16th International Congress of Metrology, 14001 (2013)
DOI: 10.1051/metrology/201314001

Discrete ordinal & interval scaling and psychometrics

Leslie PENDRILL[1,a]

[1]SP Technical Research Institute of Sweden, Measurement Technology, Box 857, S-501 15 Borås, Sweden

Abstract. Reliable decisions about conformity of product in many 'person-centred' domains of contemporary interest based on measurement require the comparability and risk assessment associated with metrological quality assurance – in terms of traceability and measurement uncertainty. But in qualitative and subjective measurement, these metrological concepts are in their infancy. Challenges include the failure of many common tools of statistics on the ordinal scales typical of 'human' measurement. Measurable constructs for perceptual characteristics such as task challenge or the quality of a service, matched by the corresponding human ability and satisfaction for these items, respectively, have also to be formulated, as is done increasingly with psychometric approaches, such as Rasch and Prospect Theory. To aid basic understanding in this relatively new field of metrology, some of the most elementary perceptions, e.g. the simple observation of clouds of dots, will be analysed in the present work where metrics for the location and dispersion of data on discrete scales, including impact, will be linked via instrumental bias and resolution, to psychometric measures of the ability to perform a series of tasks of increasing difficulty.

1 Quality-assured, 'person-centred' measurement

Demand for quality-assured measurement analysis is increasing in many 'person-centred' domains of contemporary interest − customer satisfaction, service and product quality & classification in healthcare [1], teaching, software evaluation, etc. Reliable decisions about conformity of product of any kind based on measurement require the comparability and risk assessment associated with metrological quality assurance – in terms of traceability and measurement uncertainty – but in qualitative and subjective measurement formulation and application of these concepts are in their infancy [2] compared with more traditional quantitative measurements.

2 Challenges when analysing ordinal data

Common tools of statistics, which work readily for quantitative interval and ratio scales, are unfortunately not applicable [3] to the ordinal data typical of 'human' measurement [4].

2.1 Failure of statistics on ordinal scales

Familiar statistical expressions derived for quantitative interval scales cannot of course be applied to ordinal scales since distances between different pairs of categories are not known exactly. Sum scores of multi-item assessments; the mean value; standard deviation and calculation of differences for description of change in score do not have an interpretable meaning on ordinal scales [3].

2.2 Psychometrics

As is appreciated in an increasing range of applications, traditional raw scores from e.g. questionnaires assigned to an ordinal scale can be better analysed with psychometric methods.

In such methods, measurable constructs are formulated for perceptual characteristics such as the difficulty of an exam [5] or the quality of a service [6], matched by the corresponding human ability and satisfaction for these items, respectively, as perceived by a human being and as described in an increasing number of domains with psychometric approaches, such as Item Response Theory and Rasch scaling [7].

3 Elementary observations of shape

The present work presents an approach to characterising observational data and measurement systems where a human being acts, sometimes qualitatively, as a measurement instrument in some of the most elementary perceptions, e.g. the simple observation of clouds of dots [8].

[a] Corresponding author: leslie.pendrill@sp.se

Knoblauch and Maloney [8] investigated how observers could perceive differences in stimuli using the psychophysical procedure of difference scaling [9, 10] to gauging perceptual difference judgment in the ellipticity, *r*, of clouds of dots produced with known correlation coefficient, *R*,.

Such a data set is of interest since the stimulus values are conceptually simple and known *a priori*, thus presenting the same advantages as other elementary data sets, such as numbers of dots, in making psychophysical and psychometric studies [11, 12]. Perception of the ellipticity [8] and count [13] of clouds of dots are two cases which appear to follow the psychophysical Weber-Fechner logarithmic law, relating perceptual response *P* (=*O*, output*)* to stimulus *S* (= *I*, input) of the measurement instrument through the relation:

$$O = P \propto \ln\left(\frac{S}{s_0}\right) = \ln\left(\frac{I}{s_0}\right) \qquad (1)$$

where s_0 is the threshold below which a stimulus is not perceived, in a way analogous to perception of the intensity of stimuli to the five human senses. Fechner's 'just noticeable difference'[14] and s_0 would be related to the resolution of the measurement instrument.

It should be possible to link metrics for the location and dispersion of data on discrete scales with the ability to perform a series of tasks of increasing difficulty [12], such as traditionally studied in psychometrics.

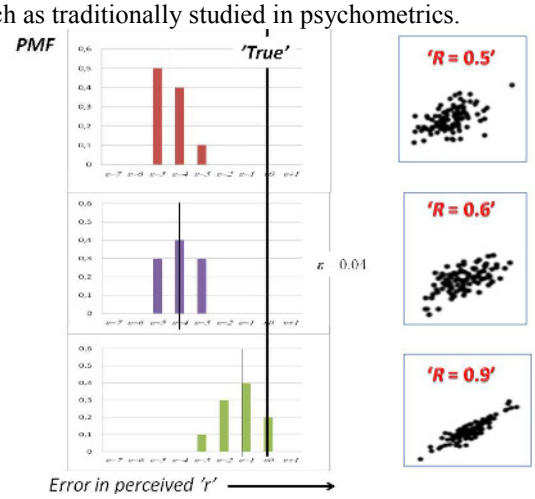

Figure 1. Scatter of perceived correlation '*r*' for three dot clouds showing increasing bias with increasing challenge.

In figure 1 is shown the scatter of perceived dot cloud correlation as a probability mass function (PMF) commonly employed for discrete data (in steps of 0.04 in *r*), where the probability, q_j, of observing a particular correlation $j = r'$ when the 'true' correlation is $i = R$, is plotted against the range of $j = 1,\ldots,K$ levels/categories.

3.1. Interval scale analysis

Assuming initially measurements on an interval or ratio scale, traditional metrological measures are the expectation and variance of a discrete quantity *X*, obtained from analysis of the probability mass function (PMF) g_{test} associated with the experimental results (such as shown in Figure 1), are respectively:

$$E(\{X\}\cdot[I]) = \sum_{j=1}^{K} q_j \cdot (\{r_j\}\cdot[I]) \qquad (2)$$

$$V(\{X\}\cdot[I]) = \frac{\sum_{j=1}^{K}\left[q_j \cdot \{r_j\}\cdot[I] - E(\{X\}\cdot[I])\right]^2}{K-1} \qquad (3)$$

q_j – probability of assigning a measurement value, r_j, to level *j*, and *K* is the number of discrete levels/categories. The unit of discrete perception of ellipticity is denoted [*I*].

As is evident from the data of Figure 1, the greater the challenge to perceiving ellipticity, the larger is the bias $E(\{X\}\cdot[I]) - R\cdot[I]$ in the perceived shape measurement but there is no obvious theory linking level of challenge to bias. A second estimate of uncertainty in perception on the interval scale complements the dispersion measure of equation (3) with the expression:

$$V(\{X\}\cdot[I]) + \frac{\left|E(\{X\}\cdot[I]) - R\cdot[I]\right|}{12} \qquad (4)$$

that is, allowing for uncorrected bias by associating it with a uniform (rectangular) distribution in a GUM Type-B approach.

3.2. Ordinal scale analysis with impact metric

The analysis in the previous section assumes an interval horizontal scale of Figure 1 along which distances are fully quantitative. The perceptual measurements at hand are however arguably more appropriately dealt with on qualitative ordinal scales where expressions of classical statistics such as in eq. 2, 3 and 4 simply do not work.

A measure of the **effectiveness of sorting**, i.e., how dispersed each classification is, has been proposed by Bashkansky *et al.* [15]:

$$Eff = 1 - \frac{EL}{EL_{WS}} \qquad (5)$$

where *EL* is the expected loss associated with incorrect classification, and EL_{WS} is the worst-case loss. In this approach an impact weighting is introduced on the ordinal scale with the expected loss calculated as $EL = \sum_i \sum_j C_{i,j} \cdot p_i \cdot P_{i,j}$ where $P_{i=R,j}$ is the (conditional) probability of assigning a measurement value, r_j, to level *j* when true level is $i = R$; the sum is over the classification levels; and weighting with a cost function *C*. The latter takes the role of a distance metric on the ordinal scale by monotonically increasing the further away (assuming proper ordering) the category *j* assignment is from the 'true' category *i*. The expected

loss reduces to $EL = \sum_j C_j \cdot q_j$ in the current case since we know *ab initio* the true value so that $q_j = p_{i=R} \cdot P_{i=R,j}$. The worst-case expected loss is taken as the loss associated with a completely random classification, i.e. $EL_{WS} = \sum_j C_j \cdot \frac{1}{K}$. Each specific case will have its own particular impact *C*. In the present case we make a common choice of $C_j = Cost \cdot \varepsilon_j^{\,2}$, i.e. a Taguchi [16] loss model where increasing bias from the true value is penalised quadratically. Other conceivable cost models include for example a balance of customer satisfaction and manufacturing costs as well as Prospect Theory[17]. Ultimately, this choice together with the choice of reference, worst-case disorder in determining the classification effectiveness will be a function of the level of challenge of the task.

Subjecting a human observer to a series of tasks over a range of difficulty is expected to result in a decrease in the 'faithfulness' (or effectiveness) with which the system (human 'instrument') converts ('interprets') the in-coming measurement information for each increase in level of challenge of the task at hand.

3.3. Rasch analysis

How well do discrete metrics on ordinal scales correspond with the parameters of psychometrics for characterising the metrological 'performance' of a human being? The main idea is to convert a deviation in measurement value on an ordinal scale to some aggregate metric of performance since it seems reasonable that the further and more spread the measurement values are from the true value, the lower the ability of Man as a Measurement Instrument. The variation in ability, θ, of the human, of discrimination, α, is to be gauged over a range of increasingly demanding tasks of challenge β (e.g. a decreasing ellipticity of cloud dots in the present case) according to the Rasch model [7]:

$$p_{success} = \frac{e^{a\cdot[\theta - \beta]}}{1 + e^{a\cdot[\theta - \beta]}} \qquad (6)$$

In evaluating the Rasch formula (6), the level of challenge, β_i, for the i^{th} level is assumed to be proportional to the square of the perceived stimulus, i.e. the same function taken as the cost function C_i for each level [§3.2]. Normalising over the range of challenge from the easiest (i = 10) to the most difficult task (i = 1), the level of challenge is then given by $\beta_i = \frac{R_i^2 - R_{10}^2}{R_1^2 - R_{10}^2}$.

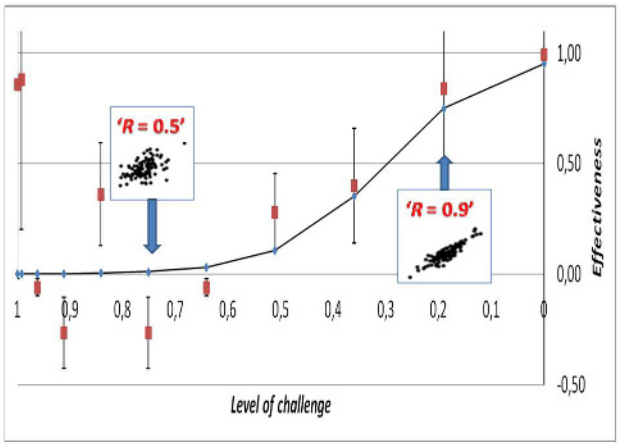

Figure 2. Variation in response of ellipticity perception with a Rasch model (eq. 6) compared with classification effectiveness (eq. 5) of the probability of success over a range of actual stimulus values from $R = 0$ to 1 and item challenge calibration $\alpha(discrimination) = 10; \theta(Ability) = 0{,}3$

We expect that the classification efficiency, *Eff,* of ordinal rating, eq. (5), to correspond closely to the probability of success according to a Rasch analysis (eq. (6), as demonstrated in Figure 2. This expectation is not surprising, noting the similarity of eq (5) with Spearman's coefficient of rank correlation mentioned by Bashkansky.

4 Conclusion

The approach presented here not only advances the treatment of measurement uncertainty for discrete measures (typically characterised by a probability mass function, PMF), for both interval and ordinal measures, but also provides new links between these discrete metrics for qualitative observations and the parameters of psychometrics.

An approach to characterising measurement systems which include a human being – either acting as a measurement instrument or as the object of measurement – is applied in studies of elementary situations where the stimulus values are conceptually simple and known *a priori*.

The location and dispersion of human perception data on an ordinal scale is interpreted in terms of instrument characteristics:

(i) bias with a loss function providing a metric

(ii) resolution.

Each of these can be related with the performance of a human as measured in a Measuring Man situation, in terms of the ability to perform specific tasks, such as traditionally studied in psychometrics.

Acknowledgments

Thanks are due to my colleagues at SP and internationally, particularly:

- Dr Bashkansky, Department of Industrial Engineering and Management, ORT Braude College, Karmiel, Israel
- William P. Fisher, Jr., Ph.D., Research Associate, BEAR Center, Graduate School of Education, University of California, Berkeley, CA (USA) & Principal, LivingCapitalMetrics Consulting

References

1. EN15224:2012 ”Health care services – Quality management systems – Requirements based on EN ISO 9001:2008” European standard
2. W P Fisher, Jr., 1997 ”Physical Disability Construct Convergence Across Instruments: Towards a Universal Metric”, *Journal of Outcome Measurement*, **1**(2), pp 87 - 113
3. E Svensson 2001, “Guidelines to statistical evaluation of data from rating scales and questionnaires”, *J Rehab Med*; **33**: 47–48
4. L R Pendrill, B Berglund *et al.* 2010 “Measurement with Persons: A European Network“, *NCSLi Measure*, **Vol. 5** No. 2 • June 2010, pp. 42 – 54
5. M Wilson 2011 ”The role of mathematical models in measurement: a perspective from psychometrics”, Joint International IMEKO TC1+ TC7+ TC13 Symposium August 31st − September 2nd, 2011, Jena (DE)
6. De Battisti, F., Nicolini, G. and Salini, S. (2005). “Rasch model to measure service quality”. *ICFAI Journal of Services Marketing,* **III**(3), 58-80
7. Rasch, G. 1961, ”On general laws and the meaning of measurement in psychology”, pp. 321–334 in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, IV. Berkeley, California: University of California Press. Available free from Project Euclid
8. Kenneth Knoblauch and Laurence T. Maloney 2008 ” MLDS: Maximum Likelihood Difference Scaling in *R*”, *Journal of Statistical Software*, **Volume 25**, Issue 2. pp. 1 – 26, http://www.jstatsoft.org/
9. L T Maloney and J N Yang 2003, ”Maximum likelihood difference scaling”, *Journal of Vision* **3**, 573 – 85
10. B Schneider 1980, “Individual loudness functions determined from direct comparisons of loudness intervals", *Perception and Psychophysics, 28*(6), 493-503. [PubMed]
11. J Z Sun, G I Wang, V K Goyal and L R Varshney 2012 ”A framework for Bayesian optimality of psychophysical laws”, *Journal of Mathematical Psychology*, http://dx.doi.org/10.1016/j.jmp.2012.08.002
12. L R Pendrill and W P Fisher Jr. 2013 ”Quantifying Human Response: Linking metrological and psychometric characterisations of Man as a Measurement Instrument”, *Joint IMEKO TC1-TC7-TC13 Symposium, Measurement across physical and behavioural sciences*, 4-6 September 2013, Genova, Palazzo Ducale (IT)
13. S Dehaene, V Izard, E Spelke and P Pica 2008, ”Log or Linear? Distinct Intuitions of the Number Scale in Western and Amazonian Indigene Cultures”, *SCIENCE*, **320**, 1217 – 20
14. B Berglund 2011, ”Measurements in psychology”, in book "*Theory and methods of measurements with persons*", Editors: B Berglund, G B. Rossi, J Townsend and L R Pendrill, Psychology Press, Taylor & Francis
15. E Bashkansky, S Dror, R Ravid, P Grabov 2007 “Effectiveness of a Product Quality Classifier”. *Quality Engineering* **19**(3): 235-244
16. G Taguchi, S Chowdhury and Y Wu 2004 ”Taguchi’s Quality Engineering Handbook”, NJ: John Wiley & Sons, ISBN-10: 0471413348
17. D Kahneman and A Tversky 1979 ”Prospect Theory: An Analysis of Decision under Risk” *Econometrica*, **47(2),** pp. 263-291