17th International Congress of Metrology, 17001 (2015)
DOI: 10.1051/metrology/20151700 1

Metrology of human-based measurements / Métrologie du facteur humain

Leslie PENDRILL[1,a] and Niclas PETERSSON[1]

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

Abstract. Demands for quality assured measurement are increasing, not only from sectors such as health care, services and safety, where the human factor is obvious, but also from manufacturers of traditional products of all kinds who need to assure the quality of their products as perceived by the customer. The metrology of human-based observations is however in its infancy – concepts such as traceability and uncertainty are poorly developed as yet. This paper reviews how this can be tackled with a measurement system analysis approach, particularly where Man acts as a measurement instrument. Connecting decision risks when handling qualitative observations with information theory, perceptive choice and generalized linear modelling – through the Rasch invariant measure approach – enables a proper treatment of ordinal data and a clear separation of person and item attribute estimates. This leads in turn to opportunities of establishing measurement references for metrological quality assurance. The measurement units associated with the Rasch attribute parameters should be intimately related to metrological traceability and measurement standards. In psychometrics, we could imagine a certified reference for knowledge challenge, for example, a particular concept in understanding physics or for product quality of a certain health care service.

1 Measurement uncertainty

Traditional handling of measurement uncertainty often starts with the evaluation of standard deviations or other statistical measures of spread [1]. Thereafter confidence intervals are formed about the measurement result to describe the extent of scatter. In subsequently introducing the impact of this scatter, one might assess the percentage risks of in-correct decisions with respect to a certain specification limit as functions of the location of the uncertainty interval to the region of permissible values of the object being measured [2].

Reasons for considering re-thinking this order – indeed, substantially reversing it and *starting instead from the 'end-user' perspective* – include several challenges. For the first, there are difficulties with less quantitative data where conventional statistical tools might not apply or where human judgment doesn't allow modelling of a probability density function [3], [4]. Secondly, the need to set proactively a 'fit-for-purpose' level of measurement quality matched to the actual needs [5] and finally, a considerable difference conceptually between the usual everyday meaning of 'uncertainty' as a degree of vagueness when making decisions, and the technical definition as a standard deviation.

Demands for quality-assured measurement analysis are increasing in many 'person-centred' domains of contemporary interest − customer satisfaction, service and product quality & classification in healthcare [6], teaching, software evaluation, etc. In response, we have embarked on a general approach, attempting to model Man as a Measurement Instrument. By exploring links between metrological (resolution, classification effectiveness) and psychometric (Rasch) characterisation of Man as a Measurement Instrument [7, 8] we want to:

- Clarify the concepts of measurement uncertainty (section 3-4).
- Show the potential of establishing measurement references for human-based (and other qualitative) observations with some analogies to reference materials (section 5).

2 Instrument engineering

Because measurement uncertainty is associated with limited measurement quality, the first step in any assessment of measurement uncertainty is to make as comprehensive description as possible of the measurement set-up.

2.1 Measurement system analysis

The Measurement System Analysis (MSA) approach is widely used, for instance in the automotive industry [9]. It is based on a model where measurement information is transmitted from the measurement object, often via an instrument, to an operator. The object, instrument, and operator are the main elements of the measurement system, but the measurement method or environment can

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

Article available at http://cfmetrologie.edpsciences.org or http://dx.doi.org/10.1051/metrology/20150017001

affect the main system elements when determining overall measurement quality.

In handling qualitative observations and decision-making, a special kind of measurement system – where the 'instrument' is a human being – appears to be a promising approach [8, 10]. A question arising in that context is then: Can the various metrological instrument performance metrics of classical engineering [11] - sensitivity; resolution; linearity; bias; environmental influence sensitivity; etc. – be applied when assessing the performance of Man as a Measurement Instrument?

The model of human perception from a psycho-physical point of view, given recently in [12], is similar to a traditional model of an engineered instrument, see figure 1. The change in *Response* (*R*) due to the change of *Stimulus* (*S*) is the *Sensitivity* (*C*) of the human (acting as an) instrument.

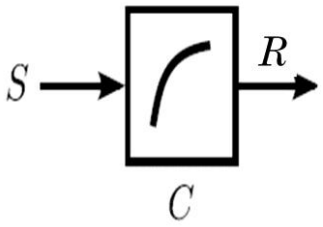

Figure 1. Simple psychophysical model of human perception.

Other recent descriptions in this area include an introduction to the measurement of psychological attributes [13, 14]. The latter draw analogies, for example, to a mechanical spring acting as a sensor.

2.2 Metrological challenges of qualitative observations

The metrological challenges when handling specifically qualitative observations and decision-making which need to be tackled are several:

Common tools of statistics, which work readily for quantitative interval and ratio scales, are unfortunately not applicable [15] to the ordinal data typical of 'human' measurement. Familiar statistical expressions derived for quantitative interval scales should not be applied to ordinal scales routinely 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 the same interpretable meaning on ordinal scales.

The concepts of metrological traceability and references for *qualitative* observations are as yet in their infancy compared to *quantitative* observations. But the need for the comparability enabled by traceability is just as relevant in fields such as healthcare [16, 17] as in other more traditional sectors where metrology is well established.

3 Decision-making and qualitative observations

Typically in a quantitative physical measurement, the metrologist will focus on how faithfully the instrument used responds to a certain stimulus value [18], particularly in terms of how much difference (measurement error) there is between, say the voltage displayed as an output (*R*) by the instrument compared with the input signal (*S*) from a voltage source (such as battery).

The response of the instrument can be parameterised in other terms apart from error, as mentioned above. To tackle qualitative observations, it appears to be particularly valuable to consider as an instrument performance characteristic, not the error in reading but a measure of *how reliably decisions* are made – such as, is the voltage source within tolerance or not?

Historically, in sectors such analytical chemistry where the quality of less quantitative observations regularly needs to be assured, an early definition of qualitative testing is: "The classification of objects against specified criteria to meet an agreed requirement" [19, 20], reflected a connection with decision-making. In the present work, we emphasise further the intimate connection [5] of qualitative observations with decision-making, relating characteristics of objects to each other or to specification limits (e.g. an upper specification limit, U_{SL}), in cases where measurement uncertainty can lead to risks of incorrect decisions.

Even in cases where the initial evidence is less quantitative, with an explanatory variable perhaps on an ordinal scale, the corresponding response variable (result of the decision) can nevertheless be quantitative, e.g., the fraction of non-conforming product is obtained just as it is in traditional acceptance sampling by attribute. Conversely, the result of a decision based on fully quantitative observations can be summarized in nominal response terms, i.e. go/no-go [21].

It is well known that any uncertainty, u_m, in an explanatory (stimulus) variable providing a basis for decisions (response) will in turn lead to certain risks (e.g. 'consumer' risk, α; 'supplier' risk, β) of incorrect decisions. This is covered by a recent document [2] that accompanies the *Guide to the Expression of Uncertainty in Measurement* (GUM) [1].

A key insight is that the human instrument is not only a sensor but additionally includes a decision-making algorithm. Here we make an additional step, connecting decision risks with dispersion in qualitative measures. Indeed, there seems to be a repeating 'loop' where measurement uncertainty in a stimulus (*S*) leads to risks of misclassification in response (*R*). This in turn can provide a new stimulus that forms the basis for subsequent decisions with commensurate decision risks, and so on.

In an introduction to the Rasch measurement approach, Mari and Wilson [14] consider factors that might lead to dispersion in the instrument output by drawing analogies between the human response (e.g. the attitude of an individual) to a test item and the 'transduction' function of the human instrument as a 'Boolean spring'. They state that probability distributions in the instrument response might reflect either – quote: "(i) the presence of an underlying unobserved ('influence') variable; (ii) a non-deterministic dependence

of the 'indication on the measurand' (i.e. attitude in this case); or (iii) that the 'measurand' is itself stochastic."

For our purpose, to clarify concepts and terminology, consider the role of Man in a measurement system in various scenarios. Studies of elementary tasks – such as counting [7] or ellipticity perception [22], an initial measurand could be, respectively, the number or degree of correlation in a clouds of dots. Man, in this case acting as a measurement instrument, yields estimates of the value of each measurand. But what is interesting is not the number or ellipticity of cloud dots since we know these already, but rather how well these measurements are performed. The ability to perform such measurements is described in terms of a decision-making process, e.g. can the human instrument resolve the difference between adjacent stimuli, such as: "Are there nine or ten dots in the cloud?" This decision-making ability can be expressed as the probability of success, $P_{success}$, vis-à-vis the risks of making incorrect decisions. As will be demonstrated below, a psychometric (Rasch) analysis can yield estimates of the new measurands, namely: (i) the ability of each human instrument and (ii) the inherent level of challenge posed by a particular object. Finally, if the ability or level of challenge is crucial in some human-factor application, then specification limits will be set on them, in which case these assessed quantities become quality characteristics, and so forth.

4 Generalized linear models and perceptive choice

So-called generalized linear models (GLM) [23] are invoked to treat decision-making scenarios where the response variable (R) cannot always be expected to vary linearly with the explanatory variable (S). There is an extensive family of GLM link functions, g, offering a linear predictor η of S,

$$\eta = S \cdot \gamma \tag{1}$$

(which will be some linear combination of unknown parameters γ), such that the expectation

$$\mathbb{E}[R] = g^{-1}(\eta) = g^{-1}(S \cdot \gamma). \tag{2}$$

This covers not only explicitly non-linear responses but also poorly known responses to the explanatory variable, perhaps allowing only less quantitative appraisals, such as on an ordinal or even a merely nominal scale. A common link function is the 'logit'

$$g(x) = \log\left(\frac{x}{1-x}\right) \tag{3}$$

which readily transforms (for instance) a probability, $P_{success} \in (0,1)$ into a number $z \in \mathbb{R}$,

$$z = \log\left(\frac{P_{success}}{1 - P_{success}}\right). \tag{4}$$

This GLM approach is known to be able to handle qualitative data, e.g. on an ordinal scale, where familiar statistical expressions derived for quantitative interval scales cannot be applied. In such methods, using logistics regression, measurable constructs are formulated for perceptual characteristics such as the difficulty of an exam or the quality of a service [24], 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 [25, 26].

Connecting the risks of incorrect decisions to GLM can be done in terms of two kinds of human-based perception (and analogous system performance metrics), namely: identification and choice as dealt with in psychophysics [27]. *Identification* involves in the dichotomous case a yes-no detection – is the stimulus within tolerance or outside a region of permissible values specification limit*?* The decision ('consumer') risk, α, is in this case estimated as the cumulative distribution function (CDF) beyond the specification limit (U_{SL}, say) on the explanatory variable, x, of the initial set of observations [5]:

$$\alpha = \mathbb{P}(x \geq U_{SL}) = \int_{U_{SL}}^{\infty} \frac{1}{\sqrt{2\pi} \cdot u_m} e^{-\frac{(x-x_m)^2}{2 \cdot u_m}} dx \tag{5}$$

If the decision-making process made by Man when acting as a Measurement Instrument can be regarded in terms of information theory, then it is possible [10] to derive the GLM link function (4) in terms of entropy and perceptive distances.

5 Man as a measurement instrument

One particular version of logistic regression (section 4) that has received considerable attention as a tool for handling ordinal data is due to the Danish statistician Rasch [25]. In response to criticism of psychometric methods of the time, Rasch explicitly attempted a separation in measured responses into a person attribute value (θ, such as ability or leniency) and an item attribute value (b, such as level of challenge or quality) by writing $z = \theta - b$. In the simplest, dichotomous case the logistic regression function is,

$$\theta - b = \log\left(\frac{P_{success}}{1 - P_{success}}\right) \tag{6}$$

The logistic regression approach to handling human-based measurement results, even those on ordinal scales, is rapidly becoming the method of choice in many areas of application, ranging from international educational studies [28], customer satisfaction surveys [24], to person-centred health care [17].

Examples of this separation include the difference between (i) the intrinsic quality of a product and the leniency of a customer; (ii) the level of challenge of a certain task and the ability of a person to tackle the challenge; (iii) the ability of an indenter and the hardness of a material, to name just a few examples.

5.1 Separating person and item attribute values in responses with the Rasch approach to logistic regression

The response of a human when encountering a particular task or feature of an item will depend on a combination of the characteristics of both the person and the item. In traditional metrology, a separation of instrument and measurement object is regularly achieved, such as when determining the mass of a weight in terms of the calibrated response of a weighing instrument. Without that separation, dispersion in the sought item attribute will be masked by instrument dispersion.

In the words of Guilford [29]:

"*... all measurements are indirect in one sense or another. Not even simple physical measurements are direct, as the philosophically naive individual is likely to maintain. The physical weight of an object is customarily determined by watching a pointer on a scale. No one could truthfully say that he 'saw' the weight…"*

"*It must be granted that, to measure such psychological attributes as appreciation of beauty,…, we must depend upon secondary signs of these attributes. The secondary signs bear some functional relationship to the thing we wish to measure, just as the movement of a pointer on a scale is assumed to bear a functional relationship to the physical phenomenon under consideration. The functional relationship may be simpler and more dependable in the latter case than in the former and the type of relationship may be more obvious. That is the only logical difference. It is admittedly a difference of some practical consequence. But it is not a difference which leads to the conclusion that measurement is possible in the one case and impossible in the other.*"

In human-based measurement, an ordinary factor analysis in traditional statistics could be attempted to separate the two attributes (person/item) in the response, but that would not necessarily work for ordinal data [30].

In a recent paper, Wilson and co-workers [31] in considering 'instrument calibration', state that "social science measurement does not usually allow for the concept of a measurement standard". They refer to a common form of standardisation in social science measurements, e.g. IQ tests, called "norm-referencing" and point out that to provide measurement standards it is necessary to extend the concept of an 'instrument' (which they regarded as a questionnaire or multi-choice examination, say) to "include the particular sample to which it is administered". But they conclude by stating that measurement standards in the social sciences will be, even in the best case, "of not much practicable use".

The key to our approach is to treat the human responder as the instrument instead.

5.2 Reliability, uncertainty and metrological traceability in human-based measurement

In its logistic regression form, the 'straight ruler' aspect of the Rasch formula, i.e. equation (6), has been described by Linacre and Wright [32] in the following terms: "The mathematical unit of Rasch measurement, the log-odds unit or "logit", is defined prior to the experiment. All logits are the same length with respect to this change in the odds of observing the indicative event."

The Rasch invariant measure approach goes further in defining measurement units [33] since it uniquely yields estimates "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." It is not simply a mathematical or statistical approach, but instead a specifically metrological approach to human-based measurement. Note that the same probability of success can be obtained with an able person performing a difficult task as with a less able person tackling an easier task. The separation of attributes of the measured item from those of the person measuring them brings invariant measurement theory to psychometrics.

In general, the measured item attribute b e.g. a level of challenge) differs, because of limited reliability, from the 'true' b', with an error ε_b:

$$b = b' + \varepsilon_b \tag{7}$$

Invariant measure theory, allowing the level challenge b for a particular task to be estimated independently of who is encountering the challenge, permits the identification of a metrological standard (an 'anchor') for item challenge if a sufficiently large and representative group of people are allowed to test the item.

The reliability – that is, the ratio of 'true' item variance to the total variance (including measurement uncertainty (often set to a target value of 0.8, corresponding to a measurement uncertainty σ_{ε_b} not larger than half the item standard deviation σ_b), is a function both of the number of test persons as well as how well their abilities match the challenges.

5.3 Standards, references and units in human-based measurement

Once an agreed definition and realisation of the standard challenge has been achieved it can subsequently be used reproducibly as a reference for new challenges. As in traditional metrology, this traceability enables all the advantages commensurate with objectively comparable measurement. For instance, having access to a psychometric barrier challenge standard would allow an estimate of each person's ability θ to negotiate a range of barriers of different challenge to be metrologically calibrated by measuring a task of known challenge. This procedure determines the measurement error ε_θ in person ability:

$$\theta = \theta' + \varepsilon_\theta \tag{8}$$

Inserting the corrected item b and person θ attribute values in equation (6) allows a more correct estimate of the accessibility score, $P_{success}$.

These metrological references – which might be termed 'performance' measures – refer to a different class of quantities than those familiar from traditional metrology and care has to be taken not to assume that the new references follow exactly the same rules as say physical quantities and units [34, 35]. Perhaps the closest analogies to references in psychometrics can be found with reference materials. We seek, when formulating examples of measurement references in psychometrics, an agreed upon, standardized measure of, for instance, the level of challenge posed by a particular task or barrier. In psychometrics, we could imagine a certified reference for knowledge challenge, for example, a particular concept in understanding physics or for product quality of a certain health care service.

6 Conclusion

The cumulative probability of a 'correct' decision is related to the change of entropy on information transmission. This can be interpreted as the combined process of observation (measurement) of an explanatory variable and response (decision-making), linearized using logistic regression. This covers not only explicitly non-linear responses but also poorly known responses to the explanatory variable, perhaps allowing only less quantitative appraisals, such as on an ordinal or even a merely nominal scale.

The explicit separation of person and item attribute estimation made possible with the psychometric (Rasch) invariant measure approach, appears well suited for introducing metrological traceability to human-based measurement. Thus, the measurement units associated with the Rasch attribute parameters θ and b, should be intimately related to metrological traceability and measurement standards.

Acknowledgments

This work was performed as part of the EMRP NEW04 project which belongs to the European Metrology Research Programme (EMRP, FP7 Art. 185), jointly funded by the EMRP participating countries within EURAMET (www.euramet.org) and the European Union.

Thanks are due, particularly, to:

- 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. JCGM, *Guide to the expression of uncertainty in measurement (GUM)*. 2008.
2. JCGM, *Evaluation of measurement data – The role of measurement uncertainty in Conformity Assessment*, in *Joint Committee on Guides in Metrology (JCGM)*. 2012.
3. T Gadrich, E Bashkansky, and R Zitikis, *Assessing variation: a unifying approach for all scales of measurement.* Quality & Quantity, 2014. **49**(3): p. 1145-1167.
4. J C Helton, et al., *Survey of sampling-based methods for uncertainty and sensitivity analysis.* Reliability Engineering and System Safety, 2006. **91**: p. 1175 - 209.
5. L R Pendrill, *Using measurement uncertainty in decision-making & conformity assessment.* Metrologia, 2014. **51**: p. S206.
6. L R Pendrill, et al., *Measurement with Persons: A European Network.* NCSLI Measure J. Meas. Sci., 2010. **5**(2): p. 42-54.
7. L R Pendrill and W P Fisher Jr, *Counting and Quantification: Comparing Psychometric and Metrological Perspectives on Visual Perceptions of Number.* Measurement, 2015. **71**: p. 46–55.
8. L R Pendrill, *El ser humano como instrument de medida.* e-medida, 2014.
9. AIAG, *Measurement Systems Analysis Reference Manual*, in *Chrysler, Ford, General Motors Supplier Quality Requirements Task Force*. 2002, Automotive Industry Action Group.
10. L R Pendrill, *Man as a Measurement Instrument.* NCSLI Measure J. Meas. Sci., 2014. **9**: p. 24 – 35.
11. J P Bentley, *Principles of Measurement Systems*. 4th ed. 2005, London: Pearson Education Limited.
12. J Sun, et al., *A framework for Bayesian optimality of psychophysical laws.* J. Math. Psychol., 2012. **56**(6): p. 495-501.
13. K Sijtsma, *Introduction to the measurement of psychological attributes.* Measurement, 2011. **44**(7): p. 1209–1219.
14. L Mari and M Wilson, *An introduction to the Rasch measurement approach for metrologists.* Measurement, 2014. **51**: p. 315–327.
15. E Svensson, *Guidelines to statistical evaluation of data from rating scales and questionnaires.* J. Rehabil. Med., 2001. **33**: p. 47-48.
16. W Fisher Jr, *Invariance and traceability for measures of human, social, and natural capital: Theory and application.* Measurement, 2009. **42**(9): p. 1278–1287.
17. EN 15224, *Health care services – Quality management systems – Requirements based on EN ISO 9001:2008*. 2012.
18. C D Ehrlich. *Traceability Considerations for the Characterization and Use of Measuring Systems*. in *NCSLi Workshop and Symposium*. 2015. Dallas, TX, USA.
19. W Hardcastle, *Qualitative Analysis: A Guide to Best Practice*. 1998, Cambridge, UK: Royal Society of Chemistry.
20. S Ellison and T Fearn, *Characterising the performance of qualitative analytical methods: Statistics and terminology.* TRAC-Trend Anal. Chem., 2005. **24**(6): p. 468–476.
21. L R Pendrill, *Uncertainty & risks in decision-making in qualitative measurement: An information-theoretical approach*, in *Advanced*

Mathematical and Computational Tools in Metrology and Testing, S.o.A.i.M.f.A. Sciences, Editor. 2012, World Scientific.
22. L R Pendrill, *Discrete ordinal & interval scaling and psychometrics*, in *Métrologie 2013 Congress*, CFM, Editor. 2013: Paris.
23. P McCullagh, *Regression models for ordinal data.* J. Roy. Stat. Soc., 1980. **42**: p. 109 - 42.
24. F De Battisti, G Nicolini, and S Salini, *The Rasch model to measure service quality.* The ICFAI Journal of Services Marketing, 2005. **vol. III**(3): p. 58-80.
25. G Rasch, *On general laws and the meaning of measurement in psychology*, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. 1961, University of California Press: Berkeley. p. 321-334.
26. J Weller, et al., *Development and Testing of an Abbreviated Numeracy Scale: A Rasch Analysis Approach.* J. Behav. Decis. Making, 2013. **26**(2): p. 198–212.
27. G Iverson and R Luce, *The representational measurement approach to psychophysical and judgmental problems*, in *Measurement, Judgment, and Decision Making*. 1998, Academic Press.
28. OECD, *The Rasch Model*, in *PISA Data Analysis Manual*, SAS, Editor. 2009, OECD. p. 79–94.
29. J P Guilford, *Psychometric Methods*. 1936, McGraw-Hill, Inc. p. 1-19.
30. B Wright, *Comparing factor analysis and Rasch measurement.* Rasch Measurement Transactions, 1994. **8**(1).
31. M Wilson, et al., *A comparison of measurement concepts across physical science and social science domains: instrument design, calibration, and measurement.* Journal of Physics: Conference Series, 2015. **588**.
32. J Linacre and B Wright, *The 'Length' of a Logit.* Rasch Measurement Transactions, 1989. **3**(2): p. 54-55.
33. S Humphry, *The Role of the Unit in Physics and Psychometrics.* Measurement: Interdisciplinary Research and Perspectives, 2011. **9**(1): p. 1-24.
34. J de Boer, *On the History of Quantity Calculus and the International System.* Metrologia, 1994/95. **32**: p. 405-429.
35. G B Rossi, *Measurability.* Measurement, 2007. **40**: p. 545 - 562.