Proceedings of the RaPID-4 @LREC 2022, pages 17–21
Marseille, 25 June 2022

A Novel Metrological Approach to a More Consistent Way of Defining and Analyzing Memory Task Difficulty in Word Learning List Tests with Repeated Trials

Melin J [1], Pendrill L [1]
[1] RISE Research Institutes of Sweden, Division Safety and transport, Department Measurement Science and Technology
Jeanette.melin@ri.se, Leslie.pendrill@ri.se

Abstract

New candidate diagnostics for cognitive decline and dementia have recently been proposed based on effects such as primacy and recency in word learning memory list tests. The diagnostic value is, however, currently limited by the multiple ways in which raw scores, and in particular these serial position effects (SPE), have been defined and analyzed to date. In this work, we build on previous analyses taking a metrological approach to the 10-item word learning list. We show i) how the variation in task difficulty reduces successively for trials 2 and 3, ii) how SPE change with repeated trials as predicted with our entropy-based theory, and iii) how possibilities to separate cohort members according to cognitive health status are limited. These findings mainly depend on the test design itself: A test with only 10 words, where SPE do not dominate over trials, requires more challenging words to increase the variation in task difficulty, and in turn to challenge the test persons. The work is novel and also contributes to the endeavour to develop for more consistent ways of defining and analyzing memory task difficulty, and in turn opens up for more practical and accurate measurement in clinical practice, research and trials.

Keywords: Cognition, Word recall, Item response theory, Entropy, Metrology

1. Introduction

Measurement of the memory ability of persons has a long tradition in neuropsychological assessment. Tests used to measure a person's memory ability typically include language- and cultural-free blocks and digits recall as well as more complex word recalling sequences. Recently, improved diagnostics for cognitive decline and dementia, particularly when including serial position effects (SPE), have been sought when measuring memory abilities based on word learning lists (see summary by Weitzner & Calamia (2020)).

SPE address the relationship between the ordering of symbols (in the present case, words) in a list and the likelihood of them being recalled. Specifically, when a test person is asked to freely recall as many words as possible from a word list, SPE mean that the first (primacy region, *Pr*) and the last (recency region, *Rr*) words are easier to remember than items in the middle (middle region, *Mr*) (Murdock, 1962). In a recent review, Weitzner & Calamia (2020) conclude that: '*The analysis of SPE 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 SPE are defined and analyzed*.' Despite the limitations, they found that individuals with MCI and AD showed reduced primacy and intact recency, with primacy being more reduced in AD.

In line with that, there are, to our best knowledge, few studies which properly handle the ordinal response of a test person taking a word learning list test, making any claim of a new diagnostic questionable. Our previous analyses of the *Rey's Auditor Verbal Learning List Test* (RAVLT) trial 1/immediate recall (IR) have challenged previous claims of disease-related changes in serial positions effects (SPE), in particular putting those claimed changes in relation to measurement uncertainty (Melin, et al., 2021a; Pendrill et al., 2021). Our analyses of word learning list tests so far have focused on the first trial, while the present work extends our study to include more trials repeated directly after each other, including learning effects, as well as delayed recall (DR).

An important part of ensuring construct validity and predictability is to explain how the difficulty of recalling is caused by a number of effects, particularly how the word list items are structured. A major result of our research so far, both of non-verbal, culture-free tests such as block or digit sequence tests, as well as the verbal lists studied here, has been to explain recall difficulty in terms of informational entropy (see section 2.2). It should be easier to recall a more ordered sequence of less entropy.

Our previous studies of IR have included *frequency* (i.e., how frequently each word occurs in its language) as an explanatory variable, although it is found to contribute little to item task difficulty compared with the major contributions from the sequence length, i.e., the number of symbols (words) in each list (Melin et al., 2021a; Pendrill et al., 2021). The minor contribution from word frequency might however be due to the fact that the words in the list studied are all very short and common in everyday language, and therefore not expected to lead to any significant variation in recall difficulty.

In contrast to RAVLT with 15 words with a fixed order on repeated trials 1 - 5, the word learning list (WLL) test included in the CERAD test battery has only 10 words and the word order changes with each of the three repeated trials. With only 10 words, SPE are expected to be less pronounced (Murdock, 1962) but repeated trials may

include learning effects similar to RAVLT (Goldberg et al., 2015; Zhan et al., 2018).

The European NeuroMET2 18HLT09 project has brought together clinicians, academics, metrologists and industry to address measurement challenges in current neurodegenerative diseases. Our part in NeuroMET includes how to properly handle cognitive data and in this paper we will present how task difficulty and SPE change with repeated trials in word recalling tests, as predicted with our entropy-based theory.

2. Methods

2.1 Participants and data collection

The NeuroMET cohort has been recruited and tested bi-annually from 2016 to 2022 at Charité hospital in Berlin. Measurements administered include neuropsychological assessments with a battery of legacy cognitive tests, clinical laboratory data for protein biomarkers and ultra-high field magnetic resonance imaging and spectroscopy (Quaglia et al., 2021).

For this work, data have been included from baseline and follow-up visits from the WLL CERAD cognitive tests (German) from 214 individual assessments of healthy controls (HC, n=73), persons with subjective cognitive decline (SCD, n=44) as well as patients with mild cognitive impairment (MCI, n=43) and suspected dementia due Alzheimer's Disease (AD, n=54).

In trial 1 of WLL CERAD, the test person is asked to freely recall as many as possible of the 10 common but unrelated words read by the test leader. In the second trial, the same 10 words are repeated but in a different order and the person is again asked to freely recall as many as possible. This is then repeated in a third trial, again with a different word order.

The study was approved by the Ethics Committee of the Charité - Universitätsmedizin Berlin, Germany, and was conducted in accordance with the declaration of Helsinki.

2.2 Data analyses

The ordinal responses (raw scores) to the WLL CERAD (classification number 1 for pass or classification number 0 for fail) were restituted through a logistic regression of the data to a dichotomous Rasch (1960) model using the WINSTEPS ® 5.2.0. This restitution process yields separate and linear measures for each memory task difficulty, δ, and individual person memory ability, θ, and compensates for ordinality:

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

The focus of this study is primarily on measures of memory task difficulties, δ.

Secondly, a state-of-the-art multivariate formulation is made of a construct specification equation (CSE) (Pendrill, 2019) for the quantity Z of the construct (in this case memory task difficulty, δ), expressed as a sum of a number of covariates, $\boldsymbol{X_k}$ (explanatory variables) in the causal associative relation: $Z = \sum_k \beta_k \cdot X_k$.

Explanatory variables $\boldsymbol{X_k}$ were identified in line with our previous work on RAVLT IR (Melin et al., 2021a; Pendrill et al., 2021) based on information theoretical entropy. In this case, the amount of information in these messages (G symbols with N repeats) according to the well-known Shannon (1948) expression of 'surprisal' in the work of Brillouin (1962), is given by:

$$I = M \cdot \left[ln(G!) - \sum_{j=1}^{N} ln(N_j!) \right]$$

where the normalisation constant, $M = \frac{1}{ln(G)}$

This general expression gave us the following definitions for explanatory variables for the different contributions to memory IR task difficulty for each word, j:

$$\delta_{Mr,j} = 2 \cdot M \cdot ln(G_j!);\ G = L/2$$

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

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

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

Finally, formulation of a CSE for overall task difficulty (Pendrill 2019) for each trial included three steps in a principal component regression (PCR):

i. A PCA amongst the set of explanatory variables, $\boldsymbol{X_k}$, using the entropy-based estimates of δ given above
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, $\boldsymbol{P}$; and
iii. A conversion back from principal components to the explanatory variables, $\boldsymbol{X_k}$

3. Results

3.1 Overall task difficulty

Figure 1 presents how task difficulty for individual items is found empirically to change over the three trials. On the y-axis lower values imply an easier task and vice versa, and the x-axis represents each item in order of appearance in each trial.

Blue dots represent trial 1 with a clear parabolic fit line, indicating easier tasks in the beginning and at the end, i.e., the SPE for *Pr* and *Rr*, and qualitatively similar to our earlier RAVLT observations. The variation in task difficulty with order is successively reduced for trials 2 and 3 (orange and grey dots in figure 1). Overall task difficulty also decreases with the three repeated trials.

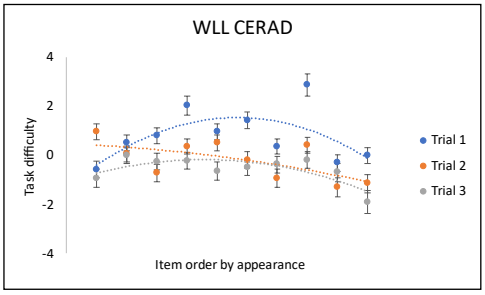


Figure 1. Empirical task difficulty values on the y-axis (lower values implies an easier task), and the x-axis represent each item ordered by appearance in each trial. Error bars show measurement uncertainties with coverage factor *k*=2.

Smaller contributions from SPE to task difficulty in trials 2 and 3 are also confirmed by the CSEs formulated as described in section 2.2, yielding the following expressions for the three different trials:

$$zR_{WLL1,j} = 6(5) + 0.8(6) \times \delta_{Pr,j} + 1.2(1.2) \times \delta_{Rr,j} - 0.2(1) \times \delta_{freq,j} \quad (1)$$

$$zR_{WLL2,j} = 1(4) + 0.3(8) \times \delta_{Pr,j} + 0.1(6) \times \delta_{Rr,j} - 0.1(1) \times \delta_{freq,j} \quad (2)$$

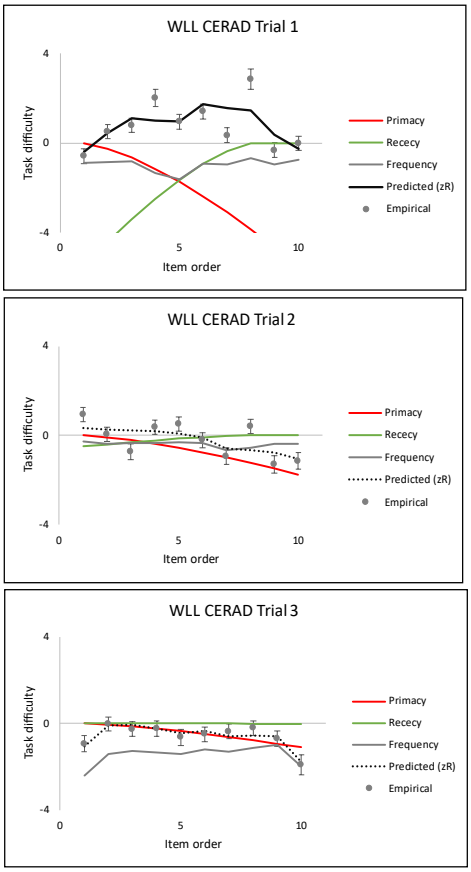
$$zR_{WLL3,j} = 1(1) + 0.2(3) \times \delta_{Pr,j} - 0.0(3) \times \delta_{Rr,j} - 0.26(4) \times \delta_{freq,j} \quad (3)$$

Figures 2a-c illustrate contributions from each explanatory variable according to eqs. 1-3 across the ten items, clearly showing how primacy disappears in the later trials, while there seems to be a small contribution from recency remaining also in trial 2 and 3.

Because of the rapidly diminishing SPE due to learning effects on repeated trials of relatively short word lists, in contrast to our previous studies on RAVLT IR/trial 1, by WLL trial 3 frequency has become the dominating explanatory variable. One must remember, however, that the variation in empirical task difficulty values is small; in fact, only the second and last items can be separated from the others by amounts significantly larger than the measurement uncertainties.

Furthermore, in figures 2a-c the predicted (zR) can graphically be compared with the empirical task difficulties (same as Figure 1). Pearson correlation coefficients were for trial 1: 0.70, trial 2: 0.65 and trial 3: 0.93, which are of comparable strength to the results for RAVLT (Melin, et al., 2021n) but not as strong as for the block and number recalling tests (Melin et al., 2021b)

In the figures, error bars show measurement uncertainties with coverage factor *k=2* for each memory task's difficulty, U(δ), which propagate through the PCR (section 2.2). In turn the U(δ) have implications for U(β) and UzR together with uncertainties in the fit itself, which is an issue of sample size, collinearity and measurement disturbance. In the present case when comparing the less cognitive able patients (MCI and AD) with the more cognitive able persons (HC and SCD), the un-even sample sizes may bias the interpretations. However, this can indicate that there are sources of dispersion when making the multivariate regression which are not yet accounted for.



Figures 2a-c. Corresponding plots presenting the contribution from each explanatory variable as well as the empirical and predicted task difficulty values for all three WLL trials. Task difficulty values on the y-axis (lower values implies an easier task), and the x-axis represents each item ordered by appearance in each trial. Error bars show measurement uncertainties *k*=2.

3.2 Differences between sub-groups

For trial 1, the "intercept" value + 6(5) (first term on the right-hand side (RHS) of each CSE for task difficulty) can be compared with $\delta_{Mr,j} = 2 \cdot M \cdot ln(5!) = +4.2$ logits for the whole cohort. Our model for the learning effects observed for the 5 RAVLT trials (Melin et al., 2022), where the intercept value decreases in inverse proportion to the root of the number of trials performed, would predict that the intercept value would be $\frac{4,2}{\sqrt{2}} = 3\ logits$ at trial 2 and $\frac{4,2}{\sqrt{3}} = 2.5\ logits$ at trial 3. These predictions are within measurement uncertainties of the observed intercept values given in equations (1), (2) and (3).

When comparing CSEs for the two groups of cohort members, for the second trial the intercepts were found to differ slightly (albeit with large uncertainties):

$$zR_{WLL2\ HC+SCD,j} = 1(1) + 0.0(5) \times \delta_{Pr,j} - 0.1(2) \times \delta_{Rr,j} - 0.4(2) \times \delta_{freq,j} \quad (4)$$

$$zR_{WLL2\ MCI+AD,j} = 2(1) + 0.3(4) \times \delta_{Pr,j} + 0.2(4) \times \delta_{Rr,j} - 0.2(0) \times \delta_{freq,j} \quad (5)$$

In line with what one may expect, this difference in intercept might indicate a faster learning for the more cognitive able cohort members. Further, a difference between the cohort groups was observed in terms of the contributions to task difficulty from primacy and recency; for the more cognitive able cohort members, the contributions from primacy and recency are negligible already at the second trial.

4. Conclusion

Our entropy-based theory earlier developed for RAVLT was successfully replicated for WLL CERAD trial 1 in the present study, although the effects of SPE are not as pronounced with repeated WLL trials. This may be explained by the fact that WLL CERAD comprises only 10 words in contrast to RAVLT as well as a different word ordering per trial.

In the present work we have shown i) how the variation in task difficulty reduces successively for trials 2 and 3, ii) how SPE change with repeated trials as predicted with our entropy-based theory, and iii) how possibilities to separate cohort members according to cognitive health status are limited.

These findings depend mainly on the test design itself: A test with only 10 words, where SPE do not dominate over trials, requires more challenging words to increase the variation in task difficulty, and in turn to challenge the test persons.

In the present case of WLL CERAD, the 10 words are all common but unrelated. Thus, it was no surprise that frequency provided relatively little explanation in the CSE, particularly in the first trial where SPE dominate. However, including less common, i.e., less frequently used, words is expected to make a greater contribution to recall difficulty from frequency. Moreover, other related aspects to consider could be: word length (Surprenant et al., 2011), semantics (Earles & Kersten, 2017; Hyde & Jenkins, 1973), phonetics (Rezvanfard et al., 2011).

The work here is not only novel, but also necessary for more consistent ways of defining and analyzing memory task difficulty, and in turn opens up for more practical and accurate measurement in clinical practice, research and trials.

The observed response in any word learning list test, as well as for other tests of human abilities, typically gets classification numbers. As in the present case, 0 for fail and 1 for pass (section 2.2). Such observed response constitutes raw data, $x_{i,j}$, for test person i and item j, which is characterized by ordinality and is not a measure of the person's memory ability nor the memory task difficulty. We have previously shown that metrological methods to simple syntax studies provide opportunities for more practical and accurate measurement in clinical practice, research and trials (Melin et al., 2021c). In this work, together with ongoing work on word learning list test (Melin et al., 2022; Melin et al., 2021a; Pendrill et al., 2021) we advance a novel metrological approach to cover more consistent ways of defining and analysing memory task difficulty.

9. Acknowledgements

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

10. Bibliographical References

Brillouin, L. (1962). *Science and Information Theory* (Second Edition). https://www.amazon.com/Science-Information-Theory-Second-Physics/dp/0486497550

Earles, J. L., & Kersten, A. W. (2017). Why Are Verbs So Hard to Remember? Effects of Semantic Context on Memory for Verbs and Nouns. *Cognitive Science*, *41 Suppl 4*, 780–807. https://doi.org/10.1111/cogs.12374

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., & Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *1*(1), 103–111. https://doi.org/10.1016/j.dadm.2014.11.003

Hawkins, K. A., Dean, D., & Pearlson, G. D. (2004). Alternative Forms of the Rey Auditory Verbal Learning Test: A Review. *Behavioural Neurology*, *15*(3–4), 99–107. https://doi.org/10.1155/2004/940191

Hyde, T. S., & Jenkins, J. J. (1973). Recall for words as a function of semantic, graphic, and syntactic orienting tasks. *Journal of Verbal Learning & Verbal Behavior*, *12*(5), 471–480. https://doi.org/10.1016/S0022-5371(73)80027-1

Melin, J., Regnault, A., Cano, S., & Pendrill, L. (2021a). *Neuropsychological assessments: Word learning tests and diagnostic potential of serial position effects*. The International Metrology Congress, Lyon, France.

Melin, J., Cano, S. J., Flöel, A., Göschel, L., & Pendrill, L. R. (2021b). Construct specification equations: 'Recipes' for certified reference materials in cognitive measurement. *Measurement: Sensors*, *18*, 100290. https://doi.org/10.1016/j.measen.2021.100290

Melin, J., Cano, S., & Pendrill, L. (2021c). The Role of Entropy in Construct Specification Equations (CSE) to Improve the Validity of Memory Tests. *Entropy*, *23*(2), 212. https://doi.org/10.3390/e23020212

Melin, J., Kettunen, P., Wallin, A., & Pendrill, L. (2022). *Entropy-based explanations of serial position and learning effects in ordinal responses to word list tests*.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488. http://dx.doi.org.ezproxy.ub.gu.se/10.1037/h0045106

Pendrill, L. (2019). *Quality Assured Measurement: Unification across Social and Physical Sciences*. Springer International Publishing. https://doi.org/10.1007/978-3-030-28695-8

Pendrill, L, Melin, J., & Cano, S. J. (2021). *Entropy-based explanations of multidimensionality in ordinal responses*. MSMM.

Quaglia, M., Cano, S., Fillmer, A., Flöel, A., Giangrande, C., Göschel, L., Lehmann, S., Melin, J., & Teunissen, C. E. (2021). The NeuroMET project: Metrology and innovation for early diagnosis and accurate stratification

of patients with neurodegenerative diseases. *Alzheimer's & Dementia*, *17*(S5), e053655. https://doi.org/10.1002/alz.053655

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

Rezvanfard, M., Ekhtiari, H., Noroozian, M., Rezvanifar, A., Nilipour, R., & Javan, G. K. (2011). *The Rey Auditory Verbal Learning Test: Alternate Forms Equivalency and Reliability for the Iranian Adult Population (Persian Version)*. 6.

Shannon, C. E. (1948). *A Mathematical Theory of Communication*. *1948*, 55.

Surprenant, A. M., Brown, M. A., Jalbert, A., Neath, I., Bireta, T. J., & Tehan, G. (2011). Backward recall and the word length effect. *The American Journal of Psychology*, *124*(1), 75–86. https://doi.org/10.5406/amerjpsyc.124.1.0075

Weitzner, D. S., & Calamia, M. (2020). Serial position effects on list learning tasks in mild cognitive impairment and Alzheimer's disease. *Neuropsychology*, *34*(4), 467–478. https://doi.org/10.1037/neu0000620

Zhan, L., Guo, D., Chen, G., & Yang, J. (2018). Effects of Repetition Learning on Associative Recognition Over Time: Role of the Hippocampus and Prefrontal Cortex. *Frontiers in Human Neuroscience*, *12*, 277. https://doi.org/10.3389/fnhum.2018.00277