Uniting Psychometric Modelling and Poisson Distributions: A Metrological Study of Elementary Counting

Leslie R Pendrill * and William P Fisher Jr.

Posted Date: 30 April 2026

doi: 10.20944/preprints202604.2173.v1

Keywords: quality assurance; metrology; probability mass function; analytical; clinical; ordinal; discrete; Poisson; counting; psychometric

 doi:10.20944/preprints202604.2173.v1

Article

Uniting Psychometric Modelling and Poisson Distributions: A Metrological Study of Elementary Counting

Leslie R Pendrill [1] **and William P Fisher Jr.** [2,*]

[1] RISE, Research Institutes of Sweden, Division Measurement Science and Technology, Gothenburg, Sweden
[2] Living Capital Metrics LLC, Sausalito, CA, USA; BEAR Center, Berkeley School of Education, University of California, Berkeley, CA, USA
* Correspondence: leslie.pendrill@ri.se; Tel.+46-0703-94-5444

Abstract

A study of elementary counting (of simple clouds of dots by the Munduruku indigenous people of Brazil) is reanalysed in order to compare and contrast three kinds of probability mass functions (PMFs): (i) quantitative response to a discrete range of counts, (ii) the classic Poisson distribution of miscounts, and (iii) psychometric (Rasch) distributions of counting task difficulty and person counting ability. PMFs provide a means of defining — for discrete and qualitative data — the basic metrics, viz. location and dispersion, of metrology — quality-assured measurement, as increasingly required since the turn of the millenium in topical and challenging quality-assurance applications, amongst others, in the human sciences and in Artificial Intelligence. PMF-based metrics, useful in 'clinical' and other applications where meaning and value are sought, complement the traditionally dominating role played by the corresponding probability density functions (PDF) in 'analytical', quantitative and continuous Metrology in Physics. New insights are provided when benchmarking the Rasch Poisson Counts Model, which has received less attention in modern metrology, against full psychometric Rasch modelling.

Keywords: quality assurance; metrology; probability mass function; analytical; clinical; ordinal; discrete; Poisson; counting; psychometric

1. Introduction

Probability mass functions (PMFs) (Section 2.1) provide a means of defining — for discrete and qualitative data — the basic metrics, viz. location (Equation (5)) and dispersion (Equation (6)), of metrology when assuring the quality of measurement in terms of traceability and uncertainty. Quality-assurance of those metrics, following satisfaction of constraints concerning statistical sufficiency and additivity, in turn supports, respectively, interoperability and defining limits on the quality of products and services of all kinds [1–3].

Deploying PMFs for discrete and qualitative data thereby complements the dominating role for these location and dispersion metrics traditionally played by PDFs, the corresponding probability density functions, in quantitative and continuous Metrology in Physics.

Location and dispersion metrics for discrete and qualitative data are increasingly required since the turn of the millennium in topical and challenging quality-assurance applications, amongst others, in the human sciences [4], sustainability [5] and in Artificial Intelligence [6].

The metrology of PMFs remains so far relatively undeveloped, and the present work presents a reanalysis of a study of elementary counting (of simple clouds of dots by the Munduruku indigenous people of Brazil, [7,8]) as a means of comparing and contrasting the metrology of three kinds of probability mass functions (PMFs) for the four scenarios shown in Table 1: (i) quantitative response to a discrete range of counts (Section 3.3), [9], (ii) the classic Poisson distribution of miscounts (Section

 doi:10.20944/preprints202604.2173.v1

4.3.3), and (iii) psychometric (Rasch) distributions of counting task difficulty and person counting ability, (Section 5) for one of the conceptually simplest cases of measurement. The analyses illustrate the typical challenges normally faced when determining quality-assured metrics with the different kinds of PMF.

Table 1. Discrete and continuous ranges versus qualitative and quantitative scales

	Discrete	**Continuous**
Qualitative	Instrument response: P_{success} per category/class	Clinical performance (point 2): Instrument ability, θ, $u(\theta)$ Task difficulty, δ, $u(\delta)$
Quantitative	Analytical counting (Figure 1) How many dots in object? Counting errors Limit of detection	Analytical measure (point 1): How much of a quantity in object? Measurement errors and uncertainties Trueness & precision

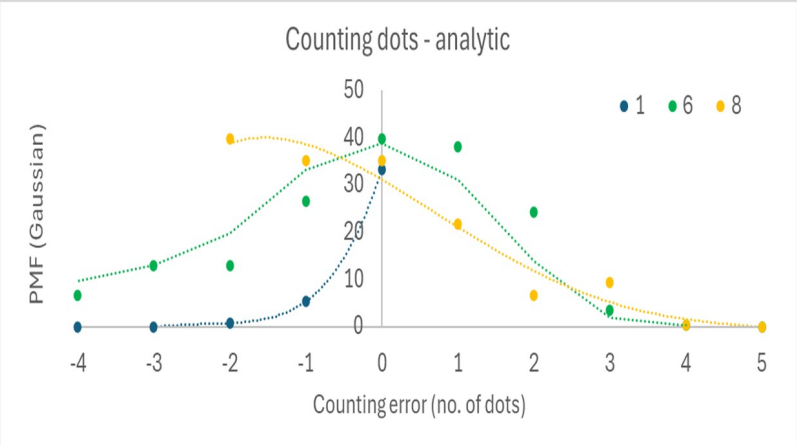

Figure 1. Analytical PMFs of distributions of perceived counting errors for three counts: (1 dot, 6 dots and 8 dots) in the study [7] of elementary counting tasks ('items', j) of sets of clouds of increasing number of dots and task difficulty. Lines for each count PMF assume a Gaussian distribution of counting error, $p_x = N((\mu - \mu_0), \sigma^2)$, $\mu =$ perceived count, Equation (5); $\sigma = 1$ dot, perceived counting dispersion, Equation (6); $\mu_0 =$true count.

The Discussion (Section 6) provides the motives for choosing an elementary construct. Classification, such as counting dots, despite sometimes having relatively large uncertainties and being considered 'trivial', is one of conceptually simplest metrological cases with a well-defined measurand, as will be illustrated. In a summary of this and other limitations in the present study, perspectives will also be given on future work including consideration of other measurement systems where a Human as a Measurement Instrument is replaced by for instance an AI agent and multidimensional measurement theories. This includes proper account of 'anti-racist', 'culturally responsive' and 'culturally specific' measurement applications being developed, for example, by Mallinson [10] and Sul [11–13].

While the metrology of the 20th Century was dominated by measurement uncertainty studies in the bottom-right entry in Table 1, based on 'analytical' measurement of physical properties in trade and industry [14], the first quarter century of the millennium has seen an emergence [3] of metrology in all the other entries, discrete and (for want of a better term) 'clinical' measurements, Section 3.2 . This reflects a growing trend towards more accountability in terms of meaning, value and effectiveness [15], and not mere numbers and efficiency.

2. PMFs and Measurement

Examples of PMFs abound in the **human sciences**, education, sustainability and the health sciences (in surveys, sensory panels and educational examinations) and by extension increasingly in the area of Artificial Intelligence as well as other applications outside of traditional physical metrology

 doi:10.20944/preprints202604.2173.v1

[3]. Models used when quality-assuring PMFs, particularly those based on qualitative observations, such as item response theory and log-odds Equation (17) in the educational, medical and social sciences, are not exclusive to human agents but apply in many cases equally well (Section 4.3.1) to more technical and physical measurement systems where they have so far not been regarded as belonging to mainstream Metrology.

2.1. Probability Mass Functions

The discrete counterparts of PDFs—probability mass functions (PMF)—are distributions of probability density (*ordinate*) over one or more discrete ranges of a variable (*abscissa*).

Let X be a discrete random variable with a range of C categories:

$$R_X = x_1, x_2, ..., x_C \tag{1}$$

An event $A = X = x_c$ is defined as the set of outcomes, s, in the sample space, S, for which the corresponding value of X is equal to x_c for a response in category c. In particular, the event:

$$A = \ s \in S | X(s) = x_c \tag{2}$$

The probabilities of events $X = x_c$ are formally shown by the **probability mass function (PMF)** of X:

$$P_X(x) = \begin{Bmatrix} P(X = x_c) & if\ x \in R_X \\ 0 & if\ x \notin R_X \end{Bmatrix} \tag{3}$$

for the range $R_X = x_1, x_2, ...$(finite or countably infinite).

The probabilities for the different categories, c, $P(X = x_c) = N_c/N_S$, that is, the occupancy, N_c, of a bin for category c as a fraction (or corresponding probability) of the sample size, N_S, must be non-negative and for PMFs are normalised to sum up to 1, respectively:

$$P_X(x) \geq 0$$

and

$$\sum_x P_X(x) = 1 \tag{4}$$

In the examples referred to in this paper, mention will be made of:

- Probabilities
  The probabilities can be either derived from actual frequencies (occupancy per bin) or estimated in terms of prior knowledge or even judgment when scores are set. Figure 2 shows several Poisson PMFs (Section 4.2) as an example. When setting response scores as probabilities P, choices have to be made about how to define the scale-end scores of 0 and 100. Inferentially stable measurement also requires that the score varies monotonically across the scale in a manner allowing summarization which exhausts the data of all available information in a minimally sufficient statistic, [16–18]. (These requirements, [19], are subsequently checked (Section 5.5)).
- Ranges
  - **intrinsically discrete ranges,** R_X (Equation (1)) (such as when counting dots (Section 3.3)).
  - **discrete ranges chosen for convenience** (such as in sensory panel responses) where the observed variable is on a continuous scale, but the large uncertainties make it practical to round off the score, say to the nearest integer.
  - ranges can be as short as two, such as for the binomial distribution from binary Bernoulli trials (Section 4.1) as well as long, polytomous ranges spanning several categories (Section 4.3.2).

 doi:10.20944/preprints202604.2173.v1

- Scales
  In different application areas, the scales associated with both X (abscissa axis) and P (ordinate axis) of a PMF can be:
  - either fully **quantitative** (Section 3.3),
  - or more **qualitative** (Section 5)

  ranging, respectively, from ratio and interval scales to the ordinal and nominal scales.

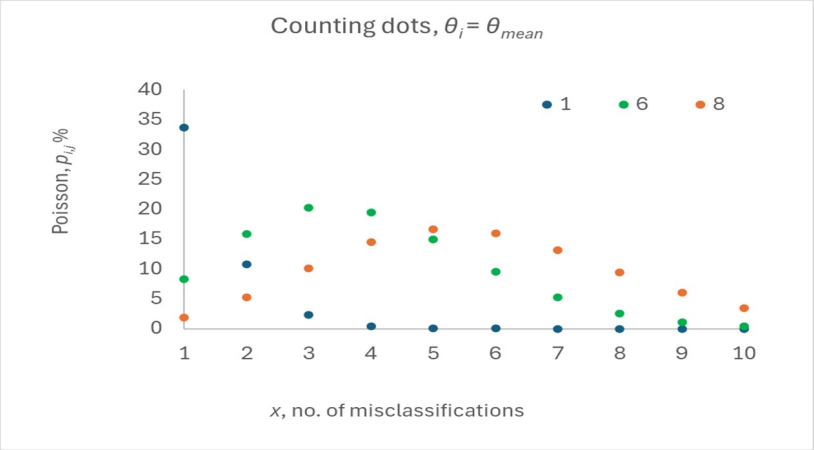

Figure 2. Poisson PMFs (Equation (21) according to [51]) for three counting tasks: one dot, 6 dots and 8 dots of increasing levels of difficulty (Equation (26). Counter ability taken as the cohort mean, θ_{mean}

There is no obvious one-to-one correlation between discrete and continuous ranges versus qualitative and quantitative scales [20]. Table 1 shows a matrix of possible correlations.

2.2. Accuracy of Classification

The accuracy of classification, in terms of both precision and trueness ISO 5725 [21], involves, respectively, differences between observed and 'correct' in terms of the location, Equation (5) and the dispersion of PMFs, Equation (6) [2] (Section 4.1.1): where (provided distances between different categories are known, Section 4.1.2)

$$\hat{p}_x = \sum_x x \cdot p_x; \tag{5}$$

where $p_x = P(X = x_c)$.

$$\sigma(\hat{p}_x) = \frac{1}{C-1}\sqrt{\sum_x ((x \cdot p_x - \hat{p}_x)^2)}. \tag{6}$$

Here, trueness is a measure of closeness of the mean to the 'true' value while precision is the dispersion of repeated responses (without reference to a 'true' value).

The concept of Accuracy can be considered for both analytic and clinical performance metrics (Section 3.2):

- analytical accuracy.
  In an 'analytical' scenario 1, the 'accuracy' estimation of the 'measurand', the quantitative number of each set of discrete objects (with the number of dots increasing from 1 to 10 in the present counting case), can be expressed in terms of:
  - (i) trueness—estimated as the difference between the perceived and true dot count, $x - \mu$.
  - (ii) precision—estimated in terms of dispersion, such as σ dots, (Figure 1).
- clinical accuracy
  In a 'clinical' scenario 2, distances between different categories of classification on the *abscissa* of a PMF (such as when measuring the difference before or after an intervention, or when

 doi:10.20944/preprints202604.2173.v1

estimating dispersion measures) may not be fully known or even meaningful (such as on nominal scales (Section 4.1)). Appropriate methods for dealing with such scales include log-odds ratio transformations, including GLMM and the Rasch psychometric theory with which 'measurands' such as counter ability and task difficulty can be identified, as dealt with later in the paper in our clinical counting example (Section 5).

Alongside concepts about PMFs found in the wider literature, such as so-called 'measure theoretic formulations', a proper *metrological* approach describes how PMFs should be associated with various stages in the *measurement* process (Section 3.1).

3. Analytical and Clinical Performance Metrics

3.1. PMF and Measurement System Analysis (MSA)

PMFs are a common tool of statistics, but here there are also **metrological** aspects. Uncertainty in counting, for instance, arises both from the identification of what is being counted and from the experimental counting procedure [22].

A proper metrological approach describes how PMFs should be associated at various stages in the measurement process, corresponding to the propagation stage of establishing a measurement result. This means consideration of how measurement information is 'transmitted' (as in any communication system) through a **measurement system** [object – instrument – observer – environment and method, ([24], Figure 14) and Figure 3, ([2] p.198)] with a particular focus on an **object** as 'probed' with an **instrument** of some kind. This MSA analysis will guide the formulation and propagation stages for PMFs in establishing a measurement result from initial observations through to restitution of the measurand(s) [23]. For the present case of measurement systems involving discrete PMFs, inputs, such as X_1, when restituting the measurands, can be characterised in terms of informational entropy ([2] Figure 5.4), assigned on the basis of available knowledge (Section 5.3).

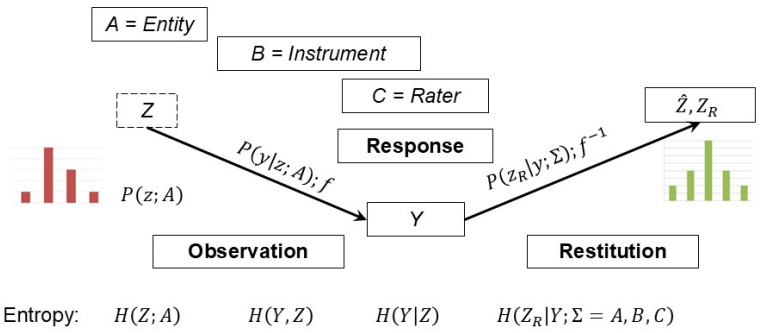

Figure 3. Probabilistic and entropy models of the measurement system and processes, inspired ([2] p.198), in part by the probabilistic model of Rossi ([23] Figure 5.5). Z—measurand; Y—response; R—restitution via uncertainty model (3.1); P—probability, and H—entropy. The measurement process is described in terms of a metrological measurement system analysis (MSA) with measurement object (entity, A), instrument (B), operator (C). The concept of entropy (H), evaluated, Equation (24), for the PMF at each stage of the measurement process, is invoked to aid explanation of how information is lost, distorted or gained from observation of the measurand (Z), via response (Y) to restitution (Z_R). The calibration function, f, relates the instrument response to the input from the measurement object.

3.2. Analytical and Clinical Performance Criteria

Different attributes associated with various elements (*MSA object – instrument – observer – environment and method*) of the measurement system (Section 3.1) can be analysed in terms of PMFs. For each MSA element attribute, two principal kinds of performance criteria — 'analytical' and 'clinical' —

 doi:10.20944/preprints202604.2173.v1

can be identified, which Hofmann [25] places as essential intermediaries between, on the one hand, *epistemic* relevance and, on the other, *meaningfulness* (ultimately for patient health). This can be topically exemplified in the context of regulation and conformity assessment of *in vitro* devices, specified in EU regulations [26,27], although equally applicable to our elementary case of counting dots (Figure 4):

1. '**Analytical**' performance criteria for determining, e.g., how much (*quality characteristic*: concentration) of a particular analyte (*MSA object*) is present in a sampled object (by "variable"), such as, first, analytical method accuracy (trueness and precision, [21]) (Section 2.2) and, second, sensitivity, such as instrument limit of detection (*MSA measurement instrument*), as exemplified in the analytical interpretation of the elementary counting case of the present study, 3.3. According to [26], "...Analytical performance focuses on the gathering of evidence that the measurement instrument in question reliably, accurately and consistently measures and or detects an analyte". This is closely related to terminology in acceptance sampling standards [28], §3.1 where **Inspection by variables** is inspection by measuring the magnitude(s) of a characteristic(s) of an item.
2. '**Clinical**' performance , according to [26], "aims to demonstrate that the measurement instrument can achieve clinically relevant outputs through predictable and reliable use by the intended users". This is closely related to terminology in the definition §3.1.3 of [29], where Inspection by attribute is "inspection whereby either the item is classified simply as conforming or nonconforming with respect to a specified requirement or set of specified requirements, or the number of nonconformities in the item is counted". Commonly used clinical performance metrics of the measurement systems include: selectivity, Equation (22) and specificity, Equation (23), which are plotted against each other on receiver operating characteristic curves, [30], when sampling by 'attribute'. A psychometric treatment of these clinical performance metrics [31,32] yields quantitative estimates for *quality characteristics* such as task difficulty (*MSA object*) and agent ability (*MSA measurement instrument*) as attributes of the different elements of the measurement system illustrated in Figure 3, corresponding to the top-right entry in Table 1.

3.3. PMFs for Counting and Related Tasks. 'Analytical' and 'Sampling by Variable'

The **elementary counting** of dots by less numerate agents (e.g., children or indigenous people [7,8], Figure 4) presents a case of the metrology of discrete counting as a typical 'analytical' or 'sampling by variable' procedure, point Section 1 and bottom-left entry in Table 1.

This study of elementary counting has several advantages, including the opportunity of analysing the measurement response where the analytical measurand (quantity intended to be measured) — the number of dots (1...10) — is:

- (i) in fact known exactly
- (ii) conceptually simple

In fact, this elementary case can exemplify each of the elements of Table 1, as will be shown throughout this paper.

 doi:10.20944/preprints202604.2173.v1

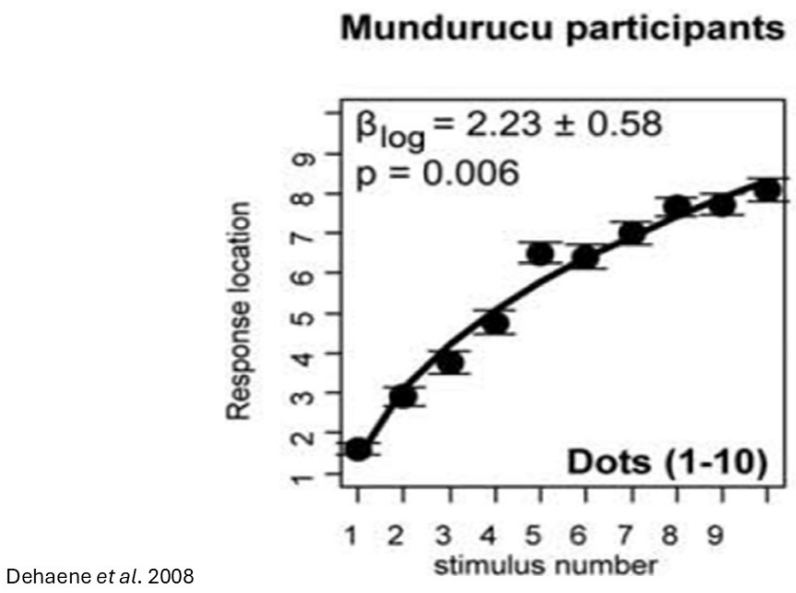

Figure 4. Perceived counts ('Response location') from a study [7] of elementary counting tasks ('items', *j*) of a set of clouds of increasing ('stimulus') number and counting task difficulty of dots

As the number of dots to be counted exceeds about 5 (the number of digits on one hand traditionally used by the agents for counting), each estimated count becomes successively more under-estimated [7], Figure 4. Similar observations of limitations in counting ability can be found in children and educated adults when they are given only a limited time to perform each counting task. An initial 'analytical' scenario 1 is the estimation of the measurand: the quantitative number of a set of discrete objects (with the number of dots increasing from 1 to 10) in the data shown in Figure 4. A count of 5 dots, where perceived counts deviate substantially from the known, true number of dots, could be considered a 'limit of detection' as a analytical performance metric.

These successive count under-estimations can be described in terms of PMFs distributed over an increasing number of counting categories and shifted towards successively underestimated values, shown in Figure 1. Analytical PMFs of distributions of perceived counting errors can be plotted, for example, for three counts: (1 dot, 6 dots and 8 dots) in the study [7] of elementary counting tasks ('items', *j*) of sets of clouds of increasing number of dots and task difficulty. The lines drawn for each count PMF assume a Gaussian distribution of counting error, $p_x = N((\mu - \mu_0), \sigma^2)$, μ = perceived count, Equation (5); $\sigma = 1$ dot, perceived counting dispersion, Equation (6); μ_0 =true count.

4. Case Studies of PMFs: Quantitative Statistical Process Control

PMFs have been invoked for many decades for a number of well-known sampling distributions in the context of trials for statistical process control (SPC, [33]) as one major area of application. In general, the parameters of a production process are unknown and can change with time. Procedures to estimate the parameters of probability distributions and solve other inference or decision-oriented problems related to them need to be developed. In this section, we will motivate how SPC PMFs can be deployed when tackling the metrology of clinical performance 2, with a particular focus on the elementary counting example, Figure 4.

4.1. Binomial Distribution and Dichotomous Bernoulli Trials in SPC

A basic kind of PMF is the **dichotomous** Bernoulli variant shown in Figure 5. For an elementary counting case, a counter may initially estimate the number of dots as one or other of two adjacent integers, say '9' or '10' when counting a cloud of ten dots, corresponding to the top-left entry in Table 1.

 doi:10.20944/preprints202604.2173.v1

For a **dichotomous** [35] 'trial' , where a 'success' is scored 1 and a 'failure', 0, let the probability of categorising a 'successful' response be p. Values on the ordinate (vertical) axis of the binomial PMF shown in Figure 5 are probability density values:

$$P_X(x) = \left\{ \begin{array}{ll} P(X = x_c) = \frac{count_c}{N_{sample}} = \frac{\sum_{i=1}^{N_{sample}} [z_i = c]}{N_{sample}} & if\ x \in R_X \\ 0 & if\ x \notin R_X \end{array} \right\} \quad (7)$$

On a **quantitative scale on the abscissa**, the PMF plotted in Figure 5 (based on Monte Carlo (MC) simulations of sampling trials [34]) has the following means and standard deviations: $\hat{p} = 0.5; \sigma(\hat{p}) = 0.28$, while equations (5) and (6) yield:

$$\hat{p} = \sum_{c=0}^{c=1} c \cdot p_c = 0.5 \quad (8)$$

$$\sigma(\hat{p}) = \frac{\sqrt{\sum_{c=0}^{c=1} (c \cdot p_c - \hat{p})^2}}{C - 1} = 0.3. \quad (9)$$

The binomial distribution is a limiting form of the Poisson distribution (Section 4.2) for an infinite number of Bernoulli trials and an infinitely small probability of non-conforming product [33].

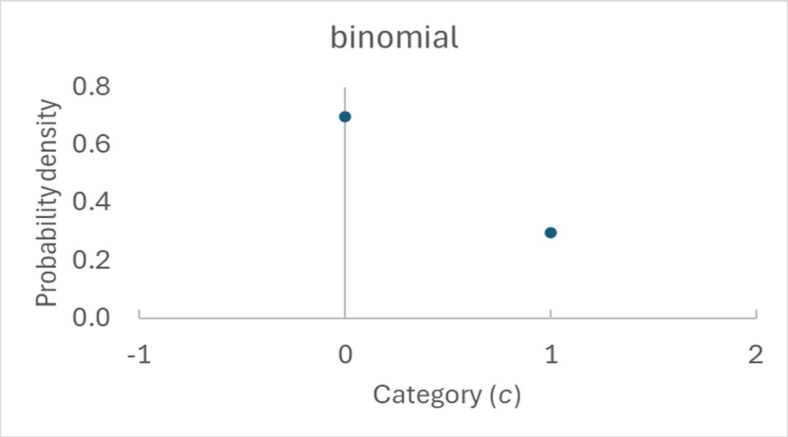

Figure 5. PMF for simulated **binomial** dichotomous distribution (probability of success $P_{success} = 0.3$, Monte Carlo sample size 10^6) [34]

4.1.1. Dichotomous Decision-Making with Uncertainty

A typical decision of conformity is whether the attribute of an entity (e.g., a product or sample) can be classified as above ('positive') or below ('negative') a specification limit, SL, for that quality characteristic. However, measurement uncertainty leads generally to the risk of incorrect decisions and classifications in such conformity assessments, as described in [14] JCGM 106:2012.

An observed response in a performance test is typically scored binarily (binomial PMF, Figure 5) in one of two categories on the abscissa: as $c = 1$ for correct and $c = 0$ for incorrect and presented on the PMF ordinate as $P_{success,i,j}$ for each instrument, i, to a specific object, j, when examined in terms of Measurement System Analysis (Section 3.1). This simplest binary classification can be extended to polytomous decision-making over a range of classification categories 4.2. In the elementary counting case, this would correspond to 'guessing' the count to lie in a range of integer numbers of dots for a given cloud as an analytic measurand, as is evident in the results shown in Figure 1. A corresponding clinical performance study (2) would identify two measurands: the (i) ability of the counter and the (ii) difficulty of each counting task derived from the probability of correct classification, corresponding, respectively to the top-right and top-left entries in Table 1. In any case, the decision quandary reflects uncertainty.

 doi:10.20944/preprints202604.2173.v1

4.1.2. Distances on Categorical Scales: Counted Fractions

Distances on categorical scales, such as on a binomial PMF, Figure 5, (but also PMFs for other distributions, such as the Poisson and psychometric) can be challenging to evaluate correctly in the case of hidden, unrecognised non-linearity or simply a non-numerical scale (2.1). This 'scale ordinality' obliges relatively more emphasis to be placed on the PMF ordinate values compared with more quantitative cases with PDFs such as the analytical (1). Typical clinical performance (2) cases include assessing the effects of an intervention or the relative performance of two counters (Figure 1) in terms of distances between PMFs.

As defined in Equation (4), the probabilities for the different categories of any PMF, c, $P(X = x_c)$, must be non-negative and sum up to 1: $P_X(x) \geq 0$ and $\sum_x P_X(x) = 1$.

A general feature of categorical data in which fractions on a finite scale are counted on any bounded scale is referred to as the '**counted fraction**' effect, [36]. Put simply, any score on a fixed scale on bounded by 0 and 100% becomes increasingly non-linear at either extreme of the scale. A difference of, for example, 1% percentage points in score $P_{success}$ of successful classification is a different amount at mid-scale (50%) compared with scores at 5% or 95%. This effect will be explained and compensated for as follows:

An early statement by Pearson [37]: "*Beware of attempts to interpret correlations between ratios whose numerators and denominators contain common parts*".

The counted fraction effect has been explained in terms of the occurrence of the same term, Z_j, in the numerator and denominator of the score $Z_j = Z_j/(\sum_{k=1}^{K} Z_k)$.

Tukey [36] wrote further: "*As a general consequence we should expect that scales which have a finite range are likely to give us trouble unless all our observations tend to be safely away from any ends which are present. Hence the fact that percentages go only from one end (at 0%) to another (at 100%) suggests that, whenever even moderately extreme percentages are likely to occur, we are likely to have to 'stretch the tails', while, if really extreme percentages occur, we may have to stretch hard enough so that there are no ends (at any finite values).*"

Filzmoser [38] stated for basically the same concept in the context of mineral compositional data studies.: "*If not all variables or components have been analysed, this constant sum property*, (i.e., $\sum_x P_X(x) = 1$ (Section 2.1)) *makes the relation between variables not 'real' but 'forced'...For example, if the concentrations of chemical elements are measured in soil samples, and if an element like SiO_2 has a big proportion of, say, 70%, then automatically the sum of the remaining element concentrations is at most 30%. Increasing values of SiO_2 automatically lead to decreasing values of the other elements, and even if not all elements of the soil have been measured, the correlations will be mainly driven by the constant sum constraint.*"

Non-linearity in the raw response scale, such as arises from the counted fraction effect [39] can be readily revealed by plotting, as in Figures 6 and 7, the marginal probabilities for the different items j (cloud of dots to be counted, for example), given by: $\frac{\sum_{i=1}^{N_{TP}} y_{i,j}}{N_{TP}}$ against the more quantitative scales from Rasch Measurement Theory (RMT) [40] for the task difficulties, δ_j, (logits), as will be described below.

Scale non-linearity caused by the counted-fraction effect is ubiquitous to all PMFs because of the limited percentage scale, Equation (4), but if not recognised (as in Classical Test Theory, CTT [41]) can lead to serious underestimation of true scores at either end of the response scale, as shown in Figures 6 and 7, compared with the substantially linear region close to mid-scale, Equation (10) (Section 4.1.3).

 doi:10.20944/preprints202604.2173.v1

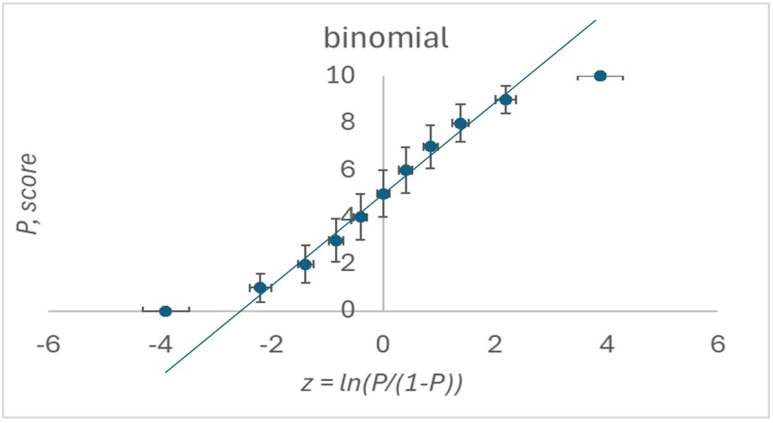

Figure 6. Binomial ogive curve showing counted-fraction non-linearity at either end of the response scale (Section 4.1.2).

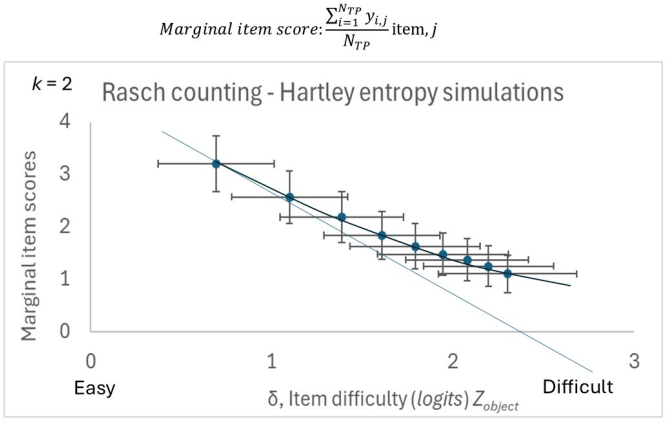

Figure 7. Marginal response scores versus log-odds for the [7] dot counting data (Section 5) showing counted-fraction non-linearity (Section 4.1.2) at the lower end of the response scale. Note the opposite sign on the logit x-axis compared with the x-axis of Figure 6, which is due to the link function: $z = \theta - \delta$.

4.1.3. Compensating for Counted-Fraction Scale Non-Linearity

There is a substantial and growing body of literature ([2], p. 88 - 9) for a diversity of applications, such as psychometry [40] and compositional data analysis [38,42] where the counted fraction effect is recognised. The 'counted fraction' effect can be compensated for using a log-odds (GLMM) approach, Equation (11):

Ordinate values compensated for the counted fractions effect are given as a straight line $y = m{\cdot}z + c$ as plotted in Figures 7 and 6 where two statistics are known:

- the intercept $c = 0.5$ at $z = 0$
- the slope

$$m = \frac{dp}{dz} = p(z) \cdot (1 - p(z)) = 0.25 \tag{10}$$

at $z = 0, p = 0.5$

These two parameters, which apply to **all bounded-scale fractions**, define the anchor (at $P_{success} =$ 50 %) and the span (the slope) of the linear, quantitative scale resulting from a log-odds transformation Equation 11, thus giving the **corrections to raw performance data for the counted-fraction effect for any PMF**.

 doi:10.20944/preprints202604.2173.v1

Ordinality in responses is generally compensated for the effects of counted fractions by taking log-odds ratios (*OR*) [[43]] i.e.,

$$\log(OR) = \log(\frac{x_j}{1-x_j}) = z \tag{11}$$

The RHS of Equation (11) is referred to as a 'link function' in the literature in the Generalised Linear Measurement Models (GLMM) approach [44] and provides the means for common-person and common-item equating in psychometrics, following Wright's expansions of Rasch's models ([45] pp. 109-117). For current application, see [46].

4.2. Poisson Distributions in SPC

The Poisson is another widely distribution used to model count data and random occurrences over time or space, such as arrivals in a queue, radioactive decay events or rare defects in SPC production, [47]. We will also examine (Section 4.3.3) how the Poisson distribution can be useful when defining clinical performance metrics for the elementary counting case, Figure 4.

Let λ denote the expected number of events per interval, and assume λ is known. Then, the distribution of X is Poisson with parameter λ:

$$X \sim \text{Poisson}(\lambda).$$

The PMF for X is

$$p_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, \ldots, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

X is the quality characteristic, x, being classified (e.g., number of defects) in n observations.

X has expectation and variance ([33], p. 55).

$$E(X) = \lambda, \qquad V(X) = \lambda. \tag{13}$$

The Poisson and binomial distributions are closely related: The Poisson distribution can be regarded as a limiting form of the binomial distribution (Section 4.1). Allowing the binomial parameters $n \to \infty$ and $p \to 0$ in such a way that $n \cdot p = \lambda$ results in a Poisson distribution [33]. According to Meredith [48] the Poisson distribution: *"also occurs as a limit distribution for a number of discrete distributions other than the binomial. Two of the more important instances of this are the negative binomial or Pascal distribution and the hypergeometric distribution."*

As for the binomial (dichotomous Bernoulli trials) case (Section 4.1), values on the ordinate (vertical) axis of the Poisson PMF are probability density values for each x:

$$p_x = count_x / N_{sample} = \sum_{i=1}^{N_{sample}} [z_i = x] / N_{sample} \tag{14}$$

On a **quantitative scale on the abscissa**, each Poisson PMF — plotted in Figure 2 for the elementary counting case — has the mean and standard deviation given by:

$$\hat{p}_x = \sum_{x=0}^{x=X} x \cdot p_x; \tag{15}$$

$$\sigma(\hat{p}_x) = 1/(X-1) \cdot \sqrt{\sum_{x=0}^{x=X} ((x \cdot p_x - \hat{p}_x)^2)}. \tag{16}$$

 doi:10.20944/preprints202604.2173.v1

4.3. Rasch Psychometric Model. Principle of Specific Objectivity

The output of the uncertainty model (3.1) will ultimately provide (through 'restitution'), from an MSA perspective, Figure 3, our best estimates of the instrument attribute (typically an 'ability' of a rater to make a 'correct' classification), alongside the object attribute which together constitute a 'conjugate' pair of measurands. This procedure for clinical properties involving ordinal (and nominal) PMFs, has the same role as the corresponding procedure for data on fully quantitative scales, when calibrating the sensitivity of, for instance, a weighing machine in order to metrologically determine the mass of an unknown weight. Providing separate estimates of measurement system attributes from the analysis of PMFs on the less quantitative scales will need to deal with possible scale non-linearity, Figure 7. Again, a psychometric model will be useful when defining clinical performance metrics for the elementary counting case, Figure 4 and corresponding to the top-right entry in Table 1.

Amongst a family of log-odds transformation models (through restitution from the ordinal or nominal response scores, $Y = P_{success}$, of the measurement system), a particularly metrological approach is the psychometric [40] modern measurement (dichotomous) theory [49,50], as a special case of the log-odds, Equation (11):

$$y_{i,j} = P_{success,i,j} = \frac{e^{(\theta_i - \delta_j)}}{1 + e^{(\theta_i - \delta_j)}} \tag{17}$$

or in the form of the log-odds:

$$ln(\frac{P_{success,i,j}}{(1 - P_{success,i,j})} = \theta_i - \delta_j \tag{18}$$

not only compensates for counted fraction non-linearity (Section 4.1.2) but also allows separate estimates, by **logistic regression** of Equation 18 to the response data (often using a maximum likelihood criterium), of the quantitative, continuous variables δ_j: task difficulty of each of the set of objects, j, and θ_i: agent ability of each of the set ('cohort') of instruments, i as the principal clinical measurands, for instance in our elementary case. (Note that this logistic regression is made on the *whole matrix* of responses of i instruments and j objects similar to, for instance, a least-squares subdivision, as opposed to bilateral, pairwise comparisons more commonly found in metrology (section 4.1.1 in [2].)

The **principle of specific objectivity** is one major approach to providing separate estimates of these attributes as measurands [40]. Following the Principle of Specific Objectivity, MSA responses are not measures of the instrument's (classifier's or decision maker's) ability, nor the classification task (object) difficulty, but depend on both. Rasch [40] attempted to motivate his principle by drawing analogies to relations between several variables and the universal equations of Physics (such as Newton's 2nd law).

The MSA approach of engineering metrology is felt to be a closer point of departure [50] (Section 3.1). In accordance with Rasch's principle of specific objectivity, task difficulty and agent ability need to be treated separately:

- 'agent' ability, θ
- 'task' difficulty, δ

In making separate estimates of the measurands, ability and task difficulty, following the Principle of Specific Objectivity, when reliability and validity are to be assessed, the normal rules of statistics apply, as well as more metrological requirements which go beyond requiring mere numbers to making judicious and representative sampling across the full range of measurement scales (Section 5.5).

Alongside his better-known log-odds expression ((17), the Poisson PMF distribution Figure 2 of SPC was chosen by Rasch [51] when originally proposing his psychometric approach (Section 4.3.3) to determining the ability of individual cohort members to perform psychometric tests such as reading. Misreadings in such tests correspond with the typical application of the Poisson distribution (Section 4.2) in quality control as a model of the number of defects or non-conformities that occur in a 'unit of product' [33]. In Rasch's psychometric models [40,51], the Poisson rate factor $\lambda = k/h$, where 'task' difficulty, $\delta = log(k)$ and 'agent' ability, $\theta = log(h)$ [52,53] is recognized as "*a necessary condition for*

 doi:10.20944/preprints202604.2173.v1

specific objectivity of comparisons between tests and between students", with "*the result that for the Poisson distribution the multiplicative structure (IX:2) of its parameter A; is both necessary and sufficient for obtaining specific objectivity can be extended to the whole class of additive exponential models*...." ([40], pp. 92-92). How the Poisson distribution can be useful when defining clinical performance metrics for the elementary counting case, Figure 4, is further examined in (Section 4.3.3).

4.3.1. Agnostic Rasch Models

Most applications of Rasch's measurement models [40,51] have to date been made in education, health care, and survey research, involving persons who act as instruments [50], such as in the elementary counting example (Section 5). But that model is not mathematically limited to psychometrics or human respondents, but should also be applicable to more technical agents, as Rasch [54] recognized in his retirement lecture, which he closed by saying that the potential for use of the models he developed "stretches to all sciences where the subjects are comparisons that must be objective".

Diverse examples, such as:

- of a **material hardness indenter** [55] to make an impression (JCGM VIM [56] EXAMPLE 1 *Rockwell C hardness*)
- of a **hospital** to provide a service (A & E [57], surgery [58])
- in **psychophysics**, that is, where various physical fields and forces impinge on the five senses of an agent [50,59].

are amenable to an analysis in terms of the measurands (i) an 'ability' of an agent and the 'difficulty' of a task, or alternatively, (ii) in usability studies [60], the 'leniency' of an agent and the 'quality' of a product or service, by exploiting the agnostic aspect of the Rasch measurement model (Section 4.3).

4.3.2. Polytomous Measurement Models

Rasch's dichotomous measurement model version, Equation 17, has subsequently been complemented by the corresponding **polytomous** measurement model variants as well as multilevel [61,62], multidimensional [63], multifaceted [64,65], mixture [66], and other models. As an example, JCGM VIM [56] EXAMPLE 4 *Subjective level of abdominal pain on a scale from zero to five* belongs here.

Registering the response, $P_{success}$, of the measurement system when **instrument resolution is limited or levels of uncertainty are high** in the measurement process, is often made more practically in terms of a Probability Mass Function (PMF) with discrete categories rather than as a continuous response scale (so-called 'visual (analogue) score') by rounding off to the nearest integer, even when the object quantity varies on a continuous scale.

Polytomous classification is the term used when more than two categories are assigned. In the elementary counting case, Figure 4, this would correspond to 'guessing' the count to lie in a range of integer numbers of dots for a given cloud, as a clinical performance study 2.

Approaches to **polytomous** response analysis include the so-called 'partial credit model' (PCM, [67,68]) in which the binary (dichotomous) log-odds ratios (OR) models are extended to polytomous cases, as summarised as follows:

Probability, $q_{i,j,k}$ of response $x_{i,j}$ of instrument (agent) i; object (item) j, over k categories for PCM:

$$q_{i,j,k} = \frac{\exp \sum_{c=0}^{k} (\theta_i - \delta_{j,c})}{\exp \sum_{k=0}^{C_j} \cdot (\sum_{c=0}^{k} (\theta_i - \delta_{j,c}))} \tag{19}$$

An alternative to the PCM is the Rasch Rating Model [69]:

$$Pr(Y_{i,j} = q) = q_{i,j,k} = \frac{\exp \sum_{c=0}^{k} (\theta_i - \delta_{j,c} - \tau_c)}{\exp \sum_{k=0}^{C_j} \cdot (\sum_{c=0}^{k} (\theta_i - \delta_{j,c} - \tau_c))} \tag{20}$$

where τ_c is the threshold of category c.

 doi:10.20944/preprints202604.2173.v1

Even the present case of elementary counting where, for each task of counting a number of dots, an individual cohort member is making a polytomous decision: "how many dots do I see?" among the set of potential dot numbers (Section 4.3.3).

4.3.3. Poisson PMFs for Counting

At about the same time as Rasch [40] formulated his psychometric model, Equation (17), he also considered [51] an alternative formulation referring to a probabilistic Poisson distribution (Section 4.2) well known from SPC as a model of the number of defects or nonconformities that occur in a unit of product when classifying it:

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}; x = 0, 1, \ldots, n \tag{21}$$

where x is the 'quality characteristic' being classified (number of defects) in n observations.

The rate parameter λ is equal directly both to the mean and variance of the Poisson distribution ([33] p. 55 and (Section 4.2)). In [51] a Rasch Poisson PMF was given as a model for the number of 'misreadings' by a child (i) for a text (j) of a particular level of difficulty. As quoted in [53], and references therein: "*Although it is one of the earliest models that Rasch has developed, the Rasch Poisson Counts Model has received less attention than other binary or polytomous Rasch models*".

A much earlier observation by Meredith [48] noted: "*a thorough understanding and application of such a particularly simple [Poisson] model is often the key to further development of suitable stochastic models in scientific work*".

The fundamental value obtained via the Poisson counts model was only belatedly realized by Rasch ([70] p. 66) when, after a 1959 conversation with a colleague, he suddenly realized "that the possibility of separating two sets of parameters must be a fundamental property of a very important class of models," and that his study of children's misreadings had led him to devise a new "class of probability models [having] the property in common with the Multiplicative Poisson Model, that one set of parameters can be eliminated by means of conditional probabilities while attention is concentrated on the other set, and vice versa." Andrich [71] then showed that the Poisson distribution is necessary and sufficient for constructing measures from discrete observations.

In the present work, the example of the elementary counting of dots, akin to the Rasch [51] counting study, not only provides an 'analytical, sampling by variable' PMF (Section 3.3) but will also exemplify how Poisson PMFs can be benchmarked against full psychometric studies as 'clinical, sampling by attribute' PMF studies (Section 5).

In the case study of dot counting by the Munduruku (indigenous people of Brazil), [7,8], Poisson PMFs can be derived with Equation (21) since task difficulty can be calculated from first principles, Equation (26), from the number of dots being counted, thanks to the conceptual simplicity of the task [72], pp.66ff. Table 2 shows the PMF values from the Poisson rate, λ, calculated by entering the simulated counting difficulty δ into Equation (21), where $k_j = e^{\delta_j}$ and taking a cohort-average ability, θ_{mean}, to yield $h_i = e^{\theta_i}$.

Figure 2 shows the resulting plots of PMFs for three cases (true counts of dots: (1), (6) and (8)). These PMFs show the distribution of probabilities for each object count (no. of dots), with a shifting maximum — e.g. at $x = 5$ mis-classifications for 8 dots — corresponding to the most likely value of x, the quality characteristic being classified (number of mis-classifications) in n observations.

- In contrast to Rasch's [40] psychometric model, it is important to note that Poisson [73] placed important restrictions on the applicability of his model: particularly that the misclassification events were to be "very rare", as should apply to the PMFs shown in Figure 2. Although there appears in the literature to be no exact threshold where the Poisson approximation breaks down, a 'rule of thumb' quotes: $\lambda = n \cdot p$ is moderate (typically $n \cdot p \leq 5 - 10$ for good accuracy) ([33], Figure 2-24). Our psychometric simulations (Section 5) of task difficulty and counter ability do not suffer from the same restrictions and allow the choice of response level, $P_{success}$, over a complete range Figure 8: from the most mis-classifications (for a low-ability counter attempting a difficult

 doi:10.20944/preprints202604.2173.v1

counting task) to the least number of mis-classifications where the Poisson approximation should be valid in the latter case (i.e., for a high-ability counter performing an easy counting task). The Poisson rule-of-thumb range can be seen to be comparable with the mis-classification rates shown in Figure 2. However, to make a full comparison between the classic 'rule of thumb' limit to the Poisson distribution and the present psychometric simulations requires a proper account of the applicability of the ergodic principle: that is, to what extent the classic group-statistic Poisson approach [33,73] corresponds to the individual statistics of Rasch [51] psychometric modelling. For instance, the Poisson rate $\lambda = n \cdot p$ has in some way to be interpreted in psychometric cases where $n = 1$, i.e., just one counting agent (*MSA: Instrument*) while, at the same time, the overall number of degrees of freedom needs to be sufficiently large to achieve adequate reliability and validity (Section 5.5).

- The Poisson PMF number of mis-classifications shown in Figure 2 has no information about which numbers are perceived correctly or not, but should in principle correspond — for each item (j) — to the occupancies of the analytical PMFs shown in Figure 1. Similarly, the Poisson PMF number of mis-classifications has no obvious relation to the corresponding 'clinical' metrics, that is, the counting task difficulty and the counter ability. (In principle, one can estimate the expected count by inverting Equation (21) since — in the present case of an elementary construct — the relation between task difficulty and the number of dots is known, Equation (26)).

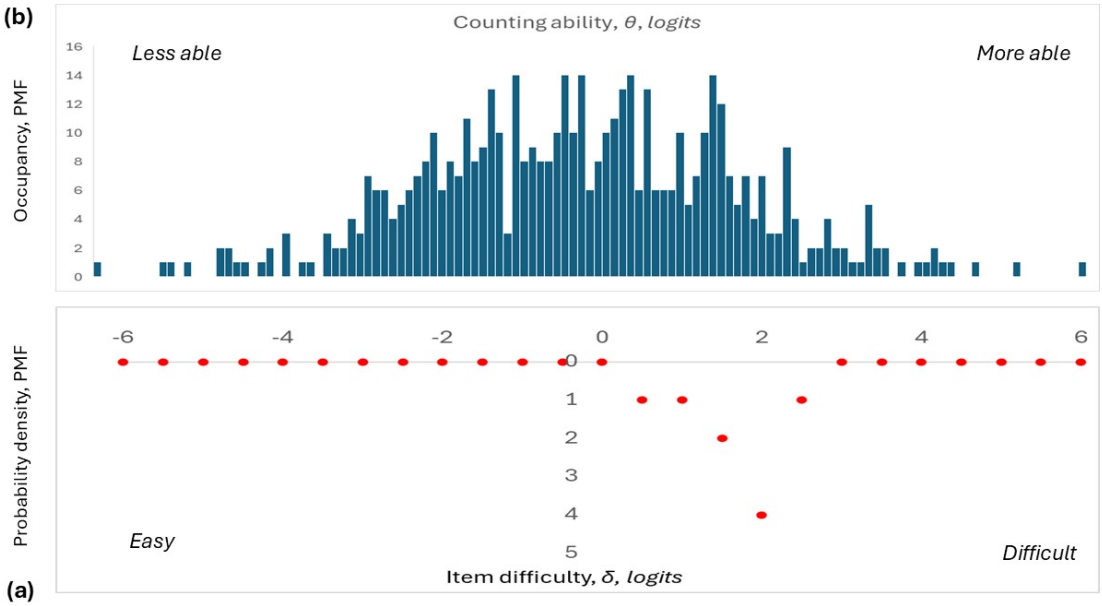

Figure 8. PMFs for simulated data from psychometric analysis (Rasch 17) for a set of clouds of dots (of increasing **(a)** (red PMF) task ("object") counting difficulty, δ_i based on the Hartley entropy theory Equation (26) and **(b)** (blue PMF) counter ability, θ_i, assuming a Normal distribution Section 5.4.2 [7,8]

Table 2. Empirical PMFs for three counting tasks: 1 dot, 6 dots and 8 dots

x	1	2	3	4	5	6	7	8	9	10
$P(X = 1)$	33.72	10.77	2.29	0.37	0.05	0.0	0.0	0.0	0.0	0.0
$P(X = 6)$	3.09	7.89	13.43	17.15	17.53	14.92	10.89	6.95	3.95	2.02
$P(X = 8)$	1.83	5.27	10.10	14.51	16.68	15.97	13.12	9.42	6.02	3.46

 doi:10.20944/preprints202604.2173.v1

5. 'Clinical' Psychometric Study: Counting Dots

The **elementary counting** of dots by less numerate agents (e.g., children or indigenous people [7,8]) also presents a case typical of the metrology of counting as a 'clinical' or 'sampling by attribute' procedure (2) corresponding to the top row of Table 1, as follows.

5.1. Dichotomous and Polytomous Mis-Classification Probabilities. CTT and Rasch Measurement Theory

Regularly used clinical performance metrics (2) of measurement systems include: 'selectivity', Equation (22) and 'specificity', Equation (23), which are plotted against each other on receiver operating characteristic curves, [30], when sampling by 'attribute', and are often analysed with CTT despite its known limitations (4.1.2).

Dichotomous mis-classification probabilities [74,75] due to binary decision quandary arising from uncertainties are:

- $FAP = P[Y = 1|X = 0]$; False Acceptance (or positive, FPR) Probability
- $FRP = P[Y = 0|X = 1]$; False Rejection (or negative, FNR) Probability

These two probabilities lie on the off-diagonal of a 2 x 2 **Decision matrix** for a binary test, where the true positive and true negative classification probabilities lie on the diagonal. Each pair of probabilities — positive (TP, FP) and negative (FN, TN) — can be plotted as binary PMFs like the binomial plots shown in Figure 5.

Performance metrics can be defined for the case of a binary decision A or B about an entity when assessed for conformity with respect to a specification limit, SL, for the quantity, z, of interest, based on a signal (SN with noise) in the presence of noise, N, [30]:

True positive rate (TPR) aka Selectivity:

$$P(A \mid SN) = \frac{\sum_{z>SL} N_{z,\text{detected}}}{\sum_z N_{z,\text{detected}}} \tag{22}$$

True negative rate (TNR) aka Specificity:

$$P(B \mid N) = \frac{\sum_{z<SL} N_{z,\text{not detected}}}{\sum_z N_{z,\text{detected}}} \tag{23}$$

The elementary counting case Section 5 corresponds to a polytomous extension (Section 4.3.2) of each of these classic clinical performance metrics.

Recall that "clinical performance aims to demonstrate that the measurement instrument can achieve clinically relevant outputs through ***predictable and reliable use*** by the intended users", 2, [26]. Considering the known (but not always recognised) limitations in CTT (4.1.2), a more reliable approach to clinical performance assessment [32,76–78] is therefore to deploy Rasch Measurement Theory (Section 4.3) which:

- (i) compensates for counted-fraction scale non-linearity (Section 4.1.3)
- (ii) exchanges the analytical measurands (such as the number of dots, 3.3) for the clinical measurands: counting task difficulty and counter ability.

In the 'clinical' scenario 2, the level of difficulty attributed to each object counted and the ability attributed to each counter (the person or, in MSA terms 3, the 'instrument'), both contribute in some way to the probability of successful counting. For a proper metrological analysis of the clinical analysis of counting, PMFs (occupancy: probability of responses in each category (number of dots) for a range of tasks of increasing levels of difficulty can be analysed psychometrically where the quantities (measurands) of interest are primarily the ability of each counter and the level of difficulty of a certain number of dots (rather than the actual number of dots – which of course can be determined by any capable counter).

 doi:10.20944/preprints202604.2173.v1

5.2. Counted-Fraction Scale Non-Linearity When Counting Dots. ICC

Figure 9 exemplifies for the elementary counting example (Section 5) so-called ICC (Item Characteristic Curves, [79]) for a particular item, calculated for a selected counting task (difficulty δ_j, $j = 4$, using PCM modelling of polytomous data distributed across a cohort of test agents (persons counting) calculated as follows:

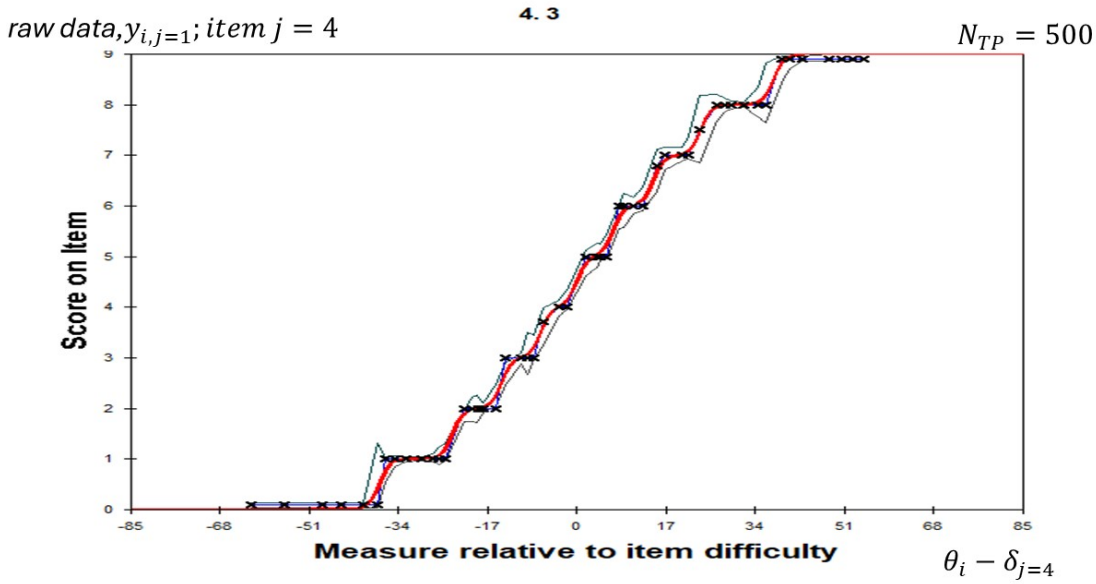

Figure 9. ICC from logistic regression of partial credit model (Equation 19, to polytomous data, 10 categories $0, ..., 9$) to raw scores y-axis; calculated with the Rasch formula for $P_{success,i,j}$, Equation 17) across a simulated cohort of varying (x-axis) counting ability, θ_i for a selected counting task (difficulty δ_j, $j = 4$ shown in Figure 8) of elementary counting tasks ('items', j) of a set of clouds of increasing number of dots $1, ..., 10$ [7] ($N_{TP} = 500$)

For an item with categories $c = 0, 1, 2, ..., k$ and Andrich thresholds τ_c Equation 20, with τ_0 chosen to be 0, then: At location x on the latent variable (relative to the item difficulty), the probability of observing category c is $P_c(x) = \frac{e^{c \cdot x - \sum_{c=0}^{k} \tau_c}}{sump(x)}$ where $sump(x) = \sum_{c=0}^{k} e^{c*x - sum(\tau_c)}$ for $c = 0$ to k.

The expected score y at location x is given by: $y = \sum_{c=0}^{k} (c \cdot P_c(x))$

Further publications referring to the simulation of psychometric ICC models can be found in, for example: [80,81]. In the present work, the simulated values of as described in (Section 4) simply use the usual dichotomous Rasch expression Equation 17 (which for benign data sets should give at least as good estimates as rounding to the nearest category, c, as described by [81]).

5.3. Communication of Measurement Information Throughout the Measurement Process. Amount of Entropy

To meet the challenge posed by PMF axes being often on less quantitative or even nominal scales where even the most basic arithmetic operations cannot be assumed to always work (Section 4.1.3) a viable alternative to calculating distances (even on ordinal and nominal scales) of PMF plots is to take differences in informational (Shannon) **entropy**.

The amount of measurement information on the categorical scales of signals in each classification category, c, at any one point and state in the measurement process (from object, through instrument and operator to restitution, Figure 3) can be expressed as a [82] 'surprisal' $-ln(q_c)$, while the relative contribution to the total entropy is weighted with the relative occupancy, q_c ([2], chapter 5).

The propagation of measurement information depicted in Figure 3, where information can be lost, distorted or gained, is described more readily with sums and differences of informational entropy than with the corresponding combination of probabilities and PMFs. An example of additions of entropy is found in the so-called *Mutual information*: $I(Z;Y) = H(Z) - H(Z|Y)$ which has been used

 doi:10.20944/preprints202604.2173.v1

by Benish [83] where an individual's disease state is denoted Z while the diagnostic test result is Y. The well-known Kullback-Leibler (K-L) metric $d_{KL}(P,Q)$ for a response Y, PMF $Q(y)$, to an object attribute Z, PMF $P(z|A)$, is:

$$d_{KL}(P,Q) = \int -z \cdot dP_{success} = -[P_{success} \cdot log(P_{success}) + (1 - P_{success}) \cdot log(q - P_{success})] = H(P,Q) - H(P). \quad (24)$$

The K-L parameter. Equation (24), is admittedly a valid metric only for infinitesimally small distances and is only one of several candidates for PMF distance metrics widely used, for example, in automatic image classification [84]. It can also be shown (section 5.5.3 in [2]) that log-odds transformed data correspond to entropy, Equation 11.

The start of the measurement process has a deficit in overall entropy (that is, the amount of 'useful' measurement information initially available) which for a discrete PMF is $\Delta H(P) = -\sum_k p_k \cdot ln(p_k)$, where p_k is the occupancy of category k from the actual quantity attributed to the measurement **object**. For conceptually simple tasks—-such as the counting of dots—this calculated entropy will provide a **metrological reference for calibration of classification properties**, as illustrated in the current case study (Section 5.4.1). From a measurement system perspective, the response of the **instrument** (often a score, $P_{success}$, made by a 'rater') will provide an input, X_1, when establishing the measurement result (3.1).

5.4. Simulated PMFs for the Elementary Counting Case

PMFs are shown in Figure 8 for simulated data for a set of clouds of dots (of increasing (a) (red histograms) task ('object') counting difficulty, based on entropy theory Equation (26) (Section 5.4.1) and (b) (blue histograms) counter ability (Section 5.4.2), as the conjugate pair of psychometric measurands [7,8]:

5.4.1. Counting Task Difficulty Simulated. Object (Task) Entropy

For the present guide on PMF simulations, 'seed' values will be considered in a study of elementary counting (of dots by Munduruku indigenous people in Brazil, [7]), a set of clouds of dots was presented visually for each counter as a sequence of counting tasks ('objects'). Task difficulty is in general proportional to entropy – a more ordered (less entropy) task is easier to perform [72,85].

For the current purpose of demonstrating best practice when simulating uncertainties associated with PMFs, of particular interest will be cases where attributes – for instance, the level of difficulty of tasks – are sufficiently simple conceptually so that simulations can be "seeded" with realistic start values. Beyond numerical simulation, it will be particularly valuable if *ab initio* estimates of categorical quantities such as task difficulty are available. Recent work has demonstrated how such knowledge-informed seed values can indeed be calculated even for categorical properties in a manner analogous to establishing metrological reference values in the more quantitative metrology in Chemistry and the Material sciences, as demonstrated in the neurodegeneration studies of legacy memory tests [86] where recall task difficulty could be expressed with so-called Construct Specification Equation (CSE, [87–90])

$$\delta = \Sigma_k \beta_k \cdot Z_k \quad (25)$$

in terms of **entropy-based explanatory variables** Z_k. The PMF plotted (red histograms) in Figure 8 against the measurand, counting task difficulty, δ_j, are those predicted to increase discretely, reflecting the increasing number, G, of integer dots in successive clouds according to the basic Hartley entropy [91]:

$$\delta_j = ln(G_j) \quad (26)$$

assuming a random distribution of identical dots in each cloud counted. (The addition of some symmetry to a cloud of dots would make the counting task easier, in terms of a lowering in the overall task entropy, as has been studied for elementary memory recall tests [72], p. 68.)

 doi:10.20944/preprints202604.2173.v1

5.4.2. Counting Classifier Ability Simulated

- Excel:

$$NORM.INV(RAND();0;2) \tag{27}$$

Returns a vector of random numbers having the Normal distribution of the measurand counter abilities across the cohort shown with the blue histograms in Figure 8.

5.5. Reliability and Validity

In making separate estimates of the conjugate attributes, δ_j and θ_i — which are not merely the measurands, but also for 'clinical' purposes the *quality characteristics*, when **conformity assessment** is made — the normal rules of statistics of course apply – including making sufficient numbers of observations of items and instruments – to ensure minimum levels of reliability and power as these are relevant in fit-for-purpose designs distinguishing between the varying demands of screening, diagnostic, accountability, and research applications. Additional, more metrological requirements go beyond requiring mere numbers to making judicious and representative sampling across the full range of measurement scales. For metrological purposes — concerning both precision and trueness (for metrological invariance, comparability and interoperability) — the underlying scale needs to be unidimensional, linear, quantitative, well targeted, free of differential item functioning, etc., and there is a whole battery of conventional tests of model validity [85].

The standard uncertainty in the 'height' (i.e., occupancy, $P_{success}$) of a PMF column is:

$$u(P_{success}) = \sqrt{\frac{P_{success} \cdot (1 - P_{success})}{n}}, \tag{28}$$

derived from the underlying binomial distribution Equation 7. One should however always respect the inherent non-linearity of the scale, such as due to the counted-fractions effect (Section 4.1.2). A more proper estimation of this uncertainty is to first calculate the uncertainties in the link function, as given by the GLMM log-odds approach, Equation 11 .

Uncertainties (so-called Standard Errors, SE, i.e., standard uncertainties) in Rasch analyses with the logistic regression, Equation 18 are those quoted by the WINSTEPS® program (section 21.120), where according to that program's manual (p. 805):

$$SE(\theta_i, \tilde{\delta}) = \sqrt{\frac{1}{\Sigma_{j=1}^{N_{items}} P_{success,i,j} \cdot (1 - P_{success,i,j})}} \tag{29}$$

and

$$SE(\tilde{\theta}, \delta_j) = \sqrt{\frac{1}{\Sigma_{i=1}^{N_{TP}} P_{success,i,j} (1 - P_{success,i,j})}} \tag{30}$$

Thereafter, one can apply the inverse of the Rasch transformation in the case the enduser wants to relate to the original PMFs. Uncertainty intervals in $P_{success}$) calculated in that way will become increasingly asymmetric, the closer the score is to either end of the scale.)

Reliability means assessing the influence of measurement uncertainty as an estimate of limited measurement quality. In psychometrics [45], a reliability coefficient, R_β, for a Rasch variable $\beta = \beta' + \epsilon_\beta$, (for either Rasch attribute: $\beta = \theta$ or δ) including an error term, ϵ_β, is defined as:

$$R_\beta = \frac{\text{var}(\beta')}{\text{var}(\beta)} = \frac{\text{var}(\beta) - \text{var}(\epsilon_\beta)}{\text{var}(\beta)} \tag{31}$$

where the 'true' variance is $var(\beta')$ and the observed variance is $var(\beta)$.

The consequences of the decisions to be made determine what the actual limits of reliability are, for instance, by setting a maximum permissible uncertainty. Traditional psychometric limits for

 doi:10.20944/preprints202604.2173.v1

so-called 'high-stake' decisions are typically $R_\beta > 0.8$ (Equation (31)) which corresponds to at most half of the observed dispersion being assigned to measurement uncertainty.

5.6. Validation of Simulated LOGISTIC regressions

The results of logistic regressions of partial credit model (Equation 19) to raw scores (y-axes of Figure 9; calculated with the dichotomous Rasch formula for δ_j, Equation 17) across a simulated cohort of varying (x-axes) counting ability, θ_i Figure 8(b) for a pair of selected counting tasks (difficulties δ_j shown in Figure 8(a)) of elementary counting tasks. Using the WINSTEPS ® program for this logistic regression, the raw scores were coded into categories by integer rounding of the % scores, as is particularly evident in the steps shown in Figure 9.

The validity of this procedure is investigated in Figures 10 and 12 by comparing the restituted estimates (y-axes) of δ and θ with the original simulated values (x-axes) plotted, respectively.

A second validation was made by comparing (Figure 11) simulated (y-axis) and original experimental data (x-axis) counting task difficulties from psychometric analysis (Rasch Equation (17)).

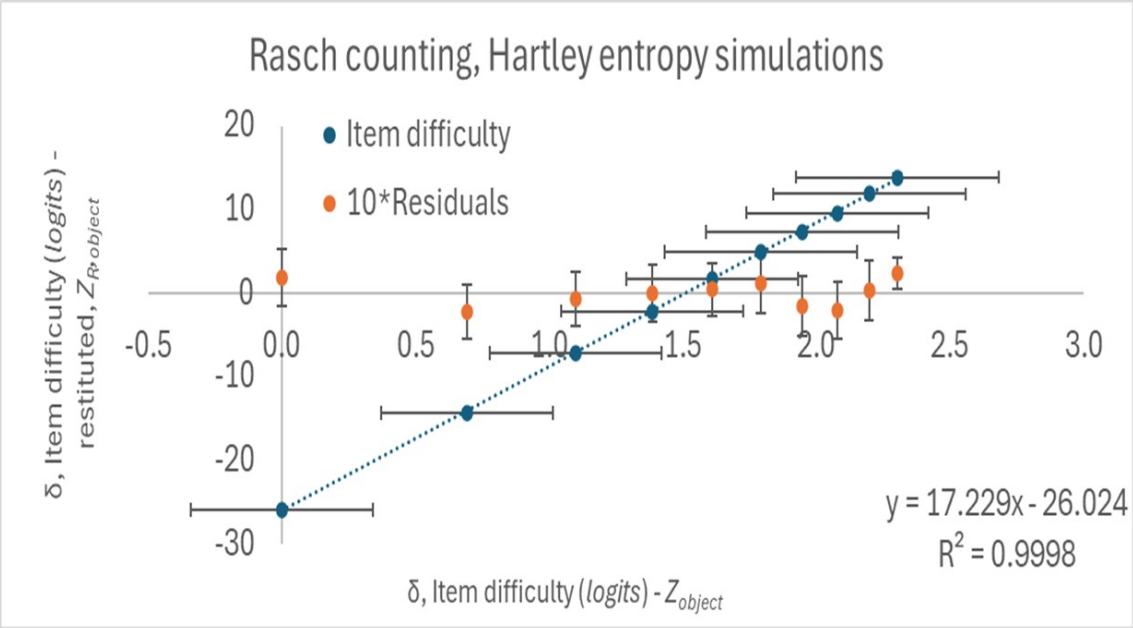

Figure 10. Comparison of simulated (y-axis) and calculated (x-axis) counting task difficulties from psychometric analysis (Rasch 17) of a cohort **(a)** (blue) Task difficulty and **(b)** (red) 10 x residuals of differences between simulated and calculated difficulties. Uncertainty coverage factor, $k = 2$

 doi:10.20944/preprints202604.2173.v1

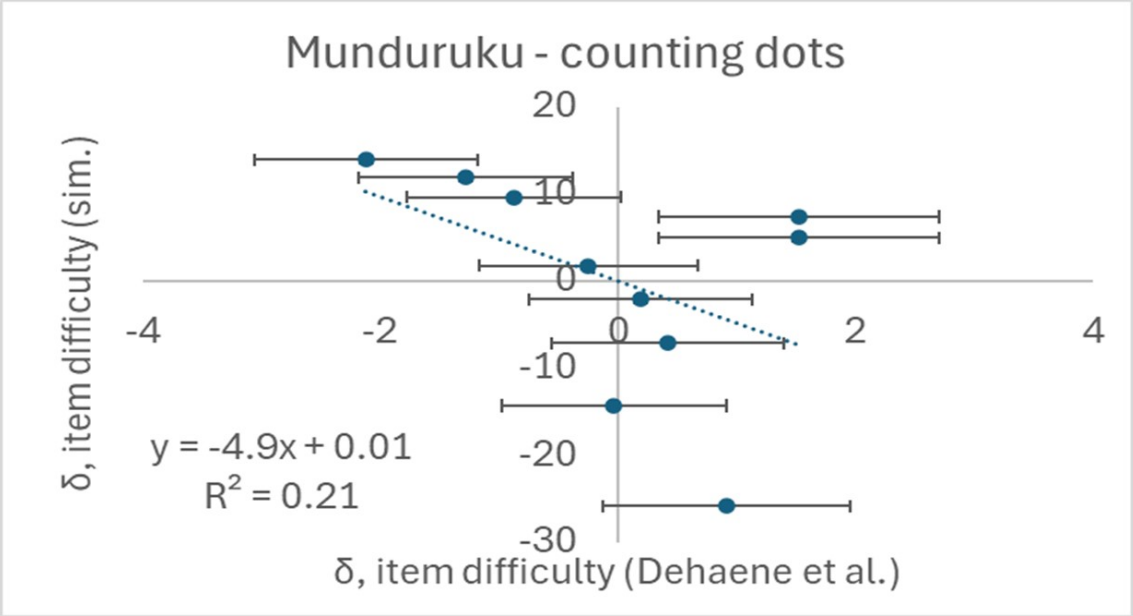

Figure 11. Comparison of simulated (y-axis) and original experimental data (x-axis) counting task difficulties from psychometric analysis (Rasch Equation 17). Uncertainty coverage factor, $k = 2$

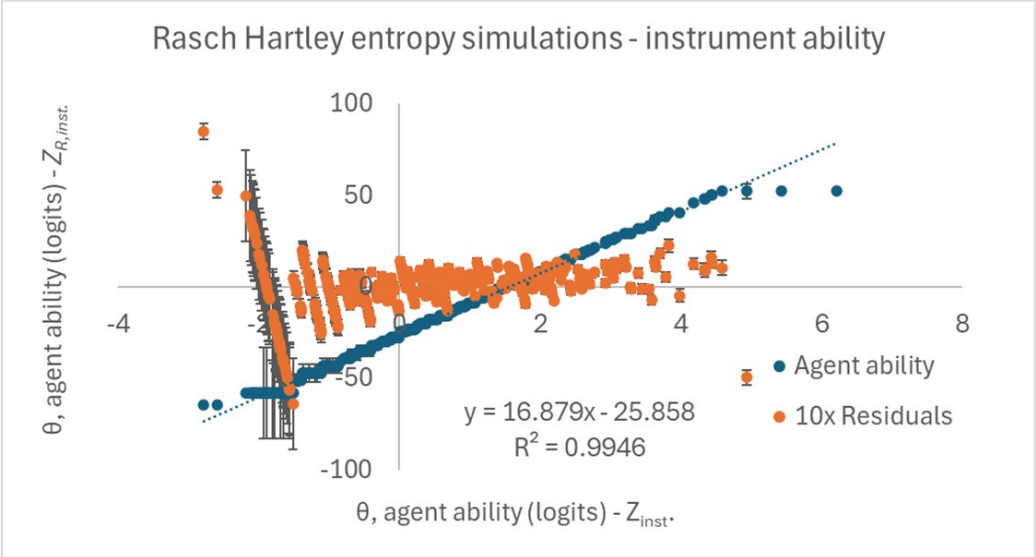

Figure 12. Comparison of simulated (y-axis) and calculated (x-axis) counting agent abilities from psychometric analysis (Rasch 17) of a cohort **(a)** (blue) Agent ability and **(b)** (red) 10 x residuals of differences between simulated and calculated abilities. Uncertainty coverage factor, $k = 2$

6. Discussion and Conclusions

The choice of construct — quantities attributed to an elementary counting process — in the current case study is made deliberately. The conceptual simplicity ensures the unidimensionality required in metrological psychometrics [40,50,81]. Additionally, it offers the opportunity to define — from first principles — metrological references ('gold standards') [72] for counting task difficulty in terms of Construct Specification Equations [87–90,92] with explanatory variables based on informational entropy [93,94], as demonstrated in the simulation of counting task difficulty (Section 5.4.1). Such

 doi:10.20944/preprints202604.2173.v1

metrological references for calibration of categorical properties enable in turn interoperability and defining limits on quality of products and services of all kinds. (Other explanatory approaches beyond CSE include [95,96].)

This conceptual simplicity facilitates contrasting the metrology of three kinds of probability mass functions (PMFs) (Table 1): (i) quantitative response to a discrete range of counts, Figure 1, (ii) the classic Poisson distribution of miscounts, Figure 6, and (iii) psychometric (Rasch) distributions of counting task difficulty and person counting ability, Figure 8. New insights are provided when benchmarking the Rasch Poisson Counts Model (Section 4.3.3) which has received less attention, against full psychometric Rasch modelling (Section 5). Limits (traditionally, rules-of-thumb on small probabilities and large number of observations) on the validity of the Poisson distribution can be tested with access to psychometric simulations across the full range of task difficulty and counter ability.

A human being acting as a counter (*MSA: A human as a Measurement Instrument* [50]) is a prototype for more advanced applications, including counting, decision-making and similar tasks performed by Artificial Intelligence agents [6,72,97,98]. Logistic regression (Section 4.3) provides an essential complement when handling discrete and sometimes qualitative observations, for example, to a typical use of Machine Learning in metrology where traditionally regular regression tasks are made on a continuous quantity [99]. The proposed methods will be useful when the traditional role of Physics in metrology is to be complemented by a typical 'turn-of-the-millennium' focus on the new metrology of 'clinical' and other applications where meaning and value, and not just numbers and analytics, are sought.

Future work is needed to extend the present studies to include less elementary and more complex constructs. There is much current research addressing multidimensional measurement theory and related topics in this field [32,100–103].

Acknowledgments: The authors thank members of the Joint Committee for Guides in Metrology (JCGM), Working Group 1 Guide to the Expression of Uncertainty in Measurement for useful discussions.

Funding: RISE, through its internal programme supporting a Competence platform for categorical measurements, has provided funding for this work. Earlier research from European projects has also contributed, with support from the EMPIR programme, co-financed by the Participating States and from the European Union's Horizon 2020 research and innovation programme, under grant numbers 15HLT04 NeuroMET and 18HLT09 NeuroMET2.

Conflicts of Interest: The authors declare no conflicts of interest. No financial or non-financial interests have arisen from this work. The authors declare no conflict of interest.

Data Availability Statement: All original measurement data come from [7]. Simulated data are produced in this work as described.

Author Contributions: Conceptualization, L.R.P and W.P.F.; methodology, L.R.P and W.P.F.; validation, L.R.P and W.P.F.; writing—original draft preparation, L.R.P; writing—review and editing, L.R.P and W.P.F.; Both authors have read and agreed to the published version of the manuscript.

 doi:10.20944/preprints202604.2173.v1

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
CTT	Classical Test Theory
EMPIR	European Metrology Programme for Innovation and Research
EPM	European Programme for Metrology
FAP	False acceptance probability
FRP	False rejection probability
GLMM	Generalised Linear Measurement Model
GUM	Guide to the expression of uncertainty of measurement
ICC	Item characteristic curve
IRT	Item Response Theory
JCGM	Joint committee for guides in metrology
K-L	Kullback-Leibler
MSA	Measurement System Analysis
PMF	Probability Mass Function
PDF	Probability Density Function
RISE	Research institutes of Sweden
RMT	Rasch Measurement Theory
SE	Standard Error
VIM	International Metrology Vocabulary

References

1. Bureau International des Poids et Mesures. *The International System of Units (SI)*, 9th ed.; BIPM: Sèvres, France, 2019. SI Brochure.
2. Pendrill, L.R. Chapter 2. In *Quality Assured Measurement – Unification across Social and Physical Sciences*; Springer, 2019. https://doi.org/10.1007/978-3-030-28695-8.
3. Fisher Jr., W.P.; Pendrill, L., Eds. *Models, Measurement, and Metrology Extending the SI: Trust and Quality Assured Knowledge Infrastructures*; De Gruyter Series in Measurement Sciences, De Gruyter Oldenbourg: Berlin/Boston, 2024. https://doi.org/10.1515/9783111036496.
4. Fisher Jr., W.P.; Massengill, P.J., Eds. *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement: Selected Papers by A. Jackson Stenner*; Springer: Singapore, 2023. https://doi.org/10.1007/978-981-19-3747-7.
5. Fisher Jr., W.P. Measure and manage: Intangible assets metric standards for sustainability. In *Business Administration Education: Changes in Management and Leadership Strategies*; Marques, J.; Dhiman, S.; Holt, S., Eds.; Palgrave Macmillan: New York, NY, 2012; pp. 43–63. https://doi.org/10.1057/9781137087102_3.
6. Barney, M.; Barney, F. Transdisciplinary Measurement through AI: Hybrid Metrology and Psychometrics Powered by Large Language Models. In *Models, Measurement, and Metrology Extending the SI: Trust and Quality Assured Knowledge Infrastructures*; Fisher Jr., W.P.; Pendrill, L.R., Eds.; De Gruyter Series in Measurement Sciences, De Gruyter Oldenbourg: Berlin/Boston, 2024; chapter 3, pp. 103–132. https://doi.org/10.1515/9783111036496-003.
7. Dehaene, S.; Izard, V.; Spelke, E.; Pica, P. Log or Linear? Distinct Intuitions of the Number Scale in Western and Amazonian Indigene Cultures. *Science* **2008**, *320*, 1217–1220.
8. Pendrill, L.R.; Fisher Jr., W.P. Counting and Quantification: Comparing Psychometric and Metrological Perspectives on Visual Perceptions of Number. *Measurement* **2015**, *71*, 46–55. https://doi.org/10.1016/j.measurement.2015.04.010.
9. Fisher Jr., W.P. Bateson and Wright on number and quantity: How to not separate thinking from its relational context. *Symmetry* **2021**, *13*. https://doi.org/10.3390/sym13081415.
10. Mallinson, T. Extending the justice-oriented, anti-racist framework for validity testing to the application of measurement theory in re(developing) rehabilitation assessments. In *Models, Measurement, and Metrology Extending the SI*; Fisher Jr., W.P.; Pendrill, L., Eds.; De Gruyter, 2024; pp. 401–428.

 doi:10.20944/preprints202604.2173.v1

11. Sul, D.; Dominguez, D.G. Culturally responsive evaluation with Latinx communities through culturally specific assessment: Building the Latinx immigration trauma assessment. *New Directions for Evaluation* **2024**, pp. 103–112. https://doi.org/10.1002/ev.20593.
12. Sul, D. Situating culturally specific assessment development within the disjuncture-response dialectic. In *Models, Measurement, and Metrology Extending the SI*; Fisher Jr., W.P.; Pendrill, L., Eds.; De Gruyter, 2024; pp. 475–500.
13. Sul, D.; Blackmon, A.T. Enacting culturally specific assessment by constructing a STEM leadership assessment framework. *Journal of Educational Measurement* **2026**. In press.
14. JCGM. Evaluation of measurement data – The role of measurement uncertainty in conformity assessment. Technical Report JCGM 106:2012, Joint Committee for Guides in Metrology, Sèvres, France, 2012.
15. Mitcham, C., Ed. *Encyclopedia of Science, Technology, and Ethics*; Macmillan Reference: New York, 2005.
16. Andersen, E.B. Sufficient statistics and latent trait models. *Psychometrika* **1977**, *42*, 69–81. https://doi.org/10.1007/BF02293746.
17. Andrich, D. Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika* **2010**, *75*, 292–308. https://doi.org/10.1007/s11336-010-9154-8.
18. Fischer, G.H. On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika* **1981**, *46*, 59–77. https://doi.org/10.1007/BF02293919.
19. Andrich, D. Distinctions between assumptions and requirements in measurement in the social sciences. In *Mathematical and Theoretical Systems: Proceedings of the 24th International Congress of Psychology of the International Union of Psychological Science, Vol. 4*; Keats, J.A.; Taft, R.; Heath, R.A.; Lovibond, S.H., Eds.; Elsevier Science Publishers, 1989; pp. 7–16.
20. Bhatt, U.; Antorán, J.; Zhang, Y.; Liao, Q.V.; Sattigeri, P.; Fogliato, R.; Melançon, G.G.; Krishnan, R.; Stanley, J.; Tickoo, O.; et al. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21). ACM, 2021, pp. 401–413. https://doi.org/10.1145/3461702.3462571.
21. ISO. ISO 5725-1:2023(en) Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions. International Standard, 2023.
22. Brown, R.J.C.; Güttler, B.; Neyezhmakov, P.; Stock, M.; Wielgosz, R.I.; Kück, S.; Vasilatou, K. Report of the CCU/CCQM Workshop on 'The Metrology of Quantities Which Can Be Counted'. *Metrology* **2023**, *3*, 309–324. https://doi.org/10.3390/metrology3030019.
23. Rossi, G.B. *Measurement and Probability: A Probabilistic Theory of Measurement with Applications*; Springer: Dordrecht, The Netherlands, 2014. https://doi.org/10.1007/978-94-017-8825-0.
24. JCGM GUM. *Joint Committee for Guides in Metrology – Part 6: Developing and Using Measurement Models*; Number JCGM GUM-6:2020, JCGM: Sèvres, France, 2020.
25. Hofmann, B. "My Biomarkers Are Fine, Thank You": On the Biomarkerization of Modern Medicine. *Journal of General Internal Medicine* **2025**, *40*, 453–457. https://doi.org/10.1007/s11606-024-09019-8.
26. European Commission. Factsheet for Manufacturers of In Vitro Diagnostic Medical Devices, 2020. Accessed 2020.
27. European Union. Regulation (EU) 2017/746 on In Vitro Diagnostic Medical Devices, 2017.
28. ISO. ISO 3951-1:2022 Sampling procedures for inspection by variables. International Standard, 2022.
29. ISO. ISO 2859-1:2026 Sampling schemes indexed by acceptance quality limit (AQL) for lot-by-lot inspection. International Standard, 2026.
30. Birdsall, T.G. *The Theory of Signal Detectability: ROC Curves and Their Character*; University of Michigan Library: Ann Arbor, MI, USA, 1973.
31. Linacre, J.M. Bernoulli Trials, Fisher Information, Shannon Information and Rasch. *Rasch Measurement Transactions* **2006**, *20*, 1062–1063.
32. Pendrill, L.R.; Melin, J.; Stavelin, A.; Nordin, G. Modernising Receiver Operating Characteristic (ROC) Curves. *Algorithms* **2023**, *16*, 253. https://doi.org/10.3390/a16050253.
33. Montgomery, D.C. *Introduction to Statistical Quality Control*, 3 ed.; Wiley: New York, 1996.
34. NIST. Uncertainty machine https://uncertainty.nist.gov/.
35. Bernoulli, J. *Ars Conjectandi*; Thurneysen Brothers: Basel, 1713. Opus posthumum; OCLC 7073795.
36. Tukey, J.A. Data Analysis and Behavioural Science. In *The Collected Works of John A. Tukey, Volume III*; Jones, L.V., Ed.; Chapman and Hall, 1984.
37. Pearson, K. Mathematical contributions to the theory of evolution: on a form of spurious correlation which may arise when indices are used in the measurements of organs. *Proc. Roy. Soc.* **1897**, *60*, 489–98.

 doi:10.20944/preprints202604.2173.v1

38. Filzmoser, P.; Hron, K.; Reimann, C. Principal Component Analysis for Compositional Data with Outliers. *Environmetrics* **2009**, *20*, 621–632. https://doi.org/10.1002/env.966.
39. Wright, B.D. Thinking with Raw Scores. *Rasch Measurement Transactions* **1993**, *7*, 299–300.
40. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Danmarks Paedagogiske Institut, 1960.
41. Andrich, D. Rating Scales and Rasch Measurement. *Expert Review of Pharmacoeconomics & Outcomes Research* **2011**, *11*, 571–585. https://doi.org/10.1586/erp.11.59.
42. Aitchison, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society* **1982**, *44*, 139–177.
43. Bland, J.M.; Altman, D.G. The Odds Ratio. *British Medical Journal* **2000**, *320*, 1468. https://doi.org/10.1136/bmj.320.7247.1468.
44. McCullagh, P. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society* **1980**, *42*, 109–142.
45. Wright, B.D.; Stone, M.H. *Best Test Design: Rasch Measurement*; MESA Press: Chicago, 1979.
46. Student, S.R.; Briggs, D.C.; Davis, L. Growth Across Grades and Common Item Grade Alignment in Vertical Scaling Using the Rasch Model. *Educational Measurement: Issues and Practice* **2025**, *44*, 84–95. https://doi.org/10.1111/emip.12639.
47. Montgomery, D.C.; Runger, G.C. *Applied Statistics and Probability for Engineers*, 5 ed.; John Wiley & Sons: Hoboken, NJ, 2011.
48. Meredith, W.M. The Poisson Distribution and Poisson Process in Psychometric Theory. ETS Research Bulletin Series RB-68-42, Educational Testing Service, Princeton, NJ, 1968.
49. Fisher Jr., W.P. Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement* **2009**, *42*, 1278–1287. https://doi.org/10.1016/j.measurement.2009.03.014.
50. Pendrill, L.R. Man as a Measurement Instrument. *NCSLI Measure* **2014**, *9*, 24–35. https://doi.org/10.1080/19315775.2014.11721702.
51. Rasch, G. On General Laws and the Meaning of Measurement in Psychology. In Proceedings of the Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Medicine; Neyman, J., Ed., Berkeley, CA, 1961; pp. 321–333.
52. Bashkansky, E.; Turetsky, V. Ability Evaluation by Binary Tests: Problems, Challenges and Recent Advances. *Journal of Physics: Conference Series* **2016**, *772*, 012012.
53. Liu, R.; Liu, H.; Shi, D.; Jiang, Z. Poisson Diagnostic Classification Models: A Framework and an Exploratory Example. *Educational and Psychological Measurement* **2022**, *82*, 506–516, [https://doi.org/10.1177/00131644211017961]. https://doi.org/10.1177/00131644211017961.
54. Rasch, G. Retirement Lecture of 9 March 1972: Objectivity in Social Sciences: A Method Problem. *Rasch Measurement Transactions* **2010**, *24*, 1252–1272. Originally presented in 1972; translated by Cecilie Kreiner.
55. Stone, M.; Stenner, J. From Ordinality to Quantity. *Rasch Measurement Transactions* **2014**, *27*, 1435–1437.
56. Joint Committee for Guides in Metrology (JCGM). International Vocabulary of Metrology — Basic and General Concepts and Associated Terms (VIM), 3rd edition, 2008. JCGM 200:2008 with minor corrections (2012).
57. Pendrill, L.R.; et al. Reducing Search Times and Entropy in Hospital Emergency Departments with Real-Time Location Systems. *IISE Transactions on Healthcare Systems Engineering* **2021**. https://doi.org/10.1080/24725579.2021.1881660.
58. Pendrill, L.R. Category-based interlaboratory comparisons: Psychometric Rasch analyses defining reference values and statistical weighting in a clinical example. *Educational Methods & Psychometrics* **2026**, *4*, 26. SAMC 2024 Special Issue, https://doi.org/10.61882/emp.2026.5.
59. Massof, R.W.; Fisher Jr., W.P. Psychophysics and the Measurement of Sensory Magnitudes. *Measurement* **2026**. Manuscript in review.
60. Rice, S.; Pendrill, L.R.; Petersson, N.; Nordlinder, J.; Farbrot, A. Rationale and Design of a Novel Method to Assess the Usability of Body-Worn Absorbent Incontinence Care Products by Caregivers. *Journal of Wound, Ostomy and Continence Nursing* **2018**, *45*, 456–464. Open access, https://doi.org/10.1097/WON.0000000000000462.
61. Adams, R.J.; Wilson, M.; Wu, M. Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics* **1997**, *22*, 47–76. https://doi.org/10.3102/10769986022001047.
62. Beretvas, S.N.; Kamata, A. Part II. Multi-level Measurement Rasch Models. In *Rasch Measurement: Advanced and Specialized Applications*; Smith, Everett V., J.; Smith, R.M., Eds.; JAM Press, 2007; pp. 291–470.

 doi:10.20944/preprints202604.2173.v1

63. Briggs, D.C.; Wilson, M. An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement* **2003**, *4*, 87–100.
64. Linacre, J.M. *A User's Guide to FACETS Rasch-Model Computer Program, Version 4.4.5*. Winsteps.com, 2026. Accessed 2026-04-09.
65. Linacre, J.M.; Engelhard, G.; Tatum, D.S.; Myford, C.M. Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research* **1994**, *21*, 569–577. Special Issue: Applications of Probabilistic Conjoint Measurement, https://doi.org/10.1016/0883-0355(94)90011-6.
66. von Davier, M.; Carstensen, C.H. *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications*; Springer, 2007. https://doi.org/10.1007/978-0-387-49839-3.
67. Masters, G.N. A Rasch model for partial credit scoring. *Psychometrika* **1982**, *47*, 149–174. https://doi.org/10.1007/BF02296272.
68. Masters, G.N.; Wright, B.D. The Partial Credit Model. In *Handbook of Modern Item Response Theory*; van der Linden, W.J.; Hambleton, R.K., Eds.; Springer: New York, 1996; pp. 101–121.
69. Andrich, D. A Rating Formulation for Ordered Response Categories. *Psychometrika* **1978**, *43*, 561–573. https://doi.org/10.1007/BF02293814.
70. Rasch, G. On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy* **1977**, *14*, 58–94.
71. Andrich, D. Models for measurement: Precision and the non-dichotomization of graded responses. *Psychometrika* **1995**, *60*, 7–26. https://doi.org/10.1007/BF02294426.
72. Pendrill, L.R. Quantities and units: order amongst complexity. In *Models, Measurement, and Metrology: Extending the SI – Trust and Quality Assured Knowledge Infrastructures*; Fisher Jr., W.P.; Pendrill, L.R., Eds.; De Gruyter, 2024; chapter 2. https://doi.org/10.1515/9783111036496-002.
73. Poisson, S.D. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*; Bachelier: Paris, 1837. Original work in French; foundational in probability theory.
74. Akkerhuis, T. Measurement system analysis for binary tests. PhD thesis, University of Groningen, 2016.
75. Akkerhuis, T.; de Mast, J.; Erdmann, T. The statistical evaluation of binary test without gold standard: Robustness of latent variable approaches. *Measurement* **2017**, *95*, 473–479. https://doi.org/10.1016/j.measurement.2016.10.043.
76. Linacre, J.M. Evaluating a Screening Test. *Rasch Measurement Transactions* **1994**, *7*, 317–318.
77. Cipriani, D.; Fox, C.; Khuder, S.; Boudreau, N. Comparing Rasch analyses probability estimates to sensitivity, specificity and likelihood ratios when examining the utility of medical diagnostic tests. *Journal of Applied Measurement* **2005**, *6*, 180–201.
78. Fisher Jr., W.P.; Burton, E. Embedding measurement within existing computerized data systems: Scaling clinical laboratory and medical records heart failure data to predict ICU admission. *Journal of Applied Measurement* **2010**, *11*, 271–287.
79. Linacre, J.M. Expected Score ICC, IRF (Rasch-half-point thresholds), n.d. Accessed 2026-03-27.
80. Baker, F.B.; Kim, S.H. *Item Response Theory: Parameter Estimation Techniques*, 2 ed.; CRC Press: Boca Raton, 2004.
81. Linacre, J.M. How to Simulate Rasch Data. *Rasch Measurement Transactions* **2007**, *21*, 1125.
82. Weaver, W.; Shannon, C.E. *The Mathematical Theory of Communication*; University of Illinois Press: Champaign, 1963.
83. Benish, W.A. A Review of the Application of Information Theory to Clinical Diagnostic Testing. *Entropy* **2020**, *22*, 97. https://doi.org/10.3390/e22010097.
84. Pele, O.; Werman, M. The Quadratic-Chi Histogram Distance Family. In Proceedings of the European Conference on Computer Vision (ECCV), 2010, pp. 749–762.
85. Melin, J.; et al. NeuroMET Memory Metric: Traceability and Comparability through Crosswalks. *Scientific Reports* **2023**, *13*, 5179. https://doi.org/10.1038/s41598-023-32208-0.
86. Melin, J.; Cano, S.J.; Flöel, A.; Göschel, L.; Pendrill, L.R. The Role of Entropy in Construct Specification Equations (CSE) to Improve the Validity of Memory Tests. *Entropy* **2022**, *24*, 934. https://doi.org/10.3390/e24070934.
87. Stenner, A.J.; Smith, M. Testing Construct Theories. *Perceptual and Motor Skills* **1982**, *55*, 415–426. https://doi.org/10.2466/pms.1982.55.2.415.
88. Stenner, A.J.; Smith, M.I.; Burdick, D.S. Toward a Theory of Construct Definition. *Journal of Educational Measurement* **1983**, *20*, 305–316. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x.

 doi:10.20944/preprints202604.2173.v1

89. Fisher Jr., W.P.; Stenner, A.J. Theory-based metrological traceability in education: A reading measurement network. *Measurement* **2016**, *92*, 489–496.
90. Fisher Jr., W.P.; Stenner, A.J. Theory-based metrological traceability in education: A reading measurement network. In *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement: Selected Papers by A. Jackson Stenner*; Fisher Jr., W.P.; Massengill, P.J., Eds.; Springer, 2023; pp. 275–293. Reprint of Fisher and Stenner (2016), https://doi.org/10.1007/978-981-19-3747-7.
91. Klir, G.J.; Folger, T.A. *Fuzzy Sets, Uncertainty, and Information*; Prentice Hall: New Jersey, 1988.
92. Melin, J. Pendrill, L.R.; et al. Construct Specification Equations: 'Recipes' for Certified Reference Materials in Cognitive Measurement. *Measurement: Sensors* **2021**, *18*, 100290. https://doi.org/10.1016/j.measen.2021.100290.
93. Shannon, C.E. A Mathematical Theory of Communication. *Bell System Technical Journal* **1948**, *27*, 379–423.
94. Brillouin, L. *Science and Information Theory*, 2 ed.; Academic Press, 1962. https://doi.org/10.1063/1.3057866.
95. De Boeck, P.; Wilson, M., Eds. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*; Statistics for Social and Behavioral Sciences, Springer-Verlag, 2004.
96. Embretson, S.E., Ed. *Measuring Psychological Constructs: Advances in Model-Based Approaches*; American Psychological Association, 2010.
97. Ru, D.; et al. RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation, 2024, [arXiv:cs.CL/2408.08067]. arXiv preprint, version 2, 17 Aug 2024.
98. Pendrill, L.R. Using Measurement Uncertainty in Decision-Making & Conformity Assessment. *Metrologia* **2014**, *51*, S206. https://doi.org/10.1088/0026-1394/51/5/S206.
99. Thompson, F.N.; et al. Trustworthy Artificial Intelligence **2024**. [arXiv:cs.AI/2406.10117].
100. Letizia, F.N.; et al. Copula Density Neural Estimation. *IEEE Transactions on Neural Networks and Learning Systems* **2025**.
101. Decruyenaere, A.; et al. Debiasing Synthetic Data Generated by Deep Generative Models, 2024, [arXiv:stat.ML/2411.04216]. arXiv preprint, version 1, 6 Nov 2024.
102. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. In Proceedings of the Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS); Gordon, G.; Dunson, D.; Dudík, M., Eds., Fort Lauderdale, FL, USA, 2011; Vol. 15, *Proceedings of Machine Learning Research*, pp. 315–323.
103. Sklar, M. Fonctions de répartition à N dimensions et leurs marges. *Annales de l'ISUP* **1959**, *8*, 229–231. HAL Id: hal-04094463.