Journal of Physics: Conference Series **459** (2013) 012057 doi:10.1088/1742-6596/459/1/012057

Quantifying Human Response: Linking metrological and psychometric characterisations of Man as a Measurement Instrument

L R Pendrill[1]
SP Technical Research Institute of Sweden, Measurement Technology, Box 857, SE-50115 Borås (SE), phone: +46 767 88 54 44,
mailto:leslie.pendrill@sp.se

William P. Fisher, Jr.,
BEAR Center, Graduate School of Education, University of California, Berkeley, CA (USA) & Principal, LivingCapitalMetrics Consulting
mailto:william@livingcapitalmetrics.com

Abstract. A better understanding of how to characterise human response is essential to improved person-centred care and other situations where human factors are crucial. Challenges to introducing classical metrological concepts such as measurement uncertainty and traceability when characterising Man as a Measurement Instrument include the failure of many statistical tools when applied to ordinal measurement scales and a lack of metrological references in, for instance, healthcare. The present work attempts to link metrological and psychometric (Rasch) characterisation of Man as a Measurement Instrument in a study of elementary tasks, such as counting dots, where one knows independently the expected value because the measurement object (collection of dots) is prepared in advance. The analysis is compared and contrasted with recent approaches to this problem by others, for instance using signal error fidelity.

1. Introduction

Advances in technology and information processing are allowing the enhancement of various human functions, be it machine learning [1] to assist in mining the ever increasing amounts of information available in society or aiding a disabled, ill or elderly person to cope better with care as well as everyday tasks [2]. To ensure the quality and promote innovation of these diverse technological services to Mankind, better understanding is needed of how a human being perceives quantities such as information content and service quality, not merely in technical terms but in even more 'human' quantities such as comfort, pleasure, beauty [3] and so on.

When characterising the human measuring instrument [4], be it with the five senses or even taking account of the full physiological, mental, cognitive and behavioural richness of human perception, formulation of the commensurate metrological concepts as established in traditional engineering is as yet in its infancy for qualitative [5,6] and subjective measurement [7,8].

It is still common, for instance, to find incorrect analysis of the scores obtained with questionnaires and similar instruments often used to measure human response, where the challenge [9] is that several of the most common tools of statistics – such as calculation of a mean or standard deviation - cannot be used to characterise the location and dispersion of qualitative measurements on ordinal scales [10] typical in such measurements.

Additionally, some method of metrological traceability for measurements based on ordinal observations is needed when the ability of a person to perform a task of classifying an entity of given reference level is to be determined. In some areas of rating person ability in this way – for instance, in the medical sector – it may be possible to promote the same kind of quality-assurance as is routine in

[1] To whom any correspondence should be addressed

more quantitative measurement, that is, in terms of metrological traceability and with declared levels of measurement quality as uncertainty. Patient health, for example, is increasingly rated by health clinics on ordinal scales linearized via a log-odds transformation, and appropriate treatment is decided on by the doctor by comparing the actual ratings with corresponding pattern of 'typical' health ratings for similar patients from earlier studies. Obviously, the comparability of such ratings has to be reliable to a sufficient degree of accuracy if the patient is to be treated appropriately. Given that such accuracy can be be regularly obtained, the observational framework could be redesigned to omit the observed ordinal scores and to incorporate a metrologically traceable unit reported alongside an uncertainty term [7,11,12].

2. Rationale for a potential expansion of the metrological framework

As an example of earlier work, Fisher [7] assesses the feasibility of developing and deploying a universal metric through studies of the quantitative stability of a physical disability construct across a range of measurement instruments and samples. One seeks estimates of the person and instrument measures that are "not affected by the abilities or attitudes of the particular persons measured, or by the difficulties of the particular survey or test items used to measure". Fisher draws parallels with traditional metrological concepts in engineering and physics with corresponding application of these concepts in psychometrics. He argues that this reflects an on-going transition in health care from "local economies of disease-crisis management to regional, national, and international economies of population-based, preventive health management". He points out that the development and commensurate demands for scale-free measurement providing accountability and comparability in healthcare are being enabled by readily accessible computational and communicational resources through internetworked personal computers.

The measurement system concept from the perspective of psychometrics where a human being is, in one way or another, a critical element of the measurement system has recently been considered [13]. The operation of any measurement system includes three steps, recalled by Mari [14]: Calibration; Data acquisition; Data presentation. Examples of recent studies of psychophysical scaling [15] where perceptual intensity is related to stimulus intensity include an attempted explanation of the Weber-Fechner law in terms of signal error fidelity [16].

The present work offers an approach which attempts to link a traditional engineering metrological characterisation of measurement systems to commonly used techniques in behavioural and psychometric studies, where the ability of a human being to perform tasks is characterised over a range of levels of challenge employing the well-established approaches in psychological measurement developed by Rasch and his students [17-19]. Studies of situations are made where a human acts as an instrument for elementary tasks – counting dots [20-21] – where a key observation will be that human performance, described in terms of person ability and level of challenge for different tasks, can in fact be related to invariant measures of location and dispersion on an interval scale when Man acts 'metrologically' as a Measurement Instrument [22, 23]. In such elementary cases, one knows independently the expected value since the measurement object (collection of dots) is prepared in advance, thus allowing calibration and assessment of measurement uncertainty which can then be compared with the ability of the human to perform the task.

3. Grounding measuring in counting

The Mundurucu are an Amazonian indigenous people with little access to Western-style educational resources [21]. Though they lack such technologies as rulers, graphs, weight scales, and numbers greater than five, the Mundurucu have sophisticated, though approximative and nonverbal, conceptions of space and number. Research investigating the conceptual link between number and spatially distributed dots suggests that the Mundurucu intuitively employ a logarithmic transformation

of impressions of varying amounts, meaning that "larger numbers require a proportional larger difference in order to remain equally discriminable" (Weber's law) [21].

This difference between intuitive sense impressions and measurement results was first noted by Fechner for the human senses, and has been documented in recent neurological research [24]. This "Gaussian tuning curve" functions like an internal biological slide rule, compactly collapsing several orders of magnitude into a portable system of relatively constant imprecision [21]. In showing the logarithmic proportions between sensations and stimuli, Fechner provided a basis for Thurstone to shift the focus of human measurement from psychophysical to psychological, economic, and social phenomena [25]. Rasch later independently developed ideas along the same lines [26].

Further, real things, from the sides of triangles to rocks to performances, are not identical, and so cannot conform completely to the parameters in a scientific law or measurement model. The measurement of counting ability, like the measurement of weight or length, requires the definition of an invariant unit amount that will not correspond directly with empirically observed counting numbers or things counted [22]. Again, the natural logarithm plays an important role in providing experimentally verifiable criteria for demonstrating linear proportional relationships among parameters in a model.

4. Human performance related to invariant measures on interval scale

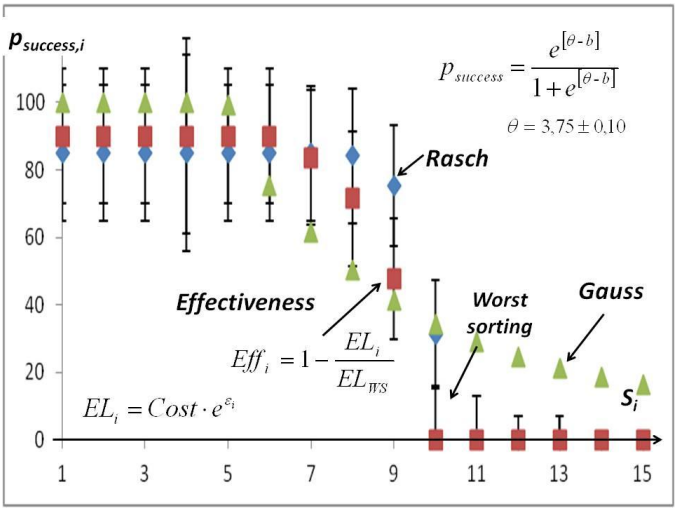

Fig. 1. Comparison of effectiveness (4), Rasch (1) and Gaussian (3) modelling of counting range 1 – 10 for the Mundurucu (w = 0,17)

Two alternative metrological approaches to deriving how the probability of successful counting depends on measures of location and dispersion of perceptive judgments over a range of stimulus values will be compared here in terms (resolution and bias) as one would use when characterising a measurement instrument (in this case, the human counter) (Fig. 1).

4.1 Probabilistic approaches to subjective differences

The Rasch approach [17-19] models individual probabilities of success in counting as the difference between person ability θ and level of challenge b::

$$p_{success} = \frac{e^{[\theta - b]}}{1 + e^{[\theta - b]}}. \qquad (1)$$

No distributional assumptions are required, counts of observed success in the assigned task are sufficient statistics [28-29], and individual ability measures are expressed as the average difference

between the person's location and each task's location on the log-odds continuum. For dichotomous observations, the model can be written to include an error term as follows:

$$X_{n,i} = P_{n,i} \pm \sqrt{P_{n,i} \cdot \left(1 - P_{n,i}\right)} \qquad (2)$$

Here $X_{n,i}$ is the scored response of person n to item i, and $P_{n,i}$ is the probability of success given in (1). The binomial error distributions for dichotomous scores approximate Gaussian distributions when accumulated across all the observations, as they are in the estimation process [30].

The second approach expresses the ability to distinguish ('resolve') two counts with stimulus values, $S_i = n_1$ and n_2, in terms of the distance, $D(P,Q)$, between two distributions, $P = N(\mu_1, \sigma_1^2)$ and $Q = N(\mu_2, \sigma_2^2)$, each centred at the respective mean count and the standard uncertainty in the difference is given[2] by $u = w\sqrt{\frac{(n_1^2 + n_2^2)}{12}}$. This approach has been used in studies of the acuity of "number sense" where values of the Weber constant, w, range from 1 to 0.1 for the acuity of "number sense" for people of different counting abilities (from infants to educated adults) [27]. The counting efficiency, $p_{success}$, is then expressed in terms of an error rate defined as the area under the overlap in the two count distributions. Assuming Gaussian distributions:

$$p_{success} = 1 - erfc\left[\frac{|n_1 - n_2|}{\sqrt{2} \cdot u}\right] \qquad (3)$$

4.2 Classification effectiveness in terms of counting bias on ordinal scale

Perceived values, P_i, such as counts, normally can only be referred to an ordinal scale, since exact perceptive distances between different points are not known. An approach to overcoming the traditional limitations of statistical measures of location and dispersion on such scales is by the inclusion of a cost function as a distance metric. Following this approach, a measure of the effectiveness, *Eff*, of sorting on an ordinal scale has been proposed [31-32] as: $Eff_i = 1 - \frac{EL_i}{EL_{WS}}$ (4)

where EL_i is the expected loss associated with incorrect classification at level i and EL_{WS} is the worst-case loss. In this second approach, the expected loss, and thereby the classification efficiency, associated with incorrect counts are expressed as functions of the 'bias' $\varepsilon_i = S_i - P_i$ for each count (rather than using the difference in adjacent counts used in the first approach above).

5. Results

Despite the differences in these approaches, reasonably good agreement (Fig. 1) is obtained between estimates of the probability of success (at counting in the present example) if both the expected loss in expression (4) and the level of challenge, b_i, for each count in expression (1) are modelled with exponential functions of, respectively, the count bias, ε_i, and stimulus value, S_i. In both approaches, the worst-case loss (most challenging task) is taken ('anchored') to be the count i = 10. The Gaussian psychophysical modelling, for its part, does not rely on these assumptions, so the discrepancies between the three curves of Figure 1 could indicate an increase slightly milder than exponential in the level of challenge across the range of counts from 1 to 10.

[2] Assuming a uniform (rectangular) distribution of the difference in means

We will also compare our results with other recent approaches [16], where the logarithmic relationship in the Weber–Fechner law [33-34] is couched in terms of perception characterised in information theory by 'quantization' followed by efficient coding where fundamental limits of compressibility are expressed in terms of entropy [35]. Fechner's law can be derived by cumulating a measure of dissimilarity of a stimulus from its "immediate" neighbour value where, according to [36], the distance, *D(P,Q)*, can be expressed more generally in terms of the symmetric Kullback-Leibler divergence information measure.

6. Discussion

Historically, psychometrics extended the Weber-Fechner law from the constructs of psychophysics to the measurement of psychological and social constructs observable only in ordinal terms. Linear comparisons effected via application of the log-odds transformation to observed scores unexpectedly offered several additional highly practical advantages. Foremost among these were new capacities for qualitative meaningfulness and for systematically accounting for missing data.

The first of these followed from the reproducible and invariant hierarchies in the questions asked [37]. Numbers ceased being mere digits but were instead strongly associated with substantive variation in the measured construct, such that the increasing difficulty of the questions asked could be understood in terms of developmental sequences. Comparisons of individual measures were now not expressed only in numeric terms, but in terms of performances relative to invariantly ordered learning progressions [38]. Additional new efficiencies are being realized within psychometrics as empirical estimation gives way to strongly predictive theories of the measured constructs [39-40].

The second new practical advantage is the equating of measures made from different sets of items from tests, surveys, and assessments measuring the same thing [39]. Just as new capacities for qualitative meaningfulness were realized from the repeated emergence of the same item hierarchies across samples, so, too, did identifiable commonalities in the item hierarchies emerge across item sets. Data-based equating methods are giving rise to theory-based methods as understanding of the cognitive and performative bases of these hierarchies improves [40]. Enhanced interpretability and the equating of different instruments bode well for future metrological standards and traceability.

7. Conclusion

In well-designed experimental contexts, counts of observed behaviours or decisions may provide evidence that invariant unit amounts exist, and this evidence may be useful in calibrating instrumentation and in devising a theory of the construct. The value of theory in this context is realized to the extent that the construct can effectively be synthesized in the laboratory. Given a theory capable of predicting 90% or more of the observed variance in the counts providing empirical evidence for experimental evaluation, a provisionally satisfactory understanding of the construct can be claimed. Following Feynman's point that, "What I cannot create, I do not understand" [41], judicious use of a combination of empirical, instrumental, and theoretical techniques may ultimately lead to new horizons in psychology and the social sciences [42]. Metrological traceability to reference standard units will be of central importance in this process.

5. References

[1] Zhang W, Yang Y, Wang Q, Shu F 2011 *Proceedings of the 23rd International Conference on Software Engineering & Knowledge Engineering* (**SEKE'2011**), Miami Beach, USA, July 7-9 2011 , http://www.ksi.edu/seke/Proceedings/seke11/51_Wen_Zhang.pdf

[2] Farbrot A, Abbas S, Nihlstrand A, Dagman J, Emardson R, Kanerva S, Pendrill L R 2013 *The Simon Foundation for Continence's Innovating for Continence Conference Series*, Chicago (US), April 2013

[3] Schmidhuber J 2009 *J SICE*, **48**(1):21-32, PDF
[4] Berglund B, Rossi G B, Townsend J, Pendrill L R 2011 *Theory and methods of measurements with persons* (New York: Taylor & Francis)
[5] Mari L 1999 *Measurement* **25** 183 – 92
[6] Pendrill L R 2011 *AMCTM 2011 International Conference on Advanced Mathematical and Computational Tools in Metrology and Testing*, Göteborg June 20-22 2011, http://www.sp.se/AMCTM2011
[7] Fisher W P Jr 1997 *J Outcome Meas* **1**(2) 87-113
[8] Pendrill L R, Berglund B *et al.* 2010 *NCSLi Measure* **5** 42-54
[9] Svensson E 2001 *J Rehab Med* **33** 47-48
[10] Stevens S S 1946 *Science* **103** 2684 677-680
[11] Wright B D 1997 *Physical Med & Rehabil State of the Art Reviews* **11** 261-288
[12] Davis A M, Perruccio A V, Canizares M et al. 2008 *Osteoarthritis Cartilage* **16** 551-559
[13] Wilson M 2011 *Joint International IMEKO TC1 + TC7 + TC13 Symposium*, August 31st − September 2nd, Jena, Germany, urn:nbn:de:gbv:ilm1-2011imeko-005:8
[14] Mari L 2000 *Measurement* **27** 71-84
[15] Berglund B 2011 Chapter 2 in ref [4]
[16] Sun J Z, Wang G I, Goyal V K, Varshney L R 2012 *J Math Psychol* http://dx.doi.org/10.1016/j.jmp.2012.08.002
[17] Wright B D 1999 pp 65-104 in *The new rules of measurement: What every educator and psychologist should know*, ed S E Embretson and S L Hershberger (Hillsdale, New Jersey: Lawrence Erlbaum Associates) pp. 65-104
[18] Rasch G 1961 pp. 321–334 in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, IV. (Berkeley, California: University of California Press)
[19] Andrich D 2004 *Medical Care* **42** 1-16
[20] Burro R, Sartori R, Vidotto G 2011 *Qual Quant* **45** 43-58, http://dx.doi.org//10.1007/s11135-009-9282-3
[21] Dehaene S, Izard V, Spelke E, Pica P 2008 *Science*, **320** 1217 – 1220
[22] Cooper G, Fisher W P 2011 Session on Fundamentals of measurement science. International Measurement Confederation (IMEKO). Jena, Germany, August 31 to September 2. http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24494/ilm1-2011imeko-019.pdf
[23] Wright B D 1989 *Rasch Measurement Transactions* **3** 62 http://www.rasch.org/rmt/rmt32e.htm
[24] Nieder A, Miller E K 2004 *Proc. Natl. Acad. Sci. U.S.A.* **101** 7457
[25] Masin S, Zudini C V, Antonelli M 2009 *J History Behavioral Sciences* **45** 56-65
[26] Andrich D 1978 *Ap Psychol Meas* **2** 449-460
[27] Halberda J, Feigenson L 2008 *Dev Psychol* **44** 1457-1465
[28] Andersen E B 1977 *Psychometrika* **42** 69-81
[29] Fischer G H 1981 *Psychometrika* **46** 59-77
[30] Linacre JM 2010 *Rasch Measurement Transactions* **23** 1238
[31] Bashkansky E, Dror S, Ravid R, Grabov P 2007 *Quality Engineering* **19** 235-244
[32] Bashkansky E, Gadrich T 2010 *Accred Qual Assur* **15** 331-336 DOI 10.1007/s00769-009-0620-x
[33] Dehaene S. 2003 *Trends in Cognitive Sciences* **7** 145–147
[34] Nieder A, Miller E K 2003 *Neuron* **37** 149–157
[35] Cover T M, Thomas J A 1991 *Elements of information theory* New York, John Wiley & Sons
[36] Dzhafarov E N 2011 Chapter 9 in ref [4]
[37] Wright B D, Masters G N 1982 *Rating scale analysis* Chicago, MESA Press
[38] Wilson M R 2009 *J Res Science Teaching* **46** 716-730
[39] Wright B D, Stone M H 1979 *Best test design* Chicago, MESA Press
[40] Stenner A J, Stone M H 2003 *Rasch Measurement Transactions* **17** 929-930
[41] Hawking S W 2001 *The universe in a nutshell* New York, Bantam Books
[42] Stenner A J, Fisher W P, Stone M, Burdick D 2013 Causal Rasch models *Frontiers in Psychology* (in review)