Measurement: Sensors 25 (2023) 100658

Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

Metrological advancements in cognitive measurement: A worked example with the NeuroMET memory metric providing more reliability and efficiency☆

J. Melin [a,*], S.J. Cano [b], A. Flöel [c,d], L. Göschel [e,f], L.R. Pendrill [a]

[a] RISE, Research Institutes of Sweden, Division Safety and Transport, Division Measurement Science and Technology, Gothenburg, Sweden
[b] Modus Outcomes Ltd, Suite 210b, Spirella Building, Letchworth Garden City, SC6 4ET, UK
[c] Department of Neurology, University Medicine Greifswald, Greifswald, Germany
[d] German Center for Neurodegenerative Diseases (DZNE), Standort Rostock/Greifswald, Germany
[e] Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Neurology, Augustenburger Platz 1, 13353, Berlin, Germany
[f] Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, NeuroCure Clinical Research Center, Augustenburger Platz 1, 13353, Berlin, Germany

ARTICLE INFO

Keywords:
Cognition
Entropy
Metrology
Rasch
Person ability
Task difficulty

ABSTRACT

Better metrics of cognition can be formed by carefully combining selected items from legacy short-term memory tests so as to enhance coherence in item design while not jeopardizing validity. In this paper, we report on how Rasch Measurement Theory and Construct specification equations (CSE) have been brought together when composing the NeuroMET Memory Metric (NMM). The NMM is guided by: i) entropy-based equivalence criteria; ii) a comprehensive understanding of the construct purported to be measured; and iii) how a collection of items works together. CSEs play a major role in ensuring the metrological legitimacy of the NMM in a way analogous to certified reference materials in more established areas of metrology. The resulting NMM for short-term memory recall has up to a five-fold reduction in measurement uncertainties for memory ability compared with an individual legacy test, and the entropy-based CSEs should enable more efficient and valid assessment.

1. Introduction

Dementia is one of the most pressing public health issues in modern time [1]. Despite this, patients remain under-diagnosed or are diagnosed 'too late'. This may be due, in part at least, to a lack of metrological quality assurance of cognitive measurements in detecting cognitive decline [2,3]. In turn, this deficiency is based on the fact that typical human responses *have no numerical meaning and only serve to … indicate … ordered categories* [4 p. 2], and are therefore not amenable to even the most basic statistical operations [5]. Comparability, through metrological traceability to the International System of Units (SI), and risk assessment based on uncertainty analyses are, as yet, unmet requirements for regulatory approval of measures of cognition [6,7].

In addition to a lack of metrological quality assurance, the most commonly-used legacy cognitive tests (e.g., Mini Mental State Examination [8] and Alzheimer's Disease Assessment Scale-Cognitive Behavior (section [3]) do not have sufficient accuracy to be able to distinguish between patients (especially in early stage disease), and are not metrologically legitimated. Specifically, person-to-item targeting is often poor, owing to skewed distributions of both task difficulty across the test items and person ability across the cohort, which leads to large measurement uncertainties, particularly for those with early memory decline [9,10].

Metrological quality assurance (i.e., comparability and declared uncertainties) of such cognitive measures is essential if reliable decisions about diagnoses, management, and treatment are to be made throughout the healthcare system [11]. Thus, building on our previous

☆ Extended version of: J. Melin, S.J. Cano, A. Flöel, L. Göschel, L.R. Pendrill, Construct specification equations: 'Recipes' for certified reference materials in cognitive measurement, Measurement: Sensors, Volume 18, 2021, 100290, ISSN 2665–9174, https://doi.org/10.1016/j.measen.2021.100290 and presentation at the IMEKO 2021 World Congress (30 August – 03 September 2021, Japan/virtual).

* Corresponding author.
E-mail addresses: jeanette.melin@ri.se (J. Melin), stefan.cano@threadresearch.com (S.J. Cano), agnes.floeel@med.uni-greifswald.de (A. Flöel), laura.goeschel@charite.de (L. Göschel), leslie.pendrill@ri.se (L.R. Pendrill).

https://doi.org/10.1016/j.measen.2022.100658
Received 25 February 2022; Received in revised form 7 November 2022; Accepted 23 December 2022
Available online 24 December 2022

Abbreviations

AD	Alzheimer's Disease
CBT	Corsi Block Test
CERAD	Consortium to Establish a Registry for Alzheimer's Disease
CRM	Certified reference materials
CSE	Construct Specification Equations
DST	Digit Span Test
EMPIR	European Metrology Programme for Innovation and Research
HC	Healthy control
IRT	Item response theory
MCI	Mild cognitive impairment
MMSE	Mini Mental State Examination
NMM	NeuroMET Memory Metric
NDD	Neurodegenerative disorders
PCA	Principal component analysis
PCOM	Person centered outcome measures
PCR	Principal component regression
RAVLT	Rey's Auditory Verbal Learning Test
RMP	Reference measurement procedure
RMT	Rasch Measurement Theory
SCD	Subjective cognitive decline
SPE	Serial position effect
WLL	Word Learning List

work [5,7,11–14], to minimise risks of incorrect medical decisions about diagnosis and therapy [14], we argue in this paper for the following key points needed to ensure measurement quality assurance for person centered outcome measures (PCOMs).

- Conceptually, we place the person responding to the test as the instrument at the heart of the measurement system, which corresponds more directly with the traditional approach in engineering science and technology [5], as opposed to considering test items as "instruments", as is commonly done in the social sciences [10]. The benefits of this approach will be further developed in section 4.3.
- A measurand restitution based on Rasch Measurement Theory (RMT; [15]) in particular, ensures a separation of task and person attributes as well as a compensation for ordinality (section 3.1). In this respect, it is worth highlighting that RMT "*is not simply a mathematical or statistical approach, but instead a specifically metrological approach to human-based measurement*" [5].
- Construct specification equations (CSE) for both task difficulty and person ability can be developed to explain each construct (task difficulty and person ability) based on best understanding (section 3.2). CSEs appear to provide metrological references for calibration and subsequent intercomparability of measurements. Regarding CSE as 'recipes for certified reference materials' (CRM) enables metrological traceability in the human sciences analogous to the role of CRMs in more established areas such as chemistry and material science [11, 13].
- As in more quantitative metrology in engineering science and technology, there are pragmatic reasons – such as simplicity and robustness – for choosing a task attribute as a metrological reference [12] rather than an attribute associated with the instrument (person, in the present case) which is both more complex and less robust.

This paper brings these key points together and describes research emerging from the EMPIR NeuroMET 15HLT04 and NeuroMET2 18HLT09 projects [16] which aim to provide more sensitive and metrologically legitimized measurements of memory ability. In this paper, we present how the NeuroMET Memory Metric (NMM) is composed, comprising carefully selected legacy short-term memory test items, links language- and cultural-free items (blocks, digits) to more complex word recalling items. In addition, we are extending the significant role CSE play when establishing a metrological traceability pyramid and reference measurement procedures (RMP) (section 4).

2. Material and methods

The NeuroMET projects have brought together clinicians, academics, metrologists and industry to address measurement challenges associated with the Alzheimer's spectrum. The overall goal is to contribute to building the infrastructure required to translate research into clinical (or pharmaceutical) settings, thereby overcoming specific metrological barriers in diagnosis and treatment.

A cohort has been recruited and tested with neuropsychological assessments with a battery of legacy cognitive tests, clinical laboratory data for protein biomarkers and ultra-high field magnetic resonance imaging and spectroscopy. A cohort comprising 30 persons from each group (Healthy controls (HC), Subjective cognitive decline (SCD), Mild Cognitive Impairment (MCI); and suspected dementia due Alzheimer's Disease) was targeted and recruited at Charité hospital, with assessments at baseline and follow-up visits at 12 months, 36 months and 48 months during 2016–2022. Inclusion criteria were 55–90 years of age, normal vision with or without aid and ability to consent. Exclusion criteria were stroke, Morbus Parkinson, untreated or severe depressive episodes, newly initiated therapy with Acetylcholinesterase (AChE) inhibitors/memantine, pregnancy, other neurological disorders, history of drug or alcohol abuse, and non-suitability for MRI (e.g., persons with claustrophobia, active implants or ferromagnetic implants such as pacemakers). During the project time, several participants have dropped out (e.g., due to too severe AD, not willing to continue or death) and recruitment has been delayed due to the COVID pandemic. At the same time the NeuroMET cohort has been complemented with participants from the SmartAge study [17], also recruited at Charité hospital.

Most of the legacy tests included in the NMM were performed on the first day of each assessment (Corsi Block Test (CBT), Digit Span Test (DST) Word Learning List (WLL) from the CERAD test battery and Mini Mental State Examination (MMSE)), while one test was made on the second day (Rey's Auditory Verbal Learning Test (RAVLT)). From the CBT and DST, only forward sequences were included, i.e., items $n = 14$ and $n = 12$, respectively. From RAVLT (Versions A, C and D) and WLL only the first trial of each was included, i.e., items $n = 15$ and $n = 10$, respectively. Hence, all items aimed to measure specifically short-term memory. In line with that, only the memory items (immediate recall $n = 3$ and delayed recall $n = 3$) from MMSE were included. The complete item-bank for the NMM comprised of 87 items.

In this work, a total of 360 individual assessments were included of HC ($n = 92$), persons with SCD ($n = 160$) and patients with MCI ($n = 50$) and AD ($n = 58$). From the SCD group, 88 participants stem from the SmartAge study. The assessments were almost equally distributed between men ($n = 182$) and women ($n = 178$) and ages ranged from 55 to 87 years.

3. Calculation

To compose the novel NMM there are two main steps.

- Step A, *Compensation of the ordinality in raw scores and a separation of person and item attributes*, which is a prerequisite to proceed with
- Step B, *Formulation of construct specification equations (CSE) for item task difficulty*.

The steps have been made iteratively during the formulation of the NMM.

3.1. A. Compensation for the ordinality in raw scores and a separation of person (θ) and item (δ) attributes

Person responses, $P_{success}$, to the memory items of any of the legacy tests used here are dichotomous data, that is, the probability of correct classification as either pass (classification number = 1) or fail (classification number = 0). Such response scores are ordinal, which implies a non-linearity (particularly so-called counted fraction effects well known on such percentage scales). In turn, this will invalidate even the most basic statistical operations on the raw data, such as summations, calculations of means and standard deviations. RMT provides an appropriate treatment of ordinal data, however it can be noted that there are still many publications in the literature which do not apply RMT, thus leading to unnecessarily large uncertainties.

Rasch analysis not only compensates for the ordinality, but also provides for a separation of task δ and person θ attributes. We applied a dichotomous Rasch-model using RUMM2030 and analyses were repeated for the individual tests, combination of tests and for the full NMM. Our analyses focus on targeting and reliability, as well as conventional tests of model validity in terms of goodness of fit.

Basic assumptions of the Rasch model, such as unidimensionality and person-item separability, are tested using classic tools such as construct alleys and principal component analysis (PCA) including loading plots of logistic fit residuals.

As part of the overall work of validating the composite NMM, a potential breakdown in specific objectivity of the Rasch model due to multidimensionality has been investigated, as might be introduced by serial position effects (SPE) (primacy and recency) in for example the word-list test RAVLT. The identification of factors common to two types of PCA – for logistic fit residuals and for CSE formation (see section 3.2) – together with our entropy-based theory can provide explanations for instance for the additional terms for response bias (b) and person discrimination (ρ) as included in the extended three parameter item response theory (3 PL IRT) model such as $P_{success} = b + (1-b) \bullet \frac{e^{\rho \bullet (\theta-\delta)}}{1+e^{\rho \bullet (\theta-\delta)}}$ [18,19]. At the same time, measurement uncertainties to date are large, making these potential breakdowns in the Rasch model as yet not significant enough to limit the validity of composing the NMM as described here. This work has also raised questions about the validity of claims in the recent literature (review of Weitzner & Calamia [20]) that SPE can offer better diagnostics of cognitive degeneration than legacy tests. Details of these analyses can be found in our recent publications and presentations [18,19,21].

3.2. Formulation of construct specification equations (CSE) for item task difficulty

We make a pragmatic choice to focus, in the first instance, on establishing metrological references for task difficulty (δ), since a task is generally conceptually simpler and more robust than the corresponding person (instrument) abilities (θ), as discussed further in section 4.2. Estimates of item task difficulty, derived from measurand restitution with the Rasch analyses, were then explained in terms of a CSE and our best understanding of the construct. CSE formulation is a process in which the quantity Z of the construct (in this case task difficulty) is expressed as a sum of a number of covariates, X_k (explanatory variables) in the associative relation: $Z = \sum_k \beta_k \bullet X_k$ [22,23].

State-of-the-art multivariate formulation of CSE which determines the coefficients β is based on principal component regression (PCR) [11]. First, a PCA for the set of explanatory variables, X_k is conducted. This step investigates the degree of correlation between the X_k and establishes an orthogonal set of principal components of variation as eigenvectors, P, of the covariance matrix $cov(X)$. (This PCA is distinct from the PCA of the item fit residuals in RMT, but the two PCAs are related and can provide complementary information when investigating construct validity and potential multidimensionality – see section 3.1.) Thereafter, a linear regression of the RMT-analysed memory task difficulty values δ_j from experiment is made against $X' = X \bullet P$ in terms of the principal components, P. This regression is general unweighted since we assume homoscedasticity, informed by the PCA, dimensionality and uncertainty analyses. In a final step, a conversion back from principal components to the explanatory variables, X_k, is made in order to express the construct specification equation for the item attribute in terms of the covariates. Full details are available in our publications [11,13,21].

Our previous work has shown that informational entropy – a measure of order – turns out to be a dominant explanatory variable for task difficulty for several of the most elementary memory tests with language- and cultural-free items (such as CBT and DST included here) [11,13,21] as well as the word-learning list tests (such as RAVLT and WLL CERAD included here) [18,19]. The basic idea is that more ordered sequences – with lower entropy – are easier to recall, and vice versa. Our work builds on the classic Brillouin expression [24] for the entropy of $ln(G!)$, for G distinguishable symbols, which we equate to the difficulty δ of the task, j:

$$\delta_j = M \bullet \left[ln\left(G_j!\right) - \sum_{a_j} ln\left(N_{a_j}!\right) \right] \quad (1)$$

where N is the number of repeated symbols in the recall sequence of any of the legacy memory tests, and where M is a normalisation coefficient. Application of this expression to memory tests is described in detail elsewhere [11,13,19,21].

4. Results and discussion

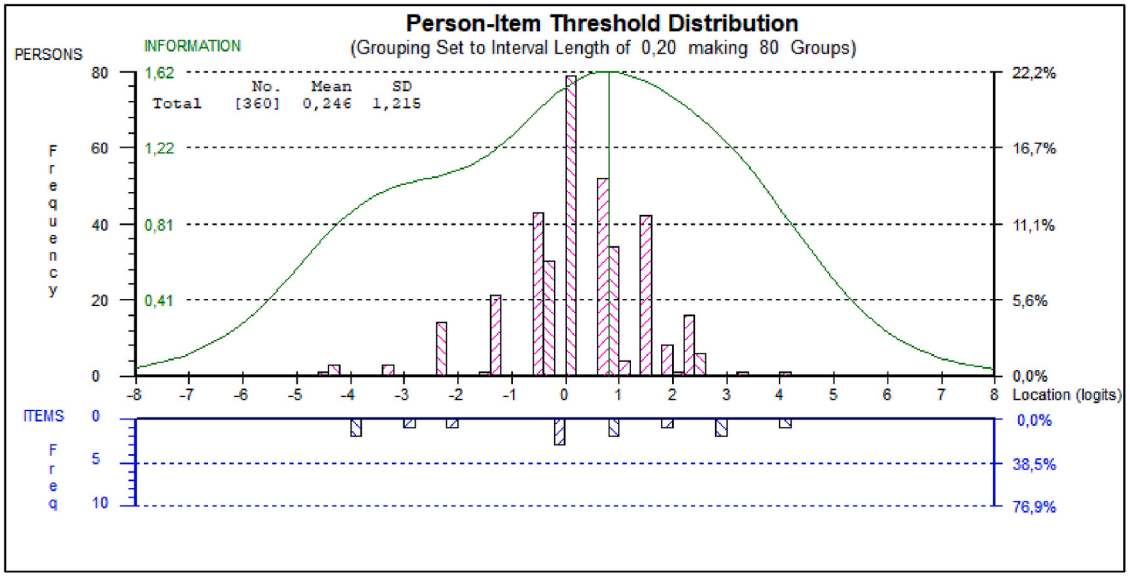
This combined result and discussion section is divided into three parts. First, we will report and discuss on the process for building up to the NMM. This will be followed by presenting the CSE for the legacy tests included in the NMM and a discussion of those CSE in terms of similarities and differences. Finally, in our third part, we discuss the role and added value of CSE in terms of validity and how they can be regarded as prototype CRM, as in chemical and materials metrology.

4.1. Filling the (memory) gaps for a well-targeted scale

With a cohort clinically rated as ranging from HC to AD, we expected a comprehensive range spanning higher to lower memory abilities. In order to maximise reliability and minimise measurement uncertainties, our initial analyses focused on ensuring satisfactory person-to-item targeting. Any skewness and gaps in the distributions of either task item difficulty or person ability will increase measurement uncertainties; we thus took several steps towards forming the NMM, to 'fill in the gaps' on the new scale and build a better 'story of memory measurement'.

We started by emulating previous research [22,23] by analyzing our data on the CBT in which the test person is asked to recall tapping sequences of blocks of increasing levels of difficulty. A first analysis showed that CBT person abilities were reasonably well matched to the item difficulties. There were, however, substantial gaps between the items (Fig. 1a), which in turn, led to high measurement uncertainties. Therefore, in our next step to improve the reliability, we added items on the new memory scale from the DST to the existing CBT items. In the DST, participants are asked to recall a series of three-to-eight-digit sequences of increasing level of difficulty. Although reliability was partially improved with the addition of DST items, it remained too low and measurement uncertainties remained too large. This led us to add further additional items on the new scale from a couple of word list tests: immediate recall trial 1 of the German version of the RAVLT and WLL

a



b

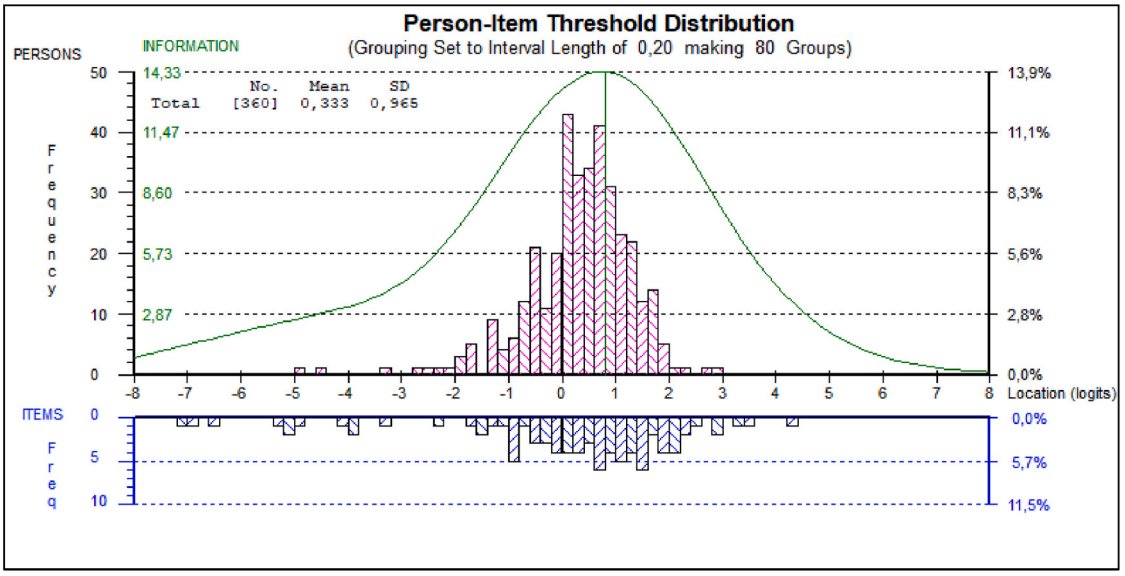


Fig. 1. a–b RMT histograms with person ability, θ, distributions above and item task difficulty, δ, distributions on a conjoint logit scale with only CBT (1a) and the full NMM (1b). From left to right: least able persons to most able persons in pink upper bars and easiest memory tasks most difficult memory tasks in lower blue bars. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

from the CERAD test battery (Figs. 1b and 2). In those tests, persons are asked to recall as many as possible of the 15 and 10 words read, respectively, by the test leader in each test. By adding these word-list items to the new scale, a person reliability of 0.83 (corresponding to a separation ratio index of 2) was finally achieved. This level of reliability is traditionally regarded as indicating a satisfactory measurement uncertainty for high-stakes testing, at which level the uncertainty is not larger than the attribute standard deviation [25], thus ensuring that the associated risks of incorrect decisions of conformity should be acceptably small [26]. The improved targeting leads, thanks to RMT, in turn to reduced measurement uncertainties in the person ability estimates. Fig. 3 shows how the measurement uncertainties, expressed as 2 standard errors (2SE), in the person abilities decrease substantially when progressing from a single legacy test (CBT) to the full NMM.

Obtaining better reliability of measures of person ability θ is of course the main focus of cognitive measurements in detecting cognitive decline. The results shown in Fig. 3a demonstrate the reduction achieved in measurement uncertainties in the person ability measures by a factor of about 5 obtained with the new composite metric NMM compared to CBT. When comparing the NMM with RAVLT A-list (Fig. 3b), a reduction in measurement uncertainties in the person ability measures is also seen, but not as significant as for CBT. The reduced measurement uncertainties are thanks to the careful selection from the considerably larger total number of items in the original legacy tests, thus leading to a substantial improvement in testing efficiency with the new metric.

The final NMM was completed by adding a further set of six recalling items from MMSE to the new scale. This did not affect the targeting or reliability noticeably but was done for purely practical reasons as MMSE is the most widely used cognitive test. By including these MMSE items and using them as anchors, we can then compare results of the MMSE with the NMM using cross-walks [27]. This procedure also, of course, applies to the other legacy tests used in the NMM. A conversion table which allows to link these legacy memory tests to a metrologically

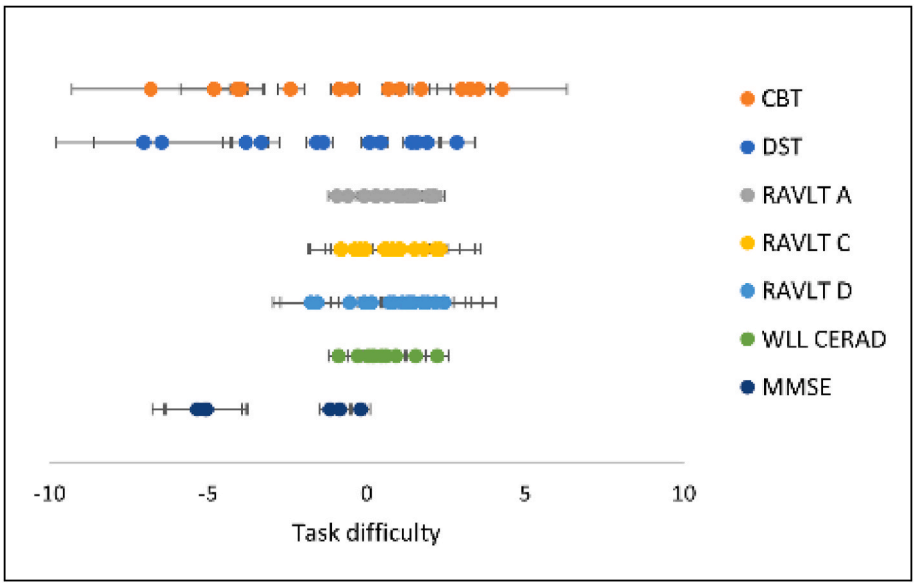


Fig. 2. Each dot corresponds to each item's task difficulty location (δ, x-axis) for the including items from the legacy tests. Easiest items are located to the left and the most difficult to the right. Measurement uncertainties with coverage factor $k = 2$.

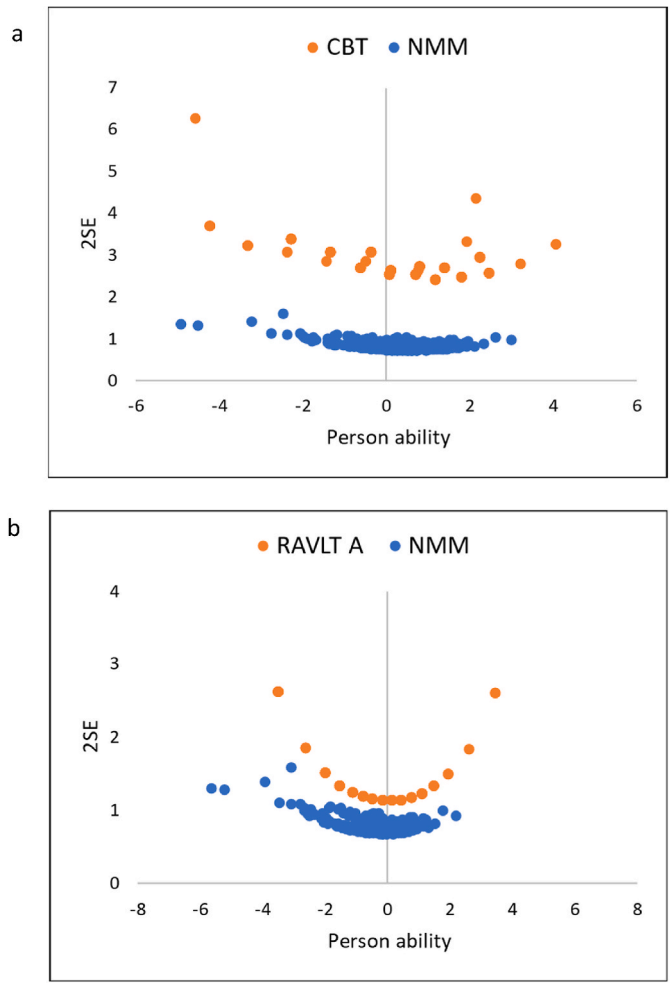


Fig. 3. a–b Measurement uncertainties, 2SE, on the y-axis and person ability measures (θ) on the x-axis compared for measures based on only CBT and the full NMM (a) and RAVLT A-list and the full NMM (b).

improved scale viz., the NeuroMET memory metric, thereby addressing metrological traceability, will be provided in forthcoming work.

4.2. CSE to explain task difficulty (δ)

In addition to the separation of item (δ) and person (θ) attributes onto a linear scale obtained with RMT, the metrological legitimacy of the NMM is further ensured by exploiting CSEs as 'recipes for certified reference material'. Table 1 shows CSE for task difficulty for each of the individual legacy tests the CBT, DST, RAVLT and the WLL from the CERAD test battery.

Composite measures have been formed in cognitive studies previously. Combining items to form a new metric – and better short-term memory tests – as done in the NMM – needs a careful selection of items so as not to jeopardize the validity, to reliably predict task difficulty and make a more efficient test [21,28,29]. This careful selection of items is guided in our work by the following principles:

- Conceptual: all items are designed and purport to measure short-term memory, and
- Explanatory: the equivalence of different items is indicated by equal entropy

Table 1
CSE and Pearson correlation coefficients for empirical task difficulty values (δ) and predicted (zR) task difficulty values. Measurement uncertainties with coverage factor $k = 2$ are given in brackets.

Memory test	CSE	Pearson correlation coefficient
CBT	$zR_{j,CBT} = -6(3) + 1.2(6) \times Entropy_j - 0.3(1.1) \times Reversals_j - 0.02(0.25) \times AveDistance_j$	0.97
DST	$zR_{j,DST} = -6(3) + 1.0(2) \times Entropy_j + 0.01(36) \times Reversals_j - 0.2(1.4) \times AveDistance_j$	0.96
RAVLT A list	$zR_{j,RAVLT\_A} = 5(3) + 0.7(5) \times Primacy_j + 0.8(5) \times Recency_j + 0.25(20) \times Frequency_j$	0.81
RAVLT C list	$zR_{j,RAVLT\_C} = 3(5) + 0.5(2) \times Primacy_j + 0.6(1) \times Recency_j + 0.44(97) \times Frequency_j$	0.70
RAVLT D list	$zR_{j,RAVLT\_D} = 10(4) + 1.1(4) \times Primacy_j + 1.2(3) \times Recency_j - 0.2(5) \times Frequency_j$	0.92
WLL CERAD	$zR_{j,WLL\_CERAD} = 5(5) + 0.7(5) \times Primacy_j + 1(1) \times Recency_j - 0.13(8) \times Frequency_j$	0.71

An observation from Table 1 is the striking similarity of the CSEs for the different legacy tests studied: irrespective of whether sequences of blocks (CBT), digits (DST) or words (RAVLT and WLL CERAD) are being recalled, the dominant variable explaining task difficulty is entropy as calculated for all tests with the same formula (eq. (1)). For the conceptually simplest recall tests (CBT and DST), a sequence of G blocks or digits has the same basic entropy expression: $entropy = -M \bullet ln(G_j!)$. For the verbal tests ($L$ words) RAVLT (A, C or D lists) and WLL CERAD, entropy can even explain the so-called serial position effects (SPE) in the first trial, that is, that it is easier to recall words from the start and end of a list, respectively, for item j:

$$Primacy_j = -M \bullet ln\left(G_j!\right);\ G = item\ order$$

$$Recency_j = -M \bullet ln\left(G_j!\right);\ G = L - 1 - item\ order$$

The first, so-called "intercept" term on the RHS of has been interpreted in terms of a discrimination factor in our work on RAVLT [18,19]. The various entropy-based contributions to task difficulty occurring in the CSEs listed in Table 1 are more prominent than the other explanatory variables – such $Reversals_j$: the number of sequence reversals (e.g., 1:3:2); $AveDistance_j$: the average distance between objects (e.g., spatially between blocks or numerically between digits) in item sequence, j; $Frequency_j = -M \bullet ln\, f_j$ for a word frequency (from e.g., a thesaurus).

4.3. Certified reference materials and reference measurement procedures for memory measurements

Others have argued that the CSE represent the highest level of construct theory [30]. In addition to the depth and specificity of the measurand given by a CSE, from the metrological point of view, we have proposed CSE for task difficulty as constituting metrological references analogous to 'recipes' for CRM in chemical and materials metrology which enable instrument calibrations for viz., reliable and traceable measurements. Analogously, in memory measurements, CRM recipes for metrological references (viz., measures of memory task difficulty) can be formulated in terms of causality as a CSE. Such CSE provide objective and scalable metrological units for traceability also for memory measurements.

In early papers, Rasch [31] argued that RMT was agnostic concerning the choice between regarding a task or a person as an instrument. As mentioned in the introduction, we however refrain from considering test items as "instruments" (as often done in the social sciences) [9], but rather prefer to argue in favor of a proper picture of the measurement system where the person is the instrument, which allows full advantage to be taken of MSA procedures well-established for decades in engineering science and technology [4]. Measurement objects are typically robust and simple, and therefore a natural choice for measurement references [11]. In much the same way as in weighing, where a mass standard is a primary choice of metrological standard in preference to a relatively sensitive and complex weighing instrument, in contrast to the robustness and simplicity of memory items, persons are more complex and more sensitive to environment, context, and method [32,33].

Amongst the advantages of using CSEs to define metrological references instead of specifying individual item difficulties is that new items can be designed to (i) have a certain task difficulty on the "fly" and (ii) crosswalk for instance item difficulty references across different cultural realisations. For instance, in the current NeuroMET project, CSEs for the RAVLT are being compared and contrasted across different languages (German, English, Swedish, …). Identification of dominant explanatory variables – particularly entropy – enables grounds for criteria for item equivalence.

To the pragmatic advantages of choosing task difficulty instead of person ability as a first choice of metrological reference is the ease with which particular 'linking' items can be included when transferring traceability from one test realisation to another. Transferability of traceability – within quoted uncertainties – is ensured by the conceptual and explanatory criteria mentioned above. Attempting to ensure comparability by including instead a reference test person as a standard for ability in different test realisations would obviously be more challenging.

Of course, although not the first choice, a calibrated measurement instrument can in some cases function as a metrological reference. For memory measurement, this would involve formulation of a CSE for person memory ability, Θ_j. Indeed, such a CSE for person ability – expressed causally for instance as a function of explanatory variables such as disease biomarkers – is of course the final aim of cognitive measurements in detecting cognitive decline [34]. Developing CRMs for person ability before a valid CSE for task difficulty has been acceptably demonstrated should, however, be done with care [28].

To be qualified as a "certified" reference material, the CSE presented here must of course be subject to requirements analogous to those stipulated for CRM and RMP in analytical chemistry and materials sciences. Currently we are doing a gap-analysis based on ISO 15193:2009 including what requirements are directly applicable (and potential also not applicable) to memory measurements and what requirements needs modifications in memory measurements. As a part of this work, based on ISO 17511:2003 we propose a traceability pyramid for memory measurements (Fig. 4) inspired from other cases where neither reference measurement procedure nor reference materials for calibration are, as yet, available.

The work reported in this paper corresponds to the "CRM manufacturer's" measurement procedure in Fig. 4, yielding the CSE for task difficulty to serve as calibrators to be included in measurement procedures in clinic applied to their samples. Thus, despite different clinical samples – anywhere around the world – their measurement results for person abilities should be traceable to the same metrological references, and in turn comparable. To the left there is an arrow pointing upwards, this implies that the metrological traceability is increasing higher up the pyramid, while on the left you see how measurement uncertainties decrease higher up the pyramid. A dotted line has been included where measurement uncertainties are stable, which might be the case for memory measurements as the procedures may not vary between our study cohort and clinical practice, but this needs further evaluation.

When there is, as in the present case, as yet no SI etalon at the top of the pyramid and neither reference measurement procedure nor reference materials for calibration are available, it is always a question what comes first, the procedure or the material? At this stage we have tentatively placed the reference procedure at the top of the pyramid, and we include persons and task here. To remedy the lack of SI etalon, we propose to start with the procedure based in the present case on the CSE for task difficulty, developed for a specific cohort and set of legacy tests as part of the procedure of developing a calibrator. In line with other metrological reference procedures, our reference procedure can be considered as defined as a set of operations, described specifically used in the performance of particular measurement according to a given methods, thus warranting an extension of the traceability pyramid to be fit-for purpose for memory measurements (and other kinds of PCOM).

The choice of task difficulty (attributed to the measurement object) rather than person ability (attributed to the human responders as measurement instruments) as a metrological reference for cognition is made mainly on practical considerations, in much the same way as one usually chooses a simple and simple mass object as a standard as opposed to the delicate and complex weighing instrument in a traditional weighing system. Each legacy test can be seen as specifying the recommended way of combining a set of symbols and sequences to create a task of a specific level of difficulty. The 'materials' themselves – the symbols and sequences – are available of course to anyone wanting to create a metrological reference, to be composed according to the reference procedure in much the same way as a reference material in Chemistry or Materials Science is to be mixed together of locally available materials according to a defined recipe.

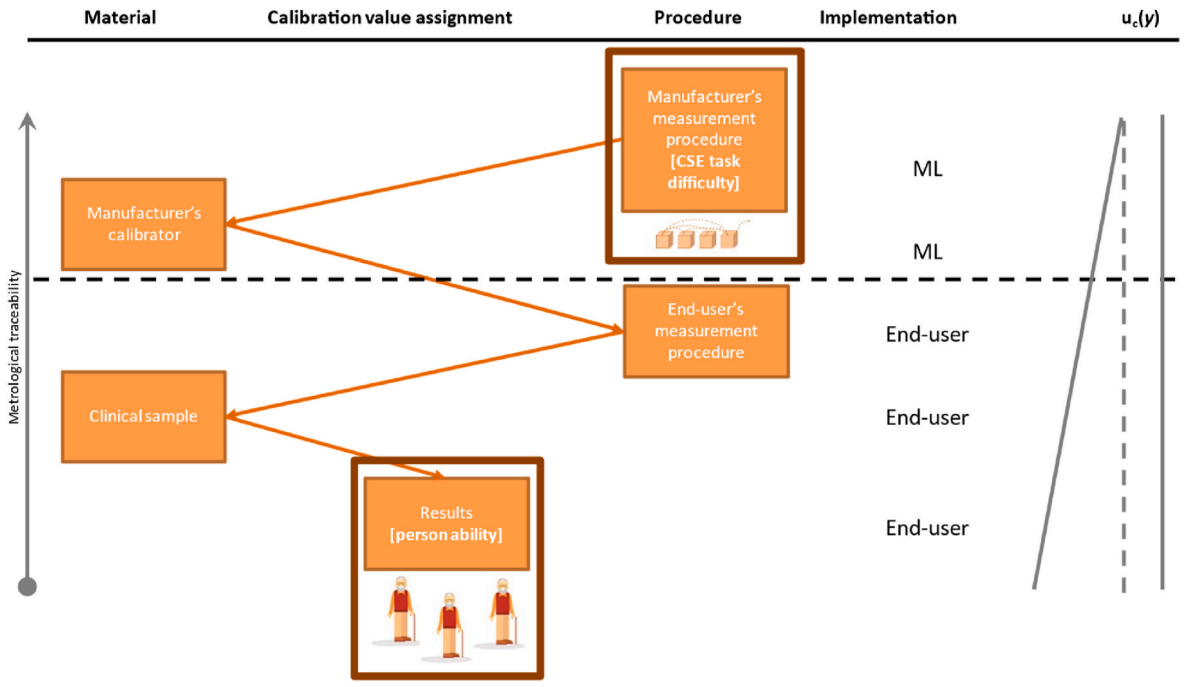


Fig. 4. A proposed traceability pyramid for memory measurements based on ISO 17511:2003.

The initial study here – based on the NeuroMET cohort – requires of course extension to include other test persons in order to make the reference procedure as representative as possible. Potentially, in the future, data from cohorts across the world might be combined into a reference sample and the production of a global CSE in much the same way as traditional references are tested in inter-laboratory comparisons. A first step in this direction has recently been taken where the reproducibility of the task difficulty CSE between the German and Swedish versions of RAVLT has demonstrated satisfactory comparability between the cohorts in Berlin and Gothenburg [35]. However, while inter-laboratory studies are well-established as a means of evaluating measurement accuracy and ensuring metrological traceability, for instance in chemical and materials metrology, in the human and social sciences such routines are less developed.

5. Conclusions

A careful selection – based on entropy equivalence – of items from legacy short-term memory tests has been demonstrated to provide more sensitive and metrologically legitimized measurements of memory ability. The NMM presented here shows up to a five-fold reduction in measurement uncertainties for memory ability compared to individual legacy tests used in NMM without jeopardizing validity.

The NMM will soon be included in an app used to deliver memory tests, where the metrologically validated NMM presented here will play a major role in making a validated app. With the app, clinicians and researchers will be able to select either sets of items or the full NMM. Patient responses will be transformed via a scoring algorithm into memory ability measures in a common frame of reference. The app will be GDPR compliant, where access to the NMM is deployed in a structured and controlled manner to ensure data and scoring integrity. In forthcoming research, we will extend the metrological validity of the NMM to also include external validity and a kind of inter-laboratory study for memory tests. Future work will aim to qualify the CSE, used here to compose the NMM, as certified reference materials.

In this paper, we have summarized a metrological approach to ensure validity for the NMM, but also a methodology to advance measurement quality in the human sciences. In line with the more quantitative metrology in engineering science and technology, the simplicity and robustness of objects call for choosing a task attribute as a metrological reference rather than an attribute associated with the instrument – typically a person in the human sciences – which is both more complex and less robust. Thus, CSE for task difficulty appear to provide metrological references for object calibrations and subsequent inter-comparability of person abilities.

The NMM will have the potential to provide more fit-for-purpose measurements, which could be used for better clinical assessments, early detection of cognitive decline, diagnosis and monitoring disease progress and evaluation of treatment effects.

CRediT authorship contribution statement

J. Melin: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. **S.J. Cano:** Methodology, Writing – review & editing. **A. Flöel:** Data curation, Writing – review & editing. **L. Göschel:** Data curation, Writing – review & editing. **L.R. Pendrill:** Conceptualization, Formal analysis, Methodology, and, theory development, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This project 18HLT09 NeuroMET2 has received funding from the EMPIR programme co-financed by the Participating States and from the European Union's Horizon 2020 research and innovation programme.

References

[1] European Brain Council, Driving Policy to Optimise Care for People with Alzheimer's Disease in Europe Today and Tomorrow [Internet], 2018. Available from: https://www.braincouncil.eu/wp-content/uploads/2018/11/Driving-policy-to-optimise-care-WEB.pdf.

[2] J. Hobart, Putting the Alzheimer's Cognitive Test to the Test I: Traditional Psychometric Methods, vol. 6, 2013.

[3] J. Hobart, S. Cano, H. Posner, O. Selnes, Y. Stern, R. Thomas, et al., Putting the Alzheimer's cognitive test to the test II: Rasch measurement theory, Alzheimers Dement 9 (1S) (2013 Feb) S10–S20.

[4] V. Turetsky, E. Bashkansky, Ordinal response variation of the polytomous Rasch model, METRON 80 (2022) 305–330. https://doi.org/10.1007/s40300-022-00229-w.

[5] L. Pendrill, Man as a measurement instrument, NCSLI Meas 9 (4) (2014 Dec) 24–35.

[6] S.J. Cano, T. Vosk, L.R. Pendrill, A.J. Stenner, On trial: the compatibility of measurement in the, J. Phys. 7 (2016).

[7] S.J. Cano, L.R. Pendrill, S.P. Barbic, W.P. Fisher, Patient-centred outcome metrology for healthcare decision-making, J. Phys. Conf. Ser. 1044 (2018 Jun), 012057.

[8] L.F. Hughes, K. Perkins, B.D. Wright, H. Westrick, Using a Rasch scale to characterize the clinical features of patients with a clinical diagnosis of uncertain, probable, or possible Alzheimer disease at intake, J. Alzheimers Dis. 5 (5) (2003 Nov 17) 367–373.

[9] J.C. Hobart, S.J. Cano, A.J. Thompson, Effect sizes can be misleading: is it time to change the way we measure change? J. Neurol. Neurosurg. Psychiatry 81 (9) (2010 Sep 1) 1044–1048.

[10] L. Mari, M. Wilson, An introduction to the Rasch measurement approach for metrologists, Measurement 51 (2014 May) 315–327.

[11] L. Pendrill, Quality Assured Measurement: Unification across Social and Physical Sciences [Internet], Springer International Publishing, 2019 [cited 2021 Feb 21]. (Springer Series in Measurement Science and Technology). Available from: https://www.springer.com/gp/book/9783030286941.

[12] S.J. Cano, L.R. Pendrill, J. Melin, W.P. Fisher, Towards consensus measurement standards for patient-centered outcomes, Measurement 141 (2019 Jul) 62–69.

[13] J. Melin, L.R. Pendrill, S.J. Cano, EMPIR NeuroMET 15HLT04 consortium. Towards patient-centred cognition metrics, J. Phys. Conf. Ser. 1379 (2019 Nov), 012029.

[14] L.R. Pendrill, Assuring measurement quality in person-centred healthcare, Meas. Sci. Technol. 29 (3) (2018 Mar 1), 034003.

[15] G. Rasch, Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests, Nielsen & Lydiche, Oxford, England, 1960, p. 184, xiii (Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests).

[16] M. Quaglia, S. Cano, A. Fillmer, A. Flöel, C. Giangrande, L. Göschel, et al., The NeuroMET project: metrology and innovation for early diagnosis and accurate stratification of patients with neurodegenerative diseases, Alzheimers Dement 17 (S5) (2021), e053655.

[17] M. Wirth, C. Schwarz, G. Benson, N. Horn, R. Buchert, C. Lange, et al., Effects of spermidine supplementation on cognition and biomarkers in older adults with subjective cognitive decline (SmartAge)-study protocol for a randomized controlled trial, Alzheimer's Res. Ther. 11 (1) (2019 May 1) 36.

[18] L.R. Pendrill, J. Melin, S.J. Cano, Entropy-based Explanations of Multidimensionality in Ordinal Responses, 2021.

[19] J. Melin, A. Regnault, S. Cano, L. Pendrill, Neuropsychological Assessments: Word Learning Tests and Diagnostic Potential of Serial Position Effects, France, Lyon, 2021.

[20] D.S. Weitzner, M. Calamia, Serial position effects on list learning tasks in mild cognitive impairment and Alzheimer's disease, Neuropsychology 34 (4) (2020 May) 467–478.

[21] J. Melin, S. Cano, L. Pendrill, The role of entropy in construct specification equations (CSE) to improve the validity of memory tests, Entropy 23 (2) (2021 Feb) 212.

[22] A.J. Stenner, M. Smith, D.S. Burdick, Toward a theory of construct definition, J. Educ. Meas. 20 (4) (1983) 305–316.

[23] A.J. Stenner, M. Smith, Testing construct theories, Percept. Mot. Skills 55 (2) (1982 Oct) 415–426.

[24] L. Brillouin, Science and Information Theory [Internet], second ed., 1962. New York, https://www.amazon.com/Science-Information-Theory-Second-Physics/dp/0486497550 [cited 2021 Feb 24]. Available from.

[25] B.D. Wright, Reliability and separation, Rasch. Meas. Trans. 9 (4) (1996) 472.

[26] L.R. Pendrill, W.P. Fisher, Counting and quantification: comparing psychometric and metrological perspectives on visual perceptions of number, Measurement 71 (2015 Jul) 46–55.

[27] T. Salzberger, S. Cano, L. Abeth-Webb, E. Afolalu, C. Chrea, R. Weitkunat, et al., Addressing Traceability of Self-Reported Dependence Measurement through the Use of Crosswalks, 2021 May, 109593. Measurement.

[28] J. Melin, S.J. Cano, A. Flöel, L. Göschel, L.R. Pendrill, Construct specification equations: 'Recipes' for certified reference materials in cognitive measurement, Meas Sens. 18 (2021 Dec), 100290.

[29] J. Melin, S. Cano, A. Flöel, L. Göschel, L. Pendrill, E.M.P.I.R. NeuroMET, NeuroMET2 consortiums, More than a memory test: a new metric linking blocks, numbers, and words, Alzheimers Dement 17 (S6) (2021), e050291.

[30] A.J. Stenner, H. Burdick, E.E. Sanford, D.S. Burdick, How accurate are lexile text measures? J. Appl. Meas. 7 (3) (2006) 307–322.

[31] Rasch G. On Objectivity and Models for Measuring. :11.

[32] L. Pendrill, Quantities and units in quality assured measurement [Internet], in: Pacific rim objective measurement symposium 2021, 2021 Dec 6. Available from: https://proms.promsociety.org/2021/.

[33] Melin J. Neurogenerative disease metrology and innovation: the European metrology programme for innovation & research (EMPIR) and the NeuroMET projects [internet]. Conference Presentation Presented at: PACIFIC RIM OBJECTIVE MEASUREMENT SYMPOSIUM 2021; 2021 Dec 6. Available from: https://proms.promsociety.org/2021/.

[34] J. Melin, S.J. Cano, L. Göschel, A. Fillmer, S. Lehmann, C. Hirtz, et al., Metrological references for person ability in memory tests, Meas Sens. 18 (2021 Dec 1), 100289.

[35] J. Melin, P. Kettunen, A. Wallin, L. Pendrill, Entropy-based Explanations of Serial Position and Learning Effects in Ordinal Responses to Word List Tests, 2022.