entropy

MDPI

Article

The Role of Entropy in Construct Specification Equations (CSE) to Improve the Validity of Memory Tests: Extension to Word Lists

Jeanette Melin [1], Stefan Cano [2], Agnes Flöel [3,4], Laura Göschel [5,6] and Leslie Pendrill [1,*]

[1] Department of Measurement Science and Technology, Research Institutes of Sweden (RISE), AWL Sven Hultins Plats 5, vån 4, 412 58 Göteborg, Sweden; jeanette.melin@ri.se
[2] Modus Outcomes, Spirella Building, Letchworth Garden City SG6 4ET, UK; stefan.cano@threadresearch.com
[3] Department of Neurology, University Medicine Greifswald, 17475 Greifswald, Germany; agnes.floeel@med.uni-greifswald.de
[4] German Center for Neurodegenerative Diseases (DZNE), Standort Rostock/Greifswald, Germany
[5] Department of Neurology, Charité—Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, 10117 Berlin, Germany; laura.goeschel@charite.de
[6] NeuroCure Clinical Research Center, Charité—Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, 10117 Berlin, Germany
* Correspondence: leslie.pendrill@ri.se

Abstract: Metrological methods for word learning list tests can be developed with an information theoretical approach extending earlier simple syntax studies. A classic Brillouin entropy expression is applied to the analysis of the Rey's Auditory Verbal Learning Test RAVLT (immediate recall), where more ordered tasks—with less entropy—are easier to perform. The findings from three case studies are described, including 225 assessments of the NeuroMET2 cohort of persons spanning a cognitive spectrum from healthy older adults to patients with dementia. In the first study, ordinality in the raw scores is compensated for, and item and person attributes are separated with the Rasch model. In the second, the RAVLT IR task difficulty, including serial position effects (SPE), particularly Primacy and Recency, is adequately explained (Pearson's correlation $R = 0.80$) with construct specification equations (CSE). The third study suggests multidimensionality is introduced by SPE, as revealed through goodness-of-fit statistics of the Rasch analyses. Loading factors common to two kinds of principal component analyses (PCA) for CSE formulation and goodness-of-fit logistic regressions are identified. More consistent ways of defining and analysing memory task difficulties, including SPE, can maintain the unique metrological properties of the Rasch model and improve the estimates and understanding of a person's memory abilities on the path towards better-targeted and more fit-for-purpose diagnostics.

Keywords: entropy; information; metrology; measurement system analysis; Rasch; cognition; memory; task difficulty; person ability; neurodegenerative diseases; cognitive neuroscience

Citation: Melin, J.; Cano, S.; Flöel, A.; Göschel, L.; Pendrill, L. The Role of Entropy in Construct Specification Equations (CSE) to Improve the Validity of Memory Tests: Extension to Word Lists. *Entropy* **2022**, *24*, 934. https://doi.org/10.3390/e24070934

Academic Editor: Pentti Nieminen

Received: 13 May 2022
Accepted: 28 June 2022
Published: 5 July 2022

1. Introduction

When searching for novel diagnostics for cognitive decline, serial position effects (SPE) in memory recall tests have recently come to the fore [1]. A SPE refers to the relationship between the placement of a symbol (e.g., word) in a list and its likelihood of being recalled. For free recall in word learning list tests, SPEs mean that the first words (primacy region, Pr) and the last few words of a list (recency region, Rr) are easier to remember than items in the middle region (Mr) [2]. SPEs in word learning tests, such as Pr and Rr, have been studied since the 1950s [1–9].

SPEs were evident in several memory recall studies, as noted in the review of Hurlstone et al. [10]: "*Forward accuracy serial position curves exhibiting effects of primacy and recency are not confined to verbal memoranda. The forward serial position curves associated with the recall of sequences composed of various types of nonverbal stimuli have been shown to exhibit an extensive primacy effect accompanied by a one-item recency effect. These stimuli include visual–spatial locations, visual–spatial movements, auditory–spatial locations, visual matrix patterns, and unfamiliar faces*."

Of particular interest is to metrologically validate and verify a recently proposed novel diagnostic for cognitive impairment based on the observation that less cognitively able individuals cannot benefit from the SPE of Primacy [1], which, for more cognitively able individuals, makes it easier to recall the first few words in a word learning list test such as Rey's Auditory Verbal Learning Test RAVLT.

Weitzner & Calamia [1], in their recent review of the diagnostic potential of SPE, concluded that: "*The analysis of SPEs has demonstrated some utility as a marker of cognitive impairment associated with MCI, AD, and other dementias; however, research is limited by the multiple ways in which SPEs are defined and analysed*."

The overall aim of the present work is to highlight entropy-based explanations when developing metrological methods to word learning list tests, including a better definition and analysis of SPEs, by extending earlier simple syntax studies [11,12]. As recounted in Appendix A, the concept of entropy can be usefully deployed when explaining such effects throughout the measurement process.

In the next section, we introduce the participants and memory test used in this paper. These are applied in three case studies: (I) compensating for the ordinality of raw scores and separating item and person attributes, (II) linking entropy to CSE, and (III) evaluating the multidimensionality introduced by SPE. Each case study on the word recall test RAVLT exemplifies the specific application of a general, entropy-based theory presented in the set of Appendices A–C.

The paper concludes with a consideration of how the present research contributes to developing a more consistent way of defining and analysing SPE when aiming for better-targeted and more fit-for-purpose diagnostics, as well as introducing the metrological quality assurance of ordinal properties more generally.

2. Materials

The present work analyses data from the European EMPIR 19HLT04 NeuroMET2 [13] and SmartAge [14] cohorts, comprising a total of $N_{TP} = 225$ participants, including $n = 66$ healthy controls (HC), $n = 99$ with subjective decline (SCD), $n = 27$ with mild cognitive impairment (MCI), and $n = 33$ with dementia due to suspected Alzheimer's disease (AD). Further details about the cohorts can be found elsewhere [13,14]. There are a number of commonly used neuropsychological tests for assessing different cognitive functions, such as memory, learning, speed, attention, visuospatial, language, and executive functions. Measuring memory abilities based on the recall of word list items has a long tradition in neuropsychological assessments and has recently become a research topic for improved diagnostics for cognitive decline and dementia.

One such "legacy memory test" is the RAVLT [15], where a list of 15 semantically unrelated words is orally presented in an initial trial to the test person, who is required to freely recall the words (*immediate recall (IR)*). Subsequently, four more identical trials are performed, making a total of five IR (or learning trials). After these trials, a distractor list of 15 different and unrelated words is presented, and the test person is asked to recall them. This distraction list is directly followed by asking the test person freely to recall again the words from the first presented list, and finally, after a delay of approximately 20 min, they are asked to freely recall the words once again (*delayed recall (DL)*). In addition, there is a list of 50 words, of which the test person is asked to recognise the 15 initial words (*recognition*).

3. Case Study I: Analyses of Word-List Task Difficulty

This first case study focuses on the difference between traditional analyses with classic test theory (CTT) and modern metrological analyses—Rasch generalised linear modelling (GLM)—of the observed response data from RAVLT IR, particularly when allowing for SPE when estimating both person memory abilities and task difficulties (Appendix A.1).

3.1. Classic Analyses in Word Learning List Tests

An observed response in a word learning list test is typically scored binarily in one of two categories: as $c = 1$ for correct and $c = 0$ for incorrect, and presented as $P_{success}$. As mentioned in Appendix A.1, these observed responses constitute the raw data, $x_{i,j}$, for test person i and task item j, which are characterised by ordinality and are not measures of the person's memory ability or the memory task difficulty.

CTT, which has been the dominating approach in previous analyses of word learning tests reported in the literature [1], simply sums up the number of correct recalled words; therefore, for instance, an average score for task item j is: $CTT_{proportion,j} = \frac{1}{N_{TP}} \cdot \sum_{i=1}^{N_{TP}} x_{i,j}$, where N_{TP} is the number of test persons in the cohort.

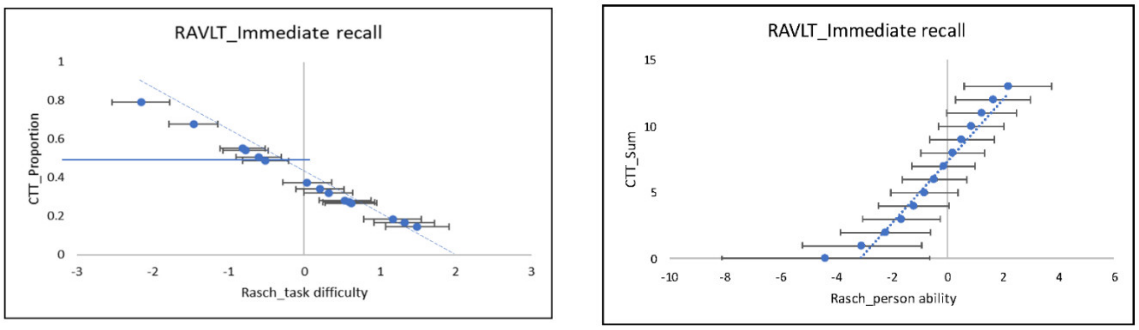
3.2. Modern Analyses in Word Learning List Tests

Compensating for ordinality and providing separate estimates of task difficulty, δ, and person ability, θ, (so-called specific objectivity) is done in modern analyses by making a logistic regression, based on Equation (A2), with the Rasch GLM [16] of the log-odds ratio to the raw binary response data, $P_{success,i,j}$for person i and item j:

$$ln\left(\frac{P_{success,i,j}}{1 - P_{success,i,j}}\right) = (\theta_i - \delta_j) \quad (1)$$

Such a logistic regression, providing a restituted estimate of the object variable Z from the measured response Y (Equation (A1)), can be readily done with, for example, the WINSTEPS ® software application (version 3.80.1 is used here).

Reflecting the nonlinearity and ordinality due to the counted fractions mentioned in Appendix A.1, typically, S- or ogive curves (Figure 1) are expected from Equation (1) when correlating the observed responses $P_{success}$ with the test item difficulty, δ, and person ability, θ. The increasing nonlinearity of the curves as one approaches each end of the scales is apparent in Figure 1, despite the large uncertainties. The straight line plotted in Figure 1a indicates the slope of the Rasch formula (Equation (1)) at its maximum value—that is, at $P_{success} = 50\%$, where $\delta = \bar{\theta} = -0{,}58\ logits$, the average person's ability across the cohort. From Figure 1a, task difficulties at the low end of the scale (the easiest tasks) are visibly underestimated, while, at the high end, the more challenging tasks are overestimated with CTT.



(**a**) (**b**)

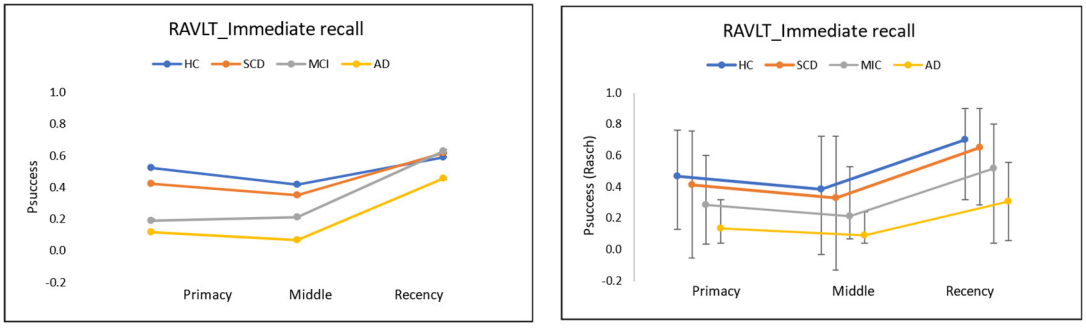
Figure 1. Correlation plots between classic scoring (y-axis; CTT proportion in (**a**) and CTT sum score in (**b**) of our actual data for RAVLT IR and corresponding Rasch restituted values (Equation (1)): (**a**)

task difficulty, δ, (x-axis), where a lower value indicates a lower task difficulty, and (**b**) person ability, θ, (x-axis), where a lower value indicates a higher ability. Uncertainty coverage, $k = 2$.

A well-designed test with good targeting, where the span of a person's abilities matches well (both in terms of centring and width) the corresponding range of task difficulties, will admittedly, in general, have small nonlinearities, as is the case for the present work with RAVLT IR. These effects nevertheless have to be investigated, however large or small they might turn out to be. Note that, in any case, the Rasch model is essential for metrological quality assurance, since it provides, through measurand restitution, separate values for task difficulty and person ability.

The classic diagnostics of SPEs, such as Pr and Rr, are traditionally plotted as curves (Figure 2a) that attempt to visualise the trends in the raw scores for the test items between the three main regions. Typically, in a word learning test with 15 words: $k = 1, \ldots, 15$, such as the RAVLT, Pr is normally found to be significant only for the first four words: $k = 1,\ldots,4$, while Rr is found in the last four words: $k = 12,\ldots,15$ of the recalled list [7]. Traditional calculations using CTT for SPE, as reported by Talamonti et al. [9], consist of simply summing up the number of correct recalls in the Pr and Rr, respectively. For comparison, we applied both methods—CTT and Rasch—to our data set of the NeuroMET cohort.

Figure 2a shows the proportion, $P_{success}$, of immediately recalled (trial 1) words for Pr, Mr, and Rr based on the CTT raw scores, while Figure 2b shows corresponding analyses from Rasch person attributes converted back to $P_{success}$, where the latter includes compensation for the scale nonlinearities shown in Figure 1. Although ordered according to the expected cognitive health status of each clinical subgroup, when considering the actual measurement uncertainties (shown in Figure 2b), there appear to be no detectable differences between the four clinical subgroups (HC, SCD, MCI, and AD). Note that we refrain from plotting uncertainty intervals on the CTT (Figure 2a), since—without an honest assessment of the effects of ordinality (which we show to be significant in Figure 1a)—it is difficult to give uncertainty estimates objectively and credibly. When uncertainty intervals are plotted in published CTT versions such as Figure 2a, those intervals are often shown (e.g., [5]) incorrectly as symmetric about the mean CTT raw score, which is inconsistent with the inherent nonlinearity of the y-axis due to the well-known counted fractions effect, as evident in the asymmetric uncertainty intervals shown in Figure 2b.



(**a**) (**b**)

Figure 2. The proportion, $P_{success}$, of recalled words in RAVLT IR (trial 1) for Pr, Mr, and Rr based on the raw scores (**a**) and Rasch attributes converted back to the $P_{success}$ scale in (**b**). AD = Alzheimer's disease; HC = healthy controls; IR = immediate recall; MCI = mild cognitive impairment; SCD = subjective cognitive decline. Uncertainty coverage, $k = 2$.

As noted above in relation to Figure 1a, because of the counted fractions effect (Appendix A.1), the task difficulties will be increasingly overestimated with CTT (Figure 2a) as one approaches the low end of the scale (the easiest tasks, such as Pr and Rr). In particular, the apparent prominence in traditional CTT plots (Figure 2a) of the SPE in the

Rr, especially for the AD cohort (yellow line), is not maintained when allowance is made for the scale nonlinearity and uncertainties (Figure 2b).

3.3. Potential, Limitations, and Implications with a Corrected Analysis of SPEs in RAVLT

This case study has provided both a theoretical justification and experimental demonstration for a proper analysis of RAVLT with a modern metrological approach, the Rasch GLM [16]. With a correct analysis of the SPEs in RAVLT, we thereby move away from the limitations of traditional CTT and the associated lack of metrological invariance and considerable risks of incorrect conclusions based on the cognitive assessment data, which would otherwise impact both clinical decision-making and the assessment of the significance of correlations between cognitive functioning and brain structure and biochemistry [17].

In fact, our present analysis indicates, by considering the actual measurement reliability, that there is hardly any detectable difference between the four clinical subgroups, and the diagnostic power of the SPEs is rather weak, owing to the relatively large measurement uncertainties, even when a proper analysis of the raw scores has been made. Therefore, our main message here is to define a methodology for dealing properly with ordinal data.

To deal with cases where appreciable portions of the cohort, for some reason or other, respond differently to particular items—e.g., those items prone to SPEs such as recency and primacy—tests to confirm the unique metrological properties of the Rasch model [16] will be examined in Section 5. However, before that, in the next section, we consider how SPEs can be explained, amongst others with a causal, entropy-based theory and in support of validity.

4. Case Study II: Explaining Serial Position Effects

In case study I (Section 3 above), our experimental observations were used to obtain empirical estimates of the memory task difficulty and person memory ability (as plotted on the x-axes of Figures 1a and 1b, respectively, and given below in Table 1) by a methodological analysis of the experimental responses, with a logistic regression based on the Rasch formula [16] in Equation (1).

Table 1. The RAVLT IR explanatory variables, empirical, CSE/quasi-theoretical, and theoretical task difficulty values and associated measurement uncertainties (k = 2) for each item.

	Explanatory Variables				Empirical		CSE Quasi-Theoretical	
	Primacy (Equation (A7))	Middle (Equation (A5))	Recency (Equation (A8))	Frequency (Equation (A6))	δ	Uδ	zR (Equation (3))	UzR
Item 1	0.00	6.31	−8.33	4.80	−0.81	0.28	−0.80	5.28
Item 2	−0.26	6.31	−7.38	4.85	−0.54	0.28	−0.24	4.86
Item 3	−0.66	6.31	−6.46	4.88	0.57	0.30	0.19	4.48
Item 4	−1.17	6.31	−5.58	3.99	1.38	0.36	0.34	4.16
Item 5	−1.77	6.31	−4.73	3.25	−0.64	0.28	0.42	3.92
Item 6	−2.43	6.31	−3.92	3.56	0.32	0.30	0.63	3.76
Item 7	−3.15	6.31	−3.15	4.31	1.16	0.34	0.84	3.71
Item 8	−3.92	6.31	−2.43	3.39	−0.09	0.28	0.66	3.79
Item 9	−4.73	6.31	−1.77	4.41	1.31	0.36	0.78	3.96
Item 10	−5.58	6.31	−1.17	4.06	0.64	0.32	0.55	4.22
Item 11	−6.46	6.31	−0.66	4.27	0.74	0.32	0.33	4.56
Item 12	−7.38	6.31	−0.26	5.44	−0.72	0.28	0.20	4.97
Item 13	−8.33	6.31	0.00	3.57	0.23	0.30	−0.66	5.40
Item 14	−9.30	6.31	0.00	3.24	−1.46	0.30	−1.38	5.87
Item 15	−10.30	6.31	0.00	4.93	−2.09	0.34	−1.86	6.36

Here, we complement these analyses with ab initio theoretical and quasi-theoretical causal estimates of the memory task difficulty, thus providing evidence for construct validity, item equivalence, and enabling metrological references based on causality and best understanding. Our previous work on recalling the simplest nonverbal items—taps and numbers [11,12]—gave first principle, ab initio explanations of the task difficulty in terms of entropy: less entropy means a more ordered task, which is thus easier to perform. This entropy-based explanatory theory, summarized in Appendix B, is extended here to include causal explanations of recall in the word learning list test RAVLT IR, including SPE such as primacy and recency.

Based on this entropy approach, we introduce CSE as a quasi-theoretical method to explain the task difficulty for word recall in general (Appendix B.2) and SPE in particular (Appendix B.4).

4.1. Construct Specification Equations (CSE) to Explain Task Difficulty in Word Learning Tests

Expressing a CSE for task difficulty in the form of Equation (A3) can be done in terms of the total entropy as a prediction of the task difficulty for all the symbols (words), j. The total task difficulty will then be simply found by adding the separate entropies discussed in Appendix B (Equations (A5)–(A8)):

$$\delta_j = \delta_{j,Pr} + 2 \cdot \delta_{j,Mr} + \delta_{j,Rr} + \delta_{j,freq} \quad (2)$$

where $M = \frac{1}{ln(L)} = \frac{1}{ln(15)} = 0.369$ for a sequence of length, and $L = 15$ words.

Equation (2) is an example of a pure ab initio theoretical definition of a task difficulty in RAVLT IR, where there are potentially four separate components: the mid-range, primacy, and recency word recall difficulties, plus word frequency. It is not obvious that the chosen explanatory variables in Equation (2) will be the principal components of the variations in the empirical data. The various coefficients β_k, as they occur in Equations (A3) and (A5)–(A8), need to be determined. A PCR analysis (Appendix B.1) was therefore performed. As a result of the initial PCA, it was found that the primacy and recency explanatory variables were correlated with a Pearson's coefficient of −0.94 (where the minus sign indicates that an increase in primacy is accompanied by a corresponding decrease in recency and vice versa). Bearing in mind the similarities between the theoretical expressions of Equations (A6) and (A7), such a high correlation is to be expected. On the other hand, there was only a weak correlation (coefficient −0.18 at most) found from the PCA between recency or primacy (Appendix B.4) and word frequency (Appendix B.3).

Based on the empirical task difficulty parameters from the present data set and the four explanatory variables (Equations (A5)–(A8)), the resulting CSE from a PCR analysis of the task difficulty of item j in RAVLT IR for the whole cohort is found to be (uncertainties, with coverage factor $k = 2$, are given in parentheses):

$$zR_{RAVLT\ IR,j} = +5(3) + 0.7(5) \cdot Primacy_j + 0.8(5) \cdot Recency_j + 0.2(2) \cdot Frequency_j \quad (3)$$

as the restituted value (R) of the object attribute Z, where the explanatory variables X are: $Primacy_j = \delta_{j,Pr}$; $Recency_j = \delta_{j,Rr}$; and $Frequency_j = \delta_{j,freq}$. The CSE such as Equation (4) indicate the predominance of entropy-based syntax contributions in recall task difficulty, while additional semantics effects, such as word frequency, are found to be small and dominated by measurement uncertainties.

Table 1 provides empirical (δ, Equation (1)) and CSE/quasi-theoretical task difficulty (zR, Equation (3)) values, as well as the explanatory variables for each RAVLT IR item, and Figure 3 shows correlation plots between the empirical and quasi-theoretical task difficulty values. As an important measure of the goodness of fit, the Pearson's correlation coefficient, R, between the empirical task difficulty values (δ) and CSE/quasi-theoretical (zR) task difficulty values was found to be 0.80, which we suggest indicates a satisfactory prediction of the task difficulty.

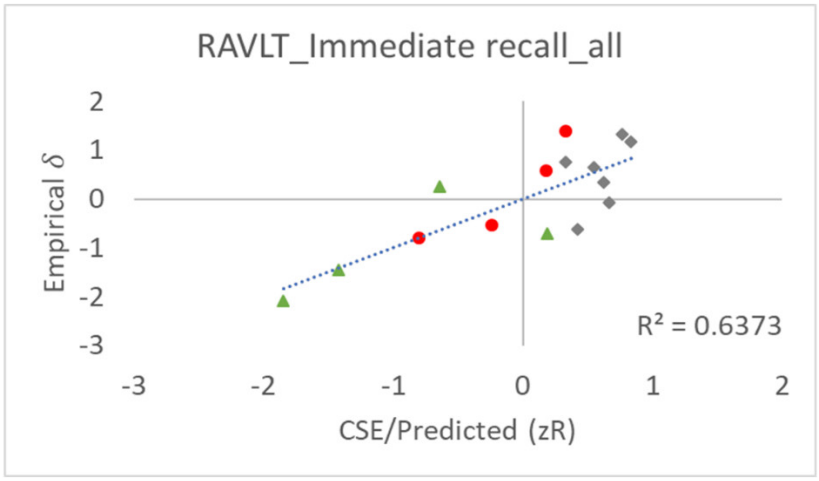


Figure 3. Correlation plot between the empirical task difficulty values ((δ, Equation (1), y-axis) and CSE/quasi-theoretical (zR) task difficulty values (Equation (10), x-axis). Item colouring and shape correspond to red dots = Pr items, grey diamonds = Mr items, and green triangles = Rr items.

The contribution from each of the explanatory variables in Equation (3) is shown graphically in Figure 4. An inspection of that plot, as well as those in Figure 5a–c, confirms the observations of the earlier researchers, who characterised Pr as a SPE of the first four words of the list, whilst Rr was a SPE of the last four words [9,18–20]. Note, however, that according to our entropy-based theory (Appendix B.4), the relative reductions in task difficulty of Pr and Rr (red and green curves in Figure 4) vary greatly across all item orders but complement each other (a large Pr is accompanied by a small Rr and vice versa, in agreement with the strong, negative correlation between these found in the PCR). The sum of these negative contributions at any one item j is substantially balanced by the average positive contribution from the mid-list term, $\delta_{j,midrange}$, as in Equation (2). (The mid-list term, the intercept, will be discussed further in the third case study, Section 5).

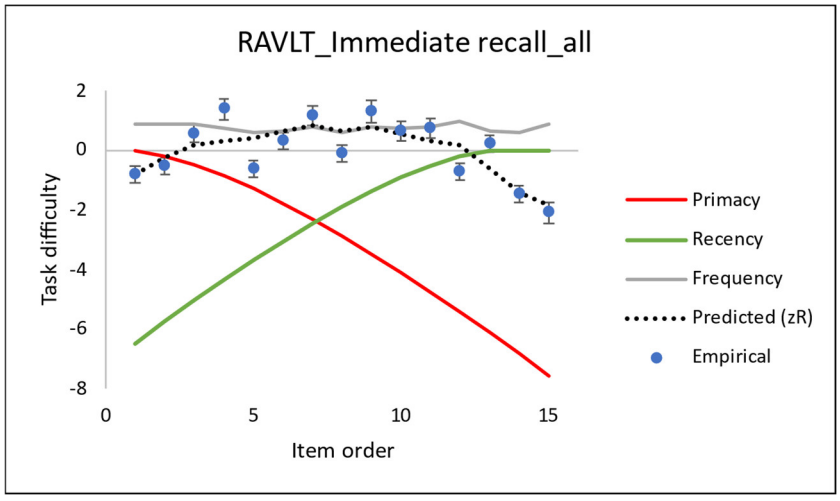
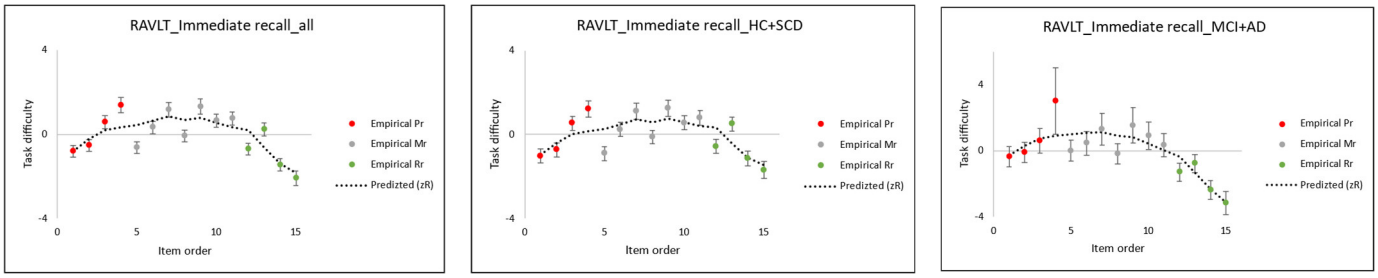


Figure 4. Contribution of the different explanatory variables for the CSE for the task difficulty in RAVLT IR from Equation (4). Uncertainty coverage factor $k = 2$.



(**a**) (**b**) (**c**)

Figure 5. Relations between the empirical values and CSE/predicted (zR) values of the task difficulty for (**a**) the whole cohort and (**b**,**c**) the two subgroups of the cognitively healthy (HD and SCD) and cognitively impaired (MCI and AD). Item colouring corresponds to red = Pr items, grey = Mr items, and green = Rr items.

The current work expands on the results of previous studies by being able to explain the various components of the task difficulty in terms of informational entropy, providing satisfactory descriptions of, in particular, SPE such as primacy and recency, simply in terms of combinatorics and Brillouin's [21] entropy expressions (Equations (A5)–(A8)).

Our theoretical explanation of the SPE task difficulty in terms of entropy is considerably simpler than earlier theories of SPEs, such as reviewed by Hurlstone [10], which invoked additional explanations in terms of memory processes, including (i) a primacy gradient of the activation levels (i.e., first item activated the strongest, and a monotonically decrease for the latter items), (ii) a cumulative match (i.e., matching of incoming sequences to previously presented sequences), (iii) a revised primacy model (i.e., a sequence is made familiar through repetition in the local chunk nodes), and (iv) an output inference model (i.e., adding increasing amounts of Gaussian noise with each successive one). Such additional effects would need to be studied further in extended investigations, hopefully with smaller measurement uncertainties than in the present work.

4.2. Construct Specification Equations for Cohort Subgroups

Having explained the task difficulty, the next step is to investigate how well RAVLT IR and, in particular, the SPE can distinguish between different cognitive health statuses in view of the purported new diagnostic tool (Section 1). With the same explanatory variables, but with different empirical task difficulties, the values for the two principal groups of the test persons portioned according to cognitive health status are found to result in slightly different CSEs for each group, although these differences are barely significant compared with the measurement uncertainties (the latter, with a coverage factor $k = 2$, are given in parentheses):

$$zR_{RAVLT\ IR\ HC+SCD} = +4(3) + 0.7(6) \cdot Primacy_i + 0.7(5) \cdot Recency_i + 0.2(2) \cdot Frequency_i \quad (4)$$

$$zR_{RAVLT\ IR\ MCI+AD} = +7(4) + 1.0(7) \cdot Primacy_i + 0.9(7) \cdot Recency_i + 0.1(3) \cdot Frequency_i \quad (5)$$

Figure 5a–c show the relation between the empirical values and CSE/quasi-theoretical (zR) values of the task difficulty for the whole cohort and the two subgroups.

As observed in the earlier literature, cognitively impaired individuals are not always able to benefit from reductions in the recall difficulty normally facilitated by SPE, particularly Pr. This health status increase in the recall task difficulty for Pr items is the basis for claiming SPE as a purported diagnostic tool for cognitive decline (Section 1). As shown in Figure 5c, these claims are confirmed here by the reduction in Pr recall task difficulty, which we find to be less pronounced for the least cognitively healthy cohort group, while the Rr effects appear largely unchanged.

When making this interpretation, it is worth recalling that a reduction in the number of test persons—as in the analyses in these cognitive healthy-specific groups—leads to increases in the measurement uncertainties, as also reflected in the CSE, in Equations (4) and (5) compared to Equation (3). Uncertainties in the β-coefficients in each CSE are a combination of the measurement uncertainties, $U\delta$, in the empirical task difficulty values (typically 0.3 logit, as shown in Figure 5a–c), which propagate through principal component regression together with the modelling uncertainties (PTC MATHCAD Prime 6.0.0.0, module *polyfitc*), which are an order of magnitude larger and arise, for example, if additional components or other sources contribute to the unexplained variance. A further point to be made on the interpretation and its relation to the number of test persons: The cognitively healthier group (HC and SCD) comprises 165 persons, while the less cognitively healthy group (MCI and AD) only comprises 60 persons. This uneven balance is reflected in the very similar Equations (4) and (5) and Figure 5a,b compared to Equation (6) and Figure 5c.

Despite the large uncertainties, reasonable agreement is found between the theory and experiment, where the Pearson's correlation coefficient R, as a measure of the agreement between the empirical task difficulty values and CSE/quasi-theoretical (zR) task difficulty values, is found to be slightly larger for the group of MCI and AD, $R = 0.85$, compared to HC and SCD, $R = 0.75$.

4.3. Insights into Serial Position Effects Based in Word Learning List Tests on Construct

This second case study started by introducing the general role of the CSE as an advanced method for testing theories of constructs, particularly task difficulty, provided by an analysis of the empirical data using the modern metrological approach, the Rasch GLM model, from the first case study. This second case study concluded with an enhanced understanding of the SPE explained in entropy terms for RAVLT IR and presented a new methodology for calculating the recall task difficulty in word learning list tests.

To identify an improved diagnostic tool for cognitive decline based on the SPE, it is important to consider that SPE are prominent contributions to task difficulties, which vary from word to word more rapidly than the average difficulty of recalling an arbitrary word anywhere on the list, as explained with the CSE.

For SPE to act as a valid diagnostic tool to identify and predict cognitive impairment, there has to be a causal effect—that is, a significant change in the metric that is related in some way to the cognitive health status of the test person. The diagnostic effects also have to be significant—that is, larger than the measurement uncertainties, where the latter have to be evaluated.

Further studies remain to be done about how the diagnostic sensitivity correlates with the cognitive health of each cohort member, which may be expressed in terms of the explanatory variables (such as biomarkers) in the CSE for a person's (instrument) ability [22,23]. Another approach—pursued in the final case study (below)—will address the evidence of (perceived) changes $\Delta\delta$ in the average task difficulty for different parts of the cohort grouped clinically according to the cognitive status and interpreted in terms of the apparent scale distortions, multidimensionality, and resultant changes (residuals) in response.

5. Case study III: Maintaining the Unique Metrological Properties of the Rasch Model in the Presence of Serial Position Effects

Despite the promising new diagnostic tool based on SPE, a caveat is that the unique metrological properties of the Rasch model [16] need to be maintained, particularly in the presence of SPEs, so that the counted fraction ordinality in the raw response scores can be adequately compensated for and unidimensional, separate estimates can be made for each item difficulty (irrespective of the test person) and each person ability (irrespective of the item), as mentioned at the end of case study I (Section 3.3).

The separation of the conjoint Rasch attributes—task difficulty and person ability—necessary for the metrological quality assurance of any indirect ordinal response is usually done by exploiting the property of the specific objectivity of the Rasch model [16] (Equation (1)) (Appendix A.1). Analysing data sets as a candidate diagnostic tool, including the SPE in word learning list tests (Section 1), can present challenges. If different cohort members experience SPE differently—for instance, according to the cognitive status (as claimed for the new diagnostic tool)—then the task difficulty might include a person-dependent factor, thereby breaking the specific objectivity. This final case study examines the tests to maintain the unique metrological properties of the Rasch model [16].

5.1. Scale Distortion and Instrument (Test Person) Discrimination in the Presence of SPE

There is some evidence in our results (Section 4.2) that the benefit from *Pr* in reducing the task difficulty is particularly diminished for the less cognitive healthy part of the cohort, apparently in line with earlier CTT-based claims about the diagnostic potential of SPE, where patients with AD do not recall the last-mentioned words. Despite large measurement uncertainties, there also appear to be some differences in the first "intercept" term on the RHS of the CSE for task difficulty—Equations (4) and (5) derived from PCR—between the two cohort groups, according to the cognitive status.

In this work, support is taken from our entropy-based theory: A theoretical explanation of the intercept in the CSE for the task difficulty for item j is that it equals twice the mid-range task difficulty:

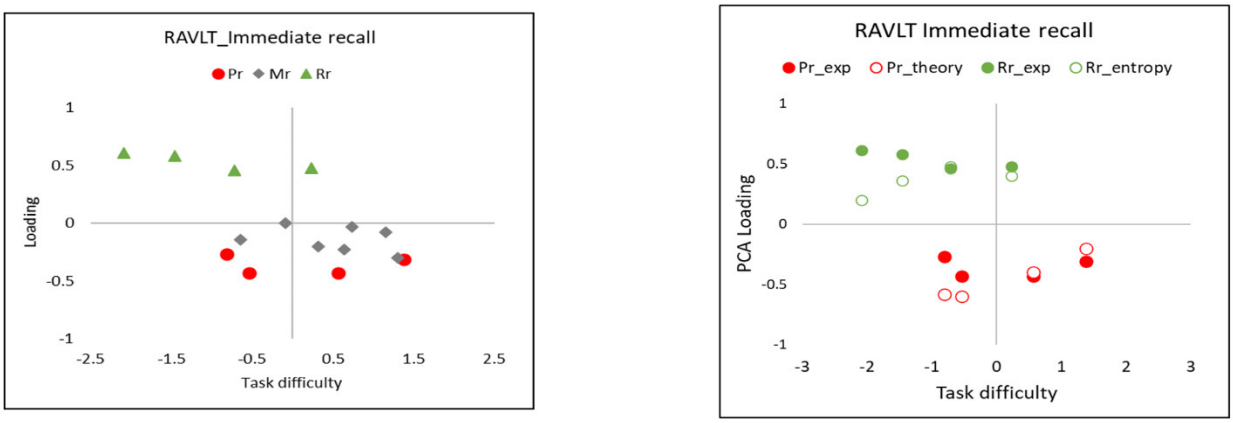
$$\delta_{j,Mr} = +M \cdot ln\left(\frac{L}{2}!\right)$$

according to Equation (2), as explained in Section 4.1. An evaluation of that term for a RAVLT list of 15 words yields $\delta_{j,Mr} = +6{,}5\ logits$, which is close to the experimentally observed intercept, i.e., +7(4) logits in the CSE for the less healthy cohort (Equation (4)). The lower intercept value for the healthy group (HC and SCD), i.e., +4(3) logits (Equation (5)), means that, apart from benefitting from the SPE, the group also has an overall reduced task difficulty compared with the less cognitively healthy group (MCI and AD) across the tasks. SPE, which are appreciable for the healthy group, thus cause an extra discrepancy $\Delta\delta_{SPE}$ of about 2 logits compared with our entropy-based theory (which is applied better in the less cognitively healthy group, since they lack SPE). Our intercept term values are comparable with the corresponding values of $Asymptote = \delta_{j=\frac{L}{2}}$ reported in earlier studies of Murdock for a wide range of word lists (L = 10,…,40) [2].

A change in the intercept value could be linked to an entropy-based change in the effective length, L, of the sequence to be recalled: a change of $-2\ logits$ between Equations (4) and (5) would correspond to an effective shortening for the cognitively healthy group (HC and SCD) of the sequence length from 15 to 10 words. For a less healthy group (MCI and AD), the effective length of the list remains 15 words. Similar effects have been found recently, where effective shortening was interpreted as an effect of learning when repeated trials of the same word list were made (as examined in detail in a separate publication [24]).

5.2. Principal Components and Multidimensionality

Loading plots of fit residuals such as shown in Figure 6 can reveal evidence—such as the clustering into two SPE groups—for multidimensionality in the principal component analysis (PCA2) of the item logistic regression fit residuals (Appendix C.2). Such multidimensionality might arise from the SPE, since they imply that each cohort individual might potentially "assess" the task difficulty in an individual way, thus breaking the conditions for what Rasch [16] called "specific objectivity" (Appendix C.1) needed to make the necessary separation of the task difficulty from the person's ability.



(**a**) (**b**)

Figure 6. PCA2 loadings (y-axes) vs. the RAVLT immediate recall task difficulty, δ, of the Rasch regression residuals (whole cohort) (x-axes). (**a**) Empirical values for all 15 items, (**b**) empirical values (Pr_exp and Rr_exp) and predictions of the effects of the scale distortion from the SPEs (Pr_theory and Rr_theory) (Equation (10), $\Delta\delta = +2$) primacy (Pr) and recency (Rr); PCA1 coefficients $a_{p,x}$ from Equation (6); distribution of a person's ability (Equation (11)): $\bar{\theta}_{all} = -0.6(1.0)$ and $\sigma = 1.0$.

The present work argues that, because we can explain a task difficulty when formulating the CSE, particularly deploying the concept of entropy (Equation (2)), it may also be possible to identify factors common to both the CSE PCA (as the first step in a PCR, Section 4.1) and the PCA of fit residuals (Figure 6). The effects of scale distortion should therefore be predictable—for instance, in loading plots, as in the present study, where different test persons have more or less discrimination in SPEs, according to their cognitive state, as is now examined empirically.

5.2.1. PCA1 for CSE Formulation

The first kind of PCA (Appendix C.1), **PCA1**, made as the first step in PCR when forming the CSE (Section 4.1), deals directly with the Rasch parameters, such as task difficulty, δ, in terms of a set of explanatory variables, X_k:

$$PC_p = \sum_{k=1}^{K} a_{p,k} \cdot X_k$$

PCA1 is expected to reveal additional components of variation (explanatory variables, X_k), for example, from the SPE when the CSE (Equation (2)) are formed for the construct task difficulty. The amount of scale distortion arising from the primacy and recency are simply fractions of the task difficulty (given theoretically term-for-term in CSE Equation (2) for each dimension in relation to the overall task difficulty).

From the PCA1 of the current empirical Rasch data, the dominant (most variance) principal component (the first column of the matrix P) is found to be:

$$PC_1 = \sum_{k=1}^{3} L_k \cdot X_k = +\mathbf{0.77} \cdot \boldsymbol{Primacy} - 0.64 \cdot Recency + 0.02 \cdot Freq \quad (6)$$

where L_k is the loading of the k^{th} explanatory variable, X_k, and where the SPE (primacy, recency, etc.) are estimated theoretically (with Equation (2)). The coefficients $a_{1,primacy} = +0.77$ and $a_{1,recency} = -0.64$ in Equation (6) will turn out to be of particular interest. The next principal components of variation are found to be:

$$PC_2 = -0.51 \cdot Primacy - \mathbf{0.60} \cdot \boldsymbol{Recency} + 0.62 \cdot Freq$$

$$PC_3 = -0.39 \cdot Primacy - 0.49 \cdot Recency - \mathbf{0.79} \cdot \boldsymbol{Freq}$$

Each PC in the present case is thus found experimentally to be related in turn to each of the three main explanatory variables for task difficulty, i.e., primacy, recency, and word frequency ("*Freq*") (respectively expressed by Equation (3)) as indicated in boldface in the three equations above. The CSE for the task difficulty are formed with PCR (Section 4.1) by regressing the Rasch estimates of δ against these three principal components of variation.

5.2.2. PCA2 Rasch Logistic Regression Residuals

In the present observations, empirical evidence was found in **PCA2** (Appendix C.2) that the SPE represented additional dimensions of the task difficulty over and above the difficulty in recalling any word in the RAVLT immediate recall, as can be seen by the clustering in the loading plot (Figure 6a).

However, the evidence was not completely clear-cut: the eigenvalue of the unexplained variance in the first contrast was <2.0 (i.e., supporting unidimensionality), while the disattenuated correlation between cluster 1 (only items from Rr) and cluster 3 (items from both Pr and Mr) was −0.60 (i.e., indicating multidimensionality). Based on Linacre's [25] further discussion about multidimensionality and PCA2 analysis:

a. "Compare the raw variance explained by the items (present case 18%) with the unexplained variance in the first contrast (present case 9%); is this ratio big enough to be a concern?" In our analysis of the data, the variance of the first Rasch dimension is about double that of the secondary dimension, but the latter is clearly noticeable.
b. "Is the secondary dimension bigger by chance?" In the present case, with the eigenvalue = 2, this is the strength of approximately two items. It is not expected to have a value of more than two items by chance [26], and at least two items are needed to think of the situation as a "dimension" and not merely an idiosyncratic item, according to Linacre.
c. "Does the secondary dimension have substance?" Looking at the loading plot in Figure 6a, the two SPE clusters of items—primacy and recency—are clearly separated vertically (the important direction) from the other (mid-list) cluster of items. Each cluster can be regarded to be important enough and different enough to be explained in terms of separate dimensions, as motivated in the Introduction.
d. One approach to handling the multidimensionality revealed by the PCA2 analyses is to make a separate fit of the Rasch formula to each cluster of items associated with a different dimension. The significance of the differences found in the test person's ability between logistic regression fits to the different item clusters can be assessed, for example, with statistical *t*-tests [27,28]. Since, in the present case, SPE such as primacy and recency only affect a few words at the start and end of a word list, this means that a separate Rasch analysis for each such small cluster would result in poor reliability in the test person's ability estimates owing to the reduced numbers of the degrees of freedom [29]. Similarly, analyses of the portions of the whole cohort grouped by health status would also have poorer reliability. Statistical significance tests or correlation plots between sets of a person's measures from separate Rasch analyses for each suspected dimension or each cohort group according to the cognitive status were therefore judged to not be useful in view of the relatively large uncertainties of the present data.

5.2.3. Combining the Two Kinds of PCA

In addition to the empirical evidence described above, the entropy-based theoretical modelling of the clusters in the **PCA2** loading plots (such as shown in Figure 6) can provide extra support for the validity for SPE multidimensionality, as follows. A change in the task difficulty $\Delta\delta$ for each item j associated with SPE (Section 5.1) can lead to a

change in the logistic fit item residual, $y_{i,j}$, of the regression of Equation (1) to raw response data where the task sensitivity $K_\delta = \frac{\partial P_{success}}{\partial \delta}$:

$$y'_{j,\delta} = y_j - \frac{\partial P_{success}}{\partial \delta} \cdot \Delta\delta \tag{7}$$

by a simple partial differentiation (Appendix C).

PCA2, based on the regression of fit residuals (Appendix C.2), can therefore contain evidence of SPE-related effects. In particular, from Equation (7), the loading of a PC in the logistic regression residuals will be proportional to the product of the sensitivity (K_δ) and a change in perceived task difficulty, $\Delta\delta$:

$$\text{PCA2 loading: } L_{p,x} \propto a_{p,x} \cdot \frac{\partial P_{success}}{\partial \delta} \cdot \Delta\delta \tag{8}$$

Term-for-term on the RHS of Equation (8):

i. The loading coefficient $a_{p,x}$ should be the same as deduced in the **PCA1** above (Section 5.2.1) when forming the CSE for task difficulty in terms of the entropy-based explanatory variables (Equation (6)).

ii. K_δ, the peculiar sensitivity of the instrument (person) in the Rasch model, will "modify" the PCA loading plots correspondingly but can be calculated from a simple differentiation of the dichotomous Rasch formula (Equation (1)) [30] to yield Equation (9). The sensitivity according to Equation (10) has a "resonance-like" behaviour, with a peak in K occurring at $\delta = \bar{\theta}$ (the 50% point $P_{success}$), and the sensitivity will fall off symmetrically on either side to approach zero at each end of the task difficulty scale [30].

$$K_\delta = \frac{\partial P_{success}}{\partial \delta} = -\frac{e^{(\theta - \delta)}}{(1 + e^{(\theta - \delta)})^2} \tag{9}$$

iii. The third term on the RHS of Equation (9) is any significant change, $\Delta\delta$, in the task difficulty. As found experimentally (Section 5.1), the task difficulty of each item deviates the most from the basic Rasch model for the healthy cohort, both owing to the SPE—such as Primacy and Recency—as well as an overall reduction in the task difficulties across all the items: $\Delta\delta = -2\ logit$. (In contrast, members of the less cognitively healthy portion of the cohort benefit less from the Primacy and have a mid-range task difficulty close to the theoretical value of the CSE intercept.)

The final step in evaluating the SPE loading according to Equation [8] and accounting for the loading plot of Figure 6 is to integrate the task difficulty sensitivity $K_\delta = \frac{\partial P_{success}}{\partial \delta}$ according to Equation (9) to allow for variations in the sensitivity at each level of task difficulty as a function of a person's ability across the distribution of the cohort. The modified expression is:

$$\textbf{PCA2 loading: } L_{p,x} \propto a_{p,x} \cdot \Delta\delta \cdot \int p(\theta) \cdot \frac{\partial P_{success}}{\partial \delta} \cdot d\theta \tag{10}$$

where a person's ability is taken to be normally distributed across the cohort "all" according to:

$$p(\theta) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \frac{(\theta - \bar{\theta}_{all})^2}{\sigma^2}} \tag{11}$$

where $\bar{\theta}_{all} = -0.6(1.0)$, the average cohort ability, and $\sigma = 1.0$, based on our Rasch analysis of the data (Section 3.2).

An inspection of the Rasch residual PCA2 loading plot (Figure 6 for the whole cohort) reveals evidence of these deviations in task difficulty across the items, as investigated for the empirical data for task difficulty plotted in Figure 5a–c and in terms of our entropy-based theory of task difficulty, leading to Equation (2).

The agreement between the experimental and theoretical values of SPE loading evident in Figure 6 is deemed satisfactory, lending support to the empirical evidence for multidimensionality associated with the SPE in the word learning list tests.

6. Conclusions

The use of entropy as a powerful explanatory variable was found to explain the difficulty of recalling word sequences as an extension of an earlier work on syntax-based and nonverbal memory tests as part of our research to provide descriptions of the intrinsic properties of a measurement system, where a test person acts as a measurement instrument.

Brillouin's classic combinatoric entropy expression (Equation (A4)) explains both the overall task difficulty, as in our earlier studies of nonverbal tests, as well as the SPE such as primacy and recency in the first trial of the word learning test RAVLT. The CSE, which explain the recall task difficulty, provide evidence for the construct validity, item equivalence, and enable metrological references based on the causality and best understanding. These results complement the conclusions of several previous studies—reviewed by Weitzner & Calamia [1]—which claimed SPEs to be significant diagnostic tools, while, at the same time, not explicitly declaring their measurement uncertainties or using the Rasch model. In response to claims in the literature that *research (into a new SPE-based diagnostic) is limited by the multiple ways in which SPEs are defined and analysed* [1], the present work proposes a more consistent way of defining, analysing, and theoretically predicting SPE multidimensionality while maintaining the unique metrological properties of the Rasch model.

Alternatives to the Rasch model, such as the so-called 2PL and 3PL item response theory (IRT) models, have claimed to handle effects such as discrimination and a number of causes of *measurement disturbances, such as start-up, guessing, plodding, carelessness, and item/person interactions* [31], each of which may plausibly depend on the state of health of each cohort member. However, such alternative IRT do not have the specific objectivity property of the Rasch model essential for its unique metrological properties. According to Wright & Stone [32], *Mathematical analysis shows that the 3PL model is a non-converging, inestimable elaboration of the Rasch model* and has traditionally has been judged to be too idiosyncratic to be useful in a general measurement system [32]. In educational contexts, one can argue, for example, that, instead of reanalysing the responses of a whole class, the one or two individuals showing another discrimination can better be dealt with separately. Such exceptions are, however, not possible in the present healthcare example, since significant portions of the cohort appear to have different levels of discrimination, depending on their cognitive status.

Multidimensionality always exists to some extent [25,33]; the question is whether it is significant enough to warrant separating groups of items into subtests (e.g., in the present case for primacy and recency) in an attempt to maintain the unique metrological properties of the Rasch model [16]. We adopted Linacre's [34] recommendation, described in Section 5.2.2, to examine the multidimensionality through studies of the residuals of fit for both the raw data and the Rasch attributes.

Our PCA-based approach has some similarities to what has been called scale alignment; in the words of Feuerstahler and Wilson [35], *"[by] projecting the dimensions from the multidimensional model onto a single reference dimension…which represents a composite of the individual dimensions…. scale alignment aims to better achieve the ideal relationship between sufficient statistics and item parameter estimates."* We consider SPEs to introduce a kind of *between-items multidimensionality*, as studied earlier, e.g., educational studies—[35,36] mention *algebra, geometry, statistics, and written mathematics*—but, in our case, with recall tasks arguably conceptually simpler and capable of being modelled theoretically from the first (entropy-based) principles in support of validity.

Two distinct but related PCAs are examined in detail in Section 5.2, where our entropy-based theory has provided support for the validity of the assessments of

multidimensionality associated with the SPE. Our studies show that the effects of scale distortion are predictable to some extent, for instance, in loading plots, where different test persons have more or less discrimination in the SPEs according to their cognitive state.

Our new methodology applies modern measurement theory but prepares us to deal with situations where SPE were to be adopted as a diagnostic tool. To date, in our study, measurement uncertainties have been relatively large, reflecting the limited sample size, collinearity, and measurement disturbances (Section 4.1), and there are sources of dispersion when making multivariate regression that are not yet accounted for. What appears to be the case is that, over and above individual variations in a person's ability, there is an overall shift in the person's ability for each clinical group. Whether one regards that as a change in ability or a change in task difficulty is a moot point. Choosing the task difficulty as a metrological reference (as proposed in terms of robustness and simplicity [37,38]) would suggest that SPE dependency would be assigned to a person's ability.

The next steps include acquiring larger samples in order to explore the multidimensionality effects beyond the measurement uncertainties, other test sequences such as different lengths, and trials, as well as evaluations with other diagnoses. This will both enhance the understanding of the multidimensionality in tests where the SPE may affect the discrimination, as well as provide clinicians with more valid and reliable tools for the diagnosis of cognitive impairment.

Here, we did not focus on the cognitive function or brain structure and networks of each test person (which are more related to explaining a person's ability, θ) but, rather, on explaining the difficulty, δ, of the task of recalling each word list, which are the objects of the measurement system. Other parts of the NeuroMET2 project are currently studying correlations between a person's ability and a number of imaging and liquid biomarkers. Forthcoming works are expected to address the brain structure and networks, also with an emphasis on the concept of entropy for cognitive processes.

From a broader perspective, the present work is part of an ongoing effort aimed at extending traditional metrological concepts to include ordinal properties in response to a widening and increasingly important set of scientific studies and the applications of these properties (such as in education, sustainability, and healthcare).

Author Contributions: Conceptualization, L.P. and J.M.; methodology, L.P.; formal analysis, J.M. and L.P.; writing—original draft preparation, L.P. and J.M.; writing—review and editing, J.M, L.P, S.C., A.F., and L.G.; visualization, J.M.; and supervision, S.C. and L.P. All authors have read and agreed to the published version of the manuscript.

Funding: Part of the work done in the 15HLT04 NeuroMET and 18HLT09 NeuroMET2 projects received funding from the EMPIR programme co-financed by the participating states (VINNOVA, the Swedish innovation agency in the present case) and from the European Union's Horizon 2020 research and innovation programme.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Charité University Hospital, Berlin, Germany (EA1/197/16 approval on October 13th 2016 and EA2/121/19 approval on October 10th 2019)

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: With regards to data availability, see: https://doi.org/10.5281/zenodo.5582001, 25 June 2022

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Entropy and the Measurement Process

Regarding measurement as a particular kind of information communication is a useful point of departure when introducing a measurement system analysis (MSA [39]) to faithfully describe an observation process—that is, how measurement information is

transmitted from a measurement object (A) via an instrument (B) to an observer (C) [11,37]. Apart from this local communication, it is, of course, in many cases, also required to communicate measurement information as globally as needed, across distances and amongst different persons, according to what is meaningful for the quality assurance in each field of application [37].

The present work shares a background common with our previous paper [11], as can be briefly summarised as follows: The amount of "useful information" in a measurement system, analogous to a certain extent with the original entropy concept as a measure of "useful energy" in steam engines [40], can be described with the well-known conditional entropy expression:

$$H(Y|Z) = H(Z,Y) - H(Z) \quad \text{A1}$$

Expression (A1) states how the amount of information changes during transmission in a measurement system in terms of the entropy in the response (Y) of the system when observing a quantity (Z) attributed to the measurement object. At the start of the measurement process, there is an initial "deficit" in the entropy (i.e., "surplus" information) $H(Z)$ coming from prior knowledge (prior distribution, P) of the measurand (attribute, Z, of the object entity, A). The losses and distortions $H(Z,Y)$ increase the entropy through measurement imperfections (including measurement system attributes, such as the sensitivity and resolution of an instrument (B)), leading, finally, to a posterior distribution (Q) in the response Y with entropy $H(Y|Z)$ as the result of the measurement process, as registered by an observer (C). The notation here is analogous to that used by Rossi [41] in a probabilistic model of the measurement process and as described in more detail in our earlier publication [11].

Informational entropy, as a measure of the amount of information, is a key concept when seeking metrological quality assurance of the measurement systems, both in terms of the measurement uncertainty and of the traceability. Two distinct contexts can be identified, where one seeks stationary (minimum or maximum, respectively) values of the entropy [37]:

i. the best units for metrological traceability are those with the most order, i.e., least entropy, as an example of the principle of least action;
ii. the change in entropy on the transmission of measurement information cannot decrease, thus allowing realistic estimates of measurement uncertainty in line with the second law of thermodynamics.

Appendix A.1. Analysing Categorical Responses

An entropy-based approach is especially useful when analysing categorical responses, such as in the cognitive memory tests studied in the present work. Such test responses (Y), typically scored as the probability, $P_{success}$, of a correct classification, are characterised in general by counted fraction nonlinearity, which needs to be compensated for to have any quantitative use for such ordinal data [42–44].

The compensation for any counted fraction nonlinearity, as will be illustrated in the first case study (Section 3), can be done by transforming the ordinal response, $P_{success}$, of a human (to a memory test in the present study) onto a linear, interval scale using the Rasch model (Equation (A2)). The latter is a particular form of a generalised linear model (GLM) [16] with special metrological properties:

$$P_{success} = \frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}} \quad \text{A2}$$

The GLM can be derived from the first principles with the method of Lagrange multipliers, where the probabilities, q_c, of classifying a response in a category c, (posterior distribution Q) are derived subject to the constraint of maximising the entropy and scores constrained to lie between 0 and 100% and the results in the exponential form [37,45]. Making a logistic regression of the Rasch model (Equation (1)) to the raw data yields

separate estimates (with so-called specific objectivity) of the ability, θ, of the instrument (B) and the level, δ, of difficulty of the task (A) in the process of restitution [41], R. These relations for a binary response can be extended to the multinomial, polytomous case, as required [37].

To the best of our knowledge, the current practice of deducing a person's ability and task difficulty in most legacy tests of memory is often limited. With a few exceptions [24,29,46], analyses of the SPE to date have, unfortunately, mainly used CTT, which is questionable for the ordinal responses typical of these kinds of tests. For example, in the memory Rey's Auditory Verbal Learning Test (RAVLT), where raw scores equal 1 for pass or 0 for fail, Moradi et al. ([47], p. 417) defined learning as *the score of trial 5 minus the score of trial 1. However, raw scores have no numerical meaning and only serve to indicate the ordered categories* ([48], p. 2). In turn, such an improper analysis will only lead to unnecessarily large uncertainties (expressed in terms of entropy according to (ii) above) and to associated increased risks of incorrect decisions.

Linear separate measures for person and item attributes, which are a prerequisite when providing quality assurance in terms of metrological traceability and measurement uncertainty, are only attainable by restitution from the raw scores through a logistic regression such as the Rasch dichotomous GLM [16] (Equation (A2)). Despite the importance of compensating for counted fraction nonlinearity and metrologically having separate person and item attributes, apart from our own work, there are only a few exceptions (particularly the works of Stenner et al. [49,50], Hughes et al. [51], and Hobart et al. [52]) where this has been achieved so far in the analysis of cognitive tests.

Appendix B. Entropy and Construct Specification Equations

Following the analysis of a measurement system, entropy as a concept can also be used to express each attribute of the various elements (object, A, instrument, B, operator, C, etc.) of the measurement system causally in terms of Construct Specification Equations (CSE). This is done based on our best understanding of each construct [23]. Since Formula A2 is derived under the constraint of maximum entropy [37,45], the various Rasch parameters, such as θ and δ, also turn out to be explainable in terms of entropy, as illustrated in the second case study (Section 4).

Specifically, a task—the object (entity, A) of the memory tests studied here—will be easier if there is some degree of order, i.e., less entropy. Similarly, the greater the degree of order in an instrument (e.g., coherency in the mental processes of a person, B), the smaller the entropy and the greater the ability to perform tasks. How much information is carried by a word (or a unit) or any other message can be measured in terms of the concept of entropy in information theory [53]. Expressions to explain both a person's ability and task difficulty can be formulated by building on the pioneering work of Brillouin [21].

The amount of information transmitted from the measurement object to the observer can range from a simple signal to increasingly "meaningful" messages, as are captured in four levels of increasing richness in information theory [53,54]: syntax, semantics, pragmatics, and effectiveness.

As demonstrated previously [11,12,55,56], the concept of entropy plays a dominant role when formulating CSEs for nonverbal memory tasks, such as recalling the sequences of blocks (Knox Cube test and Corsi Block test) and sequences of digits (Digit Span test) of the simplest syntax. The present work attempts to extend this approach to include the next level of information content—semantics—exemplified in word list memory tests such as RAVLT.

CSEs such as these play an important role in the design and implementation of memory tests, since they allow:

i. the valid and reliable prediction of the difficulty of new test items (e.g., to fill measurement gaps);

ii. provide a measure of the degree of equivalence between items associated with distinct tests (e.g., different cultures and languages) [11,23,56,57].

New memory tests that combine carefully selected items from diverse sources can thus be formulated [56,58]. The increased number and comprehensiveness of items can be expected to lead to a reduction in the measurement uncertainties much-wanted in clinical decision-making.

Appendix B.1. Construct Specification Equations (CSE): Validity and Metrology

Once separate estimates of the task difficulty and person ability have been provided by the metrologically unique Rasch model (Equation (A2)), the formulation of a CSE can be made for each construct. Each construct estimated from Rasch transformation (Equation (A2)) of the test data can be understood causally in terms of a set of explanatory variables, as expressed in a CSE. The formulation of a CSE for an attribute of interest (Z, such as task difficulty δ or a person's ability θ as dependent variables) has been defined [31] as a linear combination of a set, k, of explanatory (independent) variables, X:

$$\hat{Z} = \sum_{k} \beta_k \cdot X_k \quad \text{A3}$$

Constructs such as task difficulty and a person's ability are quantities intended by themselves to be measured (i.e., measurands) as distinct from properties specific to the measurement process that determine the performance of each test. In other words, quantities attributed to the measurement object (item) and measurement instrument (person) as the key elements of the measurement system are also quantities of interest in their own right (cf. Section 1.3 in [37]).

CSEs provide advanced methods for testing construct theories by requiring the specificity and depth of causal explanations and understandings of the construct. In turn, CSEs are predictive tools in the design of new items complementing the existing tests and scales and for demonstrating item equivalence [11,23,56,57]. The CSE articulates a theory of item score variation and simultaneously provides a vehicle for the confirmation or falsification of the theory [49]. Closely linked to test and construct validation [31], this formulation is considered the most advanced level of construct theories derived with a conjoint model that relates a person's ability and task difficulty with a common unit (proposed by Stenner et al. [50]).

Thus, the overall aim of formulating CSEs is to encapsulate our best knowledge and understanding of each test. With that understanding, it becomes possible to validate existing tests, as well as to demonstrate the equivalence of apparently distinct test items, for instance, by them having the same entropy (as a measure of the information content) [11]. Ultimately, this can enable the design of new and better tests composed of a set of items that can distinguish between test persons more sensitively, a performance metric that is metrologically legitimated and can reduce, if possible, the burden on each test person by making tests more efficient and effective. We also regard the CSE as a "recipe" for producing "certified reference materials" for calibration of both the task difficulty and a person's ability [12,22,37].

The first choice of the metrological reference based on a Rasch analysis of ordinal observations is that of a CSE causally explaining an object's attribute—in the present case, task difficulty—independent of any individual test person according to Rasch's principle of specific objectivity [23]. What parameter to choose as the metrological reference is somewhat arbitrary, considering the symmetrical occurrence of the two parameters (item difficulty, δ, and person's ability, θ) in the basic Rasch formula (Equation (A1)), apart from a minus sign. Choosing the task difficulty in the present study is done for practical reasons, analogous in the quantitative metrology to associating a standard (etalon) with an attribute of a measured object (such as the mass of a simple and robust weight) rather than a measuring instrument, since the latter is arguably less suitable as a metrological reference because of its relative complexity and lack of robustness [38].

An initial choice of explanatory variables X_k in a CSE (Equation (A3)) using ab initio theory does not necessarily coincide with the principal components of variation in the empirical data. The state-of-the-art formulation of CSEs therefore includes the three steps of a principal component regression (PCR) [37]:

i. A PCA amongst the set of explanatory variables, $\boldsymbol{X}_k$ (not to be confused with the PCA of item fit residuals commonly used to examine the assumed unidimensionality of item attributes [46,59]);
ii. A linear regression of the empirical task difficulty values δ_j against $\boldsymbol{X}' = \boldsymbol{X} \cdot \boldsymbol{P}$ in terms of the principal components, $\boldsymbol{P}$;
iii. A conversion back from the principal components to the explanatory variables, X_k.

The results of this PCR reveal how much each explanatory variable contributes to explaining and predicting the observed variations in the attribute of interest by providing estimates of the coefficients β_k in the CSE of Equation (A3).

Green and Smith [31] gave a comprehensive account of the assumptions behind a CSE formulation of the Rasch attributes (they denoted the formulation of the linear logistic test model (LLTM)):

- ***Unidimensionality*** *is an assumption of the Rasch model. A problem with the straightforward use of regression is that it will be of marginal value unless the items used form a cohesive set; that is, that the same underlying variable explains the response to every item in that set.*
- *The latent trait (such as task difficulty) can consist of* ***more than one component*** *(which may or may not belong to one and the same dimension).*
- *The latent trait identified should be a* ***valid*** *measure of the construct being assessed.*

The present case study addresses mainly the validity, while we defer considerations of dimensionality and a componential analysis to the final case study (Section 5).

Appendix B.2. Entropy and CSE for Word Learning Tests

Entropy has a broad applicability when formulating CSE in general [11,23,37] and is simply and generally described as a measure of the amount of "useful" information or "useful" energy [40]; a higher entropy implies less order, leading to a loss of information or energy and vice versa, and a lower entropy implies a higher order and less uncertainties (Appendix A). The word learning list test RAVLT IR provides a conceptually simple example with which to test our theories further but is one step more advanced semantically than our previous syntax studies of nonverbal, block, and number recalls [11,12,37,55].

The information theoretical entropy, $H_j = H(Z)$, in Equation (A1), which is a measure of the amount of information in "messages" of this kind (G symbols with n repeats), can be based on the well-known Shannon [60] expression of "surprisal" in the work of Brillouin [21], which we equate to the difficulty δ of the task, j:

$$\delta_j = M \cdot \left[ln\left(G_j!\right) - \sum_{a_j} ln\left(N_{a_j}!\right) \right] \quad \text{A4}$$

where M is a normalisation constant according to Brillouin's Section 2. Unit systems [21]. In our previous work, informational entropy using this formulation was shown to be a dominant explanatory variable for the task difficulty for several of the most elementary memory tests with language- and cultural-free items (blocks and digits) [11,12].

Based on Equation (A4), the average entropy at the middle of a list of G words (without repeats and SPE effects) is a mid-region task difficulty:

$$\delta_{Mr} = M \cdot ln\left(G_j!\right);\ G = L/2 \quad \text{A5}$$

The task difficulty would be expected, according to Equation (A5), to increase in proportion to $ln(G_j!)$, where G is the number of symbols encountered in a message, j. (In the case of an odd list length, L, the factorial expression $ln\left(\frac{L}{2}!\right)$ is evaluated by rounding $\frac{L}{2}$ to the nearest integer.)

Appendix B.3. Linguistic Effects on Task Difficulty in Word Learning List Tests

Word familiarity is another effect that may explain the difficulties of recalling words from a list. Again, for the word learning list test RAVLT IR, the task difficulty is predicted, according to the Shannon entropy model, to decrease in proportion to the lnf_j, where f is the frequency (i.e., how often) each word in (in our case, German) language occurs [61]:

$$\delta_{freq,j} = -M \cdot lnf_j \quad \text{A6}$$

that is, the more frequent a word is, the less the entropy and less difficult to recall.

Appendix B.4. Serial Position Effects and Task Difficulty

Over and above the inherent difficulty of recalling any symbol in a sequence, a peculiarity of RAVLT is that it is expected that, all other factors being equal, the initial and final symbols in the word list should be somewhat easier to recall than the symbols in the middle of the sequence, according to the well-known effects of primacy and recency.

In the present work, we propose that the same basic entropy model as in Equation (A4) can also be used to explain the SPE for the verbal test (L words) RAVLT IR; that is, the fact that it is easier to recall words from the beginning and the end of a list, j:

$$\delta_{j,Pr} = -M \cdot ln(G_j!);\ G = item\ order \quad \text{A7}$$

$$\delta_{j,Rr} = -M \cdot ln(G_j!);\ G = L - 1 - item\ order \quad \text{A8}$$

The basic task difficulty of recalling any word in the list, as given by Equation (A5), is reduced for each SPE by the corresponding reduction in entropy according to Equations (A7) and (A8) for primacy and recency, respectively. The mid-region term, Equation (A5), therefore is entered twice in the overall CSE: once for Pr and once for Rr (see Equation (2)).

Appendix C. Effects of Changes in Task Difficulty on the Rasch Model's Goodness of Fit

While one can argue that the Rasch model would be more appropriate for assuring the quality of any new diagnostic tool, the model's unique metrological properties (e.g., [62–66]) need to be confirmed in the presence of a SPE, such as primacy. The assumption of specific objectivity (Appendix A.1) might be challenged if significant portions of the cohorts appear to experience SPE differently according to the cognitive status [29,46]. Maintaining the unique properties of the Rasch model is examined in Section 5; once again, an entropy-based approach will be found useful. As part of this investigation, novel explanations of principal component analysis (PCA) loading plots in terms of scale distortion and entropy when detecting disease-related effects will be given. Investigations are made of factors, such as loading, which are common to two kinds of PCAs in formulating the CSE for recall task difficulty (Section 4) and in judging the goodness of fit of logistic regressions of the Rasch model (Section 5).

Group-dependent effects on the task difficulty of the kind mentioned in the proposed SPE-based diagnostic (Section 1) are a direct challenge to the specific objectivity assumption of the basic Rasch model, making separate estimates of each item difficulty (irrespective of the test person) and each person's ability (irrespective of item) more difficult to acheive. Maintaining the unique metrological properties of the Rasch model (Appendix A.1), and thereby of any candidate diagnostic tool analysed with it, therefore depends on testing for the effects of the SPE. In addition to the changes in the CSE for task

difficulty studied in the second case study, evidence of the SPE-related changes in response may also be sought in tests for Rasch modelling and goodness of fit, such as construct alleys and PCA loading plots.

An effect of a shift $\Delta\delta_{SPE}$ in the task difficulty, such as observed in the CSE of different cohort groups (Equations (4) and (5)), will lead, in turn, to a change, $\Delta P_{success}$, in the response—that is, a change in the probability of making a correct classification. In any measurement system (Appendix A), the instrument (B) has a certain sensitivity, $K = \frac{\partial P_{success}}{\partial \delta}$—that is, how much each test person responds to a task of a certain difficulty, δ. According to the MSA approach, the amount of response change depends on both the magnitude of the change in difficulty, as well as the sensitivity K at any given level of task difficulty. A change in the task difficulty will lead to a change in the response: $\Delta P_{success} = -\frac{\partial P_{success}}{\partial \delta} \cdot \Delta\delta$ through a simple partial differentiation.

As examined in the rest of this paper, the evidence for maintaining the unique metrological properties of the Rasch model, including the possible effects of multidimensionality, can be based on the goodness of fit—that is, when the values of the two parameters—person attribute, θ, and item attribute, δ—are adjusted conjointly in a logistic regression of Equation (1) to the score data $x_{i,j}$ for each test person (i) and item (j) on an ordered category scale by minimising the sum of the squared differences of the logistic fit residuals:

$$\sum_{i=1}^{N_{TP}} \sum_{j=1}^{L} \left(x_{i,j} - P_{success,i,j}\right)^2$$

Any change in the response, $\Delta P_{success}$, should be detectable in tests based on the logistic fit residuals.

Appendix C.1. Principal Components and Multidimensionality

In any PCA of a set of variables **X**:

- First PC: $PC_1 = \sum_{k=1}^{K} a_{1,k} \cdot X_k$, which maximises the $\mathrm{Var}(PC_1)$ subject to: $\sum_{k=1}^{K} a_{1,k}^2 = 1$.
- Second PC: $PC_2 = \sum_{k=1}^{K} a_{2,k} \cdot X_k$, which maximises the $\mathrm{Var}(PC_2)$ subject to: $\sum_{k=1}^{K} a_{2,k}^2 = 1$, etc.

The different principal components (PC) are orthogonal, i.e., $\mathrm{Cov}(PC_1, PC_2) = 0$ (rather than the explanatory variables, X, which might be correlated) and are ordered in terms of the decreasing variance: $Var(PC_2) \leq Var(PC_1)$. PCA loading is well-known to be $L_{p,k} = \mathrm{Cov}(PC_p, X_k) \propto a_{p,k}$, where:

$$\sum_{p=1}^{P} L_{p,k}^2 = \lambda_p = Var(PC_p)$$

and where opposite signs in PCA loading, $L_{p,k}$, are said to indicate opposite degrees of correlation between the two components of variation.

The two kinds of principal component analyses (PCA) relevant in the present study are:

1. "**PCA1**" performed as the first step in a PCR (Section 4.1) when forming a CSE dealing directly with the Rasch parameters, such as task difficulty, δ, and their explanatory variables X_k. The lack of orthogonality between the X_k gives rise to the loading of these in each PC variation; see L_{px} in Section 5.2.1.
2. "**PCA2**" based on the logistic regression of Equation (1) of the fit residuals (Section 5.2) dealing indirectly with the Rasch parameters, since the response (raw data) can be modified by instrument (person) sensitivity, as studied below in Appendix C.2.

The evidence for any of the effects, such as multidimensionality mentioned by Green and Smith [31] (Appendix B.1), when maintaining the unique metrological properties of the Rasch model (Appendix A.1) can be sought in such PCAs. Linacre [34], for instance,

summarised the quandary in the choice of which type of residual to employ when assessing the multidimensionality in the Rasch model.

Our approach does not put the two kinds of PCA in opposition to each other, but rather, the results from the two can be expected to yield results that are connected to a certain extent. If there is more than one explanatory variable revealed in a PCA1 as the first step in formulating a CSE with a PCR, it is also likely to be accompanied by indications of a second dimension in item residual PCA2 (where the first PC is the primary Rasch attribute).

One could surmise that, for instance, the item task difficulty might depend on more than one factor, perhaps a second "dimension" connected with another cause of difficulty, such as causes SPE, to which some cohort members—perhaps according to their cognitive status—are more discriminating against than others and that might scale differently (Section 5.1). SPE such as primacy and recency can give rise to additional dimensions and scale extensions in the response, as exemplified in the analysis of the experimental findings of the present work (Section 4.2).

Appendix C.2. PCA2 Rasch Logistic Regression Residuals

The second kind of principal component analysis (**PCA2**, Appendix C.1) consists of a factor analysis on the standardised residuals (differences between the observed and expected values divided by the square root of the variance), thereby identifying patterns in the logistic regression residuals suggestive of multiple dimensions [34], as in earlier studies—for instance, by Hoffman et al. [67].

In the Rasch approach, the values of the two parameters: person attribute, θ_i (N_{TP} test persons), and item attribute, δ_j (L items), are adjusted conjointly in a logistic regression of the Rasch formula (Equation (1)) to the raw score data $x_{i,j}$ on an ordered category scale by minimising the sum of the squared differences of the logistic fit residuals for all test persons, i, and all items, j.

Measures of the misfit are in terms of the sums of squared residuals, ${\chi_j}^2 = \sum_{i=1}^{N_{TP}} {y_{i,j}}^2$ of the logistic regression, where the residual $y_{i,j} = x_{i,j} - \mathbb{E}_{i,j}$ can be examined for either items (j) or test persons (i) [34].

For each person (i)–item (j) combination: the Rasch dichotomous expected score: $\mathbb{E}_{i,j} = \frac{1}{1 + e^{(\theta_i - \delta_j)}}$ and model variance of the expected score: $\mathbb{W}_{i,j} = \mathbb{E}_{i,j} \cdot (1 - \mathbb{E}_{i,j})$ are computed. The expected scores and associated variances are averaged, for instance, over the cohort, i, of the test persons:

- Sum of squares of the explained variance is: $\sum_i \left[\left(\mathbb{E}_{i,j} - \overline{\mathbb{E}}_\iota \right)^2 \right]$;
- Sum of squares of the unexplained variance is: $\sum_i [\mathbb{W}_{i,j}]$;
- Explained variance = $\frac{\sum_i \left[\left(\mathbb{E}_{i,j} - \overline{\mathbb{E}}_\iota \right)^2 \right]}{\sum_i \left[\left(\mathbb{E}_{i,j} - \overline{\mathbb{E}}_\iota \right)^2 \right] + \sum_i [\mathbb{W}_{i,j}]^2}$.

The relative weightings (i.e., PCA loadings) of the PC of variation should be apparent from a PCA loading plot (Figure 6), where:

- x-axis: task difficulty = entropy $H_j = H(Z; A)$ in Equation (A1), evaluated with Equation (9) or the directly measured experimental values;
- y-axis: PCA loading $L_{p,k}$, (Section 5.2).

The practical importance of these patterns can then be compared with the dominant dimensions estimated by the Rasch formula. The terminology used in a traditional PCA is exemplified in the following quote describing the bifactor model of Reise [68]: *"…it is assumed that the general and group factors all are orthogonal. Substantively, the general factor represents the conceptually broad "target" construct an instrument was designed to measure, and the group factors represent more conceptually narrow subdomain constructs*."

References

1. Weitzner, D.S.; Calamia, M. Serial Position Effects on List Learning Tasks in Mild Cognitive Impairment and Alzheimer's Disease. *Neuropsychology* **2020**, *34*, 467–478. https://doi.org/10.1037/neu0000620.
2. Murdock, B.B. The Serial Position Effect of Free Recall. *J. Exp. Psychol.* **1962**, *64*, 482–488. https://doi.org/10.1037/h0045106.
3. Bayley, P.J.; Salmon, D.P.; Bondi, M.W.; Bui, B.K.; Olichney, J.; Delis, D.C.; Thomas, R.G.; Thal, L.J. Comparison of the Serial Position Effect in Very Mild Alzheimer's Disease, Mild Alzheimer's Disease, and Amnesia Associated with Electroconvulsive Therapy. *J. Int. Neuropsychol. Soc.* **2000**, *6*, 290–298. https://doi.org/10.1017/S1355617700633040.
4. Blachstein, H.; Vakil, E. Verbal Learning across the Lifespan: An Analysis of the Components of the Learning Curve. *Aging Neuropsychol. Cogn.* **2016**, *23*, 133–153. https://doi.org/10.1080/13825585.2015.1063579.
5. Bruno, D.; Reiss, P.T.; Petkova, E.; Sidtis, J.J.; Pomara, N. Decreased Recall of Primacy Words Predicts Cognitive Decline. *Arch. Clin. Neuropsychol.* **2013**, *28*, 95–103.
6. Carlesimo, G.A.; Sabbadini, M.; Fadda, L.; Caltagirone, C. Different Components in Word-List Forgetting of Pure Amnesics, Degenerative Demented and Healthy Subjects. *Cortex* **1995**, *31*, 735–745. https://doi.org/10.1016/S0010-9452(13)80024-X.
7. Cunha, C.; Guerreiro, M.; de Mendonça, A.; Oliveira, P.E.; Santana, I. Serial Position Effects in Alzheimer's Disease, Mild Cognitive Impairment, and Normal Aging: Predictive Value for Conversion to Dementia. *J. Clin. Exp. Neuropsychol.* **2012**, *34*, 841–852. https://doi.org/10.1080/13803395.2012.689814.
8. Howieson, D.B.; Mattek, N.; Seeyle, A.M.; Dodge, H.H.; Wasserman, D.; Zitzelberger, T.; Kaye, J.A. Serial Position Effects in Mild Cognitive Impairment. *J. Clin. Exp. Neuropsychol.* **2011**, *33*, 292–299. https://doi.org/10.1080/13803395.2010.516742.
9. Talamonti, D.; Koscik, R.; Johnson, S.; Bruno, D. Predicting Early Mild Cognitive Impairment With Free Recall: The Primacy of Primacy. *Arch. Clin. Neuropsychol.* **2020**, *35*, 133–142. https://doi.org/10.1093/arclin/acz013.
10. Hurlstone, M.J.; Hitch, G.J.; Baddeley, A.D. Memory for Serial Order across Domains: An Overview of the Literature and Directions for Future Research. *Psychol. Bull.* **2014**, *140*, 339–373. https://doi.org/10.1037/a0034221.
11. Melin, J.; Cano, S.; Pendrill, L. The Role of Entropy in Construct Specification Equations (CSE) to Improve the Validity of Memory Tests. *Entropy* **2021**, *23*, 212. https://doi.org/10.3390/e23020212.
12. Melin, J.; Pendrill, L.R.; Cano, S.J.; EMPIR NeuroMET 15HLT04 consortium Towards Patient-Centred Cognition Metrics. *J. Phys. Conf. Ser.* **2019**, *1379*, 012029. https://doi.org/10.1088/1742-6596/1379/1/012029.
13. Quaglia, M.; Cano, S.; Fillmer, A.; Flöel, A.; Giangrande, C.; Göschel, L.; Lehmann, S.; Melin, J.; Teunissen, C.E. The NeuroMET Project: Metrology and Innovation for Early Diagnosis and Accurate Stratification of Patients with Neurodegenerative Diseases. *Alzheimer's Dement.* **2021**, *17*, e053655. https://doi.org/10.1002/alz.053655.
14. Wirth, M.; Schwarz, C.; Benson, G.; Horn, N.; Buchert, R.; Lange, C.; Köbe, T.; Hetzer, S.; Maglione, M.; Michael, E.; 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.* **2019**, *11*, 36. https://doi.org/10.1186/s13195-019-0484-1.
15. Rey, A. *L'examen Clinique En Psychologie. [The Clinical Examination in Psychology.]*; Presses Universitaries De France: Paris, France, 1958; p. 222. (In French)
16. Rasch, G. *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*; Danmarks pædagogiske Institut: Copenhagen, Denmark, 1960.
17. Liu, X.; Fu, Z. A Novel Recognition Strategy for Epilepsy EEG Signals Based on Conditional Entropy of Ordinal Patterns. *Entropy* **2020**, *22*, 1092. https://doi.org/10.3390/e22101092.
18. Foldi, N.S.; Brickman, A.M.; Schaefer, L.A.; Knutelska, M.E. Distinct Serial Position Profiles and Neuropsychological Measures Differentiate Late Life Depression from Normal Aging and Alzheimer's Disease. *Psychiatry Res.* **2003**, *120*, 71–84. https://doi.org/10.1016/S0165-1781(03)00163-X.
19. Hermann, B.P.; Seidenberg, M.; Wyler, A.; Davies, K.; Christeson, J.; Moran, M.; Stroup, E. The Effects of Human Hippocampal Resection on the Serial Position Curve. *Cortex* **1996**, *32*, 323–334. https://doi.org/10.1016/S0010-9452(96)80054-2.
20. Rue, A.L.; Hermann, B.; Jones, J.E.; Johnson, S.; Asthana, S.; Sager, M.A. Effect of Parental Family History of Alzheimer's Disease on Serial Position Profiles. *Alzheimer's Dement.* **2008**, *4*, 285–290. https://doi.org/10.1016/j.jalz.2008.03.009.
21. Brillouin, L. *Science and Information Theory*; 2nd ed.; Dover Publications: New York, NY, USA, 1962.
22. Melin, J.; Cano, S.J.; Göschel, L.; Fillmer, A.; Lehmann, S.; Hirtz, C.; Flöel, A.; Pendrill, L.R. Metrological References for Person Ability in Memory Tests. *Meas. Sens.* **2021**, *18*, 100289. https://doi.org/10.1016/j.measen.2021.100289.
23. Melin, J.; Pendrill, L. The Role of Construct Specification Equations (CSE) and Entropy in the Measurement of Memory. In *Person Centered Outcome Metrology*; Springer: New York, NY, USA, 2022.
24. Melin, J.; Kettunen, P.; Wallin, A.; Pendrill, L. Entropy-Based Explanations of Serial Position and Learning Effects in Ordinal Responses to Word List Tests. 2022. *accepted for IMEKO-MATHMET, Porto.*
25. Linacre, J. Dimensionality: When Is a Test Multidimensional? Winsteps Help. Available online: https://www.winsteps.com/winman/dimensionality.htm (accessed on 5 June 2021).
26. Raîche, G. Critical Eigenvalue Sizes (Variances) in Standardized Residual Principal Components Analysis. *Rasch Meas. Trans.* **2005**, *2005*, 1012.
27. Hagell, P. Testing Rating Scale Unidimensionality Using the Principal Component Analysis (PCA)/*t*-Test Protocol with the Rasch Model: The Primacy of Theory over Statistics. *Open J. Stat.* **2014**, *4*, 456–465. https://doi.org/10.4236/ojs.2014.46044.
28. Smith, E.V. Detecting and Evaluating the Impact of Multidimensionality Using Item Fit Statistics and Principal Component Analysis of Residuals. *J. Appl. Meas.* **2002**, *3*, 205–231.

29. Melin, J.; Regnault, A.; Cano, S.; Pendrill, L. Neuropsychological Assessments: Word Learning Tests and Diagnostic Potential of Serial Position Effects. In Proceedings of the International Metrology Congress, Lyon, France, 7 September 2021.
30. Pendrill, L.; Petersson, N. Metrology of Human-Based and Other Qualitative Measurements. *Meas. Sci. Technol.* **2016**, *27*, 094003. https://doi.org/10.1088/0957-0233/27/9/094003.
31. Green, K.E.; Smith, R.M. A Comparison of Two Methods of Decomposing Item Difficulties. *J. Educ. Stat.* **1987**, *12*, 369–381. https://doi.org/10.2307/1165055.
32. Wright, B.; Stone, M. *Best Test Design*; MESA Press: San Diego, CA, USA, 1979.
33. Cano, S.; Barrett, L.; Zajicek, J.; Hobart, J. Dimensionality Is a Relative Concept. *Mult. Scler.* **2011**, *17*, 893–894. https://doi.org/10.1177/1352458511406910.
34. Linacre, J.M. Detecting Multidimensionality: Which Residual Data-Type Works Best?. *J. Outcome Meas.* **1998**, *2*, 266–283.
35. Feuerstahler, L.; Wilson, M. Scale Alignment in the Between-Item Multidimensional Partial Credit Model. *Appl. Psychol. Meas.* **2021**, *45*, 268–282. https://doi.org/10.1177/01466216211013103.
36. Choi, I.-H.; Wilson, M. Multidimensional Classification of Examinees Using the Mixture Random Weights Linear Logistic Test Model. *Educ. Psychol. Meas.* **2015**, *75*, 78–101. https://doi.org/10.1177/0013164414522124.
37. Pendrill, L. *Quality Assured Measurement: Unification across Social and Physical Sciences*; Springer Series in Measurement Science and Technology; Springer International Publishing: Cham, Switzerland, 2019; ISBN 978-3-030-28694-1.
38. Melin, J. Neurogenerative Disease Metrology and Innovation: The European Metrology Programme for Innovation & Research (EMPIR) and the NeuroMET Projects. In Proceedings of the Pacific RIM Objective Measurement Symposium, Nanjing, China, 4–6 December 2021.
39. ASTM International. *E11 Committee Guide for Measurement Systems Analysis (MSA)*; ASTM International: West Conshohocken, PA, USA, 2022.
40. Carnot, L. *Principes Fondamentaux de L'équilibre Et Du Mouvement; Par L.-N.-M. Carnot;* Hachette Livre-Bnf: Paris, France, 2016. (In French)
41. Rossi, G.Battista. *Measurement and Probability [Elektronisk Resurs] A Probabilistic Theory of Measurement with Applications*; Springer Netherlands: Dordrecht, The Netherlands, 2014; ISBN 978-94-017-8825-0.
42. Pearson, K. Mathematical Contributions to the Theory of Evolution—On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs. *Proc. R. Soc. Lond.* **1896**, *60*, 489–498.
43. Jones, L.V. *The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis 1949-1964, Volume III*, 1st ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 1986; ISBN 978-0-412-74250-7.
44. Mosteller, F.; Tukey, J. *Data Analysis and Regression: A Second Course in Statistics*, 1st ed.; Pearson: London, UK, 1977; ISBN 978-0-201-04854-4.
45. Linacre, J. Bernoulli Trials, Fisher Information, Shannon Information and Rasch. *Rasch Meas. Trans.* **2006**, *20,* 1062–1063.
46. Pendrill, L.R.; Melin, J.; Cano, S.J. Entropy-Based Explanations of Multidimensionality in Ordinal Responses. In Joint Workshop of ENBIS and MATHMET, Virtual 2021. Available online: https://drive.google.com/drive/folders/1zVlzy3SKvbdxBdiCqe2AyDGLH7hfa-Ty (accessed on 1 July 2022)
47. Moradi, E.; Hallikainen, I.; Hänninen, T.; Tohka, J. Rey's Auditory Verbal Learning Test Scores Can Be Predicted from Whole Brain MRI in Alzheimer's Disease. *NeuroImage Clin.* **2017**, *13*, 415–427. https://doi.org/10.1016/j.nicl.2016.12.011.
48. Turetsky, V.; Bashkansky, E. Ordinal Response Variation of the Polytomous Rasch Model. *METRON* **2022**, 1–26. https://doi.org/10.1007/s40300-022-00229-w.
49. Stenner, A.J.; Smith, M. Testing Construct Theories. *Percept. Mo.t Ski.* **1982**, *55*, 415–426. https://doi.org/10.2466/pms.1982.55.2.415.
50. Stenner, A.J.; Smith, M.; Burdick, D.S. Toward a Theory of Construct Definition. *J. Educ. Meas.* **1983**, *20*, 305–316.
51. Hughes, L.F.; Perkins, K.; Wright, B.D.; Westrick, H. 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. Alzheimer's Dis.* **2003**, *5*, 367–373. https://doi.org/10.3233/JAD-2003-5503.
52. Hobart, J.; Cano, S.; Posner, H.; Selnes, O.; Stern, Y.; Thomas, R.; Zajicek, J.; Alzheimer's Disease Neuroimaging Initiative. Putting the Alzheimer's Cognitive Test to the Test II: Rasch Measurement Theory. *Alzheimer's Dement.* **2013**, *9*, S10–S20. https://doi.org/10.1016/j.jalz.2012.08.006.
53. Shannon, C.; Weaver, W. *The Mathematical Theory of Communication*; The University of Illinois Press: Urbana, IL, USA, 1964.
54. Klir, G.J.; Folger, T.A. *Fuzzy Sets, Uncertainty and Information*, 1st ed.; Prentice Hall: Englewood Cliffs, NJ, USA, 1988; ISBN 978-0-13-345984-5.
55. Pendrill, L.; Melin, J.; Cano, S.; the EMPIR NeuroMET 15HLT04 consortium. Metrological References for Health Care Based on Entropy. In Proceedings of the 19th International Congress of Metrology (CIM2019), Paris, France, 24–26 September 2019; Gazal, S., Ed.; EDP Sciences: Paris, France, 2019; p. 07001.
56. Melin, J.; Cano, S.J.; Flöel, A.; Göschel, L.; Pendrill, L.R. Construct Specification Equations: 'Recipes' for Certified Reference Materials in Cognitive Measurement. *Meas. Sens.* **2021**, *18*, 100290. https://doi.org/10.1016/j.measen.2021.100290.
57. Stenner, A.J.; Burdick, H.; Sanford, E.E.; Burdick, D.S. How Accurate Are Lexile Text Measures? *J. Appl. Meas.* **2006**, *7*, 307–322.
58. Melin, J.; Cano, S.; Flöel, A.; Göschel, L.; Pendrill, L.; EMPIR NeuroMET and NeuroMET2 consortiums More than a Memory Test: A New Metric Linking Blocks, Numbers, and Words. *Alzheimer's Dement.* **2021**, *17*, e050291. https://doi.org/10.1002/alz.050291.

59. Linacre, J. Rasch Analysis First or Factor Analysis First? *Rasch Meas. Trans.* **1998**, *11*, 603
60. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
61. Linguistic Data Consortium (LDC). *Web 1T 5-Gram, 10 European Languages*, Version 1; Catalog Number LDC2009T25; Linguistic Data Consortium (LDC): Philadelphia, PA, USA, 2009; ISBN 1-58563-525-1.
62. Hobart, J.C.; Cano, S.J.; Zajicek, J.P.; Thompson, A.J. Rating Scales as Outcome Measures for Clinical Trials in Neurology: Problems, Solutions, and Recommendations. *Lancet Neurol.* **2007**, *6*, 1094–1105. https://doi.org/10.1016/S1474-4422(07)70290-9.
63. Hobart, J.; Cano, S. Improving the Evaluation of Therapeutic Interventions in Multiple Sclerosis: The Role of New Psychometric Methods. *Health Technol. Assess.* **2009**, *13*, 214.
64. Wright, B.D. A History of Social Science Measurement. *Educ. Meas. Issues Pract.* **1997**, *16*, 33–45. https://doi.org/10.1111/j.1745-3992.1997.tb00606.x.
65. Andrich, D. Controversy and the Rasch Model: A Characteristic of Incompatible Paradigms? *Med. Care* **2004**, *42*, I–7. https://doi.org/10.1097/01.mlr.0000103528.48582.7c.
66. Cano, S.J.; Pendrill, L.R.; Melin, J.; Fisher, W.P. Towards Consensus Measurement Standards for Patient-Centered Outcomes. *Measurement* **2019**, *141*, 62–69. https://doi.org/10.1016/j.measurement.2019.03.056.
67. Hoffman, R.W.; Bezruczko, N.; Perkins, K. An External Validation Study of a Classification of Mixed Connective Tissue Disease and Systemic Lupus Erythematosus Patients. *J. Appl. Meas.* **2012**, *13*, 205–216.
68. Reise, S.P. The Rediscovery of Bifactor Measurement Models. *Multivar. Behav. Res.* **2012**, *47*, 667–696. https://doi.org/10.1080/00273171.2012.715555.