Abstract. Is it realistic to aspire to the same kind of quality-assurance of measurement in person-centred care, currently being implemented in healthcare globally, as is established in the physical sciences and engineering? Ensuring metrological comparability (‘traceability’) and reliably declaring measurement uncertainty when assessing patient ability or increased social capital are however challenging for subjective measurements often characterised by large dispersion. Drawing simple analogies between ‘instruments’ in the social sciences – questionnaires, ability tests, etc.– and engineering instruments such as thermometers does not go far enough. A possible way forward apparently equally applicable to both physical and social measurement, seems to be to model inferences in terms performance metrics of a measurement system. Person-centred care needs person-centred measurement and a full picture of the measurement process when Man acts as a measurement instrument is given in the present paper. This complements previous work by presenting the process, step by step, from the observed indication (e.g. probability of success, \( P_{\text{success}} \) of achieving a task), through restitution with Rasch Measurement Theory, to the measurand (e.g. task difficulty). Rasch invariant measure theory can yield quantities – ‘latent’ (or ‘explanatory’) variables such as task challenge or person ability – with characteristics akin to those of physical quantities. Metrological references for comparability via traceability and reliable estimates of uncertainty and decision risks are then in reach even for perceptive measurements (and other qualitative properties). As a case study, the person-centred measurement of cognitive ability is examined, as part of the EU project EMPIR 15HLT04 NeuroMet, for Alzheimer’s, where better analysis of correlations with brain atrophy is enabled thanks to the Rasch metrological approach.

1 Introduction

1.1 Need for quality-assured measurement in health care

Official statistics [OECD 2015] show that the apparent prevalence of dementia seems to differ by as much as a factor of ten amongst different countries. In looking into possible causes for this remarkably large variation, one might speculate that disease prevalence really does vary so much country by country, however unlikely this seems. Another potential cause could be that physicians diagnose the patient conditions differently, or maybe healthcare data systems are
interpreting diagnoses according to different codes. There may be uncertainties or different opinions about disease indicator specifications, and so on. Cases such as this underline the pressing need to provide a metrological quality infrastructure in health care. Without metrological traceability to recognised measurement reference standards and without clear statements of measurement uncertainty, it will be difficult to make the reliable decisions about care with the necessary comparability and evaluated risks needed to diagnose, treat and rehabilitate throughout the healthcare system.

The need for metrological quality assurance in care has been recognised by amongst others the OECD [2017], who in a ministerial statement recently emphasised: “health data [is] necessary to improve the quality, safety and patient-centeredness of health care services and to support scientific innovation, the discovery and evaluation of new treatments and to redesign and evaluate new models of health service delivery”. The italicised (my own) words in this quote indicate the focus of this paper.

1.2 Metrology in health care

Many different kinds of measurement are made of course in healthcare.

A familiar case is the weighing of patients. Without calibrated person scales, it will be difficult to make decisions about health-related changes in body weight. A metrological solution is to calibrate the scales with a set of masses which themselves have been calibrated at a metrological laboratory. At such a calibration, the metrologist will have to inform the owner of the mass standards that these are perhaps not perfect – maybe weighing slightly differently from nominal mass. At the same time, the owner will have to be informed – in understandable and not too technical terms – that the calibration result has a certain amount of uncertainty attached to it. This meeting of metrologist and owner is an essential event in assuring the quality of the measurements to be made, both in terms of providing for objective comparability of measurements (through traceability) and communicating the limits on measurement quality. That in turn provides for quality assurance of the products or services being offered – in this example weighing of body mass in healthcare – in corresponding terms of interoperability and quantified risks of incorrect decisions (e.g. of conformity to specification), such as whether the patient’s body mass index is within limits or not.

A second example of quality-assurance in health care could look rather similar – in terms of a meeting between two parties – but where the metrologist is replaced by a clinical physician and the owner is replaced by a patient. Instead of body mass, the quality characteristics of interest in this case could be the health of the patient. In the case of Alzheimer’s disease, an assessment of cognitive performance of the patient is an important part of the health evaluation, particularly person- and patient-centred, since, alongside various biomarkers of the disease (protein accumulation in brain regions, brain atrophy, etc), cognitive decline is arguably one of the major impacts of the disease. This article concludes with a specific example of an application of person-centred metrology to Alzheimer health care.

The challenges faced in the body mass example above will be replaced in the Alzheimer case by perhaps even more challenging aspects: Advances in information and communication technologies are being deployed with the potential to radically change how physicians – and increasingly patients themselves – assess cognition, even for prevention [PerSoCo 2016]. There is ready access to health-related information via the Internet or body-worn sensors connected through smart phones, making it increasingly likely that a typical patient will have already formed his own opinion about his or her state of health before meeting the physician. The physician in turn will have the task of informing the patient that his or her health – cognitive ability, for instance - is not perfect and – again in understandable and not too technical terms – that the calibration result has a certain amount of uncertainty attached to it. The “semantic interoperability” has to be good; one
term referring to the “precise meaning of exchanged information which is preserved and understood by all parties” [ISA 2017].

1.3 Metrology of person-centred care (PCC)

In line with its ambition to reorient health systems to become more people-centred, the OECD [2017] states the need to: “invest in measures so that health systems deliver what matters most to people. Too often, we only rely on measures of what health systems do, and how much they cost, rather than their effects on patients.” WHO [2017] emphasises the clear differences compared with more traditional healthcare, where person-centred care focuses instead on health (not illness) and considers people as partners in care. Person-centred care in turn needs person-centred metrology, as addressed in the present work.

There are a number of organisations (ICHOM\textsuperscript{a}, PROMIS\textsuperscript{b}, OMERACT\textsuperscript{c}, COSMIN\textsuperscript{d} [2017] and others) now vying to tackle quintessentially metrological challenges in person-centred care, amongst others newly created groups alongside the more established metrology and quality infrastructure bodies. The metrology of person-centred care, similarly to more established sectors of metrology, would have to “embed material measurement and traceability technologies in (i) an information infrastructure which includes literature (journals, written standards, etc.) and in (ii) social technology (peer review, professional societies, etc)” [Schaffer 1992, Cano \textit{et al.} 2017].

In laboratory medicine, the standardization of measurements is of course a high priority, with the goal of comparability of results obtained using routine procedures [Jones & Jackson 2015, Hallworth \textit{et al.} 2015]. “A determination to accept patient outcome and patient experience as the primary measure of laboratory effectiveness” is a recommendation of a recent IFCC\textsuperscript{e} Task Force on the Impact of Laboratory Medicine on Clinical Management and Outcomes. The IFCC in turn belongs to the Joint Committee for Traceability in Laboratory Medicine as a link to the metrology community.

There is a clear division in OMERACT [2017] between what is called pathophysiological manifestations, related to ICF\textsuperscript{f} body function and structure, and impact of health conditions, related to International Classification of Functioning, Disability and Health activity and participation. These are in line with the well-known WHO redefinition of health as: “a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.” Another noteworthy initiative is COSMIN [2017] which aims: “to reach consensus on which measurement properties should be evaluated of Health-Related Patient-Reported Outcomes and how they should be defined, and to develop standards for how these measurement properties should be evaluated in terms of study design and statistical analysis”. The organisational aspects of person-centred metrology have been recently reviewed by Cano \textit{et al.} [2017].

Several recent guides and others have pointed out some particular differences between measurement in person-centred care compared with more traditional professional care. Compared with traditional measures (e.g blood pressure or queueing times) in care, patient assessment often differs from professional assessment, with the patient observing more symptoms; the greater impact on daily living; and so on [COSMIN 2017]. This different content of the patient’s

\textsuperscript{a} International Consortium for Health Outcomes Measurement
\textsuperscript{b} Patient-Reported Outcomes Measurement Information System
\textsuperscript{c} Outcome Measures in Rheumatology
\textsuperscript{d} COnsensus-based Standards for the selection of health Measurement INstruments
\textsuperscript{e} International Federation of Clinical Chemistry and Laboratory Medicine
\textsuperscript{f} International Classification of Functioning, Disability and Health
own view of his or her health means that much effort has to be expended on forming fit-for-purpose measurement instruments that from the start of any measurement task faithfully capture the patient’s concerns, which means taking time to listen to those concerns and structure the issues in a systematic and comprehensive way [Browne et al. 2017].

Quality assurance of measurements in person-centred care is arguably still in its infancy. In an evidence review “Helping to measure person-centred care”, the UK’s Health Foundation [2014] stated that there is a “large number of tools available to measure person-centred care, but no agreement about which tools most worthwhile….No ‘silver bullet’ or best measure covers all aspects of person-centred care.” From a metrological point of view, in person-centred care there is a need to develop reference standards for novel kinds of quantities: one now needs to include quantities associated with patient activity and participation, social well-being etc., in addition to the traditional physical, chemical and biological quantities of clinical biomarkers. Providing quality-assurance of these person-centred factors is arguably at least as essential and pressing as other, well-publicised challenges of contemporary health care – such as posed by ‘big data’, personalised medicine or the correct handling of electronic health records. Hobart et al. [2007] have emphasised this in a review of the challenges, pointing out that: “First, the numbers generated by most rating scales do not satisfy the criteria for rigorous measurements. Second, we do not really know which variables most rating scales measure”.

While the metrological infrastructure supporting body mass has arguably been in place many years (albeit not necessarily implemented [Ferreira and Matos 2015] throughout health care), providing traceability and uncertainty statements for health-related measurands of a more person-centred character is as yet at an early stage [Cano et al. 2017]. For Alzheimer’s, there are to date few metrological references for cognitive performance [Hughes et al. 2003], Hobart et al. 2013, EMPIR NeuroMet 2017, Quaglia et al. 2017]. This development can be seen to be part of a wider context where the conceptual aspects of whether an ‘evaluation’ (including opinions and estimations, perhaps on ordinal or nominal scales) is a ‘measurement’ or not, is a current topic of discussion in the literature [Rossi 2014; Mari et al. 2016a,b; Sawyer et al. 2016] and in establishing the latest vocabularies of metrology [JCGM WG2 VIM and IUPAC VIN 2015].

Certainly, there is a considerable contrast between the relatively slow metrological development and the dramatic and impressive modernisation of innovative healthcare enabled for instance by novel digital information and communications technologies and driven by the well-known grand challenge of providing healthcare in the demographic and economic realities of contemporary and future societies.

This paper will present an account of some recent advances in providing for the metrology of person-centred care, particularly the aspects which distinguish it from more traditional metrology. Apart from a difference in content, the main focus here will be on dealing with the unique character of person-centred metrology.

2 Measurement systems in the physical and social sciences

The challenges of providing for quality-assured measurement in person-centred care can fortunately be met, but this needs a reappraisal of metrology which gives added insight not only into the new areas of PCC but also into quality-assured measurement in qualitative properties more generally, as well as more traditional areas.

One particularly suitable model – basically regarding Man as a Measurement Instrument – is the Rasch [1961] invariant measurement methodology, as has indeed already been recognised and adopted by clinicians in increasingly diverse parts of contemporary healthcare. The width of aspects of the metrology of person-centred care can be illustrated by examples such as (i) BREAST-Q, which constitutes an important complement to screening and diagnosis of breast
cancer [Cano et al. 2011] and (ii) the EMPIR 15HLT04 NeuroMet [2016] project Innovative measurements for improved diagnosis and management of neurodegenerative disease – one of the first metrological projects in person-centred care.

Two principal elements behind this approach are: (i) a correct formulation of a measurement model and (ii) proper handling of the subjective data typical of human judgment. Recalling the examples of quality-assurance in health care given in the Introduction – be it the Alzheimer patient trying to recall a particular sequence of digits or words in a cognitive test or the physician trying to interpret the patient’s symptoms – a key metric is obviously the ability to make a correct decision. Naturally, the validity of the Rasch model in this context has to be demonstrated.

The next sections review this proposed way forward, which appears to be equally applicable to assuring the quality of both physical and social measurement, including both quantitative and qualitative properties more generally. An explicit model, extending that presented earlier, will be presented as a synthesis of measurement system engineering [section 2.1], psychometrics [section 2.2] and philosophy [section 2.3]. Before concluding, we will exemplify application of the special tools of person-centred metrology of care of Alzheimer patients in the NeuroMet project, as just one case [section 3].

2.1 Measurement system engineering

Firstly, essential insight can be gained by extending traditional metrological concepts through developing an operational model of a particular kind of measurement system [AIAG 2002, Bentley 2004, ASTM 2012, Loftus and Giudice 2014, Rossi 2014]. Others have already pointed out that this perspective has not been emphasised enough, even in traditional metrology: “One major difficulty for an extensive application of GUM’s principles” has been an “almost total absence of (the notion of a measurement system) in The Guide to the evaluation of Uncertainty in Measurement (GUM) [JCGM 100]”, according to Rossi & Crenna [2016]. The notion of measurement system has also been substantially missed in representational theory [Rossi 2014]. That measurement is a “concatenation of observation and restitution” (as recalled by Bentley [2004], Sommer and Siebert [2006], and Rossi [2014]) is added here to our previous work [Pendrill and Petersson 2016], as follows.

2.1.1 Measurement equations and measurement systems. Observation

In the practical working of a measurement system [ASTM 2012 §3.1.7] set up to observe a particular measurand \(Y\), a measurement equation is often formulated expressing the system response \(X_i\) in the form:

\[
X_i = h(Y; X_2, \ldots, X_N) \tag{1}
\]

to capture the influence of factors \((X_2, \ldots, X_N)\) such as hardware, software, procedures and methods, human effort, etc which combine to cause variation among measurements of the same object that would not be present if the system were perfect.

A model of human perception from a psychophysical point of view, given recently in [Sun et al. 2012, Rossi 2014, Pendrill & Fisher Jr 2015], is similar to a traditional model of an engineered instrument, see figure 1. The change in Response \((R)\) due to the change of Stimulus \((S)\) is the Sensitivity \((C)\) of the human (acting as an) instrument.
Here a new aspect we add to our previous work is making explicit the form of function \( h \) in equation (1), relating in eq. (2) the response \( R = X_i \) to the stimulus \( S = Y \). This includes other factors – based on engineering expressions for each element of the measurement system [Bentley 2004] – which take account for instance element sensitivity, \( C \) (ratio response to stimulus); and ‘nuisance’ parameters such as: non-linearity, \( N \); bias, \( a \); modifying, \( C_m, I_m \); and interfering \( C_i, I_i \) interferences:

\[
X_i = R = O = C \cdot I + N(I) + C_m \cdot I_m + C_i \cdot I_i + a ;
\]

\[
I = Y = S
\]

Cause-and-effect relations between stimulus \( S = Y \), ‘input’ \( I \) and response \( X_i \), ‘output’ \( O = R \) ‘response’, according to the measurement equations (1 and 2) can be visualised with the help of Ishikawa or fishbone diagrams. Each element of the measurement system (object, instrument, method, operator, etc) is depicted with one major ‘bone’ of the Ishikawa diagram, as has been done since at least the 1990’s in for example the MSA approach [AIAG 2002, Pendrill 2010]. In some cases, the ‘instrument’ can be further modelled in terms of a chain of elements representing sensor; signal conversion; signal conditioning; data conversion [Bentley 2004, Sommer and Siebert 2006].

Substantially the same function and its derivatives of the measurement equations (1 and 2) can be used to describe how to calculate:

- The actual result of measurement
- Propagation of measurement bias through the measurement system
- Propagation of variances through the measurement system [Bentley 2004]

Where more than two measurements are made with a particular measurement system, it is practice [ASTM 2012 §5.5.2] to employ statistical techniques such as ANOVA to empirically assign estimates of variance associated with each element of the measurement system (hardware, software, procedures and methods, human effort, etc). Interlaboratory experiments can be used in some cases to estimate both bias and variance associated with each element of the measurement system (object, instrument, method, operator, etc), such as in proficiency testing; method accuracy experiments, etc. Such experiments in healthcare, as elsewhere, provide essential evidence of adequate quality assurance of measurement.

### 2.1.2 Measurement model and restitution

Also new in this paper is making explicit an ‘inverse model’\(^8\) or a ‘restitution process’ at which a model equation (3) is formulated for the evaluation of the quantity intended to be measured (the measurand) and its measurement uncertainty according to the GUM:

---

\(^8\) As noted in VIM 2.50, an input quantity in a measurement model is often an output quantity of a measuring system, and *vice versa.*
The measurement equation (1) gives input to this model equation, thereby establishing the relationship between the measurand \( Y \) and all relevant quantities \( X \) having an impact on this measurement result [Bentley 2004, Rossi 2014]. A prerequisite is calibration where the various measurement system attributes (such as sensitivity, bias, etc described in eq. 2) are determined by observing the response of the system to a known stimulus, such as provided by a metrological reference. Evaluation of concepts of trueness (lack of bias) and precision (low variance) in interlaboratory experiments (according to ISO 5725) is of course a key element of quality-assured measurements in metrology which aim to ensure, respectively, comparability of measurement results and explicitly declared limits to measurement quality in terms of measurement uncertainty. Implementation of such concepts in healthcare is in its infancy. Healthcare for example has been described as a “$1 trillion per year industry without a clear measure or definition of its main product” [Heinemann et al. 2006].

The suite of equations (1) – (3), covering the measurement process from observed indication, through restitution to the measurand, appears to be sufficiently general to be applicable – in one form or another – to measurement in both the hard and soft sciences.

The next section will introduce how metrological references suitable for calibration in the social sciences can be developed.

### 2.2 Psychometrics in the physical and social sciences

Attempting to explain social measurement by drawing analogies to traditional measurement instruments, e.g., a thermometer [Bond & Fox 2015], in terms of simple response error, or regarding a questionnaire as an instrument [Mari and Wilson 2014, Wilson et al. 2015] does not go far enough.

One main reason why there are so few, if any, recognised metrological standards to provide for traceability when assuring human-based quality is the need to overcome a couple of elementary challenges: (i) the correct handling of qualitative response data, where the usual tools of statistics do not always work on the categorical scales typical of such measurements [Svensson 2001], and (ii) an unclear separation of measurement and object dispersions in the results when Man acts as a measurement instrument.

Generalised linear modelling ([McCullagh 1980]) is a well-known technique to address challenge (i), i.e., handling qualitative data, e.g. on ordinal scales, such as in compositional data analysis [Aitchison 1982], by transforming ‘raw’ data from observations with the General linearised model into a new ‘space’ (called ‘simplex’) where a linear, quantitative scale can be identified.

The commonality between physical and social measurement (and qualitative estimations more generally) is first reached when one additionally recognizes that the performance metrics of a measurement system are the same concept in both [Pendrill 2014a,b]. For this, we need to explicitly include decision-making as the third and final step – together with observation [section 2.1.1] and restitution [section 2.1.2] – in a full picture of the measurement process (as will be illustrated in Figure 2 below). Performance metrics in the medical field are of course of prime importance, particularly in the context of person-centred care, in terms of the ability of the patient (to participate and be active in essential activities) and indeed of physicians and other care personnel, to act effectively as partners in care when ensuring quality, accessibility and accountability. In traditional metrology, perhaps the most familiar performance metrics are the risks of incorrect decisions of conformity (such as producer and consumer, or type 1 and type 2, risks) arising from measurement uncertainty [Pendrill 2014b]. As pointed out earlier [Pendrill 2014b, Pendrill and Petersson 2016], our approach is in
some ways a reversal of the traditional handling of measurement uncertainty – starting with the decision risks rather than evaluation of standard deviations or other statistical measures of spread in the data.

2.2.1 Performance metrics. Entropy

The measurand being measured with a measurement system set up for its observation [section 2.1.1] – in the terminology of conformity assessment, a ‘quality characteristic’ of the entity - may be, as in statistics, either:

- a measure of ‘location’, e.g. a direct quantity of the entity, such as the mass of a single object, an error in mass (deviation from a nominal value), or an average mass of a batch of objects
- a measure of ‘dispersion’, e.g. the standard deviation in mass amongst a batch of objects [ISO 10576-1:2003]

In any measurement system, not only the object but also all other elements – such as instrument, method, operator etc – can also be subject to conformity assessment.

How well the measurement system performs an assessment can also be specified in terms of both location and dispersion. Alongside classical metrics such as instrument bias or non-linearity, even decision risks arising from measurement uncertainty, for instance when assessing compliance to a specification limit [Pendrill 2014b], can be counted as a performance metric. In the new version of the key standard ISO 17025, these aspects are emphasised in more detailed regarding statements of compliance:

“When the customer requests a verification of conformity to a specification or standard for the test or calibration (e.g. pass/fail, in-tolerance/out-of-tolerance):

- The specification is clearly defined in the procedure selected;
- The decision rule for conformity, its level of risk and statistical assumptions are documented in the test method/procedure or are documented by the laboratory and communicated to the customer;
- The decision rule is agreed to by the customer” [ISO/IEC DIS 17025:2017]

A key and unifying observation is that for all scales – nominal, ordinal, interval and ratio - the probabilities of decision risks and performance success (such as rating scores) can be expressed in terms of (lost) information, i.e., (increased) entropy – a property equally applicable to qualitative and quantitative measurement results [Pendrill 2011, 2014a, Pendrill & Petersson 2016].

Recalling again the examples of quality-assurance in health care given in the Introduction – be it the Alzheimer patient trying to recall a particular sequence of digits or words in a cognitive test or the physician trying to interpret the patient’s symptoms – in the postulated measurement system, the object (‘entity’) to be assessed can be a task with attribute ‘level of difficulty’, and the instrument, a person with attribute ‘ability’. Traditional ‘instruments’ of social science – questionnaires, ability tests, etc. – provide the ‘decision-making algorithm’ in the measurement system which quantifies the performance metric from the interaction between the person (‘instrument’) and the task (‘object’). A poorly performing measurement system where measurement information from the measurement object is transmitted (illustrated in Figure 1), via an instrument, to an operator (or appraiser), will have a considerable loss of information (measured in terms of an increase in entropy [Weaver and Shannon 1963], i.e. disorder, such as loss of a pattern). A task will be easier if there is some degree of order, i.e. less entropy.

The risks of incorrect decisions, arising from finite (i.e. non-zero) measurement uncertainty with a certain measurement system, can be described with a “decision matrix” which in the simplest, binary case is:

\[
P = \begin{bmatrix}
1 - \alpha & \alpha \\
\beta & 1 - \beta
\end{bmatrix}
\]
where for instance a “producer” or “type 1” risk (i.e. false negative, incorrectly rejecting a conforming object) of an incorrect classification is denoted by $\alpha$, and the corresponding probability of a correct decision is $P_{\text{success}} = 1 - \alpha$ [Pendrill 2014a]. A higher uncertainty will lead to a more disordered response, i.e., higher risks of misclassification. The decision matrix enables the probability, $q_c$, of classifying the ‘output’ or ‘response’ state in a category $c$, to be linked to the prior probabilities, $p_k$, representing the ‘input’ or ‘stimulus’ state of the system of interest (over in general a range of categories $K$) through the relation: $q_c = \sum_{k=1}^{K} p_k \cdot P_{r,k}$ [Bashkansky and Gadrich 2012]. The act of measurement (decision or classification) will be an attempt to deduce the prior state (e.g. identify a pattern), given the results of the classification made with the measurement system.

It is straightforward to derive expressions for decisional (posterior) probabilities $q$ to prior probabilities $p$ that the measurement object is in a particular state with a generalised linear model suitable for a proper handling of the subjective data typical of human judgment, as we have summarised earlier [Pendrill 2011, 2014a, Pendrill and Petersson 2016]. The information divergence approach of Kullback and Leibler [1951] argues that the best estimate is that obtained by minimising the amount of information change during a classification, as a variant of the principle of least action. The Lagrange multipliers method, subject to a constraint maximising the (Shannon) entropy, $H = \sum_i q_i \cdot \log(q_i)$, of the posterior state (which minimises the difference in amount of information before and after a measurement amongst all possible (uncertain) outcomes), leads readily to a General linearised model expression for the ‘link’ function, $z$, which in the simplest, binary case is:

$$z = \log\left[\frac{P_{\text{success}}}{1 - P_{\text{success}}}\right]$$  

Note that this General linearised model expression is more general than, and is not primarily related to, the Weber-Fechner law, which has a different logarithmic dependence derived in the particular case where a change in the psychometric function is proportional to the fractional change in the stimulus level [Pendrill 2014a, Pendrill and Fisher 2015].

New in the present article is that we present the following fundamental expression for performance metric response, $R$, as the change in entropy, $H$, (order, e.g. in a pattern) in the output of a measurement system (eq. 2) to a change in stimulus ($S = z$), via a sensitivity $C = \frac{\partial P_{\text{success}}}{\partial z}$ reflecting a certain change of measurement information:

$$R = \frac{\partial H}{\partial z} = \frac{\partial P_{\text{success}}}{\partial z} \cdot z = C \cdot S$$  (5)

as may be readily derived with the method of Laplace multipliers.

The sensitivity, $C$, for a measurement system where the response, $R$, is a performance metric is known to have a special, ‘resonance-like’ form, illustrated in Figure 1 and derived by Pendrill & Petersson [2016]. Entropy is the key concept to handle the quality-assurance of data, be it on quantitative or qualitative scales. Expression (5) can be deployed to describe the observation and restitution processes for performance metrics, as summarised below in Figure 2. One
application of this original expression is given in a separate presentation by the author [Pendrill 2017] about limits on the reliability of the basic uni-dimensional Rasch model.

In the limit of an infinite number ($K$) of categories for classification, the multinomial discrete scale becomes a continuous interval scale familiar from ordinary quantitative measurement. A new insight is that specification limits become as “marks on a ruler”, thus unifying measurement of quantitative and qualitative properties. Measurement uncertainty on such scales is thus expressed in terms of the integrated entropy, which goes from the multinominal expression towards:

$$H_q = -\int_{-\infty}^{\infty} q \cdot \ln(q) \cdot dq = \ln\left[\sqrt{2\cdot\pi \cdot u_q}\right] + \frac{1}{2}$$  \hspace{1cm} (6)

in the case of a Normal distribution $N(q, u_q^2)$ for the continuous, quantitative scale.

Eq. 6 suggests an alternative expression of measurement uncertainty, more akin to the concepts of information theory than the classic standard uncertainties [JCGM GUM]. It is obvious by inverting eq. (6) that $u_q \sim e^{H_q}$, so the two approaches – from standard uncertainty, $u$, to decisions risks, and vice versa – can be unified. But we have a certain preference to express uncertainty in terms of entropy instead because it is conceptually closer to ‘uncertainty’ in everyday language; is also substantially distribution-free; and is indeed is accessible to treatment not only with probability theory but also possibility and plausibility theories. Examples can already be found in the literature: [Helton et al. 2006] and [Yang & Qiu 2005].

2.2.2 Rasch invariant measure theory

With General linearised modelling as the tool to handle for instance ordinal data, another main pillar supporting the metrology of person-centred care is the recognition that the specific psychometric approach of Rasch’s [1961] invariant measure theory can enable key metrological components – references for traceability – to be established when applied to our performance metric-based measurement system.

This is done in the Rasch [1961] model by writing the General linearised model link function (see eq. 4) as:

$$z = \theta - \delta, \quad \text{which allows a separation to be made in measured responses into a person, } i, \text{ attribute value (}\theta, \text{ such as ability or leniency) and an item, } j, \text{ attribute value (}\delta, \text{ such as level of difficulty or quality). In the simplest, dichotomous case of the logistic regression function:}$$

$$\log\left(\frac{P_{\text{success},i,j}}{1-P_{\text{success},i,j}}\right) = \theta_i - \delta_j$$  \hspace{1cm} (7)

In the case of classification into more than two ($k = 1, \ldots, K$) categories, the polytomous Rasch probability of response $q_{i,j}$ of person $i$ to item $j$ is given by:

$$q_{i,j,k} = \frac{e^{c(\theta_i - \delta_j)\sum_k \tau_{i,j}}}{\sum_{c=0}^{K} e^{c(\theta_i - \delta_j)\sum_k \tau_{i,j}}}$$

where $\tau_k$ denotes the threshold for the $k^{th}$ category [Andrich 1978, Masters 1982].

The ability of the Rasch approach to yield separate and objective measures of task difficulty and person ability, for instance, is essential in establishing sets (“item banks”) of metrological references based on proper use of the
psychometric Rasch model [Choppin 1968, Pesudovs 2010], as has been summarised already in our earlier paper in this journal [Pendrill and Petersson 2016].

The Rasch [1961] approach is thus not simply mathematical or statistical, but instead a specifically metrological approach to human-based measurement. Note that the same probability of success entering eq. (7) can be obtained with an able person performing a difficult task as with a less able person tackling an easier task. As with traditional measurement systems, the separation of attributes of the measured object (“item”) from those of the instrument (“person”) measuring them, achieved with the Rasch model, brings invariant measurement theory to psychometrics.

Referring to the quality-assured measurement of cognitive ability (mentioned at the start of this paper [OECD 2015]), a key work was made by Hughes et al. [2003], who were the first to do a Rasch analysis of Alzheimer disease sufferers. Their deployment of the Rasch approach demonstrated that established cognitive assessment instruments such as the MMSE have non-linear scales and the separate estimates provided of each patient’s cognitive ability and of each cognitive task’s degree of difficulty can be used to build metrological item banks which in turn can be used for future calibrations (the importance of which will be discussed further in section 3).

Summarising this section 2 so far, a full picture of the measurement process when Man acts as a measurement instrument can now be given, as in Figure 2, which complements previous work by presenting the process, step by step, from the observed indication (a performance metric, e.g. probability of success, $P_{\text{success}}$ of achieving a task [section 2.1]), and restitution with Rasch Measurement Theory [section 2.2], through to the measurand (e.g. task difficulty) in a form suitable for metrological quality assurance.

![Figure 2. Observation and restitution for performance metrics](image)

#### 2.2.3 Measurement units. Logistic measurement function

Having enabled with Rasch Measurement Theory a set of metrological references, e.g. for task difficulty, one can then proceed to set up a scale which is delineated by measurement units where any measured quantity, $\delta_j = \{\delta_j\} \cdot [\delta_j]$, is the product of a number $\{\}$ and a unit denoted in square brackets $[\]$. This step is enabled by combining a procedure to transform qualitative data to a new ‘space’ (in the present case, through restitution, to the space of the measurand, as
illustrated in Figure 2), together with ability of Rasch Measurement Theory to provide separate estimates of measurement and object dispersions in the results when Man acts as a measurement instrument. This new approach to the metrological treatment of qualitative data differs from others in that the special character of the qualitative data is assigned principally not to the measurand but to the response of the measurement system. (One can draw analogies with the common expression: “Beauty is in the eye of the beholder”.) Using Rasch Measurement Theory in the restitution process re-establishes a linear, quantitative scale for the measurand (e.g. for a property such as task difficulty) where metrological quality assurance – in terms of traceability and uncertainty - can be performed.

In order to include measurement units explicitly in the Rasch model, Humphry [2011] proposed a modified version of eq. (7), called a ‘logistic measurement function’:

$$\log \left( \frac{P_{\text{success},i,j}}{1 - P_{\text{success},i,j}} \right) = \rho_s \cdot (\theta_i^\ast - \delta_j^\ast)$$

(8)

where \(s\) indicates a classification of an empirical factor; \(\rho\) is a multiplicative constant; and the modified Rasch parameters are related to the original Rasch parameters through the expressions \(\theta_i^\ast = \theta_i / \rho\) and \(\delta_j^\ast = \delta_j / \rho\). If units are to be associated with person and item attributes, respectively, as \(\theta_i = [\theta_i] \cdot [\theta_i] \) and \(\delta_j = [\delta_j] \cdot [\delta_j] \) then assuming that item and person attributes share the same scale – a key aspect of the Rasch model - gives an expression for the ‘common unit’ of measure as: \([\theta_i] = [\delta_j]\) (denoted \([U_s]\) by Humphry [2011]).

Eq. 8 appears on first sight to be similar to Item Response Theory expressions, but there is a subtle distinction, as expressed recently: “Item Response Theory models are statistical models used to explain data, and the aim of an Item Response Theory analysis is to find the statistical model that best explains the observed data. By contrast, the aim of Rasch Measurement Theory is to determine the extent to which observed clinical outcome assessment data satisfy the measurement model” [Barbic and Cano 2016].

As pointed out by Humphry and Andrich [2008], the incorporation in an Item Response Theory model of a discrimination parameter which is estimated for each item (or person) will in general break conditions for sufficiency and specific objectivity, and thus the opportunity of establishing units and measurement scales. But this opportunity is maintained if one, as in eq. 8, associates a discrimination factor (\(\rho\)) with a set of items rather than a single item, according to Humphry and co-workers (Humphry [2011], Asril [2011]), as will be exemplified in the next section.

### 2.3 Testing limitations to the Rasch Measurement Theory

The Rasch separation assumption, essential for metrology, has of course to be tested experimentally and conceptually with every application.

#### 2.3.1 Scale and sensitivity distortions. Nuisance parameters

There is a well-established battery of consistency checks on the Rasch model which of course have to be made if one is to infer reliably metrological characteristics. A variety of analyses to detect expected types of misfit to the Rasch model has been developed over the years by the groups of Mislevy [2016, 1986, Dardick 2010]; and Wright [1995], to name a few.

The wider validity of the Rasch model in psychometrics is still under debate - many contemporary concerns about the discipline are included in a paper by Humphry [2011] and a series of comments in the same issue about his paper .
The idea of unit definitions based on Rasch models, and so also the potential of Rasch measurement to support metrological unit traceability, are controversial. A good fit to a Rasch model does not automatically confer properties of invariance, parameter separation, unidimensionality, etc. on scores or measures [Fisher Jr. 2015].

In a critique of the field, Kyngdon [2011] wrote: “False models can also fit data very well. The paired comparison Rasch model, for example, will fit data from human choices between simple lotteries. Yet this model is a descriptively false account of the utility of gains and losses under conditions of risk or uncertainty. For example, it treats preference reversals as “error” when they are experimentally robust phenomena caused by the way people weight the risk of uncertain events [Kahnemann and Tversky 1979]. If utility theorists were as content to fit probabilistic models to data as psychometricians generally are, our understanding of decision making under risk would be stuck in the mid-1960s.”

As described by Ghosh [1995], a ‘worst-case scenario’ would be where the number of nuisance parameters would “grow to infinity in direct proportion to the sample size”. This can be avoided and the viability of the Rasch model maintained, we claim, not only with statistical arguments [Freedman 2005, Stigler 2007, Spanos 2013] but also with a metrological understanding of Man as a Measurement Instrument [section 2.1]. As with any instrument, potential ‘nuisance’ parameters which might affect measurement system sensitivity and system offset [see eq. 2] can of course be compensated for by testing and calibration of the measurement system prior to use. This admittedly can be problematic the first time a new measurement system is deployed, since it is only after the first cohort of persons has been tested that metrological references can be established [see section 2.2.3]. It may also be problematic unless one can find sufficiently objective and sound measurement concepts [see further discussion in section 2.3.2].

The basic Rasch approach starts with the hypothesis that only one dimension underlies all of the items included in the analysis. Humphry’s [2011] formulation of the logistic measurement function (eq. 8) allows different scales for various sets of items or persons to be accommodated. A change in the discrimination factor, \( \rho_s \), in (eq. 8) could be associated with a set of items associated for a given cohort with a scale which is different from the scale associated with the majority of items, or equally a cohort which responds to a set of items with a scale which is different from the scale associated with the responses to these items from a majority of people.

An example of situations where the Rasch formula thus breaks down is where, because of illness, a cohort is more sensitive (e.g. for emotive reasons) to responding to questions about their performance compared with healthy individuals, as would be detected as “outliers” in goodness-of-fit tests. Pendrill [2017] treats such effects by re-examining the zstd-infit construct alleys commonly used in Rasch measurement theory [WINSTEPS]. In our opinion, both key aspects of metrology – comparability (through traceability) and uncertainty – in qualitative measurements can be accommodated by deducing the effects, respectively, of systematic distortions across the scale investigated and random noise on the response of our measurement system approach with performance metrics illustrated in Figure 2 [Pendrill 2017].

### 2.3.2 Meeting philosophical reservations

Metrological references need to be founded on objective and sound measurement. Lacking an independent objective reality, e.g. in the social sciences, might lead to measurements providing no unique ‘right’ answer. This would make metrology of such qualitative assessments challenging, since: (i) independent reference standards used in metrology to ensure the comparability of different measurements would be difficult to establish separately from the actual measurement process, and (ii) measurement uncertainty would often be very large, since each new measurement set-up would produce definitions divergent from others. There is a long and continuing debate about whether evaluations
(measurements, estimations and opinions) are indeed sound and/or objective at all in for instance the social sciences [Sawyer et al. 2016].

In considering the philosophical foundations of social measurement, Maul et al. [2016] recall various approaches, including empiricism, pragmatism and realism. The philosophical realism behind physical metrology assumes, as in physics, that there is an objective reality, which exists even when we do not perceive or have instruments to measure it. One might argue – which Mari et al [2016b] refer to in terms of the output of their evaluation process – that there is “seldom objective reality” in what is measured in social science (e.g., the challenge of a task) without our actually perceiving or measuring it. Mari et al [2016b] claim that a subjective opinion such as “I am thirty percent happier today than I was yesterday” does not “appear to deserve the trust that commonly accompanies measurement”.

From our point of view, we would provide the counterargument that, for instance, opinions about fine art – e.g. the Mona Lisa painting by da Vinci – appear to be rather constant over the centuries and across different cultures. And measures of happiness are becoming essential components of person-centred care. A Rasch approach to perceived beauty or perceived happiness (or other pleasing patterns and degrees of order or symmetry) would in fact provide separate measures of the (albeit noisy) individual preferences of different persons and the intrinsic ability of, respectively, Leonardo’s painting to stimulate pleasure or a particular activity of daily living to invoke happiness. This objectivity is perhaps not as strong as evaluations about the physical world (which would exist of course even without a human presence [Denbigh and Denbigh 1985]), but is so to say “fit for purpose” in the human-based context relevant for the present study. To use the vocabulary of the social sciences, such “fit for purpose” references provide not only objectivity but – importantly – also intersubjectivity [Gillespie and Cornish 2010].

Our approach as presented in section 2 can also be attributed to operationalism (as part of empiricism); i.e., defining a set of empirical operations performed with the measurement system. In that context, we circumvent the objections of realism since operationally it is “meaningless to ask whether something is ‘really’ being measured” [Maul et al. 2016].

In summary, the particular fusion of metrology and psychometrics proposed above, with its “fit for purpose” objectivity and operationalism, appears to go some way in countering several of the philosophical reservations that had been expressed about attempting to quality-assure measurements in the social sciences.

3 Example of metrology in person-centred care: Correlating loss of cognitive ability to brain atrophy

To illustrate the application of the above approach, recent studies of possible correlations between neurodegeneration in patients suffering from Alzheimer’s disease – specifically failing cognitive ability – and brain atrophy will be reviewed, as part of the on-going EMPIR HLT04 NeuroMet project. Tests of cognitive ability often include tasks related to everyday activities of the patient, such as naming objects, remembering and so on, and can be regarded as simple examples of person-centred measurements, in accordance with the WHO [2017] definition and in contrast with the more ‘technical’ aspects of biomarkers which are only surrogates for the impact of the disease. As described in the Introduction, this case is person-centred in at least two important aspects: (i) the patient’s health is of course in focus and not only the illness; and (ii) the physician and the patient are partners in care. The OECD [2015] data

\[v \text{‘Definitional uncertainty’ - “resulting from the finite amount of detail in the definition of a measurand” [VIM 2.27] - is of course in most cases much smaller in the strong objectivity of physics than in the social sciences.}\]
on prevalence of dementia mentioned at the beginning of the paper has to be examined from at least these two aspects of
person-centred care. Research is on-going about possible correlations between cognitive ability, measured with batteries
of neuropsychological tests, and various structures and functions of the brain, as reviewed by Klein-Koerkamp et al.
[2014]. Amongst the most common cognitive tests used by clinical specialists is the Mini-Mental State Examination
(MMSE), published by Folstein et al. [1975]. Klein-Koerkamp et al. [2014] in their earlier review cite a number of
studies where there are striking contrasts in the results of correlations analysed by the many different groups.

An example of a recent application of the MMSE test can be found of potential correlations between cognitive
ability and brain atrophy studied by Dinomais et al. [2016] - their original ("uncorrected") data is plotted in Figure 3 -
who claim that the identification of the location and nature of brain changes related to the MMSE has received little
attention to date.

![Figure 3 Correlation plots of cognitive ability (MMSE, dependent variable) versus regional grey matter volume (rGMV,
independent variable) ◆ original data [Dinomais et al. 2016]; X Rasch corrected for distortion shown in Figure 4 [Hughes
et al. 2003, and this work]. Standard uncertainties are indicated for MMSE ability scores.]

Amongst the conclusions of Dinomais et al. 2016 analysis of the (uncorrected) data shown in Figure 3, were that
“correlations between the MMSE score and the regional GM volume in right and left amygdala … persisted while
considering each disease category separately (i.e., CHI, MCI, and AD) …, which confirmed the interest of the MMSE for
this brain region in each subgroup of patients” (from the corresponding figure 3 in [Dinomais et al. 2016]).

The usual tools of statistics do not always work on the categorical scales typical of such MMSE measurements
[Svensson 2001] and an analysis with Rasch invariant measure theory described above [section 2.2], will indeed reveal
that cognitive scales, while reasonably linear at mid-range, are increasingly distorted as one approaches each scale
extremity – in the MMSE case, towards the high-end score of 30. In fact, thirteen years earlier, the linearity of the MMSE
scale was investigated thoroughly by Hughes et al. [2003] with the Rasch model. No researchers studying cognitive
correlations to date appear to have heeded the advice of Hughes et al. [2003]. The tendency of the MMSE average scores,
calculated traditionally with classical test theory (CTT) using ordinary statistics, for instance by Dinomais et al. [2016], to increasingly underestimate cognitive ability for the highest scores according to Hughes et al. [2003] is shown in Figure 4.

Figure 4  Distortion of MMSE scale where average scores calculated with classical test theory (CTT) successively underestimate cognitive ability (θ) according to Rasch analysis towards scale end (MMSE score = 30) [based on Hughes et al. 2003]

Challenges highlighted by Dinomais et al. [2016] in their studies of voxel-based morphometry included: segmentation of the brain into grey matter and white matter; mislabelling; and normalisation defects. An additional challenge we propose needs to be tackled is proper estimation of correlation between cognitive ability and regional grey matter volume, taking into account of the effects of MMSE scale distortion shown in Figure 4. These MMSE scale distortions can be corrected for in a metrological manner thanks to the Rasch invariant measure theory [section 2.2]. Such effects are over and above other potential sources of measurement error considered by Dinomais et al. [2016] who claimed to be "able to control for important characteristics that could modify the correlation, [although] residual potential confounders might still be present”.

Work is in progress in the EMPIR NeuroMet project to make a detailed re-appraisal of the MMSE, together with other cognitive tests and based on new data, which will extend the earlier Rasch analysis of Hughes et al. [2003] to include the establishment of a set – a ‘item bank’ [section 2.2.2] – of metrological references for cognitive tasks over a range of difficulty. Thus one could envisage a metrological reference scale for the calibration of cognitive assessment which typically might be spanned between the easiest and most difficult MMSE items such as immediate and delayed recall, respectively, of three words (such as ‘apple’, ‘penny’, ‘table’) as two examples of realisations of metrological references for cognitive difficulty.

The connexion between the Rasch model and entropy, in deriving equations (4) and (5) [section 2.2.1], is particularly relevant when defining the level of difficulty of different tasks used to test cognitive ability, since a lower
level of entropy of a more ordered task is easier to be performed by patients, and vice versa. Measurement units of cognitive task difficulty in the spirit of Humphry [2011] [section 2.2.3] follow once metrological references are available to define the scale. Limits on the reliability of the basic uni-dimensional Rasch model are critically reviewed in a separate presentation by the author [Pendrill 2017]. An approach to establishing metrological references with pragmatic operationalism overcomes in our opinion philosophical reservations expressed by others [section 2.3.2] and provides fitness-for-purpose adequate to the quality-assurance of person-centred care. Metrological references for cognitive task difficulty promise to improve the quality of person-centred care of Alzheimer and other neurodegenerative diseases by enabling the traceable calibration of both additional cognitive tasks as well as the cognitive ability of each individual patient.

This paper presents for the first time a unified model giving a full picture of the measurement process when Man acts as a measurement instrument, as in Figure 2, which will aid physicians (and patients) in providing quality-assured metrology of person-centred care in the case of Alzheimer sufferers as well as many other diseases. It is outside the scope of the present paper to give all details of specific application of the new model, but the reader is referred instead to up-coming publications mentioned above and results of the EMPIR NeuroMet project.

4 Conclusion

Metrological quality assurance of person-centred outcome measures is essential if reliable decisions about care needed to diagnose, treat and rehabilitate are to be made throughout the healthcare system.

A new model of the metrology of ordinal data, including performance metrics and restitution, has been presented. An example has been given of how the Rasch invariant measure theory can improve contemporary studies of possible correlations between neurodegeneration in patients suffering from Alzheimer’s disease – specifically failing cognitive ability – and brain atrophy. It is the first time to our knowledge that the known distortions - revealed by the Rasch psychometric model - of measurement scales in common clinical instruments for cognitive assessment have been corrected for in correlation studies. Possibilities of higher resolution and more reliable clinical decisions open up. The method can be extended to other qualitative measurement situations.

While of course one cannot claim that this completely explains all of the discrepancies [OECD 2015] in the apparent prevalence of dementia amongst different countries, deploying the Rasch invariant measure theory is one small step towards meeting the need for metrological quality assurance in care recognised by amongst others the OECD [2017].

The 15HLT04 NeuroMet project is currently studying a battery of cognitive tests regularly used in clinical contexts as well as correlations between cognitive ability and a range of biomarkers for Alzheimer’s disease with the overall aim of metrological quality assurance in care.

Acknowledgments

This work has been performed as part of the 15HLT04 NeuroMet project which belongs to the European Metrology Programme for Innovation & Research (Horizon 2020), jointly funded by the EMRP participating countries within EURAMET (www.euramet.org) and the European Union.

Thanks are due, particularly, to:

- Members of the EMPIR NeuroMet consortium
- Stefan Cano, Modus Outcomes, Boston (USA) & Stotford UK
References


A Asril 2011, Evaluation of the Indonesian Scholastic Aptitude Test According to the Rasch Model and Its Paradigm, Thesis for the degree of Doctor of Philosophy, University of Western Australia, Graduate School of Education


T Bond and C Fox 2015 Applying the Rasch model (New York, Lawrence Erlbaum Associates)


W R Dardick 2010, “Reweighting Data in the Spirit Of Tukey: Using Bayesian Posterior Probabilities as Rasch Residuals for Studying Misfit”, Dissertation submitted to the Faculty of the Graduate School of the University of Maryland, College Park, in partial fulfilment of the requirements for the degree of Doctor of Philosophy


M Dinomais, S Celle, G T Duval, F Roche, S Henri, R Bartha, O Beauchet and C Anweiler 2016 “Anatomic Correlation of the Mini-Mental State Examination: A Voxel-Based Morphometric Study in Older Adults”, *PLOS ONE* | DOI:10.1371/journal.pone.0162889


L F Hughes, K Perkins, B D Wright and H Westrick 2003, “Using a Rasch scale to characterize the clinical features of patients with a clinical diagnosis of uncertain, probable, or possible Alzheimer disease at intake”, *Journal of Alzheimer’s Disease*, **5**, 367 – 73


ISO 5725 1995 Accuracy – trueness and precision


ISO/IEC DIS 17025:2017, General requirements for the competence of testing and calibration laboratories, section 7.1 Review of requests, tenders and contracts


JCGM 100:2008 Evaluation of measurement data – Guide to the expression of uncertainty in measurement (GUM 1995 with minor corrections) in Joint Committee on Guides in Metrology (JCGM).

JCGM 106:2012, Evaluation of measurement data – The role of measurement uncertainty in Conformity Assessment, in Joint Committee on Guides in Metrology (JCGM).


D. Kahnemann and A. Tversky 1979, “Prospect theory: An analysis of decision under risk”, Econometrica, 47, 263-9


G.N. Masters 1982, "A Rasch model for partial credit scoring" Psychometrika, 47, 149-174
OMERACT 2017, Outcome measures in Rheumatology. https://www.omeract.org/
PROMIS 2017 Dynamic tools to measure health outcomes from the patient perspective http://www.nihpromis.org/measures/instrumentoverview


A Spanos 2013, “Revisiting the Neymann-Scott model: an Inconsistent MLE or an Ill-defined Model?” arXiv:1301.6278v1


M Wilson 2013 “Using the concept of a measurement system to characterize measurement models used in psychometrics”, *Measurement* **46** 3766-3774


