Educational Methods & Psychometrics—Vol. 4 (SAMC 2024 Special Issue), Article No. 26
February 2026
https://doi.org/10.65301/emp.2026.260

CATEGORY-BASED INTERLABORATORY COMPARISONS: PSYCHOMETRIC RASCH ANALYSES DEFINING REFERENCE VALUES AND STATISTICAL WEIGHTING IN A CLINICAL EXAMPLE

Leslie Pendrill

RISE, Research Institutes of Sweden, Division Measurement Science (NMI), Gothenburg, Sweden

The Rasch model is not mathematically limited to psychometrics or human respondents; but should also be applicable to more technical agents. This work, as a bridging exercise exploiting this agnostic aspect, explores how psychometric tools can support a major activity in metrology across the disciplines: namely, Interlaboratory Comparisons (ILCs), where one or more common objects are circulated for measurement amongst several laboratories, as a regular tool for assessing performance. ILCs are less developed for ordinal and nominal data and have been rarely studied to date in the human sciences.

The present clinical case study examines whether the performance of different surgical interventions—specifically based on carotid artery stenosis outcomes—can be interpreted in terms of metrological ILCs using Rasch psychometric modelling as an alternative to earlier Generalised Linear Mixed Modelling meta-analyses. Particular attention is given to statistical weighting and the definition of a reference value for these qualitative and categorical ILCs. Task difficulty assessed according to Rasch's principle of specific objectivity provides the ILC reference value for ordinal data.

Beyond surgical applications of the present case study, other potential applications include evaluating the performance of *in vitro* medical devices, large language models, and automated machine-readability for semantic interoperability.

Keywords: Interlaboratory comparison, quality assurance, ordinal, psychometric, clinical, surgical intervention

1. Introduction

The Rasch (1960a,b) psychometric model is not only a statistical but also a uniquely metrological approach to deducing quality-assured estimates of task difficulty and person ability (or equivalents). The special property of "specific objectivity" of the Rasch model enables not only precision but also trueness in these estimates through traceability to metrological references ("etalons") and thereby enables comparability (Pendrill 2014a).

While most Rasch studies have to date focussed on assessing human performance (in the human sciences, education and the health sciences (in surveys, sensorics and educational examinations)), there is arguably nothing in the maths of the psychometric Rasch model to restrict it solely to persons: other, more technical agents could potentially be analysed, potentially opening up a wide and topical set of applications.

Correspondence should be made to Dr. Leslie Pendrill, Research Institutes of Sweden, Division Measurement Science (NMI), Gothenburg, Sweden. Email: Leslie.pendrill@ri.se.

That “agnostic” aspect of the Rasch model has inspired a bridging exercise between the human and physical, engineering and chemical sciences, where the aim is to investigate how the wealth of psychometric tools and methodology can be potentially deployed with advantage to one of the major activities of traditional metrology namely: interlaboratory comparisons (ILCs).

ILCs where typically an object is circulated by a pilot laboratory among a number of participating laboratories – play a key role in assuring the quality of measurements in metrology, which in turn supports quality assurance of products and processes in many areas as part of a quality infrastructure. The aim of any ILC is to assess the comparative performance of the different participating laboratories (or equivalent) by collating their independent results of measurements of the common object circulated amongst them. While established in physical metrology for properties on quantitative interval and ratio scales, (CGPM 2018, Koepke et al., 2017), methods for analysis of ILCs are less well developed for categorical classifications where measurement system responses (2.3. *Measurement System Analysis*) are on the less-quantitative ordinal and nominal scales typical of certain parts of chemical and materials metrology and in the human, medical and social sciences (Bashkansky & Turetsky, 2016). The quality assurance of decision-making, for instance of conformity to specifications in the presence of measurement uncertainty also belongs to this area (5. *Conclusion*).

A human being (or other agent), apart from being a mere operator or object of measurement, is an active part of a measurement system replacing a technological instrument of measurement engineering in many situations in the human, medical and social sciences [Pendrill 2014a]. While human-based (or “person-centred”) properties (“constructs”), associated with raw response scores on ordinal and nominal scales, have not been included to any significant extent in traditional metrology, the turn of millennium has seen a substantial increase in demand for quality assurance in these areas as well the introduction of new insights and research tools to deal with categorical and classification data, as will be exemplified. The burgeoning interest in the quality assurance of ILCs in such areas is (even where the classification “agent” is not necessarily a human being) in part driven by the recent rise in AI, large language models and automatic machine-readability for semantic interoperability and can be regarded as extending the SI (*Système international d'unités*) to cover these new fields of application ([Fisher and Pendrill 2024] and section 5).

After sections in the *Introduction* describing ILCs in general and particularly for category-based properties, the methods deployed in the present study as well as previous attempts at analysing them will be presented in section 2 *Methods*. Using a case study (section 1.2) of performance metrics for carotid surgical interventions – where the “agent” is a hospital team (rather than an individual) – in a general review of ILC analysis methods made by Koepke et al. (2017), item-response scoring (section 2.2) will be compared and contrasted with earlier analysis methods (section 2.1), particularly in terms of the statistical distributions and measurement models used. Section 3 *Results* will demonstrate how data analysis with modern measurement theory – particularly Rasch psychometric modelling and Generalised Linear Mixed Modelling (GLMM) - compares in performance with previous methods for the analysis of category-based ILCs (section 2.1). How classic Forest plots can be modernised will be shown (section 2.3). The effects of statistical weighting (section 4.2) and heterogeneity will be one focus of attention. Another focus is on how the ILC reference value is defined in the case of qualitative and category-based observations. Before concluding with a perspective on future developments (section 5 *Conclusions*), in section 4 *Discussions* the paper considers a number of assumptions, reliability and validity checks to support the choice of analysis methods.

1.1 Ordinal non-linearity and specific objectivity in meta-analyses and ILCs

A meta-analysis is the statistical analysis of a collection of analytic results for the purpose of integrating findings, where information on the efficacy of a treatment is available from a number of clinical studies with similar treatment protocols. In the majority of examples considered by DerSimonian and Nan (1986) in their pioneering work on meta-analyses when judging relative treatment efficacy, the outcome measures analysed were "differences in proportions", as recalled by Koepke et al. (2017) in their ILC study.

1.1.1 Counted fractions

While guidance about how to handle the effects of differing sample sizes and patient populations as well as different levels of sampling error was provided in those earlier meta-analysis studies, there was unfortunately no consideration of the so-called "counted fraction" effect and other sources of scale non-linearity (Tukey 1984, Pendrill 2019). Ordinal "counted fraction" non-linearity is not only found with dichotomous data but is in fact a general feature of categorical data on a bounded scale, where a "*categorical variable* has a measurement scale consisting of a set of categories" [Agresti 2013]. Put simply, any score $X_j\% = \frac{X_j}{\sum_{k=1}^{K} X_k}\%$ on a fixed scale on X bounded by 0 and 100% (as in the survival rates for the surgical interventions studied here, section 1.2) becomes increasingly non-linear at either extreme of the scale (Tukey, 1984). A difference of, for example, 1% percentage points in score X is a different amount at mid-scale (50%) compared with scores at 5% or 95%. Methods for compensating for the ordinal non-linearity caused by the counted-fraction effect are well known in the literature (Pendrill 2019, pp. 88 - 9) for a diversity of applications, such as psychometry (Rasch 1960a,b, section 2.2) and compositional data analysis (Aitchison (1982).

Key performance parameters in ILCs are Degrees of Equivalence (DoEs) as defined as pairs $\{(D_i, U_{95\%}(D_i))\}$ in the CIPM [1999] (Comite international des poids et mesures) MRA (Mutual recognition arrangement). In the presence of uncompensated scale non-linearities, the DoE pairs $\{(D_i, U_{95\%}(D_i))\}$ cannot be assumed automatically to be calculable on the ordinal scales typical of percentage performance metrics:

The expression $\boldsymbol{D_i} = x_i - \hat{\mu}$, as well as the sums involved in the ILC average $\hat{\mu} = \frac{1}{N_i} \cdot \sum_{i=1}^{N_i} x_i$ over the individual scores, x_i, of the N_i "laboratories" participating in the ILC, cannot be calculated since the "distances" between different scores are not known exactly on ordinal scales.

Similarly, when calculating the $\boldsymbol{U_{95\%}(D_i)}$, the associated expanded measurement uncertainty for 95% coverage of the D_i, differences involved in calculating standard deviations in expressions such as $s = \frac{1}{(N_i-1)} \cdot \sum_{i=1}^{N_i} (x_i - \hat{\mu})^2$ are not known and not automatically calculable.

Neither trueness nor precision, as primary measures of accuracy [ISO 5725] are thus not initially calculable for qualitative ILCs.

1.1.2 Separate estimates of object and instrument attributes

Apart from these counted fractions issues and other sources of ordinality (section 1.1.1), there is also the challenge of correctly dealing with a confounding of laboratory and object. Classic test theory (CTT) not only fails to compensate for 'counted fractions' ordinality of percentage raw scores but also does not make a separation of classifier ability from task difficulty which are confounded in the raw data, as described in modern measurement theory (see sections 2.2 & 2.3). There have been a number of attempts at handling ordinal and nominal

data from ILCs, as reviewed in section 2.1. Even works which compensate for ordinality – such as the GLMM approach of Koepke et al. (2017), section 2.1 – do not make this separation of ability and difficulty which is essential for a full metrological treatment of ordinal ILCs.

Making the metrologically important separation of agent ability and task difficulty in qualitative, categorical responses of the measurement systems is dealt with below according to the Rasch (1960) principle of specific objectivity and more generally for any element (object, instrument, operator, environment and method) of a measurement system (section 2.3). It will be shown how an ILC reference value can be defined in the case of qualitative and category-based observations.

Naturally the validity of the various assumptions invoked with the Rasch model needs to be checked, as will be described in section 4.

1.2 Carotid case study

One of the examples chosen by Koepke et al.(2017) to illustrate ordinal metrological ILCs and key comparisons, the performance data from carotid artery stenosis meta-analyses of different surgical interventions, will be studied further here to illustrate handling of the mathematical limitations belonging to ordinal scales and the metrological separation of laboratory ability from task difficulty achievable with modern measurement theory (section 2.2).

Koepke et al. (2017) provide data (Table 1) for the incidence of strokes ($k_E$and k_S) among patients (n_E and n_S) undergoing two alternative carotid surgical interventions, respectively:

endarterectomy (E) and stenting (S) from which they calculate a response "score" for each of the 2 treatment types (j) and the 9 agent "laboratories" (i). $p_{i,j} = P_{success,i,j} = 1 - \frac{k_{i,j}}{n_{i,j}}$ on a categorical scale $0 \cdots 100\%$. The Forest plot shown in Figure 1 (Figure 6 in Koepke et al., 2017) appears to be based on a score as a "laboratory average" $p_{E/S} = \frac{1}{N_q} \cdot \sum_{\mathrm{q}=1}^{N_q} \frac{k_q}{n_q}$, using the raw data shown in Table 1. Table 1 also refers to the following parameters studied later psychometrically in the text (section 4.2): $\boldsymbol{w}$: weighting factor; $\boldsymbol{\theta_i}$ and $\boldsymbol{U(\theta_i)}$, as referenced in the column headings.

Forest plots, as exemplified by Koepke et al. (2017) can be part of the solution, where ordinality in ILC results is compensated for by taking log-odds ratios (OR) [Bland and Altman (2000)] of the laboratory response $y = P_{success}$, i.e., $\log(OR) = \log\left(\frac{P_{\mathrm{success,i}}}{1-P_{\mathrm{success,i}}}\right)$ for each laboratory, i, participating in the ILC.

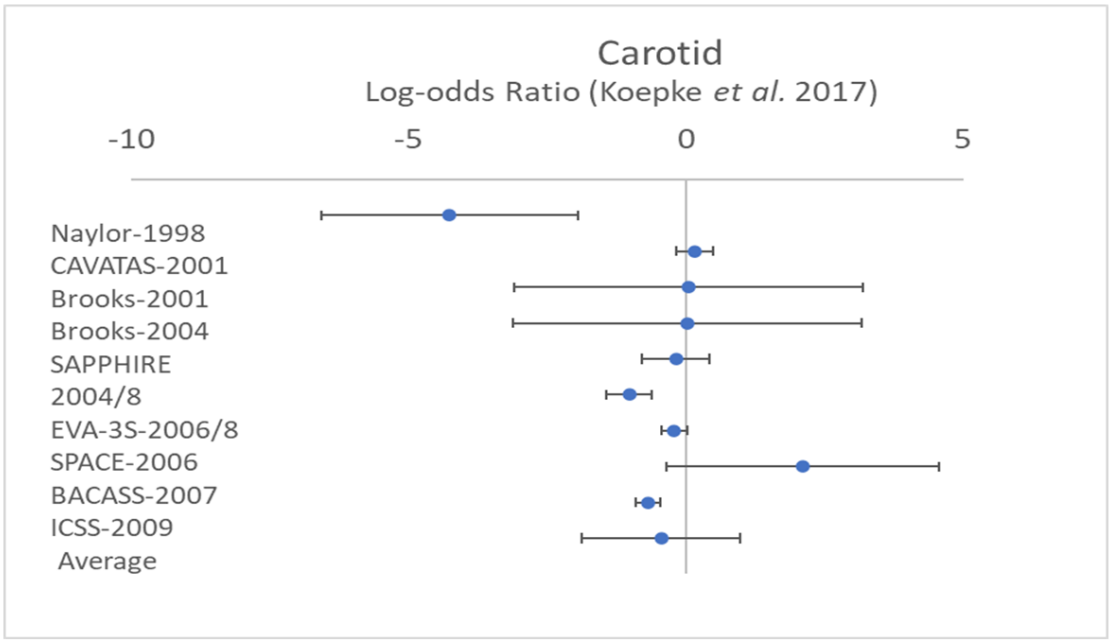

Figure 1.
Forest plot summarizing the data and results for the carotid artery stenosis meta-analysis as reported by Koepke et al., (2017)

Table 1.
Data and results for the carotid artery stenosis meta-analysis (Koepke et al., 2017). Notation explained in text after Table 1.

			Endarterectomy, E (j = 1 = E)		Rasch laboratory (i) ability (weighted, w) – *this work, section 4.2*			Stenting, S (j = 2 = S)		
Study, i	$\boldsymbol{w = \frac{n_i}{\sum_k^{N_{TP}} n_k}}$	$\boldsymbol{n = n_E + n_S}$	n_E	k_E	p_E	$\boldsymbol{\theta_i}$	$U(\boldsymbol{\theta_i})$	n_S	k_S	p_S
Naylor-1998	0.005	23	12	0	1.00	0.08	0.18	11	5	0.55
CAVATAS-2001	0.112	504	253	21	0.92	0.23	0.41	251	18	0.93
Brooks-2001	0.023	104	51	0	1.00	6.37	2.76	53	6	0.96
Brooks-2004	0.019	85	42	0	1.00	6.37	2.76	43	24	0.91
SAPPHIRE-2004/8	0.074	334	167	5	0.97	3.58	1.47	167	45	0.92
EVA-3S-2006/8	0.117	527	262	9	0.97	0.91	1.01	265	0	1.00
SPACE-2006	0.263	1183	584	36	0.94	0.67	1.02	599	65	0.92
BACASS-2007	0.004	20	10	1	0.90	2.38	1.13	10	0	1.00
ICSS-2009	0.381	1710	857	34	0.96	1.57	1.51	853	0	1.00

A particular insight gained in our present work is that Rasch's (1960) principle of specific objectivity applied to an ILC of a set of laboratories for a task of given difficulty indicates a reference (consensus) value defined by that task difficulty (Melin et al., 2021; Pendrill 2024), as discussed further below (section 2.3.2). Obviously, depending on what is most of interest, Forest plots can also be drawn for graphical comparisons of other elements of the Measurement System, viz., object, instrument, operator, environment and test method, as appropriate. Forest plots can indeed be regarded as the first stage in a Rasch analysis as part of modern measurement theory (see section 2).

2. Methods

Modern measurement theory – combining the Rasch (1960a,b) statistical transformation [section 2.2] of raw scores with measurement system analysis (MSA) [section 2.3 and Bentley 2004] – provides a contemporary alternative for the analysis of ILCs, but has been applied on only a few occasions, as will be surveyed below. At the same time, one can benefit from a wealth of recommendations and tools, many of which should be applicable even in other sciences, which have been developed in many studies applying the Rasch model psychometrically since the 1960s in situations in both education and medical sciences where typically a person acts as a measurement instrument [Pendrill 2014a]. Software – such as WINSTEPS ® and RUMM ® - have been developed over decades for Rasch analysis while open-source programs, written for instance in R and Python, are developing rapidly (Linacre 2022).

In order to illustrate a modern measurement approach to the analysis of ILCs, this paper will consider ILC case studies of performance data, particularly carotid artery stenosis meta-analyses of different surgical interventions (Koepke et al., 2017). Earlier studies in a similar vein include pregnancy testing in a recent proposal to modernise the venerable classifier performance tool called the Receiver Operating Characteristic (ROC) (Pendrill et al., 2023).

2.1 Earlier attempts at handling ordinal and nominal data from ILCs

An original approach to handling the analysis of ILCs with ordinal data was proposed in the ORDANOVA approach a variant of ANOVA (Analysis Of Variance) applicable when assessing, for example, "academic or military ranks, sports or voting results, etc" (Gadrich and Bashkansky, 2012). The approach is based on a number of metrics for distance which are claimed to apply to ordinal data. These metrics are mentioned in a recent review by Marmor and Bashkansky (2020) for not only the well-known nominal, ordinal, interval, and ratio scales but also other types of quality data, such as ranks and prioritised data, some of which were earlier considered by amongst others (Tukey, 1984). One metric in particular in their review was found in ORDANOVA to be a suitable statistic, "whose deviation above the value 1 may indicate intersample distinctions", namely (for binary classifications with squared residuals χ^2):

$$I_2 \approx \frac{\chi^2}{M-1}$$

where M samples of equal size n are drawn at different times, from different places or from different processes and so on (Gadrich and Bashkansky, 2012).

The ORDANOVA approach has been adopted in the meantime by several groups:

- Gadrich, et al. (2022) recently built on the ordinal analysis of variation (ORDANOVA), together with multinomial ordered logistic regression in their study of the quality of sausages from different producers, as an example.

- When considering the validation of binary test methods (Uhlig et al., 2013) proposed a "profile likelihood confidence interval" based on a "latent random laboratory effect" and the previous ORDANOVA approach. Typical examples included methods for correctly identifying and counting microbiological specimens and medical screening methods yielding ''yes/no'' statements.
- Nichani & Uhlig (2023) as recently as 2023 refer to their 2013 methods when reporting on so-called "non-targeted" methods (NTM) which, according to the (US Pharmacopeia., 2019), is: "A method that determines the similarity of a sample (*U*) to a reference standard or set (*Sn*). It has a binary output—the sample is atypical or typical with respect to the known sample set."

ORDANOVA statistics, such as I_2, unfortunately do not make the metrological important separation of task difficulty and agent ability, or product quality and agent leniency, and it is not sure that metrics such as χ^2 can be straightforwardly calculated on an ordinal scale where even the most basic mathematical operations can be invalid (section 1.1). Both groups of Bashkansky and of Uhlig have, in separate works, studied what can be considered essentially modern measurement theory, including the Rasch approach to analysing PTs (see section 2.2).

Uhlig et al. (2015) proposed a logit-based method to define scores in category-based ILCs reflecting both what they called the "Level of Competence of the Laboratory (LCL)" (Rasch θ_i) and the "Level of Difficulty of the Task (LDT)" (Rasch δ_j). In their model, Uhlig et al. (2015) define so-called *L*-scores, with properties "like those of z-scores", can be calculated for ILC proficiency tests for qualitative test methods:

$$L_{i,j} = \begin{cases} \Phi^{-1}\left(1 - \dfrac{0.5 \cdot e^{-LDT_j}}{1 + e^{-LDT_j}}\right); & if\ laboratory\ i\ successfully\ completed\ task\ j \\ \Phi^{-1}\left(\dfrac{0.5}{1 + e^{-LDT_j}}\right); & if\ laboratory\ i\ failed\ task\ j \end{cases}$$

where Φ denotes the cumulative distribution function of the standard Normal distribution.

A recent call for more harmonised measurement results in clinical (OMERACT) contexts (de Beurs, et al., 2022) included recommendations to express test results as what they called "*T* scores" which were z-scores $\frac{x_i - \hat{\mu}}{\sigma}$ rescaled on a convenient scale 0 – 100 and with a general population as reference group. OMERACT considered *T* scoring of both raw scores – that is, with all the limitations of Classical Test Theory (section 1.1) – as well as a more modern, measurement theory based on IRT, where we would always recommend the latter (see section 2.2).

The modelling of decisions about manufactured product in the context of industrial processes, (Akkerhuis, 2016; Akkerhuis et al., 2017) combines traditional statistical decision risk descriptions with a so-called "latent-variable" approach in an attempt at allowing for the rate of correct decision-making by a rater.

In the recent IUPAC Brief Guide to Measurement Uncertainty (Possolo et al*.,* 2023), an example (4G: Molnupiravir – maximum likelihood estimation of treatment efficacy), quotes Bernal et al. [2022] who reported the final results of the MOVe-OUT clinical trial of the efficacy of molnupiravir against COVID-19. Possolo et al. (2023) model these observations as outcomes of binomial distributions whose parameter p (the probability of hospitalization or death) depends on whether the patients were on molnupiravir (p_1) or on placebo (p_0). The corresponding odds of hospitalization or death in the two groups are $O_1 = \frac{p_1}{1-p_1}$ and $O_0 = \frac{p_0}{1-p_0}$.

Strikingly, much of the widely used GLMM methodology addressing ordinal data, as adopted by for instance Koepke et al. (2017), appears to have developed in parallel with the

Rasch approach (sections 2.2). There seems to be little if any mention of that approach in the literature (section 2.1), such as when implementing the R function *rma.glmm* defined in package *metafor* [(Viechtbauer W, 2010)] or in the keynote work on ordinality and GLMM by McCullagh (1980) and McCullagh and Nelder (1989). According to McCullagh (1980) in the GLMM approach, the proportional odds model (in his notation) is identical to the linear logistic model:

$$log\left[\frac{\gamma_j(\mathbf{x})}{\{1-\gamma_j(\mathbf{x})\}}\right] = \theta_j - \boldsymbol{\beta}^{\mathrm{T}} \cdot \mathbf{x}$$

where:

$\gamma_j(\mathbf{x}) = \pi_1(\mathbf{x}) + \cdots \pi_j(\mathbf{x})$, the sum of the probabilities π for the ordered categories of the response, Y, when the covariates (AKA explanatory variables) have the values **x**.

$\theta_j = log(\kappa_j)$; where $\kappa_j(\mathbf{x}) = \kappa_j \cdot e^{-\boldsymbol{\beta}^{\mathrm{T}} \cdot \mathbf{x}}$ are the odds that the response $Y \leq j$; $(1 \leq j \leq k)$ and **β** is a vector of unknown parameters.

The lefthand side of the GLMM equation is referred to as $link\left(\gamma_j(\mathbf{x})\right)$ by McCullagh (1980). There is a need to bring McCullagh's (1980) approach into a more metrological form, emphasising traceability (for interoperability) and uncertainty (for risk assessment).

2.2 Item-response scoring

Rasch (1960a,b), in pioneering work in psychometry, posited that the odds ratio of successfully performing a task (*j*) is equal to the ratio of an ability, h_i, ((Rasch, 1960a) used the person (*i*) attribute 'inability' instead, given by ${h_i}^{-1}$), to a difficulty, k_j:

$$\frac{P_{success,i,j}}{1 - P_{successi,j}} = \frac{h_i}{k_j} \qquad (1)$$

where the test person ("agent") ability, $\theta_i = log(h_i)$, and task ("object") difficulty, $\delta_j = log(k_j)$ (or other item:probe pairs of attributes).

The probability $P_{success,i,j}$ of a 'correct' response on item *j* (one of the 2 treatment types E and S in the present case study) by a rater *i* (each of the nine agent "laboratories", where the number of patients treated per surgical team is given in Table 1) is cumulative since easier items have already been tackled, and the cumulative distribution function of the logistic distribution for the Rasch model is an operating characteristic function (OCF, Linacre 2006). The Rasch approach essentially compensates for the 'counted fractions' ordinality (section 1.1.1).

As will be seen, Rasch's (1960a,b) principle of specific objectivity (section 1.1.2) additionally has turned out to be essential when aiming at metrological rather than mere statistical descriptions (Pendrill 2014a). While widely used in human-based systems, such as in education and health, there is nothing in the mathematics of eq. 1 to exclude other kinds of agents.

2.3 Measurement system analysis (MSA)

A first, all-important step in analysing any measurement situation with a view to quality assurance is to make as complete and correct description as possible of the actual measurement system used. In any ILC, in principle each element of the measurement system – object, instrument, operator, environment and test method – can be evaluated separately. Adopting such an MSA approach brings to classification and decision-making analyses, including

responses on ordinal and nominal scales, all the benefits of the substantial body of knowledge acquired over the decades in extensive studies, for example, in engineering measurement in manufacturing industries (Bentley, 2004), (ASTM E11 Committee). An MSA description provides invaluable support when identifying potential sources of measurement error, uncertainties, and the incorrect decisions they cause (JCGM, 2020) (Pendrill, 2019; Pendrill, 2014b). As mentioned above (section 1.1), a pre-requisite for obtaining metrological references for traceability and interoperability is to be able to make independent estimates of agent ability and task difficulty, which in the Rasch (1960a,b) model is the principle of specific objectivity, in much the same way as mass standards can be realised by weighing with a separately calibrated balance.

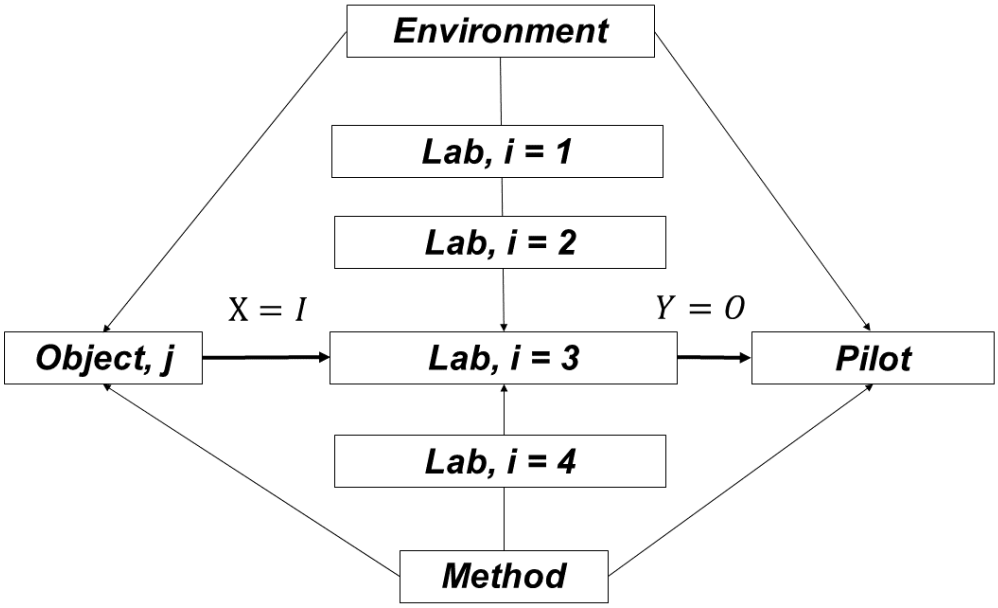

Figure 2.
MSA of an ILC

The MSA approach is not, of course, restricted to one laboratory (as a MSA instrument) but can be readily extended to describe ILCs with several instruments (i.e., laboratories), as illustrated in Figure 2. Here, X equals the input signal (*I*) to the instruments which respond with an output signal Y = *O*.

The underlying ANOVA on which MSA is based fits well with the ISO 5725 series of standards addressing accuracy experiments. A measurement model gives the following expression for the y_{jik}= "response" of a "measurement system", as registered by the Pilot (figure 2):

$$y_{jik} = x_j + B_i + \varepsilon_{jik}$$

$j = 1 \ldots q$; number of levels

$i = 1 \ldots p_j$; number of laboratories at level *j*

$k = 1 \ldots n_{ij}$; number of repeated measurements at level *j* by lab *i*

x_j = attribute (stimulus input, I) value of measurement object (circulating test object), distributed about a mean value $\bar{x} = \mu$ for each level (*j*);

B = laboratory, *i*, bias; ε_{jik}= repeatability variations.

If needed, one can single out explicit terms for each important element of the measurement system: apart from the object and instrument, also the operator, environment and test method chosen.

The classic measures of accuracy – including sub-concepts precision and trueness – are derived from repetitions under conditions of repeatability (same measurement system) and reproducibility (different measurement systems) as terms of variation in "part", "gauge", "operator", etc. (ASTM E11 Committee).

Each element of the measurement system can be evaluated in an ILC by holding the other elements constant and of known value and uncertainty. For example, an ILC intended to evaluate laboratory proficiency can involve circulation of an object of known value and uncertainty, while a test method can be evaluated in an ILC performed by laboratories of known proficiency.

The MSA approach, well-established in measurement engineering, is adapted here to include situations where an agent acts as a measurement instrument, thereby enabling the metrology of categorical properties [Pendrill 2014a, Uher 2018]. The Rasch psychometric model constitutes restitution (of the object and instrument attributes) in such measurement-specific cases (Pendrill 2019 and section 2.2, eq. 1). In the words of Rossi (2014): 'In the *restitution* phase, the indication of the measuring system is interpreted on the basis of the calibration function and the measurement value is obtained'. In psychometric studies in the human, medical and social sciences – the original context of Rasch's (1960a,b) studies – the "agent" acting as an instrument is often a human being.

In the carotid surgical example chosen by Koepke et al., (2017) to illustrate ordinal metrological ILCs and key comparisons, the "agent" (i.e. "laboratory" in MSA) will be the hospital team making each operation made at different times when performing two alternative carotid surgical interventions ("tasks" E and S, MSA objects) by the different "laboratories". The advantages of applying the principle of specific objectivity to the Carotid case study will be demonstrated in the present analysis.

2.3.1 Performance metrics for categorical classifications. Modern measurement theory

While many measurement systems deliver responses on quantitative, continuous scales, in some cases, such as the analogue-to-digital converter and decision-making algorithms, the outputs will often be on categorical scales (Pendrill, 2019).

For categorical response cases, including the important decision-making response, it is not immediately obvious whether expressions such as of accuracy can be applied at all, since the exact mathematical distances between different categories cannot be assumed to be known.

Note that making Forest plots (see section 1.2 and figure 1) of the CIPM ILC pairs $\{(D_i, U_{95\%}(D_i))\}$ does not in itself compensate for the limitations caused by ordinality mentioned above. (There is also of course no sense in making Forest plots of Rasch parameters since those parameters are already transformed onto an interval scale (Lin et al., 2021; *Raschreg*)).

For a simple binary decision, a correct decision is described as assigning the response to the category at the output of the measurement system corresponding to the 'correct' category of the measurement entity at the input to the measurement system. Analogous to the usual measurement error, the closer the categorisation, the greater the 'accuracy', measured in terms of $P_{success}$. Reliability in relation to ILC precision and ILC trueness is considered in section 4.1.

Each classification of a measurement response into a particular category is found in the approach taken here to be best treated as either identification or choice [Iverson and Luce 1998]. A related insight is that specification limits, such as dealt with in conformity assessment based on a continuous, quantitative scale, become as "marks on a ruler", thus uniting measurement of quantitative and qualitative properties [Pendrill 2019]. As pointed out already, the commonality between physical and social measurement (and qualitative estimations more

generally) is first reached when one 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 and restitution (Rossi, 2014).

For categorical responses, measurement system 'accuracy' will be identified (Pendrill, 2019 p. 50) with decision-making ability:

Accuracy (decision-making) = response categorisation – input (true) categorisation (2)

where $P_{success}$ is a metric of measurement system performance in terms of the probability of making the 'correct' decision. In many cases, Rasch modelling (eq. 1) can be done in order to restitute (Rossi, 2014) the measurands of interest from the measurement system response (Pendrill, 2019).

Specifically, the Rasch (1960a) model (section 2.2) which allows the probabilistic (dichotomous) response, $y_{i,j} = P_{successi,j}$ (eq. 1) of the agent (i) to each task (j) of each MSA to be compensated for the effects of ordinality and to be related to task difficulty (δ_j) and agent ability (θ_i) with the expression:

$$log\left(\frac{P_{success,i,j}}{1 - P_{success,i,j}}\right) = \theta_i - \delta_j \tag{3}$$

Rasch's dichotomous version (eq. 3), which has the advantage of being conceptually simpler, has subsequently been generalised to the corresponding polytomous variant, including the so-called "partial credit model" [Masters & Wright 1997] used in the present study:

Probability, $q_{i,j,k}$, of response $y_{i,j}$, of instrument (agent) *i;* object (item) j, over k categories, in the present case: $0 \cdots 100\%$:

$$q_{i,j,k} = \frac{e^{\sum_{c=0}^{k}\left(\theta_i - \delta_{j,c}\right)}}{\sum_{k=0}^{C_j} e^{\sum_{c=0}^{k}\left(\theta_i - \delta_{j,c}\right)}}.$$

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 invoked.

The next section will consider how these principles can be applied to Forest plots of ILC results.

2.3.2 Modernising Forest plots for ILC DoE

Forest plots – such as exemplified in figure 1 from the study of Koepke et al. (2017) – arguably deal adequately with counted-fraction ordinality, but at the same time do not usually apply the principle of specific objectivity – whereby independent estimates of agent ability and task difficulty are enabled – which is central to the Rasch (1960a,b) model and is seen as a pre-requisite for obtaining metrological references for traceability and interoperability. Tests of the assumptions when using the Rasch model for the present dataset which ensure the validity of the principle of specific objectivity are discussed in section 4.

As in the case of Uhlig et al. (2015) *L*-scores (section 2.1), Forest plots and GLMM analyses can be brought into a more metrological form by emphasising traceability (for

interoperability) and uncertainty (for risk assessment) by making explicit, respectively, what are the mean $\hat{\mu}$ and standard deviation σ.

Such Forest plots can be modernised by applying the Rasch (1960b) model. This allows unilateral (that is, "one at a time") DoEs (of the MRA, section 1.1) specifically for "laboratory" or "person" ability to be formulated as the pairs, analogous to Eq 1:

$$\left\{\left(D(\theta)_i, U_{95\%}(D(\theta)_i)\right)\right\}, \tag{4}$$

where $D(\theta)_i = \theta_i - \widehat{\mu(\theta)}$; and $U_{95\%}(D(\theta)_i)$ denotes the associated expanded measurement uncertainty in the Rasch parameter difference for 95% coverage of the true difference between the laboratory, *i*, ability θ_i.

One can of course calculate an ILC mean lab ability $\widehat{\mu(\theta)} = \frac{1}{N_i} \cdot \sum_{i=1}^{N_i} \theta_i$.

With a specific objectivity approach, these laboratory-ability, θ_i, estimates can be made separately from task difficulty, δ_j, (the two tasks, E and S) by combining Rasch and Forest plot approaches (Pendrill & Petersson 2016) in the expression for the bilateral difference in ILC results for two "laboratories", A and B:

$$\log(OR) = \log\left(\frac{P_{\text{success,A}}}{1 - P_{\text{success,A}}}\right) - \log\left(\frac{P_{\text{success,B}}}{1 - P_{\text{success,B}}}\right) = \theta_A - \delta_A - (\theta_B - \delta_B) \tag{5}$$

Instead of the "piecemeal" calculation of odds-ratios (OR) for each laboratory separately, the Rasch model (1960a,b) makes a "global" logistic regression to the full data set – all laboratory responses to all tasks simultaneously. That procedure provides an essential "anchoring" across the task difficulties and laboratory abilities needed when assessing the relative performance of, in the present case, different surgical interventions made at different times and operating theatres.

As will be seen in the present study, an ILC of a set of surgical interventions using the Rasch model will identify the overall task difficulty of the intervention as the attribute associated with the "circulating" object ("transfer standard") of traditional ILCs. Requirements for stability during an ILC will be analogous to those traditionally stipulated for the transfer standard CIPM [1999] MRA. At the same time, each "laboratory" will be characterized by the individual ability to perform the intervention (based on scoring in terms of surgical outcomes).

3. Results

3.1 Study population and data collection

3.1.1 Carotid

Raw data shown in Table 1 – see Koepke et al. (2017) for details.

3.2 Data analysis

3.2.1 ILCs and qualitative data

Conceptually in the context of a modernised approach combining Measurement System Analysis (MSA , section 2.3) and Rasch modelling (section 2.2), ILCs for the less-quantitative ordinal and nominal scales typical for measurement system responses in certain parts of chemical and materials metrology and in the human and social sciences and other psychometric variables should be closely analogous to traditional ILCs in physical metrology, as explained in section 1.

The "circulating objects" ("transfer standard") in an ILC have quantities attributed to them (which if known can act as metrological references, for instance "task difficulty"), thus revealing a certain (finite) performance ability of each "laboratory" (or person responding to a

memory test item). Most of the traditional statistics – such as degree of equivalence (DoE) (eq. 4) – deployed in ILCs in metrology in physics can, in principle, also be applied to category-based observations, but with the important proviso that modern measurement theory (see section 2) has been applied first to the raw response data.

3.2.2 Results of Rasch modelling

Considering the aims of an ILC (see section 1) from the point of view of Rasch modelling:

An intercomparison of measurement results for the same measurand – e.g., the level of difficulty, δ, of a given task, j, – obtained by different laboratories working independently can, amongst others, determine the measurement performance – e.g., the ability, θ_i of each lab, i participating. A consensus estimate is obtained in this way of the true value and the measurement uncertainty of each measurand attributed to the circulating ILC object ("transfer standard") of interest – in the case of (Koepke et al., 2017), the level of difficulty for the two tasks being compared: carotid artery stenosis (S) and endarterectomy (E).

A Rasch (1960a,b) analysis can readily be done based on the scoring used by (Koepke et al., 2017), that is, using a "laboratory average" $p_{E/S} = \frac{1}{N_q} \cdot \sum_{q=1}^{N_q} \frac{k_q}{n_q}$, as measures of $P_{success}$ in eq. 3. This Rasch analysis leads to laboratory ability estimates for each study (independent of specific tasks, E and S) shown in Table 1 and in the blue histogram in the upper halves of Figures 3(i) and 3(ii), using the WINSTEPS ® software, and in the Forest plots shown in Figure 4. The Rasch (1960a,b) approach is basically analysing individual scores rather than traditional group statistics, but for the carotid data (Table 1) the number of operations has varied considerably for the different hospitals and years following introduction of the novel surgical interventions. The effects of statistical weighting, to reflect these different numbers of repeated observations for each laboratory, will be considered in the *Discussion* (section 4.2). Task difficulty, δ_j, estimates for the two surgical tasks, E and S (independent of specific laboratories) for their part are shown in the red histograms in the lower halves of Figure 3 and clearly indicate no significant difference (within quoted uncertainties) in task difficulty for either unweighted or weighted Rasch analyses.

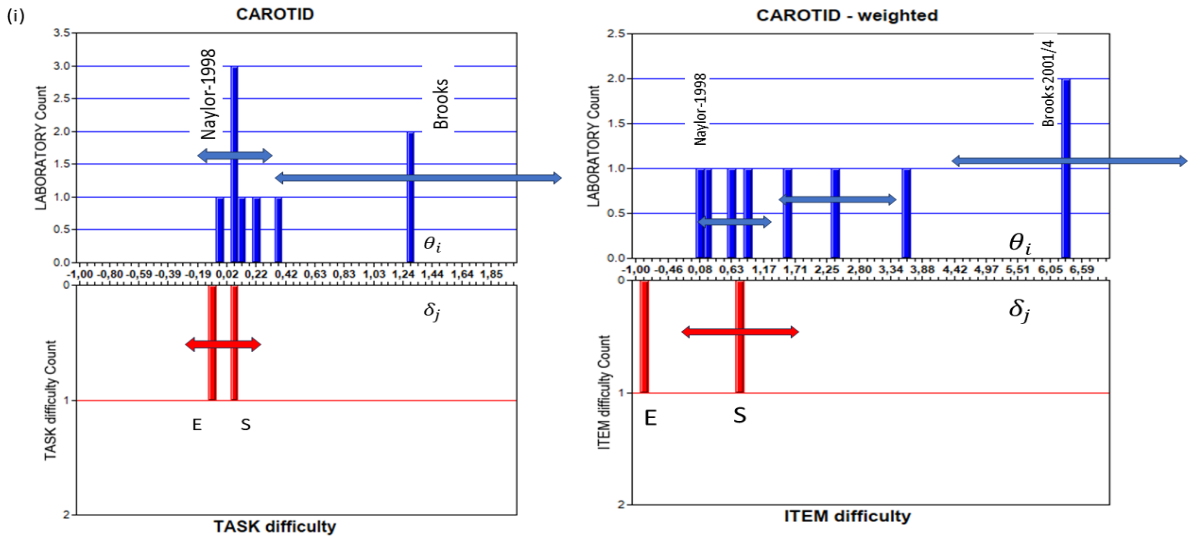

Figure 3.

(i) Unweighted and (ii) Weighted (section 4.2) Rasch plots based on scoring, $p_{E/S} = \frac{1}{N_q} \cdot \sum_{q=1}^{N_q} \frac{k_q}{n_q}$, (raw data Table 1), for the carotid artery stenosis (S) and endarterectomy (E) meta-analyses (Koepke et al., 2017) (Expanded uncertainties, coverage factor $k = 2$.)

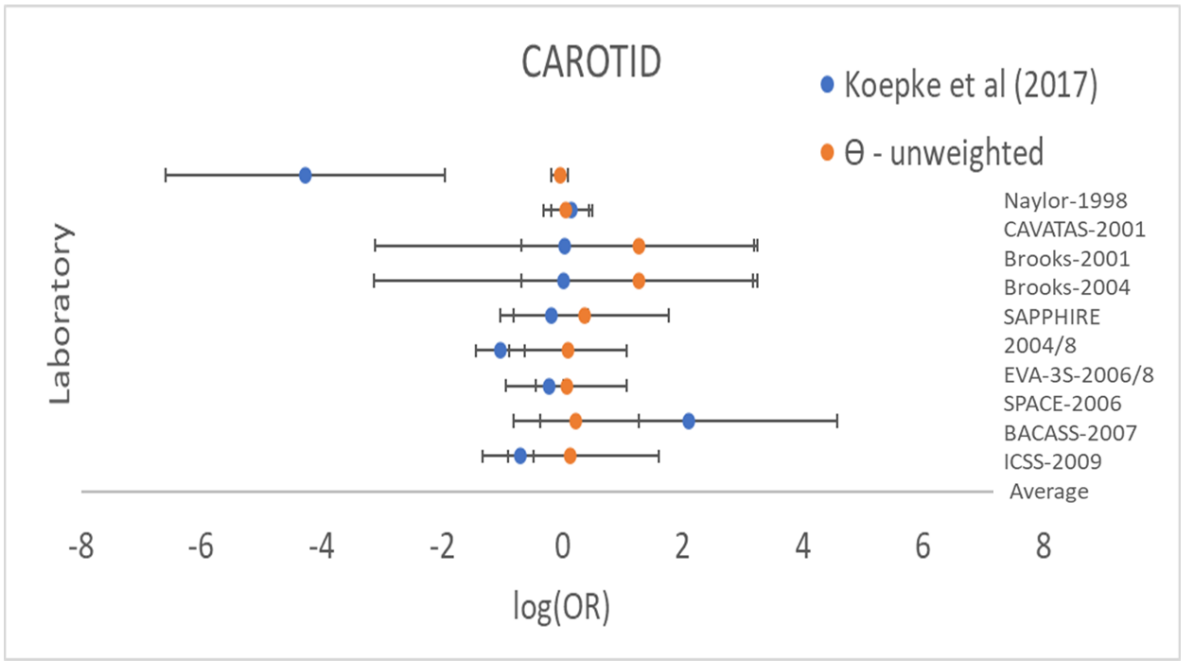

Figure 4.
Forest plots of the original Forest plots of (Koepke et al., 2017), together with the unweighted Rasch laboratory ability, Θ, estimates based on scoring, $p_{E/S} = \frac{1}{N_q} \cdot \sum_{q=1}^{N_q} \frac{k_q}{n_q}$, for the carotid artery stenosis (S) and endarterectomy (E) meta-analyses (Koepke et al., 2017) (Expanded uncertainties, coverage factor $k = 2$.)

4. Discussion

Provided one applies a modernised approach combining MSA and Rasch modelling, ILCs for the less-quantitative ordinal and nominal scales should be closely analogous to traditional ILCs in physical metrology. The "circulating objects" (“transfer standard”) in an ILC have quantities attributed to them (which if known can act as metrological references, for instance “task difficulty”), thus revealing a certain (finite) performance ability of each “laboratory” (or meter or person responding to a memory test item). Our presentation has shown how earlier ILC analyses using Forest plots (Koepke et al., 2017) can be modernised by applying the Rasch model, including compensation for counted-fraction non-linearity (section 1.1.1).

As mentioned in section 1.1.2, the MSA responses are also not measures of the instrument’s (classifier’s or decision maker’s) ability, nor the classification task (object) difficulty, but depend on both, following the Principle of Specific Objectivity. In making such estimates, the normal rules of statistics apply – including making sufficient numbers of observations of items and instruments – to ensure minimum levels of reliability and power. 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 [Melin, et al. (2023)], as follows:

4.1 Reliability

Reliability means assessing the influence of measurement uncertainty as an estimate of limited measurement quality. In psychometrics (Wright & Stone (1979)), a reliability

coefficient, R_β, for a Rasch variable $\beta = \beta' + \varepsilon_\beta$, (for either Rasch attribute: $\beta = \theta$ or δ) including an error term, ε_β, is defined as:

$$R_\beta = \frac{\text{True variance}}{\text{Observed variance}} = \frac{\text{var}(\beta')}{\text{var}(\beta)} = \frac{\text{var}(\beta) - \text{var}(\varepsilon_\beta)}{\text{var}(\beta)}$$

The consequences of the decisions to be made determine what are the actual limits, for instance on maximum permissible uncertainty. Traditional psychometric limits for so-called "high-stake" decisions are typically $R_\beta > 0.8$ which corresponds to at most half of the observed dispersion being assigned to measurement uncertainty. Arguably, in the present case study, it is in the first hand the relative performance of each surgical team which is of primary interest in the provision of healthcare services (patients will want to consider the comparability of performance metrics), and here the number of laboratories included in the Koepke et al. 2019 study should be adequate. Even though there are only 2 items (surgical tasks E and S), the reliability of task difficulty locations (shown in figure 3) is at least in part also determined by the sample size, since the two Rasch variables share a common linear (logit) scale and the logistic regressions are made for the complete matrix of raw responses for all items and instruments.

Comparing and contrasting the original Forest plots of (Koepke et al., 2017) – figure 1 – and the Rasch based plots of Figures 4 reveals some differences: Apart from obvious differences in the levels of measurement uncertainty for the different "laboratories" (i.e., surgical interventions in the Carotid case study) of the cohort, even individual results differ:

For instance, in the Rasch-based plots (figure 4), surgical interventions at Brooks are more clearly differentiated from the other instances (although admittedly the uncertainties are still large), while Naylor is less differentiated from the majority of the cohort when compared with the traditional Forest plots (figure 1). These changes in differentiation appear to be a result of applying the Rasch principle of specific objectivity, that is, where separate estimates of task and instrument attributes are made whereas in the traditional Forest plots of Koepke et al. (2017), no such separation were made.

In their study of carotid GLMM analyses, (Koepke et al., 2017 figure 1 above and section 2.1), write:

To evaluate the log-odds ratio and its standard uncertainty computed using Bayes estimates of the relevant probabilities, … a large number K of draws from beta distributions (with n_E and n_S kept fixed at the values given) were made…. finally comput(ing) the mean and standard deviation of the resulting K values of the log-odds ratio…listed under $log(OR)$ and $u\big(log(OR)\big)$ in (their) table 5.

How uncertainties were evaluated is reported by Possolo et al. (2023) in their recent COVID study:

$$u^2[ln(O_1/O_0)] = \frac{1}{n_1 \cdot p_1 \cdot (1-p_1)} + \frac{1}{n_0 \cdot p_0 \cdot (1-p_0)}$$

where notation is defined in section 2.1.

Uncertainties (so-called Standard Errors, SE, i.e., standard uncertainties) corresponding to our Rasch analyses were those quoted by the WINSTEPS® program (section 21.120), where according to that program's manual (p. 805):

$$\text{SE}\big(\theta_i, \bar{\delta}\big) = \sqrt{\frac{1}{\sum_{j=1}^{N_{items}}[P_{success,i,j} \cdot (1-P_{success,i,j})]}} \text{ and } \text{SE}\big(\bar{\theta}, \delta_j\big) = \sqrt{\frac{1}{\sum_{i=1}^{N_{TP}}[P_{success,i,j} \cdot (1-P_{success,i,j})]}}$$

4.2 Individual and group statistics

In the present study – where the hospital surgical teams replace the psychometric test persons as agents – account has to be taken of the different number, n_i, of repeated observations for each agent (surgical team) in the Carotid study (Table 1). There are procedures in standard Rasch software for weighting: In the words of Linacre and WINSTEPS® (section 19.106):

"When using significance tests with weighting, normalize the weights so that the total amount of independent statistical information in the data is not over- or under-inflated, i.e., when using PWEIGHT= with an observed sample size of N, multiply all PWEIGHT= values by N / (sum of all weights)."

As in any weighted least-squares regression or calculation of a mean, a classic choice of weighting factor is $w_i = \frac{1}{\sigma_i^2}$, that is, inverse variance – the smaller the dispersion, the more weight is apportioned. The variance associated with laboratory, *i*, can be modelled for the ILC of interest by assuming a "global" variance characterising the whole population, divided by the number of repeated observations (or, more generally, the number of degrees of freedom): $\sigma_i^2 = \frac{\sigma^2}{n_i} \Rightarrow w_i \sim n_i$, under which assumption the weighting factor simply equals the number of observations. It is usual to normalise the weighting factors in proportion to the total number of observations, as is programmed into the WINSTEPS ® datafile with the expression: $PWEIGHT_i = \frac{n_i}{\sum_k^{N_{TP}} n_k}$, as given in Table 1. These weighting factors, *w*, correlate well (Pearson coefficient, $R = 0.93$) with the DerSimonian – Laird weights quoted by Koepke et al. (2017).

The degree of agreement between Rasch (unweighted) and GLMM estimates (Koepke et al., 2017) of laboratory ability is shown in the Forest plots of Figure 5:

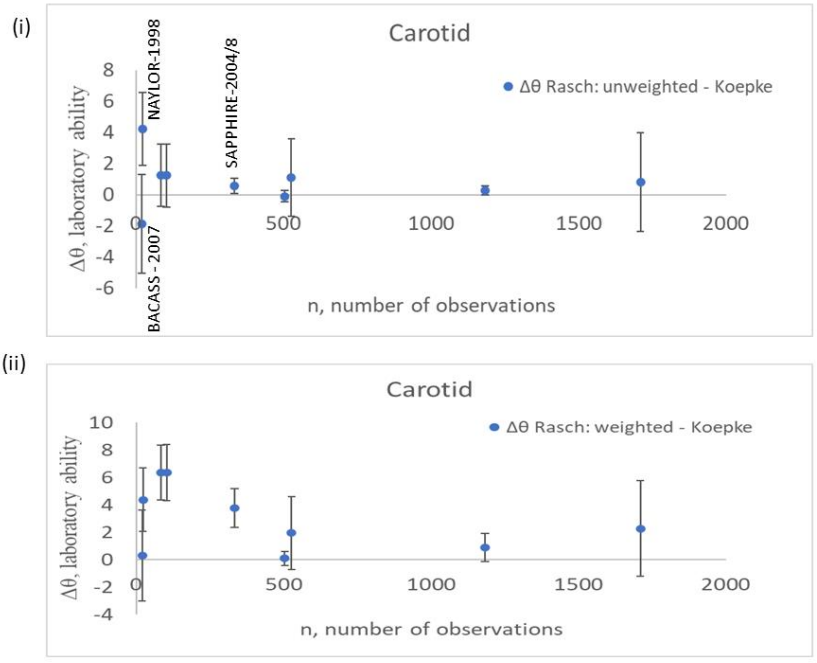

Figure 5.
Differences between surgical interventions: (i) Unweighted and (ii) weighted Rasch Θ - δ (this work) and Bayesian log(OR) (Koepke et al., 2017), based on the data and scoring, $p_{E/S} = \frac{1}{N_q} \cdot \sum_{q=1}^{N_q} \frac{k_q}{n_q}$, for the carotid artery stenosis (S) and endarterectomy (E) meta-analyses
(Expanded uncertainties, coverage factor $k = 2$.)

Comparing weighted (figure 5(ii)) and unweighted (figure 5(i)) results as differences in laboratory abilities of the different surgical interventions with respect to the original Koepke

et al. (2017) GLMM analyses, there seems to be reasonable agreement with the unweighted analyses (i), but some considerable differences for the weighted Rasch analyses (ii). The discrepancies are mostly for the laboratories with low to medium numbers (n < 500) of repeated observations; for instance, for the pair of observations labelled Brooks ($n \sim 100$).

4.3 Unidimensionality

A basic assumption of the Rasch (1960a,b) model (section 2.2) is that every item and agent should address one common dimension. When formulating the tests, this unidimensionality is intentional. Practice in psychometrics is to subsequently test the validity of that assumption with the actual test results. That is usually done in psychometrics (Linacre [2003], Melin et al. [2021]) by making a Principal Component Analysis (PCA-2) of the logistic regression residuals between the fitted and measured Rasch curves (eq. 3).

Experimentally in the present Carotid study, the PCA loadings, with just two items (E & S), are found to equal 0.23 with the WINSTEPS ® program and the data of Table 1.

The PCA results for the surgical teams ("agents") shown in Table 3 for their part can be compared with those of (Koepke et al., 2017, Possolo 2024 private communication) who report I^2 (total heterogeneity / total variability): 36.39%. This compares with the 49.4% unexplained variance by items (56.9% observed unweighted) in Table 3. Note that Koepke et al. 2017, using the classic GLMM approach (McCullagh and Nelder 1989), did not estimate task difficulty and agent ability separately.

Table 3.
Results of PCA of Rasch analyses (WINSTEPS ® Table 24.0) of the surgical teams ("agents") for the carotid artery stenosis meta-analysis data (Koepke et al., 2017)

Unweighted/weighted	Eigenvalue	Observed	Expected
Total raw variance in observations		100.0%	100.0%
Raw variance explained by measures		86.6%/87.6%	77.1%/79.0%
Raw variance explained by interventions		29.7%/38.2%	26.4%/34.4%
Raw Variance explained by items		56.9%/49.4%	50.7%/44.5%
Raw unexplained variance (total)	7 / 2	13.4%/12.4%	22.9%/21.0%
Unexplained variance in 1st contrast	7/ 0.2	13.4%/9.1%	
Unexplained variance in 2nd contrast	0 / 0	.0%/.0%	
Loading, $a_{p,x}$ + 1	NAYLOR-1998 SAPPHIRE-2004/8 EVA-3S-2006/8 SPACE-2006 ICSS-2009	NAYLOR-1998 EVA-3S-2006/8 ICSS-2009	
Loading, $a_{p,x}$ - 1	CAVATAS-2001 BACASS-2007	CAVATAS-2001 BACASS-2007 SAPPHIRE-2004/8 SPACE-2006	

Uni-dimensionality is required for satisfactory fits of the Rasch model to the data. Linacre [2003] recommends judging satisfactory uni-dimensionality by inspecting the "strength of the Rasch dimension", that is, the variance "explained by the measures", which in the present case exceeds 86% [Linacre 2003], together with the 1st contrast eigenvalue (reported in Table 3).

The eigenvalue of 7 for the 1st contrast of the initial unweighted fit is considerably larger than the largest eigenvalue (2) expected by chance [Linacre/WINSTEPS]. Note however that, after weighting [section 4.2], this eigenvalue is reduced to the threshold value of 2 considered in regular psychometric practice as a limit for acceptable uni-dimensionality, together with the relatively high strength of the Rasch dimension.

Smith's [1996] method of assessing loading patterns was also applied, where the positive and negative patterns [last two rows of Table 3] in the principal component analysis (PCA) of logistic fit residuals may indicate two or more subsets. A second form of Principal Component Analysis (denoted PCA-1 by Melin et al., 2021) can be used to quantify covariate variables when attempting to formulate a Construct Specification Equation for instance for Carotid task difficulty. To date we have no explanatory theory of the Carotid task difficulty (see 5. *Conclusions*).

4.4 Other tests of model validity

Further tests belonging to the battery of conventional tests of Rasch model validity include the following:

– *Targeting*: By inspecting the spread of agent locations (i.e., range of abilities in the sample of surgical teams) and item locations (i.e., range of the items), targeting was assessed (figures). There is no specific criterion (Cleanthous, et al., 202), but the better coverage, the better targeting and the closer the mean person location is to the mean item location indicates whether the person sample is off centred from the items. Note that merely increasing the number of repeated observations (by adding items and/or agents) does not automatically improve reliability (section 4.1), since off-targeting and scale linearity (Akkerhuis, 2016; Akkerhuis et al., 2017) are also important factors.

– *Item and agent fit*: Another test of the validity of the adopted Rasch model (eq. 3) is based on examination of so-called Construct Alleys. A classic tool for assessing potential scale distortion (Pendrill 2019, pp. 179 ff) – such as scale elongation or contraction – the so-called construct alley, is done by plotting the Wilson-Hilferty (1931) statistic, $WH = \frac{2\cdot\left(\sqrt[3]{X}-\sqrt[3]{n-\frac{2}{3}}\right)}{\sigma} = INFIT(z_{STD})$ (where $\sigma = \frac{\sqrt{2}}{3\cdot\sqrt[6]{n-\frac{2}{3}}}$ and $X = \chi^2$, the squared logistic regression residuals for n degrees of freedom versus agent ability (or task difficulty). In the present carotid study, an errant Sapphire-2004/8 in the top left sector of the construct alley could be interpreted according to our theory as a scale elongation, $\tau > 0$ while the top right sector of the errant BACASS-2007 corresponds to a scale compression, $\tau < 0$ for a scale distortion: $\Delta\theta = \tau \cdot \theta$ (Pendrill 2019, p. 186ff). Further studies are needed in order to explain these deviations and scale distortions.

– *Differential item functioning (DIF)*: The invariance and the extent to which items are stable across different subgroups – here diagnosis and gender – can be assessed by examining the estimated person ability differences between class intervals within the subgroups using analysis of variance (ANOVA) [Andrich & Hagquist 2012]. A significant *p*-value for differences between subgroups would indicate DIF.

– *Local dependency (LD)*: To assess the extent of LD among items, residual correlations are evaluated against a relative cut off. They were classified as LD if the item residual correlations were greater than 0.20 above the average correlations (Christensen et al., 2017).

5. Conclusion and future prospects

The present paper has examined ILCs where the agent of the measurement system is not a regular technical instrument as in traditional ILCs of the MRA, but a surgical intervention instead. A key observation is that Psychometrics applies equally well when characterising classification performance for an instrument and for making decisions of conformity in the presence of uncertainty [Pendrill 2024].

The results of the present work indicate good prospects of benefitting from the wealth of psychometric tools and methodology developed in recent decades mainly in the human sciences when deployed in a cross-disciplinary, bridging approach. Exploiting the agnostic character of the Rasch model allows one to tackle one of the major activities of conventional metrology in the physical, engineering and chemical sciences, namely: the analysis of Interlaboratory Comparisons (ILCs) which play a key role in assuring the quality of measurements in metrology. The unique metrological aspects of the psychometric Rasch measurement theory are particularly important in this study of ILCs for ordinal and nominal data with their focus on metrological invariance (through traceability to reference standards) and the assessment of measurement uncertainty. An interlaboratory comparison (ILC) is the metrologists' primary tool to demonstrate the accuracy (including accuracy, i.e. "trueness") of the measurement system as a performance metric. In this work, we study how ILCs should be analysed for ordinal responses by applying Rasch modelling. An exciting result is that the ILC reference values are found here to be the two item difficulty levels that only a Rasch analysis among all IRTs can deliver - thanks to the Rasch model's unique ability to deliver separate estimates of item and TP attribute values, according to the Principle of Specific Objectivity.

The present work has built on and extended earlier attempts at handling ordinal and nominal data from ILCs, as reviewed in section 2.1. In particular, the work of Koepke et al. 2017 in their Carotid surgical study has been extended using current psychometric approaches to yield separate estimates of the ability of individual surgical teams at different times and hospitals and the levels of difficulty posed by two alternative surgical carotid surgical interventions: endarterectomy (E) and stenting (S). A psychometric analysis of the Forest plots of the carotid study indicates a generic result, applicable to in principle all kinds of ordinal and nominal data ILCs, namely: that the intercomparison reference value is to be defined by the task difficulty based on Rasch’s principle of specific objectivity.

Further work is needed in the present case study, including (i) *section* 4.3: no explanatory theory yet of the Carotid task difficulty; (ii) *section* 4.4: explain deviations and scale distortions evident in construct alley plots.

Some indication of the potential gains of adopting a more modern measurement approach to ILC analysis based on ordinal or nominal data can be found in results from other studies of diverse areas, such as:

Neurodegeneration, where the patient is the measurement instrument (Göschel et al., 2024) where uncertainties in a number of legacy memory tests have been halved and some of the very first metrological references for traceability of categorical data have been proposed based on entropy-based Construct Specification Equations.

Our earlier ROC study of qualitative ILCs (Pendrill et al., 2023), modernising the venerable Receiver Operating Characteristic used regularly in assessing classifier performance in Machine Learning, showed how to deal with limited measurement quality, leading to incorrect decisions when reading the display of the measurement instrument (in that case, pregnancy meters) using modern measurement theory.

It is conceivable that the same methodology could be deployed in assessing the quality of for instance COVID *in vitro* Medical devices when setting specifications on the amount of bias

and associated risks in data which need to be tested with quality-assured measurement of device performance [Regulation (EU) 2017/746 (IVDR)].

With the rise of AI and to support legislation and conformity assessment, there is increasing need to specify how to measure qualitative and subjective concepts such as trust (Bostrom et al., 2023, Hoffman et al., 2023) in the context of large language models, and automated machine-readability for semantic interoperability, and in support of major new legislation, for example in the new European AI Act (2023), standards such as ISO/IEC 22989:2022.

Acknowledgment

The author thanks Prof. Possolo and his NIST colleagues for private communications during this study. Dr N Korsell (RISE) is also thanked for discussions and proposals to this manuscript.

Funding details and disclosure statement

RISE, through its internal programme supporting a Competence platform for categorical measurements, has provided funding for this work. No financial or non-financial interests have arisen from this work. The author declares no conflict of interest.

Data Availability Statement

No new data was measured for this study. See instead Koepke et al. (2017) who collated the original data.

How to Cite

Pendrill, L. (2026). Category-based interlaboratory comparisons: Psychometric Rasch analyses defining reference values and statistical weighting in a clinical example. *Educational Methods & Psychometrics, 4* (SAMC 2024 Special Issue): 26. https://doi.org/10.65301/emp.2026.260

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B*, *44*, 139–177.

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley. ISBN 978-0-470-46363-5

Akkerhuis, T. (2016). *Measurement system analysis for binary tests*. http://hdl.handle.net/11245/1.540065

Akkerhuis, T., de Mast, J., & Erdmann, T. (2017). The statistical evaluation of binary tests without gold standard: Robustness of latent variable approaches. *Measurement*, *95*, 473–479. https://doi.org/10.1016/j.measurement.2016.10.043

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Educational Research Perspectives*, *9*(1), 95–104.

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, *37*(3), 387–416.

ASTM E11 Committee. (2011). *Standard guide for measurement systems analysis (MSA) (E2782-11)*. ASTM International. https://doi.org/10.1520/E2782-11

Bashkansky, E., & Turetsky, V. (2016). Proficiency testing: Binary data analysis. *Accreditation and Quality Assurance*, *21*(4), 265–270. https://doi.org/10.1007/s00769-016-1208-x

Bentley, J. D. (2004). *Principles of measurement systems* (4th ed.).

Bernal, A. J., et al. (2022). Molnupiravir for oral treatment of Covid-19 in nonhospitalized patients. *The New England Journal of Medicine*, *386*, 509. https://doi.org/10.1056/NEJMoa2116044

Bland, J. M., & Altman, D. G. (2000). The odds ratio. *BMJ*, *320*, 1468. https://doi.org/10.1136/bmj.320.7247.1468

Bostrom, A., et al. (2023). Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences. *Risk Analysis*, 1–16. https://doi.org/10.1111/risa.14245

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194.

Cleanthous, S., et al. (2021). Psychometric analysis from EMBODY1 and 2 clinical trials to help select suitable fatigue PRO scales for future systemic lupus erythematosus studies. *Rheumatology and Therapy*, *8*(3), 1287–1301. https://doi.org/10.1007/s40744-021-00338-4

CIPM. (1999). *Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes* (Technical supplement revised October 2003). https://www.bipm.org/en/cipm-mra/

de Beurs, E., Boehnke, J. R., & Fried, E. I. (2022). Common measures or common metrics? A plea to harmonize measurement results. *Clinical Psychology & Psychotherapy*, *29*, 1755–1767.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

Fisher, W. P., Jr., & Pendrill, L. R. (2024). Introduction: Imagining the viability, feasibility, and desirability of extending the SI to include the psychological and social domains. In W. P. Fisher, Jr., & L. Pendrill (Eds.), *Models, measurement, and metrology extending the SI* (in press). De Gruyter.

Gadrich, T., & Bashkansky, E. (2012). ORDANOVA: Analysis of ordinal variation. *Journal of Statistical Planning and Inference*, *142*, 3174–3188.

Gadrich, T., et al. (2022). Ordinal analysis of variation of sensory responses in combination with multinomial ordered logistic regression vs. chemical composition: A case study of sausage quality. *Journal of Food Quality*, 1–12. https://doi.org/10.1155/2022/4181460

Göschel, L., et al. (2024). Plasma p-Tau 181 and GFAP reflect 7T MR-derived changes in Alzheimer's disease. *Alzheimer's & Dementia*. https://doi.org/10.1002/alz.14318

Hoffman, R., et al. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, *5*. https://doi.org/10.3389/fcomp.2023.1096257

ISO. (n.d.). *ISO 5725-1: Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions*. https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:en

ISO/IEC. (2022). *ISO/IEC 22989:2022 Artificial intelligence — Concepts and terminology*.

Iverson, G., & Luce, R. (1998). The representational measurement approach to psychophysical and judgmental problems. In *Measurement, judgment, and decision making*. Academic Press.

JCGM. (2020). *Guide to the expression of uncertainty in measurement — Part 6: Developing and using measurement models*. https://www.bipm.org/documents/20126/50065290/JCGM_GUM_6_2020.pdf

Koepke, A., et al. (2017). Consensus building for interlaboratory studies, key comparisons, and meta-analysis. *Metrologia*, *54*(3), S34. https://doi.org/10.1088/1681-7575/aa6c0e

Lin, J.-C., et al. (2021). Using forest plots to present the performance sheets provided to OSCE examinees using Rasch analysis (Preprint). https://doi.org/10.2196/preprints.35741

Linacre, J. M. (2003). Data variance: Explained, modeled, and empirical. *Rasch Measurement Transactions*, *17*(3), 942–943.

Linacre, J. M. (2006). Bernoulli trials, Fisher information, Shannon information and Rasch. *Rasch Measurement Transactions*, *20*(3), 1062–1063.

Linacre, J. M. (2022). R statistics: Survey and review of packages for the estimation of Rasch models. *International Journal of Medical Education*, *13*, 171–175. https://doi.org/10.5116/ijme.629d.d88f

Marmor, Y. N., & Bashkansky, E. (2020). Processing new types of quality data. *Quality and Reliability Engineering International*, *36*, 2621–2638. https://doi.org/10.1002/qre.2642

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. Springer. https://doi.org/10.1007/978-1-4757-2691-6_6

Melin, J., et al. (2021). More than a memory test: A new metric linking blocks, numbers, and words. *Alzheimer's & Dementia*, *17*(S6), e050291. https://doi.org/10.1002/alz.050291

Melin, J., et al. (2023). NeuroMET Memory Metric: Traceability and comparability through crosswalks. *Scientific Reports*. https://doi.org/10.1038/s41598-023-32208-0

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B*, *42*(2), 109–142.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall.

Nichani, K., & Uhlig, S. (2023). Essential terminology and considerations for validation of non-targeted methods. *Food Chemistry: X*, *17*, 100538. https://doi.org/10.1016/j.fochx.2022.100538

Pendrill, L. R. (2014a). Man as a measurement instrument. *NCSLI Measure*, *9*(4), 24–35. https://doi.org/10.1080/19315775.2014.11721702

Pendrill, L. R. (2014b). Using measurement uncertainty in decision-making and conformity assessment. *Metrologia*, *51*(4), S206–S218.

Pendrill, L. R. (2019). *Quality assured measurement: Unification across social and physical sciences*. Springer. https://doi.org/10.1007/978-3-030-28695-8

Pendrill, L. R. (2024). Quantities and units: Order amongst complexity. In W. P. Fisher, Jr., & L. Pendrill (Eds.), *Models, measurement, and metrology extending the SI* (in press). De Gruyter.

Pendrill, L. R., et al. (2021). Reducing search times and entropy in hospital emergency departments with real-time location systems. *IISE Transactions on Healthcare Systems Engineering*. https://doi.org/10.1080/24725579.2021.1881660

Pendrill, L. R., Melin, J., Stavelin, A., & Nordin, G. (2023). Modernising receiver operating characteristic (ROC) curves. *Algorithms*, *16*(5), Article 5. https://doi.org/10.3390/a16050253

Pendrill, L. R., & Petersson, N. (2016). Metrology of human-based and other qualitative measurements. *Measurement Science and Technology*, *27*, 094003.

Possolo, A., et al. (2023). A brief guide to measurement uncertainty (IUPAC technical report). *Pure and Applied Chemistry*. https://doi.org/10.1515/pac-2022-1203

Rasch, G. (1960a). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedagogiske Institut.

Rasch, G. (1960b). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Raschreg. (n.d.). https://github.com/nando11235813/raschreg

Regulation (EU) 2017/746 (IVDR) on in vitro diagnostic medical devices. https://eur-lex.europa.eu/eli/reg/2017/746/oj/eng

Rossi, G. B. (2014). *Measurement and probability: A probabilistic theory of measurement with applications*. Springer. https://doi.org/10.1007/978-94-017-8825-0

SI. (n.d.). *Système international d'unités*. https://www.bipm.org/en/measurement-units

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling*, *3*(1), 25–40.

Tukey, J. A. (1984). Data analysis and behavioural science. In L. V. Jones (Ed.), *The collected works of John W. Tukey, Volume III*.

Uher, J. (2018). Data generation methods across the empirical sciences: Differences in accessibility and data-encoding processes. *Quality & Quantity*. https://doi.org/10.1007/s11135-018-0744-3

Uhlig, S., Krügener, S., & Gowik, P. (2013). A new profile likelihood confidence interval for the mean probability of detection in collaborative studies of binary test methods. *Accreditation and Quality Assurance*, *18*, 367–372. https://doi.org/10.1007/s00769-013-0993-8

Uhlig, S., Bläul, C., & Frost, K., et al. (2015). Qualitative PT data analysis with easy-to-interpret scores. *Accreditation and Quality Assurance*, *20*, 347–353. https://doi.org/10.1007/s00769-015-1174-8

US Pharmacopeia. (2019). *USP guidance on developing and validating non-targeted methods for adulteration detection*. https://www.semanticscholar.org/paper/Appendix-XVIII

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. MESA Press.

Manuscript Received: 28 FEB 2025
Final Version Received: 14 JAN 2026
Published Online Date: 15 FEB 2026