Man as a Measurement Instrument

Leslie Pendrill

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 technical products of all kinds who realize the need to assure the quality of their products as perceived by the customer. The metrology of human-based observations is, however, in its infancy. This article 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 for establishing measurement references, and the metrological quality assurance that is urgently needed in many contemporary applications.

1. Contemporary Shift in Metrology to Include Human Beings

We are all familiar with, and mostly take for granted, how engineered instruments provide a more reliable means of probing an object than relying on our “five human senses.” But the human factor in measurements is seldom completely eliminated, no matter how advanced the technology. Thus, there is a clear need to characterize measurement systems where Man enters, not merely as a passive operator, but in two radically different and active ways (Fig. 1):

- Measuring Man: A human is the measurement object itself.
- Man as a Measurement Instrument: Instead of engineered instruments, human senses and cognition are used to measure objects [1].

This contemporary shift to include human-based measurement aims to meet the demands not only in sectors such as health care, services and safety, where the human factor is obvious, but also from an increasing numbers of manufacturers who need to assure the quality of their products as perceived by the customer.

The reliable characterization of the human measurement instrument [16], be it with the five senses or the full physiological, mental,

cognitive, and behavioral richness of human perception, is essential in many applications (Table 1). About ten years ago, one high-volume consumer electronics manufacturer reported that the percentage of products returned by dissatisfied customers with ‘no fault found’ had risen above 50 % [2]. A disabled, ill or elderly person can be helped to cope better with everyday tasks with better measurement of their abilities, the levels of challenge they face, and the performance of assistive products [4, 5]. Various human functions can be enhanced, for instance with machine learning [8], to assist in mining the ever increasing amounts of information available in society. Today’s citizens need basic mathematical and numerical skills to understand information presented by politicians, insurance companies, financial advisors, marketers, and the like [7]. However, consumers differ considerably in their ability to understand and use such information. In many of these applications, the problems are not merely technical but instead related to comfort, pleasure, beauty, anxiety, [3, 17] and other measures of human ‘capital’ [18].

Beyond ordinary descriptive approaches to human perception, other fundamental concepts also underlie predictive approaches, such as evidence theory and expert elicitation [19]; as well as prescriptive discrete choice

that aims to explain and predict, for instance, commuter travel behavior in terms of utility maximization [13, 14], prospect theory, and intuitive choice [15].

2. Demands for Quality Assurance of Human-Based Measurement

As in more traditional areas of metrology, the quality assurance of less qualitative measurements is considered essential to assuring the quality of products. However in human-based measurement, the formulation of concepts such as measurement uncertainty and traceability is as yet in its infancy [12, 18, 20]. This paper attempts to bring together recent and diverse approaches from several disciplines to help establish concepts, tools and procedures for assuring the quality of measurements in the burgeoning area of Measuring Man.

Author

Leslie Pendrill
leslie.pendrill@sp.se

SP Technical Research Institute
of Sweden
National Metrology Institute
Box 857, SE-50115 Borås
Sweden

Figure 1. Different human interventions in measurement systems.

Quality management systems have been regulated for over half a century with the help of certification systems and written standards such as the famous *ISO-9000* series [21]. A practice common to all *ISO-9000* standards for implementing quality management systems is to perform a number of steps in the Deming ‘quality loop’ assuring quality at all stages in the lifetime of any product or service, from the initial customer request to final disposal and recycling. In addition, these standards usually require quality-assured measurements.

2.1 Quality-assured measurement for human-based services

The emergence of quality assurance for predominantly human-based activities can be exemplified by recent regulations and standards in health care, be it services - with the new European norm *EN15224:2012* [22] – or claims for novel or improved medical products and devices – such as the FDA regulations based on person-reported outcomes [23] and usability engineering [24].

If the aim is to improve the treatment of patients and promote health, perhaps in line with the person-centered approach to care [25], then key metrics (step d) in the Deming loop [26]) will typically be: experience of care; shared decision making; shared goal setting; patient inclusion in the health care team; patient knowledge and understanding of the care plan; clinical communications; and support for self-care and quality of life. Alongside traditional metrics in medical physics, such as blood pressure, body temperature, etc., the new person-centered indicators might also include quantities such as satisfaction, anxiety, and so on.

<i>Response</i>	<i>Item attribute</i>	<i>Person characteristic</i>	<i>Applications (examples)</i>
Satisfaction	Quality of product	Customer leniency	Consumer electronics [2]; Cosmetics [3]; Health care [4, 5]; Services [6]
Performance of task	Level of challenge of activity	(Dis-)ability	Citizen’s understanding and information [7, 8]; Learning [9]; Psychometrics [10, 11]; Rehabilitation [12]
Accessibility (e.g. of transport mode)	Barrier hinder (or cost)	Utility (or net benefit, well-being, disability,...)	Commuter traffic [13] ; discrete choice and valued prospects [14, 15]

Table 1. Coupling item attributes to person characteristics in diverse responses for various applications.

Once these person-centered indicators are identified, the next challenge is posed at Deming loop step g) monitoring and measuring health care systems, as required e.g. in §8.2.4 of the new *EN15224:2012* [22], namely to verify that product or service requirements have been met. No decisions about the significance of apparent differences in measurements can be made without demonstrating measurement reliability and metrological traceability. Therefore, as in other *ISO-9000* like quality management system standards, the new health care service quality standard requires [§7.6 of *EN15224:2012*]:

7.6 Control of monitoring and measuring equipment

The organization shall determine the monitoring and measurement to be undertaken and the monitoring and measuring equipment needed to provide evidence of conformity of product (health care service/other health care product) to determined requirements.

...Where necessary to ensure valid results, measuring equipment shall:

a) be calibrated or verified, or both, at specified intervals, or prior to use, against measurement standards traceable to international or national measurement standards; where no such standards exist, the basis used for calibration or verification shall be recorded...

2.2 Person-Reported Outcomes

In another relatively new approach to health care in line with person-centered care, regulators such as the FDA [23] have placed increased emphasis on patient-reported outcomes. Alongside professional judgment, a patient-reported outcome can be any report of the status of a patient’s health condition that comes directly from the patient, without interpretation of the patient’s

response by a clinician or anyone else. Such outcomes can be measured in terms of specific quality characteristics, either in absolute terms (e.g., severity of symptom, sign, or state of disease), or as a change from a previous measure.

Many such person-reported outcomes – both in the quality assurance of health care services and medical products – involve Man both as a measurement instrument and as the object of measurement by various instruments. Novel methods for metrological quality assurance urgently need to be developed for the less quantitative results obtained with human-based measurement systems, not only in health care, but in many human-factor areas of application (Table 1).

3. Metrology Including Human-Based Measurement

3.1 Metrological Quality Assurance in Human Measurements

Two key defining concepts of metrology – metrological traceability and measurement uncertainty – are as yet barely established in human-based measurement. As pointed out by Fisher [12, 18]: “A metrological approach (to human-based measurement) comes at a time where trends in contemporary 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 increase the need for coherent and comparable measures. 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.”

A principal challenge is that there are few, if any, recognized metrological standards to establish traceability when assuring human-based quality, for instance to assure that patients can expect the same level of health care wherever it is provided. Health care has been described as a “\$1 trillion per year industry without a clear measure or definition of its main product” [27].

Apart from a considerable lack of measurement standards, the second important aspect of metrology – measurement uncertainty – also presents challenges in the context of human-based measurements. Measurement uncertainty in a test result is central to conformity assessment by inspection because uncertainty can, if not accounted for,

- lead to incorrect estimates of the consequences of entity error, and
- increase the risk of making incorrect decisions, such as failing a conforming entity or passing a non-conforming entity when the test result is close to a tolerance limit.

It is still common, however, to find incorrect analysis of the scores obtained with questionnaires and similar instruments that are used to measure human response. The challenge [28] is that several common statistical tools cannot be used to characterize the location and dispersion of qualitative measurements on ordinal scales [29] that are typical of such measurements.

Thus, whether the goal is to compare product or service characteristics or to set limits on decision risks in human-based measurement, it is necessary to determine the corresponding metrological traceability and measurement uncertainty of these less quantitative observations.

3.2 Measurement System Analysis Including Man

An essential first step in any quality-assured measurement is to make a complete analysis of the actual measurement situation. The Measurement System Analysis (MSA) approach is widely used, for instance, in the automotive industry [30]. It is based on a model, illustrated in Fig. 1, 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 affect the main system elements when determining overall measurement quality. In fact, establishing a measurement system which is fit-for-purpose is arguably the foremost task of an applied metrologist [31].

Metrological characterization based on a measurement system analysis includes both traceability and uncertainty, respectively:

- In principle, each element of a measurement system can be calibrated – not only the instrument, but also the method, operator, or particular object intended as a metrological standard.
- The uncertainty associated with any measurement result is an indicator of the overall quality of the measurement, which in turn is determined by the performance of various elements of the measurement system; i.e. the instrument, method, operator, and environment, as well as the measured object (‘entity’) itself.

In ‘traditional’ measurement systems, where the main human intervention is as a mere operator, there are still some human aspects, such as including prior measurement knowledge by the operator in a probabilistic approach to the expression of measurement uncertainty [32]. In the MSA approach, the contribution of the operator to the overall performance of the measurement system can be evaluated (with ANOVA for example) as a so-called appraiser variation [30].

The main focus of the present review will be on the second type of measurement system, where the instrument in the measurement system (Fig. 1) is replaced by a person playing a critical role at the heart of the measurement system. Links can be sought between psychometric characterization of Man as a measurement instrument and three major disciplines:

- classical engineering characterization of measurement instruments (in terms of resolution, sensitivity, linearity, bias, etc.) [11, 33, 34] – see Section 4.1
- measurement systems as special cases of information systems, linking loss of information on ‘transmission’ to uncertainty; entropy; and the risks of incorrect decisions [35, 36] – see Sections 4.2 and 4.3, and
- key mathematical and statistical concepts, including logistic regression [9], and generalized linear modelling – see Section 4.4.

In a third kind of measurement system - the case of Measuring Man - the measurement object (Fig. 1) is a person, with all the complexity that being a person implies. In those cases where the characterization of a human being as a measurement instrument is of prime importance, measurement instruments will then in turn be deployed to probe a human being. Examples include not only the application of sensors to the human body [1] but also other ‘instruments’ administered, such as questionnaires in opinion polls and surveys, and examinations in school.

4. Analysis of Perceptive Data

4.1 Ordinal Data

Typical of subjective scoring is where responses (e.g. from a patient assessment of health care service quality) on a five-category Likert scale [Fig. 2] vary across the range. Distances between marks on the scale do not necessarily follow an exact or known mathematical model. People can place different emphasis on the different parts of the scale: ‘monomodal’ responders tend to avoid scoring at either extreme of the scale (over-estimating grade ‘1’ and under-estimating grade ‘5’), while others will tend to avoid scores in mid-range, preferring to make strong statements (‘bimodal’). A recent example of such behavior can be found in studies [38] where responses to positive affect (e.g. feeling “enthusiastic”) tend to be monomodal, while the same person might respond bi-modally to questions of negative affect (e.g. feeling “guilty”). Miss-scoring of this kind needs to be evaluated and corrected for; otherwise the reliability of decisions based on data may be compromised. On these so-called ‘ordinal’ scales, where measurement values can be ordered but lack the exact arithmetic characteristic of higher-order scales such as the interval and ratio scales [29], the classic expressions of statistics (such as calculations of a mean or standard deviation) cannot be applied [28].

Not only are new statistical tools needed, but at the same time the traditional engineering approach has to be extended to include characterization of Man as a measurement instrument, pictured symbolically in Fig. 3. Examples of recent descriptions in this area include an introduction to the measurement of psychological attributes [10, 34]. The latter draw analogies, for example, to a mechanical spring acting as a sensor in attempting a description of a human instrument in terms of its sensitivity (C), i.e. the change in response (R) to a change in stimulus (S).

4.2 Decision-Making and Qualitative Observations

When tackling the challenge of assuring the quality of less quantitative observations, an intimate connection [40] is 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.

An early definition of qualitative testing in analytical chemistry: “The classification of objects against specified criteria to meet an agreed requirement” [41, 42], reflected this connection with decision-making. 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 [36].

A key insight is that the human instrument is not only a sensor but additionally includes a decision-making algorithm. 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 the recent document [43] that accompanies the *Guide to the Expression of Uncertainty in Measurement* (GUM). 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)

Figure 2. Example of subjective scoring [37].

Figure 3. Simple psychophysical model of human perception (adapted from [39]).

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.

The mutual link between decision-making and treating comparative or merely qualitative observations for nominal and ordinal properties can be made explicit in the framework of probability theory – see Appendix A.1 for details.

Our approach in this decision-making description of measuring qualitative quantities focuses exclusively on measurement uncertainty in the stimulus variable (associated with limited measurement information about the entity or object of measurement) that leads to decision risks in the response variable (associated with the output of the human instrument). Other factors that might lead to dispersion in the instrument output can be considered when attempting to separate person and item attributes, such as in the Rasch [9] approach described below in Section 5.6.

4.3 Loss of Information and Decision-Making, Entropy and Generalized Linear Models

An ideal tool for handling qualitative data can be found in information theory, where ‘entropy’ is the amount of information contained in a message and is a measure of the degree of order in a sign or perceived shape of any kind. An information-theoretical approach is preferred to techniques such as mismatch counts that are utilized

in areas such as DNA sequencing [44]. Another example is feature selection in machine learning, which can be utilized for the automatic classification of (non) functional requirements in high volumes of text by including semantic entropy [8].

The concept of ‘uncertainty’ in information theory [35], which refers to a loss of information in ‘transmission’ (i.e., when making decisions), enables connections to be made with the decision risks discussed above. This formulation is quite general and allows various approaches such as theories of probability, fuzzy logic, plausibility, evidence, and so on [35]. See Appendix A.2 for details of a probability based treatment.

4.4 Perceptive Choice, Logistic Regression and Generalized Linear Models

Finally, combining the above approaches of decision risks and information theoretical entropy loss with the formulation of a psychometric function [Dzhafarov, 16] – see Appendix A.3 for details – leads to a logistic regression expression (Eq. A.6) that is suitable for the proper treatment of ordinal data. This is an example of a generalized linear model. A fundamental observation is that information loss involves not only measurement uncertainty, but also contains some measure of the risks of incorrect decisions.

In this section we have shown, by connecting decision risks when handling qualitative observations with information theory, that perceptive choice and generalized linear modelling are essential tools for characterizing Man as a measurement instrument.

5. Reliability, Uncertainty and Metrological Traceability in Human-Based Measurement

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 [45] to customer satisfaction surveys to person-centered health care (Table 1).

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 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. 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 [46].

One particular version of logistic regression (Section 4.4) that has received considerable attention as a tool for handling ordinal data is due to the Danish statistician Rasch [9]. In response to criticism of psychometric methods of the time, Rasch explicitly attempted a separation in measured responses according to Eq. (A.6) into a person attribute value (θ , such as ability or leniency) and an item attribute value (β , such as level of challenge or quality, respectively) by writing $z = \theta - \beta$. In the simplest, dichotomous case the logistic regression function is [9],

$$\theta - \beta = \log \left[\frac{P_{\text{success}}}{1 - P_{\text{success}}} \right]. \quad (1)$$

The Rasch approach 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. Figure 4 exemplifies this by assessing a patient’s ability to perform a number of tasks over a range of levels of challenge, perhaps following treatment, while Table 1 summarizes the wide range of item and person attributes amenable to this approach.

5.2 Elementary Tasks of a Human Instrument

When attempting to formulate metrological concepts in perceptual contexts, a useful initial approach is to study elementary human tasks, such as counting or recognizing shapes of clouds of dots (Fig. 5). Here, the expected value is known because the measurement object (collection of dots in this case) has been prepared in advance [33, 39, 48, 49, 50]. Once the analysis of such elementary situations has been mastered, we can tackle conceptually more challenging tasks in applications that invoke human-based measurement (Table 1).

Most readers will easily be able to count three dots or five, but ten dots might present problems, especially with just a brief glance at the rightmost picture in Fig. 5(a). Particular groups of people – young children; dementia sufferers; adult Mundurucu Indians, as studied by Dehaene [48] – can have increasing difficulties counting larger numbers of dots. Similarly, even highly educated and capable people will generally have difficulty distinguishing between the ellipses of lower correlation (values of r between 50 % and 10 %) of Fig. 5(b).

A surprising result, for example, of the research of Dehaene [48] is that the familiar logarithmic dependence of human response on stimulus level seen for the human senses (such as in acoustics) seems to apply equally well to counting ability. That is, the human response is proportional to fractional, rather than to absolute, stimulus levels. Note that this logarithmic dependence, usually

referred to as the Weber-Fechner law, is not primarily related to the log term of Eq. (A.6) which is more general, but can be derived in the particular case where a change in the psychometric function is proportional to the fractional change in the stimulus level (Appendix A.3).

The probability of a correct response, P_{success} , in each case (successful counting or ellipticity estimation) depends on measures of location and dispersion of perceptive judgments over a range of stimulus values. These values are investigated in terms (resolution and bias) that would be used when characterizing a measurement instrument (in this case, the human perceiver). Three different ways of characterizing these series of observations, have been shown [33, 50] to provide comparable estimates of P_{success} over a range of actual stimulus values (e.g. from $r = 0$ to 100 %):

- i. A Rasch analysis, Eq. (1), using the WINSTEPS software [52] provides separate estimates of person ability, θ , (blue) and task challenge, β , (red) (Fig. 6 in response of ellipticity perception) on a common and quantitative scale.
- ii. The ability to resolve adjacent pairs of stimuli, calculated with Eqs. (A.1) and (A.3).
- iii. The effectiveness of classification, calculated as the ratio of the consequence cost of misclassification [33, 53]

The measurement scale in a Rasch plot, such as the one shown in Fig. 6, is established for the particular group of items and persons studied. If new items and/or test persons are added, both the span and the centering of the scale can be adjusted as necessary. If there appear to be not enough people in the study being challenged by items of corresponding difficulty, a Rasch analysis can suggest how the study may be complemented. An example is the Rasch analysis of vision tests for cataract sufferers; the original tests were conducted at a time when clinical treatment was only justified above a certain level of vision loss, and are today not considered to be sufficiently challenging. Modern treatment can be administered at significantly lower levels of vision loss because early treatment is now considered both feasible and beneficial [54].

Figure 5. Series of elementary visual tasks with known prior and increasing challenge: (a) counting m dots (adapted from [33, 48]) and (b) estimating degree (r) of ellipticity (adapted from [50, 51]).

Figure 6. Separate estimates of person ability (blue) and task challenge (red) in response of ellipticity perception with a Rasch model over a range of actual stimulus values from $r = 0$ to 100 % [50].

We now return to our quest for ways to formulate concepts for quality assurance of human-based measurements.

5.3 Elementary Tasks of a Human Instrument
Measurement uncertainty, which is needed to make statements about apparent

differences between values, such as the relative probabilities or quality scores of the successful sorting amongst different categories, can be dealt with in Rasch modelling as follows.

For dichotomous observations by person i of item k , the scored response

$$v_{i,k} = P_{i,k} \pm \sqrt{P_{i,k} \times (1 - P_{i,k})}, \quad (2)$$

where $P_{i,k}$ is the probability of success, $P_{success}$, given in Eq. (1). The binomial error distributions for dichotomous scores approximate Gaussian distributions when accumulated across all the observations, as they are in the estimation process [55].

Because of limited reliability in most measurements, the measured item attribute value β in the Rasch expression, Eq. (1), is determined as a ‘consensus’ value over a population of test persons acting as measurement instruments, and differs from the ‘true’ β' , with a measurement error ϵ_β , $\beta = \beta' + \epsilon_\beta$.

A gauge of measurement uncertainty in the estimated Rasch attribute values θ and β is expressed as a standard error, $SE(\theta_i, \beta_j) = \sqrt{\frac{1}{\hat{P}_{i,j} \times (1 - \hat{P}_{i,j})}}$ derived from the Fisher information in the context of maximum likelihood estimation, for instance for a particular item j ,

$$u(\epsilon_\beta)_{all;j} = \text{realSE}(\beta)_{all;j} = \frac{1}{\sqrt{\sum_{i=1}^{TP} \left[\sum_{k=1}^K k^2 \times q_{i,j,k} - v_{i,j}^2 \right]}}. \quad (3)$$

In traditional factor analysis there remains some contention in the literature about the best expression for reliability, such as the Cronbach α factor. In contrast, Rasch analysis includes a clear separation of person and item scatter. Therefore, an accepted reliability coefficient is simply the fraction of the total variance in an attribute associated with the actual item variance, as opposed to measurement uncertainty scatter,

$$R_\beta = \frac{\text{True variance}}{\text{Observed variance}} = \frac{\text{var}(\beta')}{\text{var}(\beta)} = \frac{\text{var}(\beta) - \text{var}(\epsilon_\beta)}{\text{var}(\beta)}. \quad (4)$$

A typical target value for this reliability coefficient is $R_\beta = 0.8$ which corresponds to a measurement uncertainty not exceeding half of the sought product variation. This kind of metrological conformity assessment in terms of a maximum permissible uncertainty is needed to limit decision risks [40].

5.4 Metrological Units, References and Traceability

Without access to recognized metrological standards, it will be difficult to judge objectively the relative levels of quality of products and services appraised by humans in the myriad of applications indicated in Table 1. Fortunately, it turns out that the Rasch [9] approach, with its explicit separation of person and item attribute estimation, is well suited for introducing metrological traceability to human-based measurement, as follows.

Since invariant measurement theory allows the level of challenge β for a particular task to be estimated independently of whoever is encountering the challenge, a metrological standard for task challenge can be identified. Once an agreed definition and realization of a set of standard tasks has been achieved, it can then be used as a reference in other situations. As in traditional metrology, this traceability allows measurements to be objectively compared. For example, a range of items, such as diabetes care clinics [56], can be compared in terms of the perceived quality of their services.

Having access to a calibrated set of psychometric item challenge standards would also allow a calibration of each person’s ability θ (acting as a ‘measurement instrument’) to negotiate items for a range of different challenge. This procedure determines the corresponding measurement error, ϵ_θ , in person ability: $\theta = \theta' + \epsilon_\theta$ in cases where Man acts as a measurement instrument. Such a calibrated instrument can be subsequently used to define new item standards, and so on.

5.5 Quantities and Units in Human-Based Metrology

What are the similarities and differences between this kind of psychometric standard and more traditional measurement standards in physics and engineering? Humphry [20], when considering units in psychometry and quantities versus numerical values, points out that we cannot raise a number to the power of a dimensioned quantity value such as done in the Rasch expression,

$$P_{success} = \frac{e^{\theta - \beta}}{1 + e^{\theta - \beta}}. \quad (5)$$

In its logistic regression form (Eq. 1), the ‘straight ruler’ aspect of the Rasch formula has been described by Linacre and Wright [57] in the following terms: “The mathematical unit of Rasch measurement, the log-odds unit or “logit”, is defined prior to the experiment. One logit is the distance along the line of the variable that increases the odds of observing the event specified in the measurement model by a factor of 2.718..., the value of “e”, the base of “natural” or Napierian logarithms used for the calculation of “log-” odds. All logits are the same length with respect to this change in the odds of observing the indicative event.”

If the Rasch attributes for persons, tasks or products are ‘quantities’ as opposed to mere numerical values, then there should be metrological ‘references’ associated with them if we are to be consistent with the definition of ‘quantity’ in the international metrology vocabulary VIM [§1.1, 58]. In a note to that definition, a ‘reference’ in this context can be a “measurement unit, a measurement procedure, a reference material, or a combination of such.” Access to metrological references for psychometric quantities would – in addition to the mathematical logit units – enable the scales of different ‘rulers’ for a given quantity, e.g. person ability or task challenge, to be objectively compared with each other.

Thus, the measurement units (discussed for instance by Humphry [20]) associated with the Rasch attribute parameters θ and β , should be intimately related to metrological traceability and measurement standards. Perhaps the closest analogies to references in psychometrics can be found with reference materials that are utilized as references for metrological traceability in chemistry. 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. 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.

One of the seven base units of the International System (SI) – the unit for luminous intensity – is related to human perception. For the SI units of vision, the human aspect of visual perception of luminous intensity is dealt with [59] by separate references for ‘items’ and ‘people’, respectively:

- the objective physical definition of a standard emitter of a certain number of watts in the SI unit for radiometry, the candela; and
- two action spectra been defined by the International Commission on Illumination (CIE): $V(\lambda)$ for photopic vision and $V'(\lambda)$ for scotopic vision.

With access to measurement references, a calibration process in education and learning can be described as follows: When applying the Rasch model, item parameters are first scaled in a process called calibration – for instance, the difficulty of an item in an educational test can initially be assessed by the teacher creating the exam by the proportion of correct responses obtained in previous classes. However, the assessment of difficulty can also include construct modelling, where the construct can be a simple linear succession of the discrete segments of a continuum, arranged in a “construct map” from the lowest to the highest level [60].

Once such references are established and accepted, then a system of calibrations would be set up to ensure that any psychometric measurement result could be traced to a reference, where metrological traceability is defined as a “property of a measurement results whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty [58, §2.41].”

In education, such traceability would ensure fairness to each new class of students, and also assure that all exams conducted nationally or internationally are comparable.

An example (Fig. 7) of an early study of the metrological aspects of Rasch analysis is that made by Fisher [12] in which estimates for up to ten different instruments (i.e. surveys, questionnaires, or examinations) rating physical disability of the probability, P_{success} , of succeeding with a number of tasks over a range of levels of challenge were compared – from the easiest (feeding) to the most challenging (climbing stairs) at the far left of the horizontal axis. The level of agreement and the amount of scatter are indications of the metrological quality of the various instruments.

5.6 Quantities and Concepts in Conformity Assessment and Metrology
Comparative studies (such as Pendrill and Fisher Jr. [33]) can support the understanding

Figure 7. Estimates of probability, P_{success} , of succeeding with a number of tasks over a range of levels of challenge for up to ten different instruments rating physical disability, adapted from [12].

and applicability of the Rasch model in various cases. As in other scientific areas, the models utilized as ‘working hypotheses’ are considered acceptable as long as empirical evidence doesn’t refute them. We should avoid too liberal elimination of outliers, and there are, of course, cases where the Rasch model in its simplest form is not applicable. The wider validity of the Rasch model in psychometrics is under debate - recommended readings include both a paper by Humphry [20] and a series of comments in the same issue about his paper which contain many contemporary concerns about the discipline. As pointed out recently [61], there are claimed to be some similarities between a Rasch model and additive conjoint measurement models [62]. While some claim that relating Rasch modelling to representational measurement theory (RMT) puts psychometric measurement on the same plane as more quantitative interval-level measurement, others point out – so-called ‘strict representationalists’ – that probabilities are not the observable empirical entities referred to in RMT [61].

The intimate connection between the treatment of less quantitative observations and decision-making (Section 4.2) implies that some care is needed in handling concepts, definitions and nomenclature at the interface

where two disciplines – metrology and conformity assessment – meet. Two distinct but closely related concepts coming from the two disciplines are:

- A ‘measurand’ is a quantity intended to be measured
- A ‘quality characteristic’ is a quantity intended to be assessed

In an introduction to the Rasch measurement approach, Mari and Wilson [34] 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: “(a) the presence of an underlying unobserved (‘influence’) variable; (b) a non-deterministic dependence of the ‘indication on the measurand’ (i.e. attitude in this case); or (c) that the ‘measurand’ is itself stochastic.”

To clarify concepts and terminology, consider how, in the three types of measurement systems illustrated in Fig. 1, different quantities are in focus. For instance, in the elementary studies of counting or ellipticity perception, an initial measurand is

the number or degree of correlation in the clouds of dots shown in Fig. 5. Man, in this case acting as a measurement instrument, yields estimates of the value of each measurand. But what is interesting is how well these measurements are performed, not the number of dots since we know this already. The ability to perform such measurements is then 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?” In such a decision-making process, an assessment of quantity is made by the human observer, so the object attributes – the number or ellipticity of clouds of dots – become quality characteristics instead. The exercise will in turn, through a Rasch analysis of the probability of success as decision risks, provide estimates of new measurands, namely: (a) the ability of each human instrument and (b) 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.

6. Conclusions

Despite advanced technology, the human factor is still present in many diverse measurement applications. Thus, there is a clear need to be able to treat measurement systems where Man enters, not merely as a passive operator but in two, radically different and active ways – either as the object of a measurement or as a measurement instrument.

Quality assurance of human-based measurements, on which the quality assurance of products and processes rests, is as yet in its infancy. The psychometric Rasch invariant measurement method, which is described here in terms of perceptive choice and information entropy, luckily opens up opportunities of introducing the metrological concepts of uncertainty and traceability. Metrological units related to certified reference materials in chemistry seem to be the closest analogy.

7. Acknowledgements

Particular thanks are due to Steve Ellison at LGC, William P. Fisher Jr., at Berkeley, and Antonio Possolo at NIST for fruitful discussions. The author is grateful for the invitation to present this paper in the Spanish-language *e-medida* and the *NCSLI Measure* journals. Part of 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.

8. References

- [1] L. R. Pendrill, R. Emardson, B. Berglund, M. Groning, A. Hoglund, A. Cancedda, G. Quinti, F. Crenna, G. Rossi, J. Drnovsek, G. Gersek, T. Goodman, S. Harris, G. van der Heijden, K. Kallinen, and N. Ravaja, “Measurement with Persons: A European Network,” *NCSLI Measure J. Meas. Sci.*, vol. 5, no. 2, pp. 42-54, June 2010.
- [2] E. Ouden, Y. Lu, P. Sonnemans, and A. Brombacher, “Quality and Reliability Problems from a Consumer Perspective: An Increasing Problem Overlooked by Businesses?,” *Qual. Reliab. Eng. Int.*, vol. 22, no. 7, pp. 821-838, 2008.
- [3] A. Klöcker, C. Arnould, M. Penta, and J. Thonnard, “Rasch-built measure of pleasant touch through active fingertip explorations,” *Frontiers in Neurorobotics*, vol. 6, no. 5, pp. 1-9, 2012.
- [4] V. Handa and R. Massof, “Measuring the severity of stress urinary incontinence using the Incontinence Impact Questionnaire,” *Neurourol. Urodynam.*, vol. 23, no. 1, pp. 27-32, 2004.
- [5] A. Farbroth, S. Abbas, A. Nihlstrand, J. Dagman, R. Emardson, S. Kanerva, and L. Pendrill, 2013, “Defining comfort for heavily-incontinent patients assisted by Health care products in several contexts,” *The Simon Foundation for Continence’s Innovating for Continence Conference Series*, Chicago, USA, April 2013.
- [6] F. De Battisti, G. Nicolini, and S. Salini, “The Rasch model to measure service quality,” *The ICFI Journal of Services Marketing*, vol. III, no. 3, pp. 58-80, 2005.
- [7] J. Weller, N. Dieckmann, M. Tusler, C. Mertz, W. Burns, and E. Peters, “Development and Testing of an Abbreviated Numeracy Scale: A Rasch Analysis Approach,” *J. Behav. Decis. Making*, vol. 26, no. 2, pp. 198-212, 2013.
- [8] W. Zhang, Y. Yang, Q. Wang, and F. Shu, “An Empirical Study on Classification of Non-Functional Requirements,” *Proceedings of the 23rd International Conference on Software Engineering & Knowledge Engineering (SEKE 2011)*, Miami Beach, USA, July 2011.
- [9] G. Rasch, “On general laws and the meaning of measurement in psychology,” *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, pp. 321-334, 1961.
- [10] K. Sijtsma, “Introduction to the measurement of psychological attributes,” *Measurement*, vol. 44, no. 7, pp. 1209-1219, 2011.
- [11] M. Wilson, “Using the concept of a measurement system to characterize measurement models used in psychometrics,” *Measurement*, vol. 46, no. 9, pp. 3766-3774, 2013.
- [12] W. Fisher, Jr., “Physical Disability Construct Convergence across Instruments: Towards a Universal Metric,” *Journal of Outcome Measurement*, vol. 1, no. 2, pp. 87-113, 1997.
- [13] M. Ben-Akiva and M. Bierlaire, “Discrete choice methods and their application to short term travel decisions,” *Handbook of Transportation Science*, Kluwer, pp. 5-34, 1999.
- [14] D. McFadden, “Economic choices,” *Nobel Prize Lecture*, Stockholm, Sweden, December 8, 2000.
- [15] D. Kahneman, “Maps of bounded rationality: a perspective in intuitive judgment and choice,” *Nobel Prize Lecture*, Stockholm, Sweden, December 8, 2002.
- [16] B. Berglund, G. Rossi, J. Townsend, and L. Pendrill 2011, eds., *Measurement With Persons: Theory, Methods, and Implementation Areas*, Psychology Press, 2011.
 - Chapter 9: E. Dzhamarov, “Mathematical foundations of Universal Fechnerian Scaling”
 - Chapter 16: L. Pendrill, “Risk Assessment and Decision-making”
- [17] J. Schmidhuber, “Simple Algorithmic Theory of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes,” *Journal of SICE*, vol. 48, no. 1, pp. 21-32, 2009.
- [18] W. Fisher, Jr., “Invariance and traceability for measures of human, social, and natural capital: Theory and application,” *Measurement*, vol. 42, no. 9, pp. 1278-1287, 2009.
- [19] J. Helton, J. Johnson, and W. Oberkampf, “An exploration of alternative approaches to the representation of uncertainty in model predictions,” *Reliab. Eng. Syst. Safe*, vol. 85, no. 1-3, pp. 39-71, 2004.

- [20] S. Humphry, "The Role of the Unit in Physics and Psychometrics," *Measurement: Interdisciplinary Research and Perspectives*, vol. 9, no. 1, pp. 1–24, 2011.
- [21] ISO, "Quality management systems – Requirements," *ISO 9001*, 2008.
- [22] EN, "Health care services – Quality management systems – Requirements based on EN ISO9001:2008," *EN 15224*, 2012.
- [23] FDA, *Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*, U.S. Department of Health and Human Services, Food and Drug Administration, 2009.
- [24] FDA, *Applying Human Factors and Usability Engineering to Optimize Medical Device Design*, U.S. Department of Health and Human Services, Food and Drug Administration, 2011.
- [25] WHO, "Primary Health Care – Now More Than Ever," *The World Health Report*, 2008.
- [26] Institute of Medicine, *Core measurement needs for Better Care, Better Health, and Lower Costs*, Washington, DC: The National Academies Press, 2013.
- [27] A. Heinemann, W. Fisher, Jr., and R. Gershon, "Improving Health Care Quality With Outcomes Management," *J. Prosthet. Orthot.*, vol. 18, no. 1S, pp. 46–50, 2006.
- [28] E. Svensson, "Guidelines to statistical evaluation of data from rating scales and questionnaires," *J. Rehabil. Med.*, vol. 33, pp. 47–48, 2001.
- [29] S. Stevens, "On the Theory of Scales of Measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946.
- [30] AIAG, *Measurement Systems Analysis Reference Manual*, 3rd ed., 2002.
- [31] P. Loftus and S. Giudice, "Relevance of methods and standards for the assessment of measurement system performance in a high-value manufacturing industry," *Metrologia*, vol. 51, pp. S219–27, 2014.
- [32] JCGM, "Guide to the expression of uncertainty in measurement (GUM)," *JCGM 100*, 2008.
- [33] L. Pendrill and W. 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), *Journal of Physics: Conference Series* **459**, 012057
- [34] L. Mari and M. Wilson, "An introduction to the Rasch measurement approach for metrologists," *Measurement*, vol. 51, pp. 315–327, 2014.
- [35] G. Klir and T. Folger, *Fuzzy sets, uncertainty and information*, Prentice Hall, New Jersey, USA, 1988.
- [36] L. Pendrill, "Uncertainty & risks in decision-making in qualitative measurement: An information-theoretical approach," *Advanced Mathematical and Computational Tools in Metrology and Testing IX*, Series on Advances in Mathematics for Applied Sciences, World Scientific, 2012.
- [37] J. Linacre, "Optimizing Rating Scale Category Effectiveness," *Journal of Applied Measurement*, vol. 3, no. 1, pp. 85–106, 2002.
- [38] M. Erbacher, K. Schmidt, S. Boker, and C. Bergeman, "Measuring Positive and Negative Affect in Older Adults Over 56 Days: Comparing Trait Level Scoring Methods Using the Partial Credit Model," *Journal of Applied Measurement*, vol. 13, no. 2, pp. 146–164, 2012.
- [39] J. Sun, G. Wang, V. Goyal, and L. R. Varshney, "A framework for Bayesian optimality of psychophysical laws," *J. Math. Psychol.*, pp. 1–7, 2012.
- [40] L. Pendrill, "Using measurement uncertainty in decision-making and conformity assessment," *Metrologia*, vol. 51, pp. S206–S214, 2014.
- [41] W. Hardcastle, ed., *Qualitative Analysis: A Guide to Best Practice*, Royal Society of Chemistry, Cambridge, UK, 1998.
- [42] S. Ellison and T. Fearn, "Characterising the performance of qualitative analytical methods: Statistics and terminology," *TRAC-Trend Anal. Chem.*, vol. 24, no. 6, pp. 468–476, 2005.
- [43] JCGM, "Evaluation of measurement data – The role of measurement uncertainty in Conformity Assessment," *JCGM 106*, 2012.
- [44] T. Schneider and K. Lewis, *A Glossary for Biological Information Theory and the Delila System*, Schneider Lab, Version 3.43, April 2014.
- [45] OECD, "The Rasch Model," chapter 5 in the *PISA Data Analysis Manual: SAS, Second Edition*, pp. 79–94, 2009.
- [46] B. Wright, "Comparing factor analysis and Rasch measurement," *Rasch Measurement Transactions*, vol. 8, no. 1, p. 350, 1994.
- [47] T. Mallinson, A. Keenan, S. Purl, and C. Velozo, "Measuring the Impact of Fatigue on Everyday Activities during Chemotherapy," *Proceedings of International Conference on Objective Measurement (COMET)*, Chicago, Illinois, USA, October 2001.
- [48] S. Dehaene, V. Izard, E. Spelke, and P. Pica, "Log or Linear? Distinct Intuitions of the Number Scale in Western and Amazonian Indigene Cultures," *Science*, vol. 320, no. 5880, pp. 1217–1220, 2008.
- [49] A. Giordani and L. Mari, "Property evaluation types," *Measurement*, vol. 45, no. 3, pp. 437–452, 2012.
- [50] L. Pendrill, "Discrete ordinal & interval scaling and psychometrics," *Proceedings of 16th International Congress of Metrology*, Paris, France, pp. 1–4, October 2013.
- [51] K. Knoblauch and L. Maloney, "MLDS: Maximum Likelihood Difference Scaling in R," *J. Stat. Softw.*, vol. 25, no. 2, pp. 1–26, 2008.
- [52] WINSTEPS, (<http://www.winsteps.com>).
- [53] E. Bashkansky, S. Dror, R. Ravid, and P. Grabov, "Effectiveness of a Product Quality Classifier," *Quality Engineering*, vol. 19, no. 3, pp. 235–244, 2007.
- [54] K. Pesudovs, "Item Banking: A Generational Change in Patient-Reported Outcome Measurement," *Optometry Vision Sci.*, vol. 87, no. 4, pp. 285–293, 2010.
- [55] J. Linacre, "Rasch Model with an Error Term," *Rasch Measurement Transactions*, vol. 23, no. 4, p. 1238, 2010.
- [56] Socialstyrelsen, Appendix 2 of *National Guidelines for Diabetic Care* (in Swedish), National Board of Health and Welfare, 2010.
- [57] J. Linacre and B. Wright, "The 'Length' of a Logit," *Rasch Measurement Transactions*, vol. 3, no. 2, pp. 54–55, 1989.
- [58] JCGM, "International vocabulary of metrology – Basic and general concepts and associated terms (VIM)," *JCGM 200*, 2012.
- [59] BIPM, *The International System of Units*, 8th ed., 2006.

- [60] M. Wilson, "The role of mathematical models in measurement: a perspective from psychometrics," *Proceedings of Joint International IMEKO TC1 + TC7 + TC13 Symposium*, Jena, Germany, pp. 1-7, August-September 2011.
- [61] D. Borsboom and A. Zand Scholten, "The Rasch Model and Conjoint Measurement Theory from the Perspective of Psychometrics," *Theor. Psychol.*, vol. 18, no. 1, pp. 111-117, 2008.
- [62] R. Luce and J. Tukey, "Simultaneous conjoint measurement: A new type of fundamental measurement," *J. Math. Psychol.*, vol. 1, pp. 1-27, 1964.
- [63] G. Iverson and R. Luce, "The representational measurement approach to psychophysical and judgmental problems," Chapter 1, in *Measurement, Judgment, and Decision Making*, Academic Press, 1998.
- [64] J. Linacre, "Bernoulli Trials, Fisher Information, Shannon Information and Rasch," *Rasch Measurement Transactions*, vol. 20, no. 3, pp. 1062-1063, 2006.

9. Appendix

A.1 Decision-Making and Qualitative Observations

In the simplest case where the prior state of an entity (the measurement object) being observed is known to be in a particular category M , $p_k = \begin{cases} 1; k = M \\ 0; k \neq M \end{cases}$, for the stimulus variable (S in Fig. 3), then the probability, q_c , of classifying the response (R) in a category c , is given by the accumulation of probabilities [53],

$$q_c = \sum_{k=1}^2 p_k \cdot P_{c,k} = 1 - \alpha = P_{\text{success}}. \quad (\text{A.1})$$

Equation (A.1) links the response variable (probability of successful categorization, P_{success} , in the decision-making) to the risks, α , of incorrect decisions arising from uncertainty in measurement of the explanatory variable. An example of such a known prior state is the prepared number of dots shown in Fig. 5 (a). These decision risks can be modelled for two kinds of human-based perception, as dealt with in psychophysics [63]:

- **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 [40, 43],

$$\alpha = G_X(U_{SL}) = \Pr(x \geq U_{SL}) = \left[\int_{U_{SL}}^{\infty} \frac{1}{\sqrt{2\pi} \times u_m} \times e^{-\frac{(x-x_m)^2}{2u_m^2}} \times dx \right]; x_m \leq U_{SL}. \quad (\text{A.2})$$

- **Choice** involves, in the dichotomous case, the pairwise discrimination of stimuli. The ability to resolve two adjacent stimuli, s_a and s_b , – e.g. counting nine dots when there are 10 – will be given by the overlap of the two distributions, where the width of each distribution is given by the perceptual uncertainty, $w \times s$ of each observation, where w is the Weber constant,

$$\alpha = 0.5 \times \operatorname{erfc} \left[\frac{|s_a - s_b|}{\sqrt{2} \times w \times \sqrt{(s_a^2 + s_b^2)}} \right]. \quad (\text{A.3})$$

This can be seen as a more general case which reduces to Eq. (A.2) when the width of one of the stimulus distributions is reduced to zero [33].

The values of the 'decision risk' in both cases follow the general shape of an ogive curve according to Eq. (A.6). This is reminiscent of the 'operating characteristic' across the range of values of the explanatory variable, but with slightly different forms depending on the underlying kinds of decision being made.

These relations can be extended to the multinomial, polytomous case, as required [40].

A.2 Loss of Information and Decision-Making. Entropy and Generalized Linear Models

Connections to the decision risks discussed above can be made with the concept of 'uncertainty' in information theory [35], which refers to a loss of information in 'transmission', i.e, when making decisions, as the dissimilarity between the amounts of information in the posterior (Q) and prior (P) states. This dissimilarity can be measured in information theoretical terms as the so-called Relative entropy or Kullback-Leibler (KL) divergence,

$$D_{KL}(P||Q) = H(P, Q) - H(P), \quad (\text{A.4})$$

as the difference in entropy (H) after making a choice ($H(P, Q)$, called the 'equivocation') and before ($H(P)$).

In probability theory, where the amount of semantic information is expressed as the Shannon entropy H , the corresponding Kullback-Leibler (KL) divergence (of semantic information) is

$$D_{KL}(P||Q) = -\sum_k p_k \times \log(q_k) + \sum_k p_k \times \log(p_k) = H(P, Q) - H(P). \quad (\text{A.5})$$

A.3 Perceptive Choice, Logistic Regression and Generalized Linear Models

Considering information transmission as a decision-making perception of pairwise discrimination of adjacent stimuli, such as in choice in cognitive psychology [63, Eq. (2)], the subjective distance $D(a, b)$ between two stimuli, a and $b > a$, is expressed according to Dzhamfarov [16] as the integral over the level, s , of stimulus of a measure (the so-called 'psychometric function') of the ability to perceive a dissimilarity $P(s, s + ds) = \Pr[b \text{ is judged to be greater than } a]$:

$$D(a, b) = \int_a^b \frac{P(s, s + ds)}{ds} \cdot ds.$$

Consider again the simplest, dichotomous case where the prior is known to be M ,

$$p_k = \begin{cases} 1; k = M \\ 0; k \neq M \end{cases}, \text{ as in the elementary dots studies. In that case, the subjective distance, identified with the } KL \text{ divergence } D_{KL}(P, Q) \text{ (Eq. A.5) for the measurement-based decision, is derived by substituting } \frac{P(s, s + ds)}{ds} = dP_{\text{success}} \text{ and } ds = -z \text{ to yield:}$$

$$D_{KL}(P, Q) = \int -z \times dP_{\text{success}} = -[P_{\text{success}} \times \log(P_{\text{success}}) + (1 - P_{\text{success}}) \times \log(1 - P_{\text{success}})] = H(P, Q) - H(P),$$

where the prior entropy, $H(P) = 0$ (since the prior state is known). This leads readily [40, 64] to the logistic regression link function:

$$z = \log \left[\frac{P_{\text{success}}}{1 - P_{\text{success}}} \right]. \quad (\text{A.6})$$

Equation (A.6) is an example of a ‘link’ function, g , in a so-called generalized linear model (GLM) employed 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 , which have the property that the expectation value $E(R) = \mu = g^{-1}(\eta) = g^{-1}(S \times \beta)$, offers a linear predictor $\eta = S \times \beta$ which will be some linear combination of unknown parameters, β , (z in Eq. A.6) even when the response itself is not linear.

In the special case where a change in the psychometric function is proportional to the fractional change in stimulus level, i.e. $P(s, s + ds) = w \times \frac{ds}{s}$, where the Weber constant of proportionality, w , is as it appears in Eq. (A.3), then the subjective distance between two stimuli a and b is: $D(a, b) = w \times \ln\left(\frac{b}{a}\right) = w \times [\ln(b) - \ln(a)]$, that is the Weber-Fechner law (Dzhafarov in [16]).

Summarizing: 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 less quantitative appraisals, such as on an ordinal or even a merely nominal scale.