Journal of the International Neuropsychological Society (2021), **27**, 396–400
Copyright © INS. Published by Cambridge University Press, 2020.
doi:10.1017/S1355617720000892

BRIEF COMMUNICATION

Breaking the Percent Memory Retention Ceiling using Bayesian Statistics

Umesh M. Venkatesan, Amanda R. Rabinowitz and Rachael M. Riccitello
Moss Rehabilitation Research Institute, 50 Township Line Road, Suite 100, Elkins Park, PA 19027, USA

(Received March 3, 2020; Final revision July 14, 2020; Accepted July 15, 2020; First Published Online October 5, 2020)

Abstract

Objectives: Neuropsychological tests of episodic memory often include a measure of memory retention to facilitate the diagnosis of memory disorders. However, the traditional percent retention (PR) score has limited interpretability when smaller amounts of information are both initially learned and later recalled, creating a pseudo-ceiling effect. To improve psychometrics of PR, we investigated a scoring procedure that incorporates levels of certainty into estimates of memory retention based on learning level. **Methods:** Word-list recall data from adults with traumatic brain injury were modeled using a uniform prior in the Bayesian framework. From the resultant posterior probability distributions, we derived a measure referred to as retention probability (RPr), which distinguishes the retention of relatively good and poor learners. PR and RPr scores were compared on their distributional properties and associations with theoretically related memory measures. Results: Significant distributional differences between PR and RPr were observed. RPr removed the conspicuous ceiling of PR, resulting in stronger correlational and predictive relationships with other memory measures. **Conclusion:** A Bayesian procedure for quantifying memory retention has psychometric advantages and potentially widespread applicability for measuring the change in behavioral features over time. Future directions are briefly discussed. A sample RPr calculator is provided for interactive exploration of the method.

Keywords: Episodic memory, Memory consolidation, Neuropsychological assessment, Psychometrics, Traumatic brain injury, Dementia

INTRODUCTION

Neurocognitive disorders often result in some form of memory disturbance, making the assessment of episodic memory an important component of clinical neuropsychological evaluations. Memory retention, or the storage of learned material over time, is of particular interest to clinicians and researchers given that impaired storage/consolidation is the key differentiator between amnestic and non-amnestic presentations (Squire, 2006) and a hallmark of medial temporal lobe (MTL) damage (Squire, Genzel, Wixted, & Morris, 2015). Because only learned material can be retained, it is important to distinguish between "good learners" and "poor learners" when making inferences about retention ability. However, these distinctions are not possible in the traditional method of estimating memory retention.

To illustrate, we consider a scenario where Person A recalls a maximum of 4 words after multiple presentations of a 10-word list (recall at learning, or initial recall), and successfully recalls all 4 of these words after a 20-min delay (delayed recall). Person B recalls all 10 words initially and at the delay. Person C learns and recalls 7 words. To quantify the retention ability of these individuals, clinical memory tests typically prescribe the use of percent retention (PR) score obtained from dividing delayed recall by initial recall and multiplying by 100. Although A was a relatively poor learner, B learned very well, and C was average, all of these individuals would be assigned a PR score of 100.

A practical limitation of PR scores is that they induce a pseudo-ceiling effect, where individuals demonstrating markedly different levels of initial and delayed recalls are assigned similarly high retention scores. This may hinder or even prevent statistical analysis due to a lack of sufficient variability at the ceiling. Their psychometric challenges are likely why PR scores often show suboptimal sensitivity and specificity for detecting memory impairment in aging and are relatively understudied compared to recall and recognition measures (Weissberger et al., 2017). Further, to

Correspondence and reprint requests to: Umesh M. Venkatesan, PhD. E-mail: venkateu@einstein.edu facilitate interindividual comparisons, PR scores standardize retention measurement across learning levels by proportionalization on initial recall; this may be inappropriate if memory retention and learning are not independent, as some have argued (Elliott, Isaac, & Muhlert, 2014; see Loftus, 1985, for original argument). In this view, retention cannot be isolated as a "pure" process within most standardized memory paradigms.

The current article addresses the limitations of the PR score by presenting a scoring procedure that incorporates an index of certainty in retention estimates conditioned on the learning level. We consider that our certainty in retention scores is lower when these scores are based on fewer observations—as is the case for poor learners. For example, Person A described above had less information to retain than Person B, thereby reducing certainty in the "perfect" retention ability suggested by their PR score. We use Bayesian statistics with individual recall data to derive an alternate measure of retention, referred to here as retention probability (RPr). Comparing psychometric properties between PR and RPr in a sample of older adults with traumatic brain injury (TBI), we hypothesize that RPr has a more statistically favorable frequency distribution than PR, affording greater sensitivity to individual differences in the upper tail (i.e., a reduction of the ceiling effect). It is also anticipated that, compared with PR, the increased measurement precision of RPr will result in stronger correlations with other putative correlates of MTL integrity, namely, recognition memory and independent measures of delayed recall.

METHODS

Participants included 110 middle age-older adult individuals with chronic TBI (sustained at least 1 year ago) enrolled in a bi-center study on aging with TBI. Participants must have sustained a TBI of at least moderate severity as evidenced by at least one of the following: Glasgow Coma Scale score <13 (not due to intoxication/sedation) on admission to emergency care, documented loss of consciousness of at least 1 h, documented post-traumatic amnesia of at least 24 h, or acute neuroimaging abnormality. Individuals were excluded if they had a history of TBI separate from the index injury, other neurological disorder, or serious psychiatric illness such as schizophrenia or bipolar disorder. Three participants were removed due to incomplete memory recall data. The remaining 107 participants had a mean age of 64.7 ± 8.2, 13.6 ± 2.6 years of education, and were 9.9 ± 6.6 years post-injury. Thirty participants were female. Study procedures were approved by the institutional review boards of Moss Rehabilitation Research Institute (Elkins Park, PA, USA) and Pennsylvania State University (University Park, PA, USA), and were in accordance with the Helsinki Declaration.

Participants were administered a neuropsychological test battery as part of a larger study on the long-term health effects of TBI. The battery included the Hopkins Verbal Learning Test-Revised (HVLT-R; Benedict, Schretlen, Groninger, & Brandt, 1998) and portions of the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS; Randolph, Tierney, Mohr, & Chase, 1998). The HVLT-R assesses verbal episodic memory and consists of a 12-word list presented 3 times, after each of which the examinee freely recalls as many words as possible (learning/immediate recall trials). Following a 20- to 25-min delay, participants freely recall the word list one time (delayed recall trial). A yes/no recognition trial is subsequently administered consisting of 12 target words and 12 foils. In addition to recall and recognition scores, a PR score is calculated from dividing delayed recall by the number of words initially recalled and multiplying by 100. Note that "initial recall" here refers to the best performance on learning trials 2 or 3 (Benedict et al., 1998), and not recall on the first learning trial. When delayed recall exceeded initial recall ($n = 5$), the latter was set to the delayed recall score under the assumption that recalled words at delay must have been learned (maximum attainable PR score = 100%). This scoring scheme was also used for the calculation of RPr (described below).

The RPr score was calculated using the Bayesian framework. Briefly, the Bayesian method characterizes uncertainty in quantities by modeling them as random variables with associated probability distributions (Gelman, Carlin, Stern, & Rubin, 2003). A distribution based on previous theoretical or empirical information (the prior) is "updated" with new data to provide a new probability distribution (the posterior).

We consider retention as a random variable in probability space, where it is assumed that recall responses follow a Bernoulli distribution (i.e., successful or failed delayed recall of an initially recalled word) with probability q representing an individual's overall retention ability. The choice of the prior is a key consideration in Bayesian analysis, and selections may vary based on clinical goals or data availability. For example, normative test data or meta-analytic studies may provide guidance on "expected" performance (van de Schoot et al., 2014). To provide a proof-of-principle demonstration of the Bayesian approach in the current study, we implement a uniform prior using the beta distribution (Bayes–Laplace prior; see Tuyl, Gerlach, & Mengersen, 2009), where all possible retention probabilities, .00 to 1.00, are equally likely. The beta prior is typically used for data expressed in proportions (Lynch, 2007), and reflects a *conjugate* prior, in that both the prior and the resulting posterior follow beta distributions. Note also that the beta prior reflects a distribution *over probabilities*. To update this prior for each individual, we use Bayes' theorem with initial and delayed recall data to obtain an individual-specific posterior probability distribution of q, or $P(q)$:

$$P(q) = q^{(\alpha-1)}(1-q)^{(\beta-1)}/B(\alpha,\beta) \tag{1}$$

where $\alpha = 1 +$ delayed recall; $\beta = 1 +$ initial recall − delayed recall; and the beta function, $B(\alpha,\beta) = (\alpha - 1)!(\beta - 1)!/(\alpha + \beta - 1)!$. The distribution over retention probabilities $P(q)$ describes an individual's probabilistic retention level based on their recall behavior; the greater a given individual's initial recall and the more extreme their delayed

https://doi.org/10.1017/S1355617720000892 Published online by Cambridge University Press

recall in relation to that initial recall, the greater the certainty in the retention estimate (i.e., the sharper the probability density function of q). This approach is intuitive, reflecting high certainty that retention is low in the case of an individual who learns perfectly and recalls nothing at the delay, and equally high certainty that retention is high when an individual learns perfectly and also recalls perfectly at the delay. From Equation 1 we derive the mean $E(q)$ (Equation 2) and standard deviation $\sigma(q)$ (Equation 3) of the posterior probability distribution, where the magnitude of the standard deviation is inversely related to initial recall:

$$E[q] = \alpha/(\alpha + \beta) \tag{2}$$

$$\sigma[q] = \surd\{\alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)]\} \tag{3}$$

Dividing $E(q)$ by $\sigma(q)$ gives a summary metric. To facilitate direct comparison with PR scores, this quantity was rescaled by dividing by the maximum possible value for $E[q]$ on the HVLT-R (i.e., when initial recall and delayed recall both equal 12) and multiplying by 100, yielding the RPr score that was used in subsequent analyses. Thus, while ultimately derived from the posterior probability distribution, the RPr score is not an individual's probability of recalling words from the list. Rather, it is interpreted as the proportion of words maintained over a delay (similar to PR) with an adjustment for each individual's starting point (learning level, which is not captured in PR). It is important to note that this latter feature can alter the rank ordering of individuals when comparing them on RPr versus PR. For example, Person X who recalls 8/12 words initially learned has a slightly lower PR score (67) than Person Y who recalls 5/7 words (71). However, Person X's RPr score (37) indicates a slightly *better* performance than Person Y (32), reflecting the former's higher learning level.

Psychometric properties of PR and RPr scores were compared using frequentist statistics, due to their interpretability and standard usage in psychological research. Paired differences in retention scores were evaluated using the Wilcoxon signed-rank test. The distributional properties of the two measures were characterized quantitatively and visually. Coefficients of variation (CoV) were computed as indices of dispersion. Relationships of PR and RPr with demographic variables were assessed with Spearman rank-order correlations. To evaluate the extent to which PR and RPr are related to theoretically convergent memory performances, we quantified their relationships to these other memory measures with Spearman rank-order correlations and tested differences in correlation strength using Steiger's Z tests (Myers & Sirois, 2004). Convergent memory measures were available for most participants ($N = 102$) and included the Recognition Discrimination Index from the HVLT-R recognition trial as well as scores from RBANS subtests: Story Recall and Figure Recall, which assess narrative verbal memory and visual memory, respectively. Convergent memory performances were age-corrected based on normative data from their respective test manuals. To characterize the extent to which RPr may have greater predictive validity than PR in the upper end of the PR distribution, the full sample was median-split based on PR scores; in the top and bottom 50% separately (n's = 51), we performed multivariate regressions with PR or RPr as predictors and convergent measures as dependents. All analyses evaluated significance at $\alpha = .05$. Z-test statistics were obtained from Lee and Preacher (2013). All other analyses were performed in jamovi version 1.0.7.0 (The jamovi project, 2019).

RESULTS

As visualized in Figure 1, both PR and RPr deviated significantly from normality but were generally mesokurtic. Expectedly, PR scores were markedly left-skewed and exhibited a clear ceiling effect, with half of the sample recording greater than 80% retention. In contrast, RPr scores were right-skewed but maintained appreciable variability near the floor. Standardized dispersion as measured by CoV was greater for RPr than PR. Figure 1 also demonstrates that even in PR scores further away from the ceiling (e.g., at 50%), RPr provided greater discrimination between individuals.

Descriptive and inferential statistics for PR and RPr are provided in Table 1. PR scores were significantly greater overall than RPr scores within individuals. As expected, PR and RPr showed a strong monotonic relationship ($\rho = .97$, $p < .001$). Neither measure was related to age (PR: $\rho = \text{-}.13$, $p = .19$; RPr: $\rho = \text{-}.13$, $p = .17$) or years of education (PR: $\rho = .10$, $p = .30$; RPr: $\rho = .15$, $p = .12$). Both PR and RPr correlated significantly with performance on all theoretically convergent measures (word list recognition, story recall, and figure recall). However, these bivariate relationships were significantly stronger with RPr than PR (Table 1, lower half).

In individuals scoring in the upper half of the PR score distribution (>78 PR), PR did not significantly predict convergent memory performances (F[3,47] = .81, $p = .49$, Wilks' $\Lambda = .95$). In contrast, RPr demonstrated significant predictive validity in this subsample (F[3,47] = 6.43, $p < .001$, Wilks' $\Lambda = .71$) for all convergent measures (univariate p's: list recognition, <.001; story recall, .002; figure recall, .04). Therefore, proximal to the PR ceiling, RPr accounted for meaningful variance in theoretically related test performances that could not be detected with PR. In the lower half of the PR distribution (<78 PR), both measures significantly predicted convergent performances, with RPr recording a numerically larger effect size ($\eta^2 = .58$ *vs.* .45). See Appendix A for full regression results.

DISCUSSION

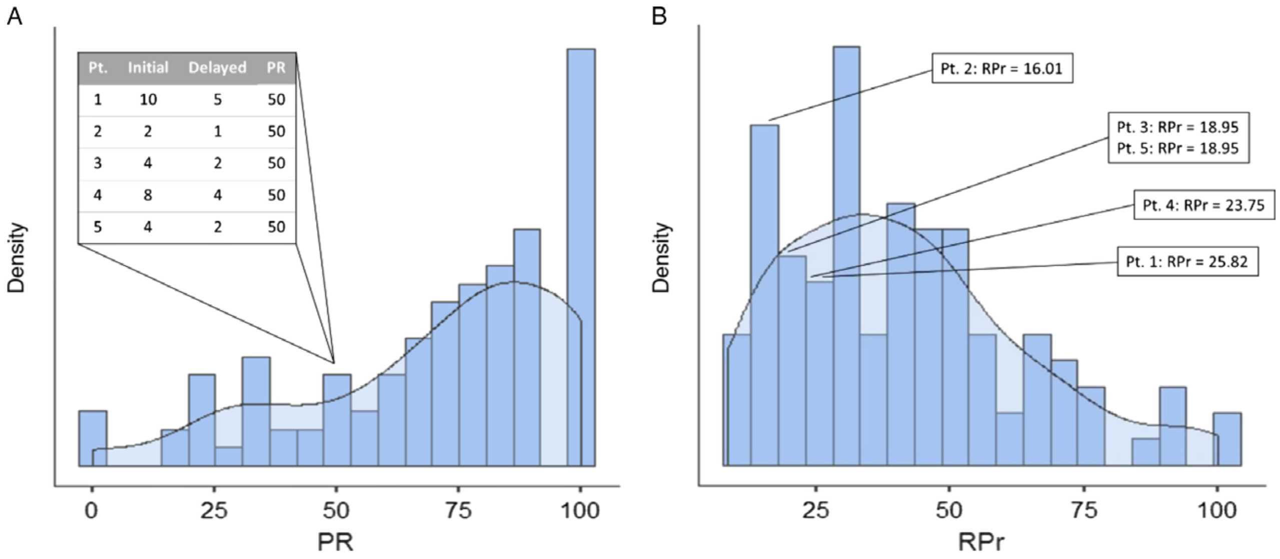
In clinical episodic memory paradigms, traditional PR scores lack utility when learning is suboptimal, resulting in pseudoceiling effects that have statistical and conceptual limitations. The present findings demonstrate that a Bayesian normalization method applied to initial and delayed recall scores

https://doi.org/10.1017/S1355617720000892 Published online by Cambridge University Press

Table 1. Descriptive and inferential statistics for Percent Retention (PR) and Retention Probability (RPr) scores. SD = standard deviation. CoV = coefficient of variation. HVLT-R RDI = Hopkins Verbal Learning Test-Revised Recognition Discrimination Index. RBANS = Repeatable Battery for the Assessment of Neuropsychological Status

$N = 107$	PR	RPr	Hypothesis tests
Mean (SD)	70.7 (26.7)	41.0 (22.0)	$W = 5745$, $p < .001$, Cohen's $d = 2.15$
Median (range)	80.0 (100)	37.2 (91.5)	
Skew	–.88	.74	
Kurtosis	–.05	.09	
Normality[1]	$p < .001$	$p < .001$	
CoV	.38	.54	
$N = 102$	Spearman ρ		
HVLT-R RDI	.414***	.531***	$Z = –5.51$, $p < .001$
RBANS story recall	.390***	.500***	$Z = –5.05$, $p < .001$
RBANS figure recall	.323***	.402***	$Z = –3.42$, $p < .001$

[1] Reflects results from the Shapiro–Wilk test. ***$p < .001$.



Pt.	Initial	Delayed	PR
1	10	5	50
2	2	1	50
3	4	2	50
4	8	4	50
5	4	2	50

Pt. 2: RPr = 16.01

Pt. 3: RPr = 18.95
Pt. 5: RPr = 18.95

Pt. 4: RPr = 23.75

Pt. 1: RPr = 25.82

Fig. 1. Frequency distributions for retention measures. Shown are (A) Percent Retention (PR) and (B) Retention Probability (RPr) score distributions. Data popouts: PR and RPr scores with respective initial and delayed recall data for the same five participants (Pt.), showing equivalence in PR and score separation in RPr ($N = 107$).

removes this ceiling effect and permits statistical analyses at the group level. The increased variability in retention scores afforded by our approach appears to be clinically meaningful.

In addition to large magnitude differences between PR and RPr within individuals, distributions of the two scores were visually differentiable, with a conspicuous ceiling effect in PR that was absent in RPr. Our prediction that RPr would increase variability in retention scores was supported by greater score dispersion in RPr than PR and stronger association of RPr versus PR to theoretically related memory measures. This has important measurement and interpretive implications.

As described earlier, the PR score artificially equates differences in learning levels across participants. We showed that this conventional approach removes potentially meaningful variance in retention ability when compared against a measure that *accommodates* the learning level. Across the entire sample, memory measures theoretically related to retention were more strongly associated with RPr than PR. Further, in participants with >78 PR (representing half of the sample), PR scores did not predict convergent memory performances while RPr continued to have significant predictive value. These results indicate that incorporating information about learning into retention estimates may have psychometric advantages that could offer a more robust approach to measuring vulnerabilities in learning and memory compared to the traditional method.

There are limitations to this work that reveal opportunities for future research. Our sample was comprised of older individuals with TBI, and extension to other clinical disorders is needed. Limited variability in age, education, and diagnosis hampered our ability to examine relationships between retention scores and clinical/demographic characteristics. Nonetheless, as noted earlier, memory retention ability is typically of special interest in older adult populations and a cornerstone of the differential diagnosis of dementia. Therefore, our patient group represents

https://doi.org/10.1017/S1355617720000892 Published online by Cambridge University Press

a population where RPr is particularly applicable, although we recommend validation in larger and more clinically diverse samples. In particular, future studies evaluating classification accuracy against diagnostic criteria in clinical groups with distinct memory presentations (e.g., neurodegenerative dementias) would be informative (Weissberger et al., 2017). Because RPr captures both learning levels and differences between initial and delayed recall, it may have more immediate utility as a summary measure of memory in research studies, potentially obviating the need for separate learning and delayed recall parameters and sparing a valuable degree of freedom in patient studies.

Although beyond the scope of the current article, a similar approach using Bayesian tools may be appropriate for other applications in the social and behavioral sciences. The method would be most relevant for describing change in behavior where initial and follow-up measurements cannot be assumed independent and effects of random error limit interpretation of change scores (i.e., regression to the mean; Barnett et al., 2005). For example, it could be used to calibrate changes in individual post-treatment cognitive or physiological responses by pre-treatment functioning.

We provide a user-friendly calculator as an interactive guide to the RPr method (see Appendix B). Readers can enter their own data and compare results with the traditional PR approach. The accessibility of the RPr method has the potential to stimulate retrospective analyses of a variety of experimental data and guide future research and practice.

CONFLICT OF INTEREST

The authors have nothing to disclose.

ACKNOWLEDGMENTS

This work was supported by the Pennsylvania Department of Health (A.R.R., grant number 4100077082). The authors thank Sean Falconer, Ph.D., and Frank Hillary, Ph.D., for methodological consultation and manuscript review.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S1355617720000892.

REFERENCES

Barnett, A.G., van der Pols, J.C., & Dobson, A.J. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Benedict, R.H.B., Schretlen, D., Groninger, L., & Brandt, J. (1998). Hopkins verbal learning test—revised: normative data and analysis of inter-form and test-retest reliability. *The Clinical Neuropsychologist*, *12*(1), 43–55.

Elliott, G., Isaac, C.L., & Muhlert, N. (2014). Measuring forgetting: a critical review of accelerated long-term forgetting studies. *Cortex*, *54*, 16–32. https://doi.org/10.1016/j.cortex.2014.02.001

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2003). *Bayesian Data Analysis* (2nd ed.) Taylor & Francis. https://books.google.com/books?id=TNYhnkXQSjAC

Lee, I.A. & Preacher, K.J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common. http://quantpsy.org

Loftus, G.R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 397–406. https://doi.org/10.1037/0278-7393.11.2.397

Lynch, S.M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York: Springer. https://books.google.com/books?id=JN0rPpEpw3IC

Myers, L. & Sirois, M. (2004). Differences between spearman correlation coefficients. *Encyclopedia of Statistical Evidence*. https://doi.org/10.1002/0471667196.ess5050

Randolph, C., Tierney, M.C., Mohr, E., & Chase, T.N. (1998). The repeatable battery for the assessment of neuropsychological status (RBANS): preliminary clinical validity. *Journal of Clinical and Experimental Neuropsychology*, *20*(3), 310–319. https://doi.org/10.1076/jcen.20.3.310.823

Squire, L.R. (2006). Lost forever or temporarily misplaced? The long debate about the nature of memory impairment. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *13*(5), 522–529. https://doi.org/10.1101/lm.310306

Squire, L.R., Genzel, L., Wixted, J.T., & Morris, R.G. (2015). Memory consolidation. *Cold Spring Harbor Perspectives in Biology*, *7*(8), a021766–a021766. https://doi.org/10.1101/cshperspect.a021766

The jamovi project. (2019). Jamovi (Version 0.9). https://www.jamovi.org

Tuyl, F., Gerlach, R., & Mengersen, K. (2009). Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters. *Bayesian Analysis*, *4*. https://doi.org/10.1214/09-BA405

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J.B., Neyer, F.J., & van Aken, M.A.G. (2014). A gentle introduction to Bayesian analysis: applications to developmental research. *Child Development*, *85*(3), 842–860. PubMed. https://doi.org/10.1111/cdev.12169

Weissberger, G.H., Strong, J.V., Stefanidis, K.B., Summers, M.J., Bondi, M.W., & Stricker, N.H. (2017). Diagnostic accuracy of memory measures in Alzheimer's Dementia and mild cognitive impairment: a systematic review and meta-analysis. *Neuropsychology Review*, *27*(4), 354–388. https://doi.org/10.1007/s11065-017-9360-6

https://doi.org/10.1017/S1355617720000892 Published online by Cambridge University Press