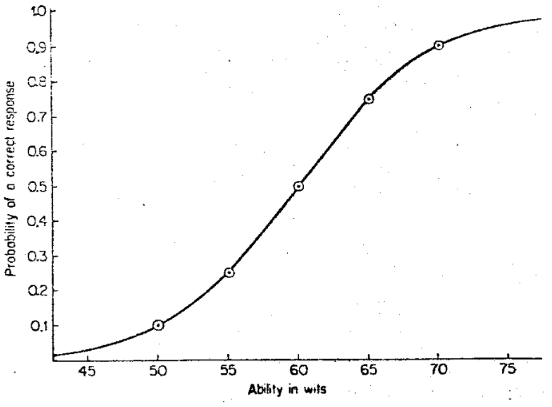

Fig. 1. Item Characteristic Curve example for the Rasch Model, by Chopping in 1983. [13].

The only groups who could efficiently deploy these methods were highly resourced assessment organizations.

4.1. New concepts: shifts in conceptual pre-conditions

The 1970s (and into the 1980s) saw a plethora of research papers and new ways of applying IRT and the Rasch model. Measurement and psychometrically aligned journals published many mathematical models which ensured that new researchers coming into the field were exposed to a variety of methods and techniques. Not only was the Rasch model becoming more accepted, it was becoming more powerful, and its domain of applicability was widening.

In 1974, Lord developed (and in 1976 released) a computer program for carrying out parameter estimation with the logistic test model. In 1977 the *Journal of Educational Measurement* ran a special issue on IRT applications. in 1979, Wright and Stone published *Best Test Design*, which described the theory underlying the Rasch model and its potential applications. In 1979, a conference on “Computerized Adaptive Testing” was held in Minnesota. The important point to make is that the field was becoming actively researched and structured, with more professionals dedicating time and energy to the pursuit of the advancement of Rasch methods.

4.2. New techniques: shifts in the domain of applicability

Statistical analyses in the social sciences took off in the 1970s. With the initial release of the SPSS software in 1968, and the first manual in 1970, researchers were given the power to manage data and undertake their own statistical analyses. As computing power increased, so too did the usability of SPSS for researchers.

However, it was not until the early 1990s that advances in computing resulted in the development of sophisticated computer software to run Rasch analysis (such as the ACER *Quest* program, later named ConQuest) [14]. These programs were developed and, to the most part, used only by major testing organisations. Software had also allowed educational researchers to develop Computer Adaptive Testing, whereby computer platforms could adapt assessment instruments based on achievement of prior items, allowing for a real-time adapted instrument based on candidate abilities.

By the mid 2000s, more ‘off the shelf’ software was available for purchase. Many of these programs, such as *Winsteps* [15] and *RUMM2020* [16] were designed to be particularly user friendly for people with little mathematical training. A basic understanding of statistics was sufficient to navigate through the interfaces of these programs.

4.3. Rasch epistemic practices today

This paper argues that these new techniques were the defining historical pre-condition for the proliferation of Rasch analysis through the social sciences, in non-achievement contexts. Rasch analysis has quickly become the standard for publishing results based on scales in the social sciences. The number of articles published based on Rasch demonstrate this transformation seen in academia. Fig 2. shows the increase in Rasch articles published from 1970-2010. Similar trends can be seen by searching on PsychInfo and Scopus.

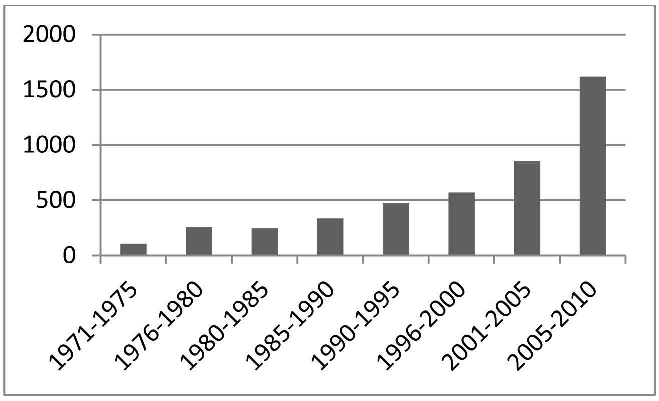

Fig. 2. Number of Rasch articles published by year (via GoogleScholar).

Importantly, the practitioners using this technique no longer need to understand the mathematical and statistical intricacies of Rasch analysis. Instead, a working knowledge of the theory, along with the ability to interpret output from these computer programs, is epistemologically sufficient.

There is a new culture of acceptance surrounding Rasch analysis. Researchers working in disparate fields—from clinical medicine to marketing—perform Rasch analysis in the process of scale development. Terms such as ‘logits’, ‘thresholds’ and ‘differential item functioning’ are used and understood by a vast array of researchers working in disparate fields.

For example, Pallant and Tennant analyzed the HADS scale mentioned [6] earlier in a 2007 paper [17]. Using Rasch analysis on the RUMM2020 platform [16], Pallant and Tennant build a case for the validity of the HADS scale by examining person separation, disordering of thresholds, and differential item functioning. Through their discussion, they identity several items which demonstrate misfit, and argue that the issue of dimensionality of the HADS remains problematic.

A more illustrative example of the way Rasch is epistemologically constituted in these non-achievement contexts (in this instance, in the field of nursing) can be found in a 2009 paper by Hagquist, Bruce and Gustavsson [18]. Using Rasch to evaluate the nursing self-efficacy (NSE) scale, the researchers used graphical explanations to show that items operated in the correct direction and that there were minor signs of multidimensionality. These authors also utilized the RUMM2020 software. Fig. 3 shows the researchers’ person-item threshold distribution (9 items with 11 categories). Fig. 4 shows an ordered set of thresholds for a particular item.

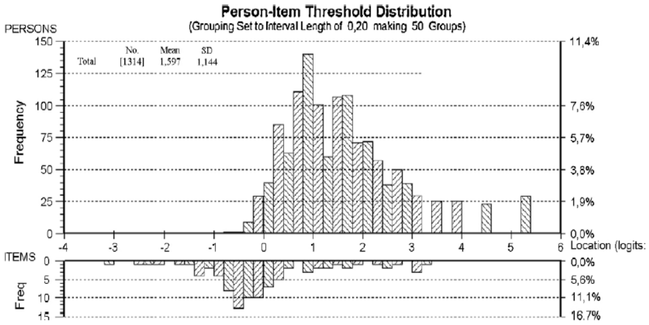

Fig.3. Example of Person-item threshold distribution, by Hagquist, Bruce and Gustavsson [X].

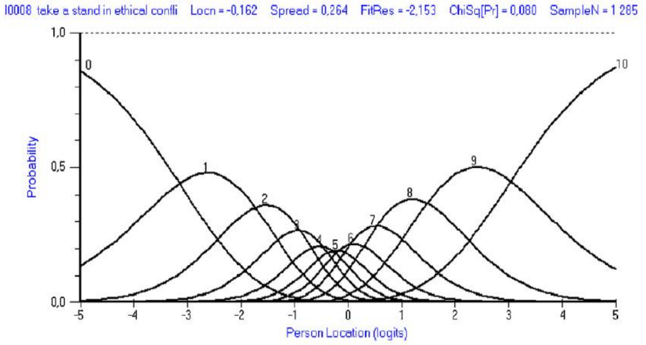

Fig. 4. Example of a category probability curve distribution, by Hagquist, Bruce and Gustavsson [X].

Short courses and training programs on Rasch measurement, and particularly how to use the software packages, are offered frequently around the world [see www.rasch.org]. It seems that the innovations in techniques (computer software) is now motivating the sets of questions that are deemed pressing in many academic pursuits. The tools of the practitioners are intricately connected to their goals.

Rasch analysis in non-achievement contexts is becoming more and more prevalent as an accepted and fruitful form of epistemic practice. There is a dynamic dialectic between the questions asked in academic inquiry and the techniques motivating such questions. Rasch measurement is no longer reserved for the realm of educational testing.

5. CONCLUSION

When considering the current state of research around scale development, and the proliferation of Rasch analysis as an accepted technique, it is illuminating to uncover the latent historical conditions which have contributed to its current realm of application. This analysis makes no claims whatsoever as to the value of Rasch over other techniques. Rather, the focus has been on tracing the shifts in the conceptual pre-conditions, and in epistemic practice, which have influenced its trajectory.

ACKNOWLEDGMENTS

I would like to thank the Australian Council for Educational Research (ACER) for supporting this research, and Kristian Camilleri for direction on many of the aspects of this historiographical approach.

REFERENCES

[1] H-J. Rheinberger, *On Historicizing Epistemology: An Essay*, trans. David Fernbach, Stanford University Press, Stanford, 2010, p. 2.

[2] A. Davidson, “Styles of Reasoning, Conceptual History, and the Emergence of Psychiatry”, In *The Disunity of Science: Boundaries, Contexts, and Power*, P Galison & D Stump (eds.), Stanford University Press, Stanford, 1996. p. 86.

[3] N. Rasmussen, *Picture Control: The Electron Microscope and the Transformation of Biology in America, 1940-1960*, Stanford University Press, Stanford, 1997, pp. 10-11.

[4] D. L. McArthur, “Educational Testing and Measurement: A Brief History”, *CSE Report number 216*, Centre for the Study of Evaluation University of California, Los Angeles, 1983.

[5] R. F. DeVellis, *Scale Development: Theory and Applications* (3rd ed.), Thousand Oaks, Sage.

[6] A. S. Zigmond, R. P. Snaith, “The Hospital Anxiety and Depression Scale”, *Acta Psychiatrica Scandinavia*, vol. 67, pp. 361-370, 1983.

[7] L. A. Clark, D. Watson, “Constructing Validity: Basic issues in objective scale development”, *Psychological Assessment*, vol. 7, nº.3, pp. 309-319, September 1995.

[8] G. Rasch, *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen, 1960.

[9] G. N. Masters, “A Rasch model for partial credit scoring”, *Psychometrika*, Vol. 47, pp. 149-174, 1982.

[10] D. Andrich, “A rating formulation for ordered response categories”, *Psychometrika*, Vol. 43, pp. 561-73, 1978.

[11] F. M. Lord, “A Theory of Test Scores and Their Relation to the Train Measured”, *ETS Research Bulletin*,Vol. 1951, pp. i-126.

[12] H. Wendt, W. Bos, M. Goy “On applications of Rasch models in international comparative large-scale assessments: a historical review”, *Educational Research and Evaluation*,Vol. 17, nº.6, pp. 419-446, December 2011.

[13] B. Choppin, “The Rasch Model for Item Analysis”, *CSE Report number 219*, Centre for the Study of Evaluation University of California, Los Angeles, 1983.

[14] M. L. Wu, R. J. Adams, M. R. Wilson, S. Haldane, *ConQuest* (Version 2.0), Computer Software. Camberwell. Australia. ACER.

[15] J. M. Linacre, *Winsteps* (Version 3.68), Computer Software. Beaverton, Oregon: Winsteps.com.

[16] D. Andrich, B. Sheridan, G. Luo, *RUMM2020: A Windows Interactive Program for Analysing Data with Rasch Unidimensional Models for Measurement.* Computer Software. RUMM Laboratory, Perth, Western Australia.

[17] J. F. Pallant, A. Tennant, “An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS)”, *British Journal of Clinical Psychology*, vol. 46, pp. 1-18, 2007.

[18] C. Hagquist, M. Bruce, J. P. Gustavsson, “Using the Rasch model in nursing research: An introduction and illustrative example”, *International Journal of Nursing Studies*, vol. 46, pp. 380-393, 2009.