Measurement 71 (2015) 46–55

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number

L.R. Pendrill [a,*], William P. Fisher Jr. [b,c]

[a] SP Technical Research Institute of Sweden, Measurement Technology, Box 857, SE-50115 Borås, Sweden
[b] BEAR Center, Graduate School of Education, University of California, Berkeley, CA, USA
[c] LivingCapitalMetrics Consulting, Sausalito, CA, USA

ARTICLE INFO

Article history:
Received 9 June 2014
Received in revised form 26 March 2015
Accepted 8 April 2015
Available online 23 April 2015

Keywords:
Human
Response
Health care
Metrology
Instrument
Psychometrics

ABSTRACT

In gaining a better understanding of how to characterise human response, essential to improved person-centred care and other situations where human factors are crucial, recent work has attempted to link metrological (resolution, classification effectiveness) and psychometric (Rasch) characterisation of Man as a Measurement Instrument. The present work offers a more detailed account of these investigations following our first preliminary conference report, continuing 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 and loss functions. Independent sources of measurement uncertainty, such as under-estimation of scores, are distinguished from separate estimates of task challenge and individual counting ability, and accounted for in estimates of reliability of the various measures.

1. Introduction

The reliable characterisation of the human measuring instrument [1], be it with the five senses or the full physiological, mental, cognitive and behavioural richness of human perception, is essential in many applications. Some of these include enhancement of various human functions, machine learning [2] 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 everyday tasks [3], to name a few. Quantities of concern are not merely technical but also more human, such as comfort, pleasure, and beauty [4].

Bearing in mind that formulation of metrological concepts commensurate with those established in traditional engineering is as yet in its infancy in perceptual contexts [5–8], we have initiated work attempting to link metrological (resolution, classification effectiveness) and psychometric (Rasch) characterisation of Man as a Measurement Instrument, as briefly reported in a conference proceedings [9]. The present work offers a more detailed account of these investigations, continuing our 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. Two key aspects of quality-assured measurement – traceability and uncertainty – must be kept in focus when metrologically characterising Man as a Measurement Instrument.

Some method of metrological traceability to invariant unit standards 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. Patient health, for example, is increasingly rated in health clinics on ordinal scales linearized

[*] Corresponding author. Tel.: +46 767 88 54 44.
E-mail addresses: leslie.pendrill@sp.se (L.R. Pendrill), wfisher@berkeley.edu (W.P. Fisher Jr.).

http://dx.doi.org/10.1016/j.measurement.2015.04.010
0263-2241/© 2015 Elsevier Ltd. All rights reserved.

via a log-odds transformation, and appropriate treatment is decided by comparing the actual ratings with corresponding patterns of typical health ratings for similar patients from earlier studies. The comparability of such ratings has to be reliable to a sufficient degree of accuracy if the patient is to be treated appropriately. When and where such accuracy can be regularly obtained, the observational framework could be redesigned to omit the observed ordinal scores and to incorporate a metrologically traceable unit [7,10,11].

The second aspect of quality-assured measurement – namely, measurement uncertainty and reliability – also presents some challenges where observations are made with Man as a Measurement Instrument. In particular the usual tools of statistics, such as for calculation of the mean and standard deviation, needed when expressing uncertainty, cannot generally be applied to scores obtained with questionnaires and similar instruments, since a lack of an invariant unit renders uninterpretable the location and dispersion of qualitative measurements on ordinal scales [12]. At the same time, measurement uncertainty, since it reflects the quality of measurement, also provides a measure of the ability of human to perform as an instrument.

2. Rationale for a potential expansion of the metrological framework

An inquiry into the viability of a uniform unit of physical functioning in medical rehabilitation compared results produced across four different measurement instruments applied to eleven different samples [7]. The heuristic model employed producing each set of results demands 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”. Fit to this model allows parallels to be drawn between psychometric concepts of invariance and equating, on the one hand, and metrological concepts of traceability to repeatable and reproducible unit standards, on the other. This need for more rigorously defined and more widely distributed measures is implicated by the trend in health care associated with a shift from “local economies of disease-crisis management to regional, national, and international economies of population-based, preventive health management”. As demand for proactive prevention displaces reactive responses, it is virtually inevitable that continuing growth in the speed and networking reach of computational tools will propel invariant measurement into significant new roles supporting accountability and comparability in health care.

Human beings inevitably play critical roles in measurement [13]. From the perspective of engineering, the operation of any measurement system requires calibration, data acquisition, and data presentation [14]. Recent studies in psychophysical scaling [15] combine psychometric and engineering perspectives, relating perceptual intensity to stimulus intensity by explaining the Weber–Fechner law in terms of signal error fidelity [16]. Other studies describe adaptation by biological sensory systems in terms of the costs of perceptual task errors [17].

Another approach to linking engineering and psychometric conceptualizations of measurement systems is suggested by psychological measurement models introduced by Rasch [18–20]. When a human is instrumental to the performance of elementary tasks – such as counting dots [21,22] – person abilities relative to the degree of challenge posed by different tasks can be expressed in terms of measures invariantly located and dispersed on an interval scale [23,24]. The ability to perform the task can be calibrated and measurement uncertainty can be assessed in this context with the special advantage of independent advance knowledge of the measurement object’s true value (a given number of dots).

3. Grounding measuring in counting

Our first preliminary brief report in a conference proceedings [9] highlighted the suitability for our research of previous studies by others [22] concerning the counting ability of the Mundurucu, an Amazonian indigenous people with little access to Western-style educational resources and where counting above the number five is often a challenge. Research investigating the conceptual link between number and spatially distributed dots had already suggested that the Mundurucu intuitively employ a logarithmic transformation of impressions of varying amounts (Fig. 1), meaning that “larger numbers require a proportional larger difference in order to remain equally discriminable” (Weber’s law) [22]. In the present work, we extend the analysis of Dehaene et al. [22], taking advantage, as they, when attempting to characterise human response, of the conceptual simplicity where one knows independently the expected value (the number of dots). Our aim is to explore further the link between metrological (resolution, classification effectiveness) and psychometric (Rasch) characterisation of Man as a Measurement Instrument.

Fechner was among the earliest to note the contrast between the linearity of measures and the nonlinearity of intuited impressions for the human senses, a contrast of ongoing interest in neurological research [25]. Because a fairly constant degree of imprecision is maintained across several orders of magnitude, the “Gaussian tuning curve” serves, in effect, as a kind of internally embodied sensory slide rule [22]. By identifying and describing the logarithmic proportionality of sensations and stimuli, Fechner connected physical experience with linear geometry in a way that set the stage for Thurstone, and, later, Rasch, to refocus human measurement away from its previous preoccupation with purely psychophysical phenomena to broader concerns with psychological, economic, and social phenomena [26,27].

It is important to note the deeper connection here that real things, like the sides of triangles, rocks, or human behaviours, are never identical, and so do not ever conform perfectly with expectations formed on the basis of a mathematical formulation of a scientific law or measurement model. Measurement, whether of counting ability or of mass or temperature, requires abstract invariant units that physically cannot correspond directly with empirical

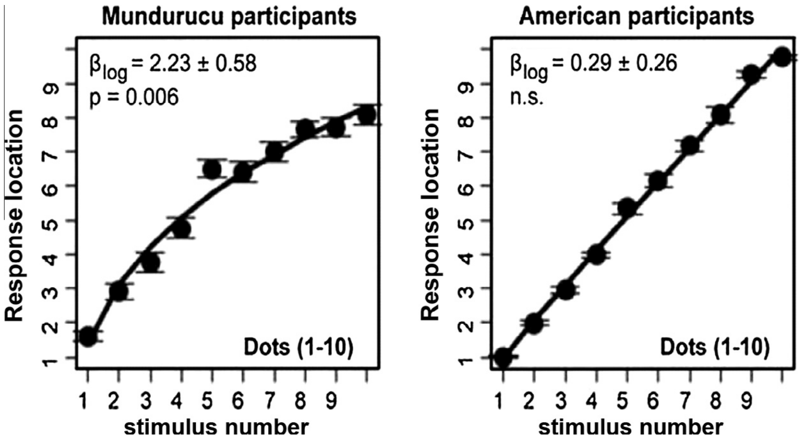

Fig. 1. Average location of numbers, *m*', on the horizontal segment, separately for Mundurucu participants (left column) and for American participants (right column) [22]. Licensed for reproduction: Licensee: Leslie R Pendrill, License Date: Apr 13, 2015, License Number: 3607091428035, Publication: Science.

observations [23]. By 'invariant unit' we mean a constant unit, 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. Here again, experimental tests as to the usefulness of approximations to the desired linear units hinge on the natural logarithm.

4. Human performance related to invariant measures on interval scale

Can the various metrological instrument performance metrics of classical engineering [28] – sensitivity; resolution; linearity; bias; environmental influence sensitivity; etc – be applied to assessing the performance of Man as a Measurement Instrument [13,14]?

4.1. Metrological characterisation

Fig. 2 shows the scatter of counting data amongst the 33 Indians (adults and children) studied by Dehaene et al. [22], portrayed as a probability mass function (PMF) commonly employed for discrete data, where the probability, q_c, of classifying a particular count, c, is plotted against the range of K levels/categories.

Limited measurement quality leads to incorrect classification, as depicted with the PMF shown in Fig. 2. This scatter is associated with the limited performance of the human instrument (the number of dots is of course fixed and known for each nominal count), in general both for each individual as well as from variations from person to person studied by Dehaene et al. [22]. In the general case, the probability, q_c, would be given by:

$$q_c = \sum_{k=1}^{K} p_k \cdot P_{c,k}$$

where the prior distribution is described by p_k and the (conditional) probability of observing a particular count c when the 'true' count is M is $P_{c,M}$. In the particular case of the Mundurucu study:

$$q_c = P_{c,k}$$

since the prior state is known exactly, i.e. $p_k = \begin{cases} 1; k = M \\ 0; k \neq M \end{cases}$

It is assumed, in characterising the Mundurucu counting data with discrete PMFs, that the human observer recognises discrete dots (as 'units' $[I]$ of counts, where I is the unit matrix) – perhaps by 'coding' [16,17] – so a non-integer response for the overall count for each stimulus value is a result of scatter in estimated *integer* counts rather than a perception of fractions.

It might be tempting to calculate simply, from the PMF of Fig. 2, a mean (expectation value, E) and variance (V) of the measured count $\{X = m'\}$ from the regular statistical expressions:

$$E(\{X\} \cdot [I]) = \sum_{c=1}^{K} q_c \cdot (\{m_c\} \cdot [I]) \tag{1}$$

$$V(\{X\} \cdot [I]) = \frac{\sum_{c=1}^{K} [q_c \cdot \{m_c\} \cdot [I] - E(\{X\} \cdot [I])]^2}{K-1} \tag{2}$$

to characterise the dispersion of counting data, where K is the number of discrete levels/categories.

This has, however, to be done with care since the perceived counts along the horizontal axis of Fig. 2 are arguably not on a regular quantitative scale, where relations between the numbers would represent (exact) relations between the objects (i.e. perceived counts of dots), but rather on a comparative 'ordinal' scale which merely allows objects to be related to each other with respect to an overall order [29]. To describe this mathematically, note firstly that if expressions (1) and (2) were to be evaluated in the state prior to measurement, which provides the 'stimulus' set of metrological references with which the human instrument is to be calibrated, then the probability for count category k is p_k and the term $m_k = k$ is simply the known integer number of dots. In that prior state, the x-axis would be a regular quantitative scale and the usual statistical tools can be applied without reservation. However, after measurement, for the 'response' results shown in Fig. 2, with count probability q_c, the corresponding simplification of m_c as the perceived number of dots

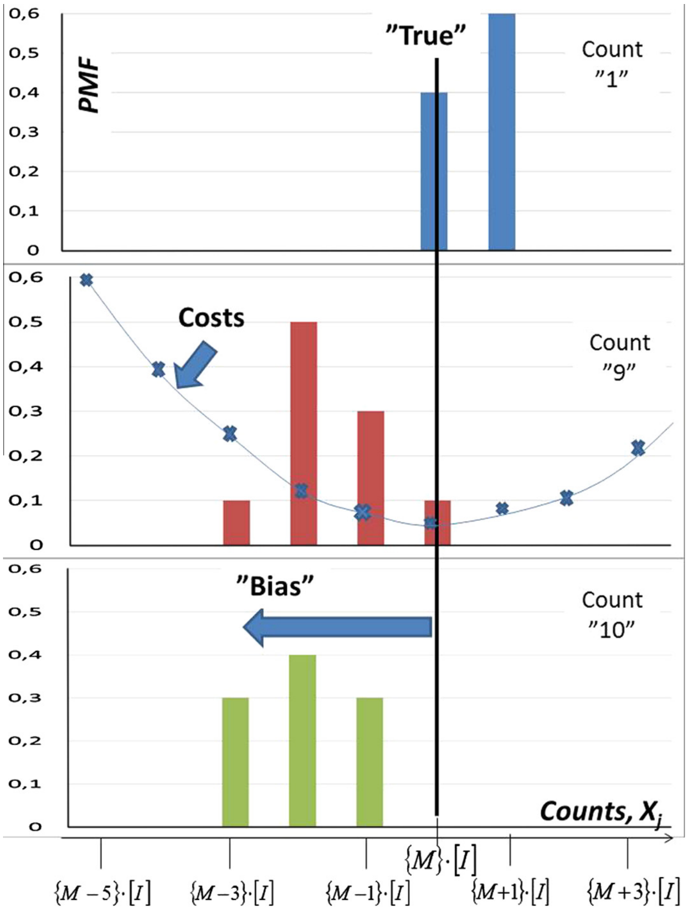

Fig. 2. Probability 'mass' distributions (PMF) of 'perceptive' counting, with 'true' value *M*. Data for Munduruсu counters from [22].

cannot be made. The perceived count numbers along the *x*-axis in this posterior case are increasingly distorted as the human instrument attempts to count larger numbers of dots. Since the *x*-axis in each plot of Fig. 2 spans a range of count values – e.g. the count of '9' dots covers the range 4–12 – account has to be taken for this ordinal scale distortion in each plot.

The *x*-axis in the present elementary case can be seen as a prototype for whole families of qualitative scales – such as responses to multiple choice questionnaires – where it is increasingly recognised that many of the regular tools of statistics cannot be applied [12]. In such wider applications, one does not in general have the advantage of having access to known stimulus values.

A simple, psychophysical model of each human perception can be formulated, considering Man as a Measurement Instrument. As indicated in Fig. 3, the instrument sensitivity, ***C*** – defined in engineering metrology as the ratio of the output ***R*** (the perceived count) to the input ***S*** (the known stimulus) – varies across the range of stimulus values: The observations (shown in Fig. 2) seem to indicate that, as the challenge of counting higher numbers of dots increases, there is a corresponding increase in bias between stimulus and response values. At the same time, the dispersion (variance) in the Dehaene et al. [22] data appears largely constant from count to count.

Similar care in treating ordinal data has to be taken when adopting the quantization (*Q*) mapping approach employed recently [16] to interpret the Munduruсu observations, since in the numerator of an expression [Eq. (3) below] for expected relative error (*ERE*), the difference between the stimulus, *S*, and its 'quantized' value, $\hat{S}$ in that approach, is arguably on an ordinal rather than a more

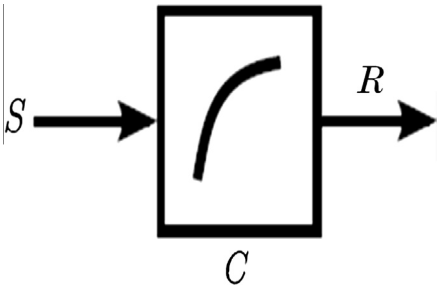

Fig. 3. Simple psychophysical model of Weber–Fechner perception (adapted from [16]).

quantitative scale defined by an invariant unit. Dividing by the ‘energy’ S^2 in expression (3) does not in our opinion guarantee the scale-invariance claimed by Sun et al. [16].

$$ERE(Q) = E\left[\frac{|S-\hat{S}|^2}{S^2}\right] \tag{3}$$

A commonly expressed reservation to modelling human response in terms of engineering metrology is that the human ‘instrument’ is notoriously variable and unpredictable. We would counter this reservation by pointing out that even a regular engineered instrument used in the field – e.g. in a harsh industrial environment – could display potentially large variability: the size of the variation in itself is no reason not to attempt, or indeed could be a strong motive, to make such a characterisation. Secondly, as will be discussed below, there are plenty of examples, for instance in psychology and the social sciences, where human ‘instruments’ are characterised in this way, notwithstanding their variability: the examination of pupils in a school class being a classic example. The perception of instability in psychological and social constructs is related in large part to uncontrolled variation in representations and experimental contrasts. With the wider and larger scale applications of probabilistic measurement modelling that have taken place over the last several decades, however, evidence of empirical stability and theoretical explanatory power has increased substantially [7,10,13,18–20].

In this section, probability terms have been usefully employed to deal with ordinal data, even when the corresponding measurement values are not susceptible to regular statistical tools. Our current work therefore adopts various probabilistic approaches to Man as a Measurement Instrument, where instead of scatter, we derive several alternative expressions for the probability of successfully counting the number of dots, as described in more detail below.

4.2. Probabilistic approaches to subjective differences

Two different 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 investigated here in terms (resolution and bias) as one would use when characterising a measurement instrument (in this case, the human counter), as briefly reported earlier [9]. These will be compared with a psychometric, Rasch analysis (Fig. 4).

4.2.1. Rasch

The Rasch approach [18–20] models individual probabilities of success as the difference between person ability θ and level of challenge β; in the dichotomous (binary decision) case:

$$P_{success} = \frac{e^{[\theta-\beta]}}{1+e^{[\theta-\beta]}} \tag{4}$$

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 of course 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. The Rasch approach organizes observations and their representations in ways that support the formation and testing of hypotheses concerning the separability of person ability and task challenge to be made in human-based measurement.

For the actual multinomial case of counting by the Mundurucu Indians, we adopt the corresponding polytomous expression for the success score,

$$P_{success,i,j} = \upsilon_{i,j,c} = \frac{e^{\left[\sum_{k=1}^{c}[(\theta_i-(\beta_j-\tau_k))]\right]}}{\sum_{j=1}^{K} e^{\left[\sum_{k=1}^{j}[(\theta_i-(\beta_j-\tau_k))]\right]}} \tag{5}$$

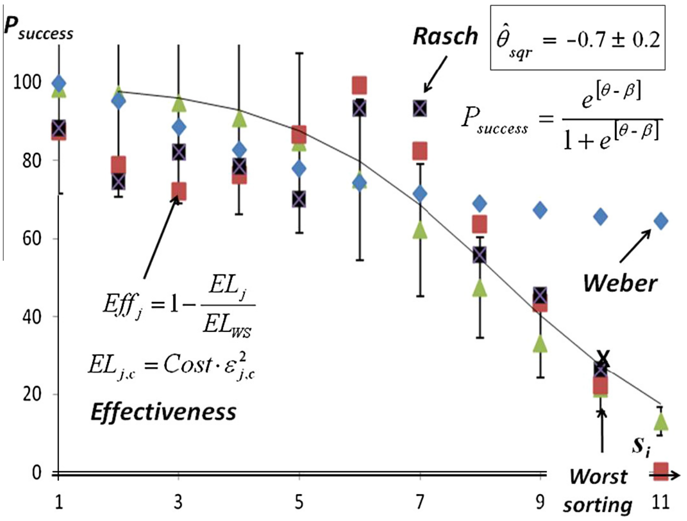

Fig. 4. Comparison of effectiveness (6), Rasch (3) and Weber (5) modelling (w = 0.17) of counting range 1–10 for the Mundurucu; squared loss function.

of person i and count j, where τ_k is the Rasch-Andrich threshold, that is, the point of equal probability on the latent variable between categories $k-1$ and k. A logistic regression is made between where observed, $v_{i,j}$, and expected scores, $v'_{i,j}$ in which the sum of squared residuals, $\Delta v'^2_{i,j,c} = (v_{i,j} - v'_{i,j})^2$, is minimised, to yield individual estimates of person ability θ and level of challenge β.

No distributional assumptions are required, counts of observed success in the assigned task are minimally sufficient statistics [30,31], 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 this analysis, a weighted score: $v_{i,j,c} = w(m_{i,j,c} - M_j)$ is calculated for each count, j, test person, i, and category c. The weighting factor, w, is some function of the distance between the observed count, m, and the (known) stimulus value (M), reflecting an increasing 'penalty' score as the bias (shown in Fig. 2) increases: for instance for a squared weighting: $v_{i,j,c} = (m_{i,j,c} - M_j)^2$. Other weightings investigated include linear and exponential functions of the distance (as also employed below in modelling classification effectiveness). Using proprietary software for Rasch analysis [32], these scores are then analysed to produce estimates of person ability θ_i and level of challenge β_j across the full set of observations of Mundurucu counting $j = 1 \ldots 10$, as shown in Fig. 5.

In order to draw significant conclusions from the analysis, it is necessary to estimate limits to measurement quality in the Rasch analysis. A so-called 'construct alley' [33] plot (Fig. 6) of the estimated Rasch level of challenge β against the INFIT z-score $\frac{\Delta v'^2_{i,j,c}}{[v'^2_{i,j} - v^2_{i,j}]}$ gives indications of the precision and trueness of the Rasch estimates.

For dichotomous observations, the model can be written to include a measurement uncertainty term in the scored response $v_{i,j}$ from Eq. (5) as follows:

$$v_{i,j} = P_{success,i,j} \pm \sqrt{P_{success,i,j} \cdot (1 - P_{success,i,j})} \qquad (6)$$

from the well-known (Bernoulli) variance of the binomial distribution. The binomial error distributions for dichotomous scores approximate Gaussian distributions when accumulated across all the observations, as they are in the estimation process [34].

In the present case, all fitted points lie well within ±2 standard deviations along the z-score axis. Measurement uncertainties are indicated along the Rasch β axes in both Figs. 5 and 6.

The corresponding reliability coefficient, for instance for Rasch β, is given by [33]:

$$R_\beta = \frac{\text{True variance}}{\text{Observed variance}} = \frac{\text{var}(\beta')}{\text{var}(\beta)} = \frac{\text{var}(\beta) - \text{var}(\varepsilon_\beta)}{\text{var}(\beta)}, \text{where } \beta = \beta' + \varepsilon_\beta \qquad (7)$$

Of course, the 'true' variance cannot be known, but it can be estimated by subtracting the estimated measurement variance from the total variance observed, as done in Eq. (7).

Widespread references in the literature of psychological research cite Nunnally's classic text [35] as a source recommending 0.70 as an acceptable minimum reliability indicating the presence of more true variance than error variance [36]. That recommendation, like the concept of reliability itself, is widely misunderstood. Reliability coefficients are not the measures of unidimensionality or internal consistency they are often assumed to be [37].

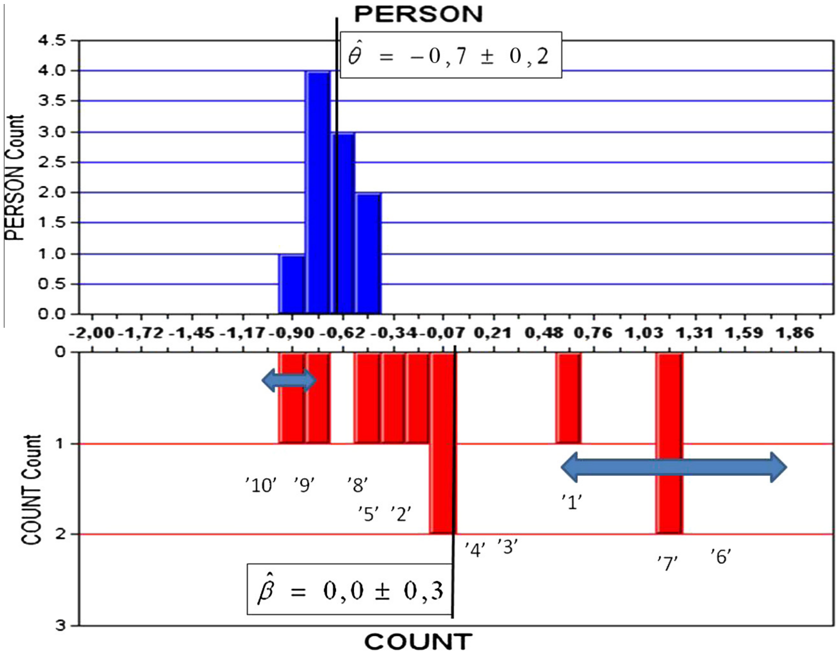

Fig. 5. Estimates of person ability θ and level of challenge β across the full set of observations of Mundurucu counting 1…10: squared weighting model.

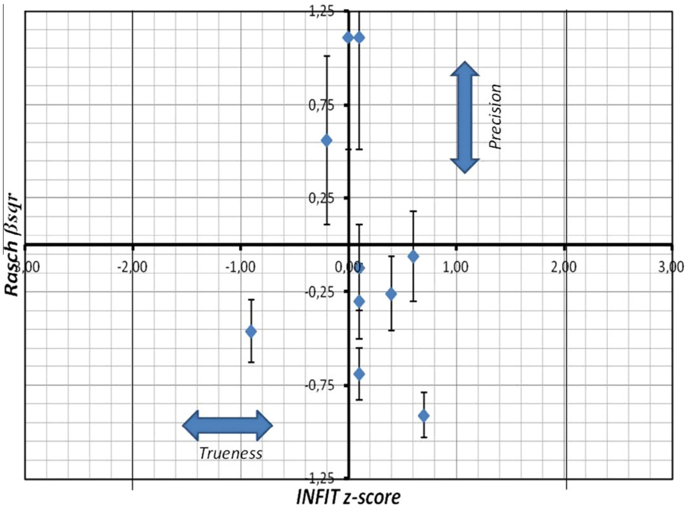

Fig. 6. ‘Construct alley’ plot of the estimated Rasch level of challenge β against the INFIT z-score; squared weighting model.

Reliability coefficients are most directly influenced by the number of score groups distinguished via the number of questions asked and the way the responses are categorised. More distinctions generally give higher reliabilities, as the model mean square error shrinks relative to the variance [38]. Internal consistency only secondarily affects reliability [39] and is more properly assessed in terms of model fit [40].

Furthermore, the usual interpretation of Nunnally’s recommendation is flawed on at least two counts. First, reliability is a property of data, not of an instrument. That is, by definition, reliability coefficients can reasonably be expected to range anywhere from 0.00 to 1.00 depending on whether the measured sample varies much less than, or many times more than, an error of measurement. This issue is, of course, not in contention in conscientiously conducted research that establishes definitive parameters and conditions for the relevant population as a whole via appropriate representative sampling. Few studies in psychology operate at this level of sophistication.

The second problem with taking 0.70 as an acceptable minimum reliability standard is that Nunnally recommended it only for early stage research in which the risk of failure is high. All that is desired in this kind of investigation is an inexpensive initial indication of potential returns on investments in research in a new direction [36]. Nunnally recommended 0.80 for widely used scales, 0.90 for high stakes decision making, and 0.95 as the desirable standard. The assumption in all of these recommendations, of course, is that the sample measured is representative of the entire range of variation in the relevant population.

A reliability of 0.80 corresponds with a separation ratio $G = 2$, meaning that measurement uncertainty is not larger than half the object standard deviation [41]. For the analysis shown in Fig. 5, reliability was $R_{\beta,sqr} = 0.73$. Here var $(\varepsilon_\beta) = realSE(\beta)^2$ and

$$SE(\theta_i, \beta_j, \tau_k) = \sqrt{\frac{1}{\sum_{i \text{ or } j}\left[\sum_{k=1}^{K}\left(k \cdot \hat{p}_{i,j,k} - \sum_{c=1}^{c} c \cdot \hat{p}_{i,j,c}\right)^2\right]}}$$

A Rasch analysis of similar counting tasks has been reported earlier by Mair [42] in the context of the development of a neuropsychological test battery for number processing and calculation in children (in a Western culture), where the counting of dots belongs to the most elementary tasks [43]. Their studies indicated that, given a limited time to count dots, faster children were found to count some of the points while estimating the rest, while slower children could only estimate. This difference in counting strategy led to indications of different dimensions in the Rasch analysis [42].

According to the present Rasch analysis, the most challenging countings for the Mundurucu are those for dot numbers higher than 7. A striking observation is that even the ‘easiest’ task – counting just one dot – seems to be a bit of a challenge, since the test persons estimate 1 dot as being closer to a count of 2. Dehaene et al. [22] discuss various effects, but our explanation is that there is scale overestimation at the low count, left hand end of the measurement scale – where the respondent refrains from marking the least point on the scale, but places the mark higher. In total, there seem to be three distinct regions when analysing 1–10: (i) low counts overestimation on the screen; (ii) middle section (3–7) fairly accurate counting; and (iii) 8–10 increasingly challenging counting.

For comparison with the following alternative probabilistic approaches, a plot is made (Fig. 4) of the probabilities of success in counting, according to Eq. (3), versus the stimulus values, using the Rasch estimates of person ability θ and level of challenge β across the full set of observations of Mundurucu counting 1,...,10.

4.2.2. Weber resolution

A surprising result of the research of Dehaene et al. [22] is that the familiar logarithmic dependence on stimulus level seen for the human senses seems to apply equally well to counting ability. That is, the human response is proportional to the fractional, rather than absolute, stimulus levels. The constant of proportionality between response and fractional stimulus is called the ‘Weber’ constant, values of which can range from 1 to 0.1 for the acuity of "number sense" for people of different counting abilities (from infants to educated adults) [44]. For the Munduruсu, the Weber constant has been previously estimated as 0.17, corresponding to a threshold of cognitive counting at about 6 or 7 dots, i.e., $w = \frac{7-6}{6} = 0.17$, as done in ref [44].

A first alternative approach to Rasch in estimating the probability of successful counting expresses the ability to distinguish ('resolve') adjacent counts with stimulus values, $s_j = j$, in terms of the distance, $D(R_i)$, between different (Normal) distributions, $P = N(s_j, u_j^2)$, centred at each respective mean count, s, and the perceptive uncertainty $u_j = w \cdot s_j$. One can have different models of how the human brain perceives each cloud of dots when counting: we have adopted a continuous Gaussian distribution, as other researchers have done previously, e.g. in Ref. [44]. This resolution approach is an example of the 'choice paradigm' as dealt with in psychophysics [45]. That is, according to the human observer, which is the most preferred or possesses most of some perceived attribute shared by a series of stimuli?

This approach is applied in the present work, where the ability to resolve adjacent pairs of stimuli (a,b) – i.e. counts – is estimated in terms of the standard uncertainty in the distance given by $u_{a,b} = w \cdot \sqrt{s_a^2 + s_b^2}$. In this dichotomous case, the counting efficiency, $P_{success}$, is then expressed in terms of an error rate defined as the area under the overlap in two count distributions [44]. Assuming Gaussian distributions, this overlap of two Gaussians will also be a Gaussian, for which the probability of successfully resolving the pair of stimuli is estimated as:

$$P_{success} = 1 - 0.5 \cdot erfc\left[\frac{|s_a - s_b|}{\sqrt{2} \cdot w \cdot \sqrt{(s_a^2 + s_b^2)}}\right] \tag{8}$$

For comparison with the other probabilistic approaches, a plot is made (Fig. 4) of the probabilities of success in counting, according to Eq. (8), versus the stimulus values, across the full set of observations of Munduruсu counting $1,\ldots,10$, where the only free parameter is the Weber constant w (= 0.17) As can be seen in Fig. 4, these resolution-based estimates of $P_{success}$ agree well with the other approaches for counts below the 'worst-sorting' threshold, but seem to over-estimate counting success for higher, more challenging counts. Future work would address an extended model where, especially for the higher counts, more than pairs of distributions (such as shown in Fig. 2) centred on adjacent counts would be accounted for in a modified ("multiresolution") version of Eq. (8). Iverson and Luce [45] point out some of the challenges of modelling choice in psychophysics in tasks more complex than pair comparisons of stimuli.

4.2.3. Classification effectiveness in terms of counting bias on ordinal scale

Our first preliminary brief report in a conference proceedings [9] also recalled an approach to overcoming the traditional limitations of statistical measures of location and dispersion on such scales by the inclusion of a cost function as a distance metric. As noted for the PMF shown in Fig. 2, limited measurement quality leads not only to measurement uncertainty but above all to incorrect classification. Following this approach, a measure of the effectiveness, *Eff*, of sorting on an ordinal scale has been proposed [46,47] as:

$$Eff_j = 1 - \frac{EL_j}{EL_{WS}} \tag{9}$$

where EL_j is the expected loss associated with incorrect classification at level j and EL_{WS} is the worst-sorting case loss. In this second alternative approach to Rasch, the probability of successful counting is simply equated with the sorting effectiveness, i.e. $P_{success,j} = Eff_j$. The expected loss, and thereby the classification efficiency, associated with incorrect counts, are expressed as functions of the 'bias' $\varepsilon_c = r_c - s_c$ for each count category, c (rather than using the difference in adjacent counts used in the Weber resolution approach above). The 'worst-sorting' case would corresponding to a completely incorrect sorting, essentially a random guesswork classification, as described in Ref. [46–47], with a uniform PMF across the different categories. A number of cost models have been investigated in the present work: linear $EL_c = Cost \cdot \varepsilon_c$; parabolic $EL_c = Cost \cdot \varepsilon_c^2$ (illustrated in Fig. 2); and exponential $EL_c = Cost \cdot e^{\varepsilon_c}$. One can calculate the expected loss as the accumulated loss per category for each count PMF, (here for the squared model) as [6]:

$$\begin{aligned} EL_j &= \sum_{c=1}^{C} P_{j,M,c} \cdot \mathrm{Cost} \cdot w_{j,c} = \sum_{c=1}^{C} P_{j,M,c} \cdot \mathrm{Cost} \cdot \varepsilon_{j,c}^2 \\ &= \sum_{c=1}^{C} P_{j,M,c} \cdot \mathrm{Cost} \cdot (\nu_{j,c} - \nu'_{j,M})^2 \end{aligned} \tag{10}$$

Referring again to Fig. 4, the probabilities of success in counting, according to Eq. (9), with a squared cost model, are contrasted with the stimulus values, across the full set of observations of Munduruсu counting $1,\ldots,10$. Using other cost models makes little difference (within uncertainties) to the classification effectiveness for these bias-based estimates of $P_{success}$ which agree well with the other approaches for counts below the 'worst-sorting' threshold, but seem to under-estimate counting success for higher, more challenging counts.

4.3. Exploring links between the probabilistic approaches to counting classification

The Rasch model [Section 4.2.1, Eq. (3)] has been connected to the Shannon entropy, H, in communicating information about a binary observation [48] through the relation:

$$\begin{aligned} H(P_{success}) = -[P_{success} \cdot \log(P_{success}) + (1 - P_{success}) \\ \cdot \log(1 - P_{success})] = \int -z \cdot dP_{success} \end{aligned} \tag{11}$$

where $z = \log\left[\frac{P_{success}}{1-P_{success}}\right] = \theta - \beta$, the familiar Rasch psychometric log-odds, is the explanatory variable in

Generalised Linear models behind the success in counting in the present study. One can regard $P_{success}$ as a measure of the dissimilarity ("psychometric function"), that is, the probability of judging that one stimulus is greater than another. The integral of z in Eq. (11) corresponds to estimating the subjective distance, $D(a,b)$ between two perceived stimuli (a,b) by cumulating the psychometric function between the adjacent stimuli in so-called Fechnerian scaling [49]. These relations can be extended to the multinominal, polytomous case, as required.

The logarithmic dependence of counting response on stimulus level, dealt with in Section 4.2.2, is a special case where a change in the psychometric function is proportional to the fractional change in stimulus level – with the Weber constant of proportionality, w [49]. Sun et al. [16] have observed independently that the change in instrument sensitivity (gradient of C (Fig. 3) – termed by them the 'psychophysical scale') – is inversely proportional to the fractional stimulus change when entropy coding is allowed.

The reasonable agreement evident from Fig. 4 between different estimates of $P_{success}$ across the range of counting challenge means that the different expressions (5), (8) and (9) can be equated, in line with our aim of gaining insight into the aspects which unite these different approaches. Each approach has its own version of divergence metric, often invoked in decision-making for instance based on maximum likelihood, and the equalities found imply equivalence between these metrics, which in informational or statistical inference go under various names such as: Information divergence/relative entropy; Kolmogorov/error divergence; χ^2 divergence; etc, but nevertheless seem to give convergent results. As an indication of future work beyond maximum likelihood, reference can be made to the modelling of the adaptation by biological sensory systems, for example, where visual perception adjusts to variations in observation environment. Grzywacz and Balboa [17] postulate that minimising the costs of perceptual task errors modelled in Section 4.2.3, and not only maximising likelihoods, is a major goal of such biological adaptation.

Invariant measure theory, allowing the level of challenge β for a particular task (such as counting a certain number of dots) to be estimated independently of who is doing the counting permits the identification of a metrological standard for counting challenge. Similarly, an estimate of each person's ability θ to count for a range of tasks of different challenge can be metrologically calibrated by measuring a task of known challenge. As in traditional metrology, this traceability allows all the advantages of objectively comparable measurement.

5. Discussion

Early psychometricians extended the Weber–Fechner law from psychophysics to psychological and social constructs. Unexpected new capacities for meaningfully interpreting measures, and for taking missing data into account, emerged in the process.

Qualitatively meaningful measures can be interpreted in terms of experimentally reproducible, invariant hierarchies formed by the difficulty or agreeability of the questions asked in an assessment or survey [50]. Numbers are no longer mere digits but instead denote consistent variation in the thing measured supported by both theory and data, allowing interpretation of the ordered hierarchy as a matter of a learning progression or developmental sequence. Individual measures could now be expressed not only numerically, but in terms of performance levels related to theoretically justified learning progressions [51]. As empirical estimation is complemented by predictive theories, measurement in psychology is advancing into previously unimagined new efficiencies in research and practice [52,53].

Another area of practical innovations concerns the equating of different instruments (tests, surveys, and assessments) intended and shown to measure the same thing [52]. In the same way that data from different sets of examinees or respondents reproduces invariant item hierarchies across samples, so, too, do item hierarchies exhibit shared, invariant features across item sets. Theory-based equating methods can complement data-based methods as predictive control of these hierarchies improves [53]. Linking different instruments measuring the same thing to a common unit of measurement that is substantively interpretable in a shared frame of reference appears to provide a basis for exploring new metrological vistas.

6. Conclusion

This study provides another source of theory and evidence in support of the idea that some form of metrological traceability may be feasible for the objects of psychological and social measurement [7,8]. The feasibility of this aim is further supported by previous work that has similarly employed necessary and sufficient estimators [30,31] of model parameters derived from Rasch's separability theorem [19] to (1) reproduce SI units from ordinal observations of length, density, and weight [54–56], to (2) demonstrate that different rating scale instruments intended to measure the same thing can converge on the same construct and could be equated [7], and to (3) illustrate theoretical control over the reproduction of the measured construct [52,53]. Research in progress is investigating whether the everyday cognitive processes adapted in scientific model-based reasoning [57] might also be identified and emphasized in psychological and social measurement.

This study shows that experimental evaluations of concrete counts have the potential of providing evidence as to the feasibility of invariant units of measurement, calibrated instrumentation, construct theories, and shared frames of reference. A central goal will be practical applications taking advantage of the efficiencies obtained from theory-based reproductions of expected response patterns, situating them in a context of known uncertainties and guidance concerning what to do in the inevitable instances in which expectations are contradicted. Judicious application of a balanced combination of theory, experiment, and calibrated instrumentation is a shared interpretive framework may enable more widespread appreciation of the value in Feynman's point that, "What I cannot create, I do not understand" [58], and may expand the horizons of research in psychology and the social sciences [59].

Acknowledgement

Financial support is gratefully acknowledged from the Swedish National Metrology Program.

Part of this research was performed in the project NEW04 Uncertainty of the European Metrology Research Programme (EMRP, FP7 Art. 185) which is jointly funded by the EMRP participating countries within EURAMET (www.euramet.org) and the European Union.

References

[1] B. Berglund, G.B. Rossi, J. Townsend, L.R. Pendrill, Theory and Methods of Measurements with Persons, Taylor & Francis, New York, 2011.

[2] W. Zhang, Y. Yang, Q. Wang, F. Shu, in: 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>.

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

[4] J. Schmidhuber, J. SICE 48 (1) (2009) 21–32.

[5] L. Mari, Measurement 25 (1999) 183–192.

[6] L.R. Pendrill, AMCTM 2011 International Conference on Advanced Mathematical and Computational Tools in Metrology and Testing, Göteborg, 2011, June 20–22, <http://www.sp.se/AMCTM2011>.

[7] W.P. Fisher Jr, J. Outcome Meas. 1 (2) (1997) 87–113.

[8] L.R. Pendrill, B. Berglund, et al., NCSLi Meas. 5 (2010) 42–54.

[9] L.R. Pendrill, W.P. Fisher Jr, J. Phys.: Conf. Ser. 459 (2013) 012057, http://dx.doi.org/10.1088/1742-6596/459/1/012057.

[10] B.D. Wright, Phys. Med. Rehabil. State Art Rev. 11 (1997) 261–288.

[11] A.M. Davis, A.V. Perruccio, M. Canizares, et al., Osteoarthritis Cartilage 16 (2008) 551–559.

[12] S.S. Stevens, Science 103 (2684) (1946) 677–680.

[13] M. Wilson, Measurement 46 (2013) 3766–3774, http://dx.doi.org/10.1016/j.measurement.2013.04.005.

[14] L. Mari, Measurement 27 (2000) 71–84.

[15] B. Berglund, Chapter 2 in Ref. [1], 2011.

[16] J.Z. Sun, G.I. Wang, V.K. Goyal, L.R. Varshney, J. Math. Psychol. (2012), http://dx.doi.org/10.1016/j.jmp.2012.08.002.

[17] N.M. Grzywacz, R.M. Balboa, Neural Comput. 14 (2002) 543–559.

[18] B.D. Wright, The new rules of measurement: What every educator and psychologist should know, in: S.E. Embretson, S.L. Hershberger, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1999, pp. 65–104

[19] G. Rasch, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV, University of California Press, Berkeley, California, 1961.

[20] D. Andrich, Med. Care 42 (2004) 1–16.

[21] R. Burro, R. Sartori, G. Vidotto, Qual. Quant. 45 (2011) 43–58. <http://dx.doi.org//10.1007/s11135-009-9282-3>.

[22] S. Dehaene, V. Izard, E. Spelke, P. Pica, Science 320 (2008) 1217–1220.

[23] G. Cooper, W.P. Fisher, Session on Fundamentals of measurement science. International Measurement Confederation (IMEKO). Jena, Germany, August 31 to September 2, 2011. <http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-24494/ilm1-2011imeko-019.pdf>.

[24] B.D. Wright, Rasch Meas. Trans. 3 (1989) 62.

[25] A. Nieder, E.K. Miller, Proc. Natl. Acad. Sci. U.S.A. 101 (2004) 7457.

[26] S. Masin, C.V. Zudini, M. Antonelli, J. History Behav. Sci. 45 (2009) 56–65.

[27] D. Andrich, Appl. Psychol. Meas. 2 (1978) 449–460.

[28] J.P. Bentley, Principles of Measurement Systems, forth ed., Pearson Education Limited, London, 2005. www.pearsoned.co.uk, ISBN 0 130 43028 5.

[29] L. Mari, A. Giordani, Metrologia 49 (2012) 756–764. <http://dx.doi.org//10.1088/0026-1394/49/6/756>.

[30] E.B. Andersen, Psychometrika 42 (1977) 69–81.

[31] G.H. Fischer, Psychometrika 46 (1981) 59–77.

[32] J.M. Linacre, A user's guide to WINSTEPS Rasch-Model computer program, v. 3.72.0. Chicago, Illinois, 2011, Winsteps.com.

[33] T. Bond, C. Fox, Applying the Rasch model, 2d ed., in: D. Andrich (Ed.), 1982 Education Research and Perspectives, Lawrence Erlbaum Associates, Mahwah, New Jersey, 2007, pp. 95–104.

[34] J.M. Linacre, Rasch Meas. Trans. 23 (2010) 1238.

[35] J.C. Nunnally, Psychometric Theory, McGraw-Hill, New York, 1978.

[36] C.E. Lance, M.M. Butts, L.C. Michels, Organ. Res. Method 9 (2006) 202–220.

[37] K. Sijtsma, Psychometrika 74 (1) (2009) 107–120.

[38] J.M. Linacre, Rasch Meas. Trans. 7 (1) (1993) 283–284.

[39] W.P. Fisher Jr., B. Elbaum, W.A. Coulter, J. Phys. Conf. Ser. 238 (1) (2010). <http://iopscience.iop.org/1742-6596/238/1/012036/>.

[40] R.M. Smith, J. Appl. Meas. 1 (2) (2000) 199–218.

[41] B.D. Wright, Rasch Meas. Trans. 9 (4) (1996) 472.

[42] P. Mair, Example analysis, Part 7: The Rasch Model in Practice, Psychometric Methods 2010/11, Vienna University of Economics and Business, 2010, <http://statmath.wu-wien.ac.at/~hatz/psychometrics/10w/RM_handouts_7.pdf>.

[43] M. Von Aster, M. Weinhold Zulauf, R. Horn, ZAREKI-R (Neuropsychological Test Battery for Number Processing and Calculation in Children), revised version, Harcourt Test Services, Frankfurt, 2006.

[44] J. Halberda, L. Feigenson, Dev. Psychol. 44 (2008) 1457–1465.

[45] G. Iverson, R.D. Luce, Measurement, Judgment, and Decision Making, in: M.H. Birnbaum (Ed.), New York, Academic Press, 1998, pp. 1–79.

[46] E. Bashkansky, S. Dror, R. Ravid, P. Grabov, Qual. Eng. 19 (2007) 235–244.

[47] E. Bashkansky, T. Gadrich, Accred. Qual. Assur. 15 (2010) 331–336, http://dx.doi.org/10.1007/s00769-009-0620-x.

[48] J.M. Linacre, Rasch Meas. Trans. 20 (3) (2006) 1062–1063.

[49] E.N. Dzhafarov, Chapter 9 in Ref. [1], 2011.

[50] Wright B D, Masters G N 1982 Rating scale analysis Chicago, MESA Press.

[51] M.R. Wilson, J. Res. Sci. Teach. 46 (2009) 716–730.

[52] B.D. Wright, M.H. Stone, Best test Design Chicago, MESA Press, 1979.

[53] A.J. Stenner, W.P. Fisher Jr, M.H. Stone, D.S. Burdick, Frontiers Psychol.: Quant. Psychol. Meas. 4 (536) (2013) 1–14.

[54] A. Stephanou, W.P. Fisher Jr., J. Phys. Conf. Ser. 459 (2013). <http://dx.doi.org/10.1088/1742-6596/459/1/012026>.

[55] T. Pelton, V. Bunderson, J. Appl. Meas. 4 (3) (2003) 269–281.

[56] S.E. Choi, Rasch Meas. Trans. 11 (2) (1997) 557.

[57] N.J. Nersessian, Philos. Sci. 73 (2006) 699–709.

[58] S.W. Hawking, The Universe in a Nutshell, Bantam Books, New York, 2001.

[59] W.P. Fisher Jr, Measurement 42 (9) (2009) 1278–1287.