
BRIEF COMMUNICATION

Retention weighted recall improves discrimination of Alzheimer's disease

HERMAN BUSCHKE,^{1,3} MARTIN J. SLIWINSKI,^{1,3,4} GAIL KUSLANSKY,^{1,3} MINDY KATZ,¹
JOE VERGHESE,¹ AND RICHARD B. LIPTON^{1,2}

¹Department of Neurology, Albert Einstein College of Medicine, Bronx, New York

²Department of Epidemiology and Social Medicine, Albert Einstein College of Medicine, Bronx, New York

³Rose F. Kennedy Center for Mental Retardation and Human Development, Albert Einstein College of Medicine, Bronx, New York

⁴Department of Psychology, Syracuse University, Syracuse, New York

(RECEIVED August 23, 2005; FINAL REVISION January 31, 2006; ACCEPTED February 6, 2006)

Abstract

Impaired recall for early items (primacy) and late items (recency) on word list recall tests are seen in Alzheimer's disease (AD). We compared conventional scoring on the Telephone Instrument for Cognitive Status (TICS) recall list with scorings based on retention-weighted recall (RWR: each item weighted by its serial position) in older adults participating in a community-based aging study. Subjects with mild AD ($N = 18$) did not differ from those without dementia ($N = 231$) with respect to recency (46% vs. 59%, $p = 0.2$), but had impaired primacy (2% vs. 39%, $p < .001$) on word recall on the TICS. RWR scoring improved the effect size (1.52 SD) compared to conventional scoring (1.08 SD). With a fixed sensitivity of 85%, specificity was lower using conventional scoring (56%) than RWR (76%) scoring. Our findings suggest that optimized RWR scoring of word list free recall can improve detection of mild AD compared to conventional scoring. (*JINS*, 2006, *12*, 436–440.)

Keywords: Dementia, Memory, Serial Position, Primacy, Recency, Aging

INTRODUCTION

The identification of mild Alzheimer's disease (AD) is an urgent public health priority since early dementia often goes undetected and untreated (Callahan et al., 1995). Memory impairment is the earliest indicator of AD, and is the only cognitive domain that must be impaired to diagnose dementia (American Psychiatric Association, 1994). Individuals with poor memory performance are at elevated risk for developing clinically diagnosable dementia (Masur et al., 1994)

Strategies for optimizing the early detection of AD by memory testing may involve using different aspects of memory (list learning, paired-associate learning, story recall, etc.), and/or using procedural manipulations to maximize the separation of intact and impaired memory performance

(e.g., the use of controlled learning and category cues to enhance normal performance) (Buschke, 1984; Buschke et al., 1997). A complementary approach is to use novel scoring procedures on free recall list learning (Meiran et al., 1996; Sliwinski et al., 1997; Shankle et al., 2005). In this article, we pursue this second strategy by developing scoring algorithms based on serial position effects to improve discrimination of older adults with and without early AD (Buschke & Sliwinski, 1999). Free recall (FR) is typically measured by counting the number of items recalled. This unit-weighted counting assumes that all items make an equal contribution to the identification of dementia. However, the probability of recalling an item varies as a function of its serial position in the to-be-learned list (Nipher, 1878). Analysis of serial position effects reveals that recall is more likely for items presented first ("primacy") or last ("recency") than items in the middle of the list (Murdoch, 1962). Primacy and recency effects persist in recall by cognitively intact older adults. The primacy effect reflects a response

Correspondence and reprint requests to: Herman Buschke, M.D., Department of Neurology, Kennedy 912, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461. E-mail: buschke@aecom.yu.edu

preference for the first few items in a list, and depends critically on the ability to rehearse early items during presentation of subsequent list items (Howard & Kahana, 1999). Several studies have demonstrated that patients with AD exhibit impaired recall for early list items and little or no impairment in recall of late list items (Spinnler & Della Sala, 1988; Gainotti et al., 1989; Gainotti & Marra, 1994). It has been reported that the difference between the first few primacy items and recall of the last few recency items was much greater in recall by subjects with AD than in those without AD (Gainotti & Marra, 1994). Recall of items given early in a word list should contribute more information to the discrimination of FR performance in AD patients from FR by their nondemented peers. A scoring strategy that weights items based on their position may improve the identification of diagnosable dementia as well as individuals at very high risk for future dementia.

We compared the performance of a sample of older adults with and without AD on the 10-item word recall list from the Telephone Instrument for Cognitive Status (TICS) (Brandt et al., 1988) using conventional unweighted free recall (FR) and an alternative item-specific serial position (retention-weighted recall: RWR) scoring procedure. As preventive interventions become available, early detection will be essential to introduce these measures early. Hence, we focused on older adults with early AD. We predicted that RWR would improve discrimination of demented from nondemented individuals compared to unweighted FR.

METHODS

Research Participants

The population for this study included 257 community-dwelling older adults seen between July 1996 and August 1997 in the Einstein Aging Study (EAS), and who were administered the TICS as part of a validation study of telephone-based cognitive screening tests (Lipton et al., 2003). Subjects in the current study included those recruited by systematic sampling from population lists ($n = 110$, 43% of sample) as well as community volunteers ($n = 147$, 57% of the sample). This sample was comprised of 163 females (63%) and 94 males (36%), 212 Caucasians (82%), 41 African-Americans (16%), and 4 of other ethnicities.

The EAS recruitment methods have been previously detailed (Buschke et al., 1997; Lipton et al., 2003). All EAS subjects receive medical, epidemiological, and behavioral questions, a neurological exam, and extensive neuropsychological testing at enrollment and at 12 to 18 monthly follow-up visits.

Dementia Diagnosis

A diagnosis of dementia based on the *Diagnostic and Statistical Manual III-Revised* (DSM-III-R) criteria (American Psychiatric Association, 1987) was assigned at consensus

case conferences attended by the study neurologist, a neuropsychologist, and the social worker, who were all blind to results of the telephone interview. Severity of cognitive impairment was rated by the study clinicians using the Clinical Dementia Rating (CDR) scale (Hughes et al., 1982). A diagnosis of possible or probable AD was assigned according to NINCDS/ADRDA criteria (McKhann et al., 1984). Twenty-six individuals were diagnosed with dementia (base rate 10.8%), and of these, 22 were assigned a subtype of possible or probable AD, 1 vascular dementia (VaD), and 2 mixed AD/VaD, and 1 frontotemporal dementia. Of the 24 individuals with AD or mixed AD/VaD dementia, 19 were assigned CDR 0.5 (questionable dementia) or CDR 1 (mild dementia). Eighteen had Blessed Information-Memory-Concentration test (BIMC) scores of <14 (range 0–32, >7 abnormal) indicating mild disease severity (Blessed et al., 1968). All analyses were restricted to the cognitively normal individuals and the 18 individuals diagnosed with possible, probable or mixed AD/VaD, and with mild dementia severity ($n = 18$).

Study Procedure

The TICS, a validated telephone-administered instrument (Brandt et al., 1988), was administered as part of a telephone interview 1 to 3 weeks before or after the EAS clinic visit. The memory task consisted of a single presentation of 10 words (2–3 seconds per word), to be recalled immediately in any order: “cabin, pipe, elephant, chest, silk, theatre, watch, whip, pillow, giant.” Single items were repeated if the subject requested.

Scoring and Data Analyses

The standard TICS FR measure of memory was the total number of words recalled (maximum 10). This scoring assumes that all items contribute equally to the measurement of memory, regardless of their length of retention (due to order of presentation or recall), or any other features that might affect retrieval. This conventional FR measure was compared with a simple retention-weighted recall (RWR) memory measure in which each item was weighted according to its relative length of *retention before recall begins*. This RWR measure weights each recalled item by its serial presentation order: In recall of a 10-item list, recall of the first presented item would score 10; recall of the second presented item would score 9, and so forth. Such RWR weights recall by retention, inversely to recency of presentation. The general formula for scoring each item by this measure is:

$$\text{RWR for each item} = \text{List Length} - \text{Presentation Position} + 1.$$

After each recalled item is weighted according to this formula, the weighted scores for all recalled items are summed

to obtain the total RWR measure of memory performance shown by:

$$FR \text{ (total RWR)} = \sum \text{RWR for all items recalled.}$$

Primacy and recency effects are calculated by determining the proportion of individuals who recalled the first three items (primacy) and the last three items (recency). Effects of AD *versus* no dementia, serial position, and the interaction between AD status and serial position on the probability of correct recall were tested using generalized estimating equations (GEE) for binomial data (PROC GENMOD in SAS). The data from serial positions 3 and 6 were omitted as none of the AD adults correctly recalled items in these positions.

Analyses of the receiver operating characteristics (ROC) for FR and RWR were conducted to examine the trade-off between sensitivity and specificity of each scoring algorithm for AD compared to no dementia. Sensitivity is the proportion of subjects with dementia (according to the gold standard diagnosis) with a positive test result (diseased with a true positive test/all with disease). Specificity is the proportion of nondemented subjects (according to the gold standard diagnosis) who will be correctly classified as nondemented (those with a true negative test/all without disease).

RESULTS

There were no significant differences between the 18 mild AD cases and the 231 nondementia controls in age (mean 80.6 ± 6.4 vs. 80.8 ± 7.8 years), educational level (12.1 ± 3.2 vs. 12.8 ± 3.8 years), or sex, (59% vs. 64% female). There were significant differences between the AD and nondemented groups in performance on mental status (BIMC) (mean 12.4 ± 4.5 vs. 2.9 ± 2.7 , $p = .01$) and WAIS-R Verbal IQ (91.6 ± 13.3 vs. 106.1 ± 13.6 , $p = .01$), and subtests of the WAIS-R performance IQ such as digit symbol (19.9 ± 11.3 vs. 35.4 ± 12.4 , $p < .001$) and block design (12.7 ± 7.7 vs. 18.5 ± 4.5 , $p = .02$) (Wechsler, 1981).

Differential Recall by Serial Position

Figure 1 presents the probability of correctly recalling each item as a function of serial position in order of presentation by nondemented older adults and individuals with dementia. Inspection of the figure reveals that the nondemented sample is more likely to recall the first 3 items (primacy) and the last several items (recency); recall is least likely for the middle items on the list. For AD, the proportion of items recalled is lower on average; the recency effect is prominent while the primacy effect is decreased. Figure 1 suggests that separation between the no dementia and AD groups is maximal for items 1 to 6.

Overall, non-AD adults had a higher probability of recall than AD participants ($z = 5.18$, $p < .05$), and both groups

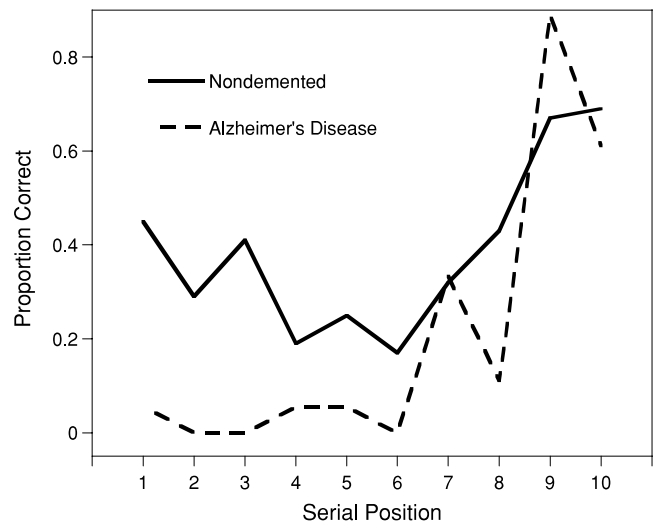


Fig. 1. Serial position curves for free recall by older adults with and without AD.

showed clear serial position effects ($z = 6.75$, $p < .05$). Recall by the nondemented older adults is characterized by *primacy* as well as *recency*, but recall by individuals with dementia is characterized only by *recency*, as evidenced by a significant group \times position interaction ($z = 4.64$, $p < .05$). Although AD subjects recalled the last two items with nearly the same probability as controls, they had much lower recall of the earlier items, indicating that most of the difference in recall by the AD group is due to decreased recall of items that must be retained longer.

Primacy and Recency

Both the AD group and the nondemented controls showed recency effects on word recall on the TICS (57% vs. 59%, $p = .2$). Participants with AD recalled significantly fewer primacy items compared to controls (2% vs. 39%, $p < .001$). The difference between primacy and recency recall by these older adults with AD is twice as large as the difference between primacy and recency recall by the nondemented group (55% vs. 21%, $p < .001$), confirming the loss of primacy in AD (Figure 1).

Free Recall

Retention-weighted scoring increased the memory measure for the nondemented group from 3.9 (FR) to 18.6 (RWR), much more than the increase from 2.2 (FR) to 5.5 (RWR) for the AD group. Effect size ('d') estimates the magnitude of the difference between mean recall by the dementia and no dementia groups in standard deviation units. Because the retention-weighted scoring algorithm produced scores that were positively skewed, a square-root normalizing transformation was applied before calculating the effect size.

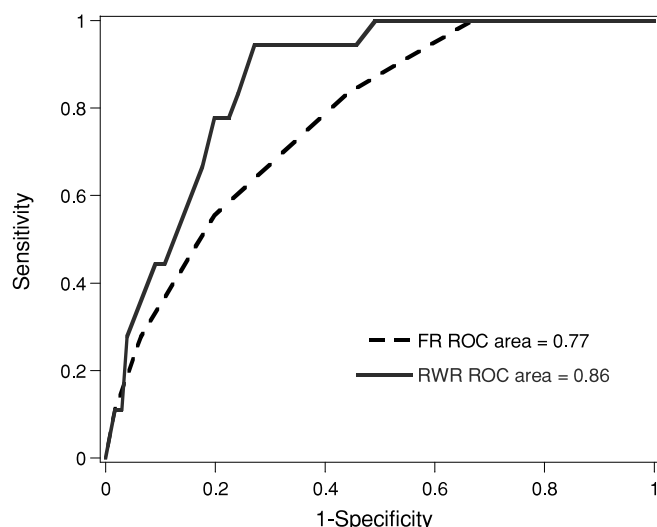


Fig. 2. Receiver operating characteristics (ROC) curves for free recall (FR) and retention-weighted recall (RWR).

For FR (counting) the effect size was 1.08 standard deviations, but for retention-weighted recall (weighting) the effect size was 1.52 standard deviations.

Discriminative Validity

Figure 2 shows the ROC curves for FR and RWR. The area under the ROC curve is an index of discriminability, that is, the larger the area the under the curve, the better the discrimination (DeLong et al., 1988). The RWR provided significantly better discrimination than the FR index [$\chi^2(1) = 9.27, p < .05$] (DeLong et al., 1988). Of additional interest are focused comparisons of the performance of the difference algorithms for fixed levels of discriminability. At a sensitivity of 0.85, the specificity for conventional FR was 0.56 compared to 0.76 for RWR.

DISCUSSION

The present results demonstrated that the discriminative validity of a simple 10-word FR test for identifying older adults with mild AD can be improved by using scoring rules that weight each item in a list with regard to its serial position. The serial position curves, primacy and recency comparisons, effect sizes, and discriminative validity shown by sensitivity and specificity in this comparison of memory measurement by weighting recall (RWR) and by counting recall (FR) provide a demonstration of how the power of memory measurement and detection of AD can be improved by weighting, and why weighting should be considered when measuring memory.

The retention-weighted-recall score capitalizes on the established empirical finding, confirmed in the present data, that in dementia (most often AD), there is selective loss of

the primacy effects seen in nondemented young and older adults (Spinnler & Della Sala, 1988; Gainotti et al., 1989; Gainotti & Marra, 1994). The primacy and recency effects shown in our controls are similar in magnitude to other studies (Spinnler & Della Sala, 1988; Gainotti et al., 1989; Gainotti & Marra, 1994). However, the earlier studies did not consider implications of these findings to the clinical detection of AD. We found that RWR improved discriminative validity over the usual unweighted scoring. These differences in the serial position curves confirm similar previous findings of impaired primacy in recall by older adults with AD (Gainotti et al., 1989; Gainotti & Marra, 1994; Spinnler & Della Sala, 1988), and provide an empirical basis for retention-weighted measurement of memory performance.

The following limitations need to be noted. A longer or shorter word list, repeated presentations or in-person *versus* telephone presentation may yield different results. Also, it is possible that sample characteristics could alter results. Our results build upon empirical findings reported in AD patients, and are explained by impaired performance in our subjects with mild AD. The gold standard used for this study was clinical diagnosis of AD. As our focus was on early detection of dementia, we restricted our focus to early stages of AD. It is likely that the differences noted in this sample may be accentuated in samples with subjects with more severe dementia. Older adults with dementia may have multiple pathologies on pathological brain examination. Hence, our findings should be verified in other samples that include non-Alzheimer dementias and possibly with pathological validation. A recent study reported improved discrimination of cognitively normal and impaired subjects using weighting of recall responses (Shankle et al., 2005). In contrast to RWR, this technique does not specifically account for serial positioning and requires computer-based weighting, which may limit its use in some settings.

The impairment in primacy and relative preservation of recency effects in AD remain in need of explanation. Since rehearsal is the key to primacy, the presentation of subsequent list items may interfere with the ability of AD patients to rehearse earlier list items. This is consistent with the view that progressive AD disrupts the ability to simultaneously encode information and process information. This explanation, though admittedly *ad hoc*, is plausible and consistent with findings showing AD deficits in working memory (Baddeley et al., 2001).

Regardless of the correct explanation of the serial position deficits observed in the FR performance on AD patients, the empirical fact of such deficits can inform scoring rules designed to optimize the detection of AD-related memory impairment using TICS or other list-based recall tests. Future research that isolates the mechanism underlying the AD deficit in primacy effects could inform the design of memory testing procedures to better assess those aspects of memory performance that are most sensitive to impairments occurring early in AD.

ACKNOWLEDGMENTS

This study was supported by National Institute on Aging grants AG03949 and AG12448, and The National Institute of Child Health and Human Development grant HD-01799. The Albert Einstein College of Medicine owns the patent rights for serial-position weighted recall and makes this test available as a service to the research community, but licenses the test for commercial use.

REFERENCES

- American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders—Revised* (3rd edition). Washington, DC: American Psychiatric Association.
- Baddeley, A.D., Baddeley, H.A., Bucks, R.S., & Wilcock, G.K. (2001). Attentional control in Alzheimer's disease. *Brain*, *124*, 1492–1508.
- Blessed, G., Tomlinson, R., & Roth, M. (1968). The association between quantitative measures of dementia and of senile changes in the cerebral gray matter of elderly subjects. *British Journal of Psychiatry*, *114*, 797–811.
- Brandt, J., Spencer, M., & Folstein, M. (1988). The Telephone Interview for Cognitive Status. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, *1*, 111–117.
- Buschke, H. (1984). Cued recall in amnesia. *Journal of Clinical Neuropsychology*, *6*, 433–440.
- Buschke, H., Sliwinski, M.J., Kuslansky, G., & Lipton, R.B. (1997). Diagnosis of early dementia by the Double Memory Test: Encoding specificity improves diagnostic sensitivity and specificity. *Neurology*, *48*, 989–997.
- Buschke, H. & Sliwinski, M.J. (1999). Item-Specific Weighted Memory Measurement. In E. Tulving (Ed.), *Memory, Consciousness, and the Brain: The Tallinn Conference* (pp. 18–27). Philadelphia: The Psychology Press.
- Callahan, C.M., Hendrie, H.C., & Tierney, W.M. (1995). Documentation and evaluation of cognitive impairment in elderly primary care patients. *Annals of Internal Medicine*, *122*, 422–429.
- DeLong, E.R., DeLong, D.M., & Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.
- Gainotti, G. & Marra, C. (1994). Some aspects of memory disorders clearly distinguish dementia of the Alzheimer's type from depressive pseudo-dementia. *Journal of Clinical and Experimental Neuropsychology*, *16*, 65–78.
- Gainotti, G., Monteleone, D., Parlato, E., & Carlomagno, S. (1989). Verbal memory disorders in Alzheimer's disease and multi-infarct dementia. *Journal of Neurolinguistics*, *4*, 327–345.
- Howard, M.W. & Kahana, M.J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *25*, 923–941.
- Hughes, C.P., Berg, L., Danziger, W.L., Coben, L.A., & Martin, R.L. (1982). A new clinical scale for the staging of dementia. *British Journal of Psychiatry*, *140*, 566–572.
- Lipton, R.B., Katz, M.J., Kuslansky, G., Sliwinski, M., Stewart, W.F., Verghese, J., Crystal, H., & Buschke, H. (2003). Screening for dementia by telephone using the memory impairment screen. *Journal of the American Geriatric Society*, *51*, 1382–1390.
- Masur, D., Sliwinski, M., Lipton, R.B., Blau, A.D., & Crystal, H.A. (1994). Neuropsychological prediction of dementia and the absence of dementia in healthy elderly persons. *Neurology*, *44*, 1427–1432.
- McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, *34*, 939–944.
- Meiran, N., Stuss, D.T., Guzman, D.A., Lafleche, G., & Willmer, J. (1996). Diagnosis of dementia. Methods for interpretation of scores of 5 neuropsychological tests. *Archives of Neurology*, *53*, 1043–1054.
- Murdock, B.B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482–488.
- Nipher, F.E. (1878). On the distribution of errors of numbers written from memory. *Transactions of the Academy of Science of St. Louis*, *3*, 210–211.
- Shankle, W.R., Romney, A.K., Hara, J., Fortier, D., Dick, M.B., Chen, J.M., Chan, T., & Sun, X. (2005). Methods to improve the detection of mild cognitive impairment. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 4919–4924.
- Sliwinski, M., Buschke, H., Stewart, W.F., Masur, D., & Lipton, R.B. (1997). The effect of dementia risk factors on comparative and diagnostic selective reminding norms. *Journal of the International Neuropsychological Society*, *3*, 317–326.
- Spinnler, H. & Della Sala, S. (1988). The role of clinical neuropsychology in the neurological diagnosis of Alzheimer's disease. *Journal of Neurology*, *235*, 258–271.
- Wechsler, D. (1981). *Wechsler Adult Intelligence Scale—Revised*. New York: Psychological Corporation.