Age differences in storage and retrieval: A multinomial modeling analysis

DAVID M. RIEFER
California State University, San Bernardino, California

and

WILLIAM H. BATCHELDER
University of California, Irvine, California

An experiment is reported that examined the memory differences between 23 college students (21-41 years old) and 23 elderly adults (64-92 years old). Both groups memorized lists of categorically related words over a series of three study-test trials in a free-recall design. Consistent with prior research on aging effects, the elderly subjects performed more poorly on the number of words recalled, the number of categories recalled, the number of items recalled per category, and the amount of category clustering. To help determine if these differences were due to storage or retrieval processes, Batchelder and Riefer's multinomial model for measuring storage and retrieval was used to analyze the data. The model suggested that age differences were primarily due to retrieval factors and not to storage factors. These results are consistent with a number of retrieval-based theories of aging effects.

The finding that memory deficits occur in the elderly has been well documented in a variety of different studies (see Burke & Light, 1981, and Poon, 1985, for reviews). But a question that still remains is which specific cognitive functions actually decline with age and which do not. Some theorists (e.g., Schonfield & Robertson, 1966; Shaps & Nilsson, 1980) have argued that older subjects suffer from impairment of retrieval processes, with no corresponding decrement in storage capacity. Experimental support for this idea comes from research showing that age differences that occur in free-recall tasks tend to disappear when memory is tested by cued recall (e.g., Laurence, 1967) or recognition (e.g., Schonfield & Robertson, 1966). This finding is consistent with retrieval explanations of aging effects, since retrieval processes are assumed to be lessened or eliminated with cued recall or recognition. In addition, Craik and Masini (1969) have shown that memory impairment in the elderly is characterized by differences in the number of categories recalled from a categorized list but not in the number of items recalled per category. This finding suggests that the elderly have poorer accessibility, or retrieval, of items but have normal storage, or availability, of items.

Other researchers, however, have questioned this interpretation of the above findings. One problem with the "retrieval hypothesis" is that many of the findings support-

ing this explanation have not always been replicated. For example, there are studies that show that age differences are still evident for cued recall (Drachman & Leavitt, 1972), recognition (White & Cunningham, 1982), and items per category (Hultsch, 1975). Also, a number of theorists (e.g., Craik & McDowd, 1987) have pointed out the methodological problems inherent in comparing free-recall performance with recognition performance. Perhaps the most comprehensive criticism comes from Smith (1980), who argues that encoding explanations of aging effects cannot totally be ruled out by examining the contrast between free recall, cued recall, and recognition. This is because differences in retrieval inherently depend on the initial levels of encoding, which may not always be equivalent in young and old adults. These considerations have led Smith (1980) and other theorists (e.g., Drachman & Leavitt, 1972) to conclude that the elderly do not suffer from retrieval deficits alone but also exhibit some impairment in their storage processes. Thus, the storage-retrieval question remains a matter of some debate.

One of the limitations of previous research on the storage-retrieval debate is that substantive models of memory have rarely been used to address this issue (see Howe, 1988, for a notable exception). Instead, most researchers have used comparisons between various ad hoc statistics, such as cued recall, item recognition, and items recalled per category, to explore the storage and retrieval bases of aging effects. But Batchelder and Riefer (1986, 1990; Riefer & Batchelder, 1988) have explicitly shown that empirical statistics such as the ones above can present a misleading picture of cognitive processing, because they often result from a combination of different cognitive events. In other words, a statistic that is designed to mea-

sure a specific cognitive process can sometimes be influenced by other processes and thus is not a pure measure of any single one. As an alternative to empirical statistics, Riefer and Batchelder (1988) have advocated a type of mathematical modeling, called multinomial modeling, for studying and measuring underlying cognitive processes. With regard to models for the measurement of storage and retrieval, Batchelder and Riefer (1986) have developed and tested a multinomial model that is capable of taking a special type of free-recall data and estimating separate storage and retrieval parameters. As such, it provides a more theoretically motivated measure of storage and retrieval that does not rely on an empirical comparison of ad hoc statistics.

The experimental paradigm behind the model involves a free-recall task in which subjects memorize a list of words containing several category pairs (e.g., oxygen and hydrogen) and singletons (unpaired category words). The recall of each category pair is scored into one of four mutually exclusive events: (1) E_1—both words recalled, adjacently, (2) E_2—both words recalled, nonadjacently, (3) E_3—only one word in the pair recalled, and (4) E_4—neither word in the pair recalled. The recall of the singletons is scored into two events: (1) F_1—recalled and (2) F_2—not recalled. N_i is defined as the frequency of occurrence for recall event E_i, and M_j is the frequency of event F_j, with $N = N_1+N_2+N_3+N_4$ and $M = M_1+M_2$. The model assumes that these recall statistics are a function of three probabilistic events: (1) c—the probability of forming and storing a cluster, (2) r—the probability of retrieving a cluster, if stored, and (3) u—the probability of recalling a nonclustered item as a singleton. In essence, c measures the storage of items, whereas r measures their retrievability. Batchelder and Riefer (1986) have provided closed-form, maximum-likelihood estimators for the three parameters, as well as information on how to conduct various hypothesis tests about differences in parameter values across experimental groups.

Thus, the purpose of this article is to apply Batchelder and Riefer's (1986) storage-retrieval model to the issue of whether memory differences due to aging are the result of storage loss, retrieval failure, or both. Specifically, the model is applied to a free-recall experiment in which groups of both young and old adults memorized lists of categorically related words over a series of three study-test trials. Many previous studies have also examined aging effects by using a serial free-recall paradigm, and so it was expected that the results of this experiment would show a memory deficit in the elderly group similar to that in the young adults. However, by using the model to measure separate storage and retrieval parameters from the data, it can be determined what combination of the two factors contribute to the differences between the young and elderly groups.

METHOD

Subjects

A total of 46 volunteers (23 young adults and 23 elderly adults) participated in the experiment. The young adults were college students from

California State University at San Bernardino. Their ages ranged from 21 to 41, with a mean of 29.9. The elderly participants were recruited from a local retirement center and a senior citizens church group. Their ages ranged from 64 to 92, with a mean of 77.0. Questionnaires were administered to all participants to ascertain their education, health, and vision. All of the elderly subjects included in the study were in good health and had good vision with glasses. Also, the majority of the elderly subjects (74%) had some college education; the rest had finished high school.

Materials

Subjects were presented with lists of 39 words, each consisting of 16 category pairs (e.g., oxygen and hydrogen) and 7 singletons (individual words that did not have a category mate). All words were selected from the Battig and Montague (1969) category norms. Howard (1980) has shown that these norms are generally appropriate for use with elderly populations, and care was taken to select categories that would be equally familiar to both age groups. Twenty-four total categories from the Battig-Montague norms were used, and two high associates from positions 2 to 5 were selected from each category. The 39 items for each list were selected from this pool of 48 words, with a different set of 39 items randomly selected for each group of subjects run in the experiment. To control for serial-position effects, 3 pairs and 1 singleton were randomly selected for inclusion in the primacy region of the list (first 7 items), and another set of 3 category pairs and 1 singleton was selected for the recency region (last 7 items). The first and last 7 items were not included in the data analysis. The remaining portion of the list consisted of 10 category pairs and 5 singletons. Each singleton was generated by first selecting a category pair and then randomly choosing one of the two exemplars from that category to serve as the singleton.

Procedure

All subjects were run in small groups—the college students within university classrooms and the elderly subjects at either their retirement center or their church. Subjects attempted to memorize and recall three word lists in a series of successive study-test trials. Words were presented one at a time on a slide projector, at a rate of 5 sec/word. The same 39 words appeared within each of the three presentations of the list but in a different order of presentation for each trial. Presentation order was randomly determined, under the constraint that a lag of between 2 to 6 items appeared between the two members of each category pair. After the presentation of each list, subjects were given 3 min for written free recall.

RESULTS

Empirical Analysis

Except where noted, all statistical tests were conducted using the .01 rejection level. Table 1 presents the proportion of words recalled, $P(c)$, for the young and old adults

Table 1
Recall Data for Young and Old Adults

	$P(c)$	$P(cat)$	IPC	RR	N_1	N_2	N_3	N_4	M_1	M_2
			Trial 1							
Young	.41	.53	1.58	.11	48	25	50	107	37	78
Elderly	.23	.33	1.53	.09	38	5	33	154	15	100
			Trial 2							
Young	.66	.76	1.77	.24	104	33	38	56	69	46
Elderly	.37	.47	1.58	.12	54	14	41	121	34	81
			Trial 3							
Young	.78	.84	1.88	.29	125	47	21	37	86	29
Elderly	.47	.59	1.68	.19	81	15	39	95	44	71

Note—$P(c)$ = probability correct; $P(cat)$ = proportion of categories recalled; IPC = items per category (out of 2); RR = ratio of repetition; N_1 = both items recalled adjacently; N_2 = both items recalled nonadjacently; N_3 = one and only one item recalled; N_4 = neither item recalled; M_1 = singleton recalled; M_2 = singleton not recalled.

across the three study-test trials. Not surprisingly, the recall for both groups significantly improved across trials [$F(2,88) = 122.59$, $MS_e = 5.78$]. Also as expected, the young adults recalled significantly more words than did the elderly across all three trials [$F(1,44) = 28.93$, $MS_e = 50.06$]. These main effects are qualified, however, by a significant age × trials interaction [$F(2,88) = 7.11$, $MS_e = 5.78$], which revealed that the difference in recall between the young and the elderly subjects became larger over trials.

In addition to the proportion of words correctly recalled, three other measures appropriate to categorized lists were examined: P(cat), the proportion of categories represented in recall; IPC, the proportion of items recalled per category; and the ratio of repetition (RR), a clustering measure used by Sanders, Murphy, Schmitt, and Walsh (1980) to examine the effects of age on organizational processes in the elderly. All of these statistics were computed for the category-pair data only, and the results are presented in Table 1. All three measures showed significant improvement over trials for both young and old subjects [P(cat), $F(2,88) = 63.74$, $MS_e = 1.47$; IPC, $F(2,82) = 17.53$, $MS_e = 0.03$; RR, $F(2,88) = 34.35$, $MS_e = 2.31$]. In addition, the college students had significantly higher scores than the elderly for P(cat) [$F(1,44) = 27.91$, $MS_e = 9.78$], IPC [$F(1,41) = 4.95$, $MS_e = 0.14$, $p < .05$], and RR [$F(1,49) = 14.23$, $MS_e = 9.20$].

Model Analysis

Although the experimental results show reliable age differences for all of the empirical measures, these analyses in themselves do not indicate whether the differences are due to storage or retrieval. To determine this, the N_i and M_j data statistics for the storage–retrieval model were tabulated and are presented in Table 1. These values were summed across subjects, and the logic behind applying the model to group aggregate data rather than to individual subject data is addressed in detail in Riefer and Batchelder (1991).

The first step in analyzing the data is to determine the goodness of fit of the model. As shown by Batchelder and Riefer (1986), this can be determined by the log-likelihood ratio statistic G^2, which is asymptotically distributed as a chi-square variable with 1 degree of freedom. The G^2 goodness-of-fit values are presented in Table 2 for each data set, and in general, the fit of the model is fairly good, with only one data set leading to rejection of the model. The overall fit of the model, computed across all six data sets in the experiment, also provides an acceptable fit to the data [$G^2(6) = 10.62$, $p > .10$].

Table 2 also presents the maximum-likelihood estimates for the three parameters and 95% confidence intervals for parameters c and r. As can be seen, there were small storage differences (as measured by c) between the young and the elderly subjects across the three trials; however, the differences were not statistically reliable [$G^2(3) = 2.41$, n.s.]. In contrast, significant differences were found between the two groups on the retrieval parameter r. Although the groups did not significantly differ for Trial 1

Table 2
Parameter Estimates and Goodness-of-Fit Measures for Young and Old Adults

	c	r	u	G²(1)
Trial 1				
Young	.47 (.35, .60)	.44 (.30, .57)	.38	5.44*
Elderly	.40 (.11, .68)	.42 (.11, .73)	.15	1.46
Trial 2				
Young	.64 (.56, .71)	.70 (.62, .79)	.61	0.23
Elderly	.56 (.43, .69)	.42 (.30, .54)	.32	1.66
Trial 3				
Young	.69 (.63, .75)	.79 (.72, .86)	.78	1.48
Elderly	.63 (.53, .73)	.56 (.45, .66)	.40	0.35

Note—c = the probability of forming and storing a cluster; r = the probability of retrieving a cluster, if stored; u = the probability of recalling a nonclustered item as a singleton. Numbers in parentheses are the 95% confidence interval for each parameter. $G^2(1)$ is the log-likelihood ratio statistic. *$p < .05$.

[$G^2(1) = 0.02$, n.s.], the college students did retrieve significantly more items on Trial 2 [$G^2(1) = 7.34$] and Trial 3 [$G^2(1) = 8.52$]. Table 2 also reveals that the young subjects had larger values of the "nuisance" parameter u than did elderly subjects. However, that parameter combines both storage and retrieval processes, so this difference is difficult to interpret. One advantage of the model is its ability to factor out this source of difference in assessing storage and retrieval differences.

One might argue that these results could be due, in part, to the difference in education level between the two groups. Some researchers (e.g., Perlmutter, 1978) have pointed out that education level is a possible confounding variable in aging research and hence is a legitimate source of concern. In an attempt to examine this factor, the 23 elderly subjects were divided into two subgroups, the 11 who attended more than one year of college and the 12 who did not. No significant differences either in recall or in the model's parameters were found between the two subgroups. Moreover, the college-educated subgroup of elderly subjects still exhibited substantially poorer retrieval scores than did the young college students. This tends to indicate that the retrieval advantage for the young college students was not merely due to their higher average education level.

DISCUSSION

The empirical results of both the experiment and the model's analysis are consistent with retrieval-based explanations of age differences in memory. The finding that elderly subjects recalled fewer categories than did the young subjects is consistent with the results of Craik and Masini (1969) and Hultsch (1975), and in itself supports the idea that category pairs that were available for recall in the experiment were not

as accessible to the older subjects. In addition, the model's analysis provides even stronger support that it is retrieval failure, and not problems with storage, that underlies the memory differences in the current experiment. Although differences between the two age groups were not evident for the first trial, the model revealed that the elderly subjects exhibited impairment of retrieval for later trials, without a corresponding loss of storage.

It is interesting to note that the elderly group also recalled fewer items per category. This is an empirical finding that has traditionally been used as an indication of storage deficits in memory. For example, Hultsch (1975), who found aging differences for *IPC* in an experiment similar to this one, concluded that the elderly subjects in his experiment exhibited a decrement in storage. If one were to examine only the empirical pattern of results in the experiment reported here, one might also be inclined to conclude that there were storage differences between the age groups. But from the perspective of the multinomial model's analysis, that would be an incorrect conclusion and would reinforce our earlier point that mathematical modeling provides a more theoretically motivated and potentially superior method for measuring underlying cognitive processes.

It should be noted that Bayen (1990) and Erdfelder and Bayen (in press) have used our storage-retrieval model in their independently reported experiments on aging in German-speaking populations. In their experiments, young and elderly subjects were presented with three presentations of a list containing category pairs separated by lags of 0 and 15 items. The results of their more extensive studies were similar to ours for Trial 1, but they found significant effects for both parameters *c* and *r* on later trials for pairs separated by Lag 0. Thus, there seems to be strong evidence for an age-related decrease in the retrieval parameter in the current paradigm; however, in some cases, there may also be small but reliable declines in the storage parameter.

Clearly, more model-based work is needed in this area to sort these issues out. Even though storage was not shown to be strongly influenced by aging in the experiment reported here, it is possible that storage deficits may be found by using other models with different memory tasks. For example, Howe (1988), using a two-stage Markov model, found reliable storage differences between young and elderly groups, although he found retrieval effects to be stronger. In addition, multinomial models exist for measuring storage and retrieval in other paradigms (e.g., Chechile, 1987; Wilkinson, DeMarinis, & Riley, 1983), and, in fact, a number of these have been used to examine differences between children and adults (see Brainerd, 1985, for a review).

In addition, even though the model tested here implicates retrieval deficits as the source of age differences in memory, the results of this analysis say nothing about which particular aspects of retrieval are impaired or why that impairment occurs. Many different hypotheses have been proposed concerning the source of retrieval problems in the elderly, and further experimental research is needed to establish the specific retrieval deficits that occur in old age. The major point of this article has been to demonstrate how a particular mathematical model of measurement can be used to augment traditional analyses in exploring these issues. It is reasonable to presume that theorists in this area will continue to be interested in the underlying cognitive bases for age differences in memory, and it is hoped that researchers will increase their use of simple, formal models to investigate these issues.

REFERENCES

BATCHELDER, W. H., & RIEFER, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical & Statistical Psychology*, **39**, 129-149.

BATCHELDER, W. H., & RIEFER, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, **97**, 548-564.

BATTIG, W. F., & MONTAGUE, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monographs*, **80**(3, Pt. 2).

BAYEN, U. J. (1990). *Zur Lokalisation von Altersdifferenzen im episodischen Gedaechtnis Erwachsener* [The localization of adults' age-differences in episodic memory]. (Report No. 16-2). Bonn: University of Bonn, Psychological Institute.

BRAINERD, C. J. (1985). Model-based approaches to storage and retrieval development. In C. J. Brainerd & M. Pressley (Eds.), *Basic processes in memory development: Progress in cognitive development research* (pp. 143-208). New York: Springer-Verlag.

BURKE, D. M., & LIGHT, L. L. (1981). Memory and aging: The role of retrieval processes. *Psychological Bulletin*, **90**, 513-546.

CHECHILE, R. A. (1987). Trace susceptibility theory. *Journal of Experimental Psychology: General*, **116**, 203-222.

CRAIK, F. I. M., & MASINI, P. A. (1969). Age and intelligence differences in coding and retrieval of word lists. *British Journal of Psychology*, **58**, 291-299.

CRAIK, F. I. M., & McDOWD, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 474-479.

DRACHMAN, D. A., & LEAVITT, J. (1972). Memory impairment in the aged: Storage versus retrieval deficit. *Journal of Experimental Psychology*, **93**, 302-308.

ERDFELDER, E., & BAYEN, U. J. (in press). Episodisches Gedächtnis im Alter: Methodologische und empirische Argumente für einen Zugang über mathematische Modelle (Episodic memory in old age: Methodological and empirical arguments for a mathematical modeling approach). In D. Frey (Ed.), *Bericht über den 37. Kongress der Deutschen Gesellschaft für Psychologie in Kiel 1990* (Vol. 2). Göttingen, Germany: Hogrefe.

HOWARD, D. V. (1980). Category norms: A comparison of the Battig and Montague (1969) norms with the responses of adults between the ages of 20 and 80. *Journal of Gerontology*, **35**, 884-890.

HOWE, M. L. (1988). Measuring memory development in adulthood: A model-based approach to disentangling storage-retrieval contributions. In M. L. Howe & C. J. Brainerd (Eds.), *Cognitive development in adulthood: Progress in cognitive development research* (pp. 39-64). New York: Springer-Verlag.

HULTSCH, D. F. (1975). Adult age differences in retrieval: Trace-dependent and cue-dependent forgetting. *Developmental Psychology*, **1**, 197-201.

LAURENCE, M. W. (1967). Memory loss with age: A test of two strategies for its retardation. *Psychonomic Science*, **9**, 209-210.

PERLMUTTER, M. (1978). What is memory aging the aging of? *Developmental Psychology*, **14**, 330-345.

POON, L. W. (1985). Differences in human memory with aging: Nature, causes, and clinical implications. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 427-462). New York: Van Nostrand Reinhold.

RIEFER, D. M., & BATCHELDER, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, **95**, 318-339.

RIEFER, D. M., & BATCHELDER, W. H. (1991). Statistical inference for multinomial tree models. In J.-C. Falmagne & J.-P. Doignon (Eds.), *Mathematical psychology: Current developments* (pp. 313-336). Berlin: Springer-Verlag.

SANDERS, R. E., MURPHY, M. D., SCHMITT, F. A., & WALSH, K. K. (1980). Age differences in free recall rehearsal strategies. *Journal of Gerontology*, **26**, 470-475.

SCHONFIELD, D., & ROBERTSON, B. A. (1966). Memory storage and aging. *Canadian Journal of Psychology*, **20**, 228-236.

SHAPS, L. P., & NILSSON, L. G. (1980). Encoding and retrieval operations in relation to age. *Developmental Psychology*, **16**, 636-643.

SMITH, A. D. (1980). Age differences in encoding, storage, and retrieval. In L. Poon, J. Fozard, L. Cermak, D. Arenberg, & L. Thompson (Eds.), *New directions in memory and aging* (pp. 23-45). Hillsdale, NJ: Erlbaum.

WHITE, N., & CUNNINGHAM, W. R. (1982). What is the evidence for retrieval problems in the elderly? *Experimental Aging Research*, **8**, 169-171.

WILKINSON, A. C., DeMARINIS, M., & RILEY, S. J. (1983). Developmental and individual differences in rapid remembering. *Child Development*, **54**, 898-911.