Longitudinal Changes in Quantitative and Qualitative Indicators of Word and Story Recall in Young-Old and Old-Old Adults

Brent J. Small, Roger A. Dixon, David F. Hultsch, and Christopher Hertzog

Department of Gerontology, University of South Florida, Tampa.
Department of Psychology, University of Victoria, Victoria, British Columbia.
School of Psychology, Georgia Institute of Technology, Atlanta.

The present study examined longitudinal changes in quantitative and qualitative measures of episodic memory. The sample, taken from the Victoria Longitudinal Study, consisted of 158 young-old adults (initially 55 to 70 years old) and 54 old-old adults (initially 71 to 86 years old) who were tested three times over six years. Average word and text recall, as well as five indicators of qualitative aspects of word recall (e.g., number of categories recalled) and one indicator of structure of text recall (i.e., levels of information) were used. For word recall, although both age groups exhibited negative longitudinal changes in quantitative performance, overall qualitative performance was generally stable. Two qualitative indicators (number of categories and intrusions) showed modest decline and one (organization at recall) showed improvement. Results for overall text recall showed significant performance increments for the young-old group, whereas the old-old group exhibited slight declines in overall performance. Analyses of qualitative measures showed stable structure of hierarchical recall, with the old-old being impaired at all levels of detail in the stories. Overall results suggest that some underlying structural characteristics of word and text recall may be maintained into late life even when significant overall decline is observed.

Much previous cross-sectional research has confirmed that older adults perform worse than younger adults on numerous episodic memory tasks (e.g., Bäckman, Small, & Larsson, in press; Craik & Jennings, 1992). Longitudinal studies provide unique information pertaining to actual changes in performance over time. Recent longitudinal investigations of episodic memory have focused on such issues as whether (a) normal aging-related change is relatively gradual or precipitous in late life, (b) aging-related change is uniform across the older adult age span, and (c) change patterns vary by episodic task and type of performance indicator. In this article we provide new evidence from the Victoria Longitudinal Study (VLS) pertaining to these issues.

In an earlier report from the VLS, no significant changes were observed over three years for two measures of episodic memory: word and text recall over three years (Hultsch, Hertzog, Small, McDonald-Misczak, & Dixon, 1992). This finding was not entirely unexpected given the relatively short three-year retest interval. Zelinski, Gilewski, and Schaie (1993) also did not find significant decline on measures of word and story recall over a brief (two-year) period. Other studies have reported declines on episodic measures over intervals from four to six years (e.g., Colsher & Wallace, 1991; Taylor, Miller, & Tinklenberg, 1992) and longer (Giambra, Arenberg, Zonderman, Kwas, & Costa, 1995; McCarty, Siegler, & Logue, 1982; Zelinski & Burnight, 1997). Using an age/cohort/period analysis for VLS data, Hultsch, Hertzog, Dixon, and Small (1998) reported differential age change patterns for word and text recall. Specifically, over a six-year longitudinal interval, we observed significant age decline for a quantitative indicator of categorical word recall (i.e., total number of words recalled), but not for a comparable quantitative indicator of gist text recall (i.e., total number of propositions recalled). However, the data also suggested the possibility of modest declines in text recall at the oldest age ranges.

If actual longitudinal decline in total recall is relatively gradual, does the same pattern apply to other memory performance indicators? Most previous research on longitudinal changes in episodic memory functioning has focused on the simplest quantitative measures of performance. Whereas cognitive products (e.g., number of words correctly recalled) are typically classified as quantitative measures, other performance- or process-related indicators may be considered as qualitative measures. In gerontological research, the potential importance of supplementing quantitative measures of memory performance with more qualitative indicators has been recognized for some time (e.g., Riegel, 1973). A wide range of qualitative indicators and interpretations have emerged (e.g., Adams, 1991; Adams, Smith, Nyquist, & Perlmutter, 1997; Dixon & Bäckman, 1993; Labouvie-Vief, 1985). For example, some observers have argued that selected indicators of qualitative aspects of cognitive processing could assume a privileged status in aging research because they are believed to represent more accurately the quality of older adults' adaptive cognitive processing. In contrast, qualitative indicators can be viewed as useful sources of information about episodic memory and processes that support total performance (e.g., Gould, Trevithick, & Dixon, 1991; Salthouse, 1991; Stine & Wingfield, 1990). A continuum of qualitative measures may be posited, with indicators ranging from those closely linked to total recall to those at greater conceptual and empirical distance. In addition, a continuum of methods for assessing qualitative performance ranges from direct experimental manipulations of specific processes to indirect assessments based on evaluations of output characteristics.

There are compelling reasons for examining changes in qualitative characteristics of episodic word and text recall in older
adults. Regarding the former, theorists have long argued that recalling supraspan word lists involves multiple processes, including those focused on the relations among items (relational processing) and those focusing on the unique characteristics of the items (item-specific processing) (Einstein & Hunt, 1980; Hunt & Einstein, 1981; Tulving, 1962). In cognitive aging, several studies have found that older adults have particular difficulties in spontaneously engaging in relational or organizational strategies (e.g., Allen & Coyne, 1989; Basden, Basden, & Bartlett, 1993; Hultsch, 1971; Luszcz, Roberts, & Mattiske, 1990; Stuss, Craik, Sayer, Franchi, & Alexander, 1996; Witte, Freund, & Brown-Whistler, 1993). Whether this reflects an impaired or deteriorating ability to employ these types of processes with aging is not clear, partly because most studies examining organizational measures have been cross-sectional. Thus, the nature and extent of changes in various indicators of relational processing in later life is not known.

Four key characteristics of the present study are unique in the literature on aging and qualitative memory performance. First, although we have a select set of qualitative performance indicators, we have examined three-occasion (six-year) longitudinal changes in these measures for young-old and old-old adults. To our knowledge, all previous research has compared the qualitative performance of younger and older adults using cross-sectional designs. Second, much previous research on qualitative memory performance has been conducted with novel tasks (e.g., interpreting stories) or materials (e.g., new or adapted stories). A central focus of the VLS was to examine quantitative cognitive change and we adopted memory tasks with this objective in mind. Although this fact limits the range of qualitative memory indicators available to us, it facilitates examining the linkage between quantitative and qualitative aspects of episodic performance. We focus only on those qualitative indicators that are conceptually closest to total performance. The third key characteristic of this study is that we examine changes in qualitative indicators in two different episodic memory tasks, namely word and story recall. Most previous research has employed only one of these traditional episodic tasks. Finally, the fourth notable aspect of this study is that we are able to compare the above three characteristics in two cohorts of older adults, a young-old and old-old group.

We obtained three classes of measures linked to relational processing of categorical word lists. First, we calculated both the number of different categories represented in each participant’s recall, and the number of words within each of these categories (Tulving & Pearlstone, 1966). Second, we tallied the number of words repeated in the protocol, as well as the number of intrusions into the protocol from categories that were or were not represented in the current lists. Third, a measure of the structural organization of the recalled words, the Adjusted Ratio of Clustering (ARC), was computed (Roenker, Thompson, & Brown, 1971). The ARC ratio may also be correlated with actual recall scores (quantitative measure), perhaps differentially for younger and older adults (Bäckman & Larsson, 1992).

Regarding text recall, the typical quantitative performance measure is number of propositions, or idea units, recalled. Our qualitative indicator reflected the structure of the recalled propositions. Models of text processing view texts as sets of connected and hierarchically organized propositions. As a result, propositional levels differ such that the higher levels represent the more important ideas of the text and the lower levels represent the details of the text (Kintsch & van Dijk, 1978).

Research consistently reveals a levels effect such that the propositions higher in the hierarchy have a greater probability of being recalled than propositions lower in the hierarchy (e.g., Kintsch, Kozminsky, Strebly, McKoon, & Keenan, 1975; Meyer, 1987). Such a pattern may be interpreted as reflecting a relatively well-structured retelling of the original story.

Both younger and older adults exhibit a pronounced levels effect in their recall of prose materials, indicating general sensitivity to text structure (e.g., Dixon, Simon, Nowak, & Hultsch, 1982; Meyer & Rice, 1989; Stine, Soederberg, & Morrow, 1996; Stine & Wingfield, 1990). Nevertheless, whether there are age differences in the levels effect function is not yet clear. Some studies report that older adults show less discrimination among levels of text than younger adults (e.g., Byrd, 1985; Dixon, Hultsch, Simon, & von Eye, 1984; Meyer & Rice, 1983), whereas other studies find no difference (e.g., Petros, Norgaard, Olson, & Tabor, 1989; Tun, 1989). However, when processing demands of the task are high, either as a function of characteristics of the task (e.g., complex syntactic structure) or the rememberer (e.g., low verbal ability), older adults may differentiate among levels of propositions less sharply than younger adults (Stine & Wingfield, 1990). To date, no longitudinal data pertaining to qualitative characteristics of text recall are available. In the present study, we examined whether there were aging-related changes in discriminability among the levels of information in text recall.

Finally, we considered one additional issue, namely, whether qualitative performance differences were related to gender. In our previous reports (Hultsch, Masson, & Small, 1991; Hultsch et al., 1992), gender effects have occasionally appeared, but have not been interpreted as playing a major role in quantitative episodic memory deficits in older adults. However, we have recently reported significant standardized effects for gender on both quantitative word and text recall (Hultsch et al., 1998). A recent article by Herlitz, Nilsson, and Bäckman (1997) presented data from a population-based sample indicating gender differences among episodic memory tasks. Across several word list tasks, small but statistically significant differences in favor of women were observed. Herlitz and colleagues tentatively speculated that the gender effect may be due more to differential problems with encoding than retrieval. In the present study, because we have performance-based qualitative indicators of encoding and retrieval processes present for word recall, we can explore the Herlitz and colleagues hypothesis that gender differences are related more to the former than the latter.

**Method**

**Participants**

The participants consisted of 242 adults who participated in three waves of the VLS. Participants in the VLS were initially aged 55 to 85 years and were examined at 3-year intervals over a period of 6 years. At baseline, 487 adults (290 women and 197 men) participated in the testing sessions. Approximately three years later (M = 2.92 years), 335 individuals returned for retesting. At the third time of testing, 250 individuals returned approximately 3 years later (M = 2.99 years). Further information regarding the design, sample, and specific reasons for attrition can be found elsewhere (Hultsch et al., 1998).

Because of missing data, the final sample in the present study consisted of 242 adults (143 women and 99 men). The sample
was divided, based on age at entry into the study, into two groups: young-old (n = 158; 55–70 years; M = 64.84) and old-old (n = 84; 71–86 years; M = 74.49). In general, the young-old and old-old age/cohort groups were well educated (M = 13.72 years, 13.45 years, respectively) and rated their health to be good to very good on a 5-point scale (MTO = .63; MOT = .70 on a scale ranging from 0 = very good to 4 = very poor). A multivariate analysis of variance (MANOVA) performed on the background characteristics revealed no overall effects of age, gender, or their interaction.

Measures

Word recall.—Six categorized lists of common English nouns were developed from the Howard (1980) and Battig and Montague (1969) norms. Each list contained six words from each of five taxonomic categories for a total of 30 words per list. Categories and exemplars were chosen to minimize potential interference effects within and between lists. In general, high-frequency exemplars ranked two through nine were chosen, but to minimize guessing, the most frequently ranked noun was not used. Each participant studied and recalled two lists at each time of measurement. The presentation of the list was counterbalanced across times of measurement to minimize the effects of time of measurement. The lists were presented in unblocked order in typed booklets for study. Participants had 2 minutes to study the words, followed immediately by a 5-minute written recall test. Participants were instructed to write down as many of the words as they could remember in any order. Parallel forms reliability of the word lists at the three times of measurement were .66, .56, and .71, respectively.

In addition to performance on the total recall of words, five other measures were derived to represent more qualitative characteristics that may underlie aging-related deficits in overall recall. First, recall was partitioned into number of categories recalled, which was used as a measure of the retrieval of superordinate organizational units (Tulving & Pearlstone, 1966). A category was considered to be recalled if at least one word from the category was remembered. Second, the average number of words recalled per category was used as an index of retrieval of subordinate units, given retrieval of the superordinate unit. Third, interference was indexed by the number of intrusions (words from the category that were not on the list). Information on two additional intrusion measures that indexed the frequency of incorrect items that were from categories other than those seen by the participant was gathered. However, because the occurrence of these intrusions was relatively infrequent (MT1 = .04; MT2 = .12; MT3 = .08), they were eliminated from the present analyses. The fourth qualitative measure of word recall measured the number of repeated words and can be considered to be a measure of online monitoring of memory functioning (Koriat, Ben-Zur, & Sheffer, 1988; Stuss et al., 1996). Fifth, the ARC score (Roenker et al., 1971) was computed. This measure provides an index of the degree to which the individual has clustered his or her recall, and is typically interpreted as an index of the degree of categorical organization that has been applied in encoding and retrieving the words. In addition, the ARC score was correlated with total recall.

Text recall.—Six narrative stories from a set of 25 parallel texts developed by Dixon, Hultsch, and Hertzog (1989, 1993) were used. Each story described an event in the life (lives) of an older protagonist, who was either a woman (two stories), a man (two stories) or a couple (two stories). The structurally equivalent stories consisted of 24 sentences and contained approximately 300 words, organized into approximately 160 propositions (Dixon et al., 1993; Kintsch, 1974). The stories were well organized, and contained the main theme of the story in the first few propositions. Parallel forms reliability of the stories at the three times of measurement were .68, .75, and .76, respectively.

Each participant studied and recalled two texts at each time of measurement. Similar to word recall, the texts were counterbalanced across time to minimize practice effects. They were presented in typed booklets for study followed by written recall. Participants were given four minutes to study each story and ten minutes to write their recall. Participants were instructed to recall as much of the substance of the story as possible, including the main ideas and details. They were told that they could recall the story in their own words, those of the story, or both.

To score text recall, the template text base is compared to the recall protocol produced by the individual (Turner & Green, 1978). Gist recall of the propositions in the texts was used as a measure of quantitative recall. Although more information on the scoring system can be found in Dixon and colleagues (1989, 1993), it should be noted that propositions were scored independent of their position in the hierarchy. Reliability estimates on this scoring system, within and across times of measurement, showed percentage agreement invariably above .90. Our procedures for scoring recall protocols produced highly reliable scores both within and across occasions of measurement.

In the present study, we focus on the one qualitative indicator available in the VLS. The propositions of the six stories were rated in terms of their hierarchical level. Two expert judges collaborated in developing the first draft of the ratings, and these initial ratings were independently evaluated by a third expert judge. This process generated ratings for each proposition of a story ranging from 1 to 7, corresponding to a transition from the main ideas (1) to specific details (7) contained in the stories. For purposes of analysis, these levels were collapsed into four categories: Levels 1–2, Level 3, Level 4, and Levels 5–7. The measures that were summed across several levels were first weighted to reflect differences in the frequency of their occurrence.

Procedure

The test battery from the VLS was administered across multiple sessions scheduled approximately one per week. The word and text recall tasks were administered in the first and second testing sessions at each time of measurement. One word list and one story were presented in each session. A complete description of the procedures is available elsewhere (Hultsch et al., 1998).

Results

The data analysis consisted of two parts. First, we examined longitudinal changes in the quantitative and qualitative characteristics of word recall performance. Second, we examined changes in the quantity of propositions recalled, as well as the extent to which longitudinal changes in performance varied as a function of level of proposition in text recall.

Word Recall

The means and standard deviations for the quantitative and qualitative indicators of episodic word recall are shown in Table
A 2 (Gender) × 2 (Age/Cohort) × 3 (Time of Testing) repeated measures MANOVA was computed on the total number of words recalled. The results indicated a significant main effect for Age/Cohort group \(F(1,238) = 17.31, \text{MSE} = 39.19, p < .001, \eta^2 = .07\), whereby the young-old group recalled significantly more words than the old-old group \(M_{\text{YO}} = 18.60; M_{\text{OO}} = 16.38\). In addition, the main effect for Time of Testing was reliable \(F(1,238) = 22.23, \text{MSE} = 4.63, p < .001, \eta^2 = .09\), indicating significant longitudinal declines in performance across the retest interval \(M_{T1} = 18.17; M_{T2} = 17.97; M_{T3} = 17.36\). None of the interactions between Age/Cohort, Gender, or Time of Testing approached statistical significance.

The next set of analyses attempted to determine whether the source of the significant effects in quantitative word recall could be attributed to performance decrements in one or all of the qualitative indices. A 2 (Gender) × 2 (Age/Cohort) × 3 (Time of Testing) MANOVA was computed on the number of categories recalled, number of words per category, number of intrusions, number of repeated words, and the ARC score. Only the main effects for Age/Cohort group \(F(1,238) = 15.32, \text{MSE} = .40, p < .001, \eta^2 = .06; \) recalled more words per category, \(F(1,238) = 12.23, \text{MSE} = 1.13, p = .001, \eta^2 = .05; \) and produced more intrusions, \(F(1,238) = 4.89, \text{MSE} = 2.36, p = .028, \eta^2 = .02.\) The marginal means for the age group effects were as follows: Categories recalled: \(M_{\text{YO}} = 4.70; M_{\text{OO}} = 4.49; \) words per category: \(M_{\text{YO}} = 3.96; M_{\text{OO}} = 3.65; \) and intrusions: \(M_{\text{YO}} = .98; M_{\text{OO}} = .75.\)

Significant longitudinal change was observed on three measures. There was a significant linear decline in the number of categories recalled, \(F(1,238) = 15.32, \text{MSE} = .14, p < .001, \eta^2 = .06, a \) significant linear increase in the number of intrusions, \(F(1,238) = 8.53, \text{MSE} = .16, p = .004, \eta^2 = .04, \) and a significant increase in ARC score, \(F(1,238) = 7.01, \text{MSE} = .66, p = .009, \eta^2 = .04.\) The marginal means for those tasks with significant time of testing effects were as follows: Number of categories recalled: \(M_{T1} = 4.69; M_{T2} = 4.62; M_{T3} = 4.58; \) intrusions: \(M_{T1} = .83; M_{T2} = .86; M_{T3} = 1.02; \) and ARC score: \(M_{T1} = .69; M_{T2} = .72; M_{T3} = .75.\)

Although the gender effect for total (quantitative) recall was not significant \(F(1,238) = 2.07, \text{MSE} = 39.19, p = .152, \eta^2 = .01, \) the size \(d; \text{Cohen, 1977}\) and direction \(M_{\text{M}} = 18.27; M_{\text{W}} = 17.29\) of the effect were very similar to those reported by Herlitz and colleagues (1997). In the present study, the gender effect in total word recall averaged .24 across the three times of measurement \((d_1 = .23; d_2 = .29; d_3 = .20).\) In a study with a larger sample, Herlitz and colleagues (1997) found a similar sized effect \(d = .25\) in the same direction for the recall of 12 unrelated words. We therefore examined the univariate effects for gender in each of the five qualitative measures. Contrary to the expectations of Herlitz and colleagues, the only significant gender effect was for the retrieval-linked number of categories recalled variable \(F(1,238) = 4.60, \text{MSE} = .40, p = .033, \eta^2 = .02.\) Women \(M = 4.68\) produced more categories than did men \(M = 4.55.\) Examination of the longitudinal and age/cohort profiles revealed that this difference was consistent across these two factors.

Finally, we computed ARC-total recall correlations. These correlations, shown in Table 1, revealed a substantial relationship between quality of organization and total number of words recalled. Moreover, nonsignificant \(p > .10\) cross-time and cross-age comparisons of the magnitude of the correlation coefficients, revealed that they were consistent across the three longitudinal occasions for both young-old and old-old adults.

**Text Recall**

The average proportion of total propositions recalled, as well as the proportion of propositions recalled by story level, are presented in Table 2 by time of measurement and age/cohort group. A 2 (Gender) × 2 (Age/Cohort) × 3 (Time of Testing) repeated measures MANOVA was computed on the total proportion of propositions recalled. The results indicated a significant main effect of Age/Cohort group \(F(1,238) = 13.39, \text{MSE} = .02, p < .001, \eta^2 = .06.\)

### Table 1. Word Recall by Age/Cohort Group and Time of Measurement

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th></th>
<th>Time 2</th>
<th></th>
<th>Time 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Young-Old</td>
<td>Old-Old</td>
<td>Young-Old</td>
<td>Old-Old</td>
<td>Young-Old</td>
</tr>
<tr>
<td>Proportion of words recalled</td>
<td>(M)</td>
<td>.63</td>
<td>.56</td>
<td>.62</td>
<td>.56</td>
</tr>
<tr>
<td></td>
<td>(SD)</td>
<td>.13</td>
<td>.13</td>
<td>.13</td>
<td>.14</td>
</tr>
<tr>
<td>Number of categories recalled</td>
<td>(M)</td>
<td>4.75</td>
<td>4.57</td>
<td>4.66</td>
<td>4.55</td>
</tr>
<tr>
<td></td>
<td>(SD)</td>
<td>.38</td>
<td>.48</td>
<td>.47</td>
<td>.53</td>
</tr>
<tr>
<td>Words per category</td>
<td>(M)</td>
<td>3.99</td>
<td>3.71</td>
<td>3.97</td>
<td>3.67</td>
</tr>
<tr>
<td></td>
<td>(SD)</td>
<td>.66</td>
<td>.69</td>
<td>.70</td>
<td>.75</td>
</tr>
<tr>
<td>Intrusions</td>
<td>(M)</td>
<td>.90</td>
<td>.69</td>
<td>.96</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td>(SD)</td>
<td>1.13</td>
<td>.84</td>
<td>1.32</td>
<td>.84</td>
</tr>
<tr>
<td>Repeated words</td>
<td>(M)</td>
<td>.14</td>
<td>.16</td>
<td>.21</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td>(SD)</td>
<td>.28</td>
<td>.40</td>
<td>.52</td>
<td>.29</td>
</tr>
<tr>
<td>Clustering</td>
<td>(M)</td>
<td>.70</td>
<td>.66</td>
<td>.72</td>
<td>.71</td>
</tr>
<tr>
<td></td>
<td>(SD)</td>
<td>2.5</td>
<td>.27</td>
<td>2.72</td>
<td>.25</td>
</tr>
<tr>
<td>Total recall/ARC</td>
<td>(r)</td>
<td>.56</td>
<td>.45</td>
<td>.39</td>
<td>.54</td>
</tr>
</tbody>
</table>
.05], with the young-old group outperforming the old-old group ($M_{young} = .36; M_{old-old} = .31$). In addition, there was a main effect of gender [$F(1, 238) = 4.27, MSE = .02, p = .040, \eta^2 = .02$], indicating that women performed better than men ($M_w = .36; M_m = .32$). Finally, there was a significant Age/Cohort × Time of Testing interaction [$F(1,238) = 8.92, MSE = .01, p = .003, \eta^2 = .03$]. The interaction reflected the tendency of the younger Age/Cohort group to exhibit a significant linear increase in performance over the 6-year follow-up interval [$F(2,312) = 3.18, MSE = .01, p = .043, \eta^2 = .06$], whereas the older group shows a slight, but not statistically reliable, decrease. None of the other interactions between Age/Cohort, Gender, or Time of Testing were reliable.

To determine whether the aforementioned effects generalized across all hierarchical levels of text information, a 2 (Gender) × 2 (Age/Cohort) × 4 (Text Level) × 3 (Time of Testing) MANOVA was computed on the proportion of propositions recalled. As expected, the effects seen previously were replicated, including main effects for Age/Cohort Group, Gender, and the Age/Cohort Group × Time of Testing interaction. As predicted, the analysis also revealed a significant main effect for levels, Wilks $\lambda = .116, F(3,236) = 600.66, p < .001$. Polynomial contrasts indicated the linear ($t = 37.96, p < .001$), quadratic ($t = 10.63, p < .001$), and cubic ($t = 7.60, p < .001$) trends were statistically reliable. In all cases, recall is highest for the main ideas of the texts (Levels 1–2), is lower at the two intermediate levels (Level 3 and Level 4), and is lowest for the details of the texts (Levels 5–7). Thus we observe a typical levels effect for these stories. None of the other variables (Gender, Age/Cohort Group, Time of Testing) interacted with text level.

**DISCUSSION**

The results from analyses on quantitative word recall revealed significant three-occasion longitudinal declines for both young-old and old-old adults. These results are comparable to a number of other studies that have focused on age-related changes over similar follow-up intervals (e.g., Colsher & Wallace, 1991; Taylor et al., 1992). For example, Colsher and Wallace (1991) reported significant 6-year longitudinal declines in both immediate and delayed word recall in their population-based sample of adults over age 65. Similarly, Giambra and colleagues (1995) reported significant longitudinal declines in visual episodic memory across a 7-year period for a group of 65- to 70-year-olds. By contrast, Schaie (1996) noted significant increments in both immediate and delayed word recall over a 7-year period in the Seattle Longitudinal Study. Note that although we failed to find significant 3-year longitudinal changes in word recall in our previous analysis of the data from the VLS (Hultsch et al., 1992), we now see significant declines 6 years after baseline testing. Thus, this result suggests that decline is relatively slow for older adults. The magnitude of change may not be sufficient to produce reliable differences until relatively long periods are examined. At present this conclusion holds at least for individuals fitting the present age ranges and selection characteristics, and for sample sizes similar to those studied here.

Notably, we demonstrated that qualitative indicators of word recall showed varying profiles of longitudinal change. We observed significant longitudinal changes in three of the five performance indices. There was an overall decrease in the number of categories recalled, together with increases in the number of intrusions that were produced. Both may be viewed as indicating decrements. In contrast, the significant increases in clustering over time can be viewed as improvement. The two measures that did not exhibit any age-related changes in performance were number of words recalled per category and the number of words repeated in the recall attempt. It should be noted that the mean decline observed for number of words per category is rather small in magnitude, with most individuals able to recall at least four of the six items. Similarly, all means for the number of categories variable hovered just below the maximum value of five.

The fact that not all of the qualitative measures of word recall exhibited significant longitudinal changes in performance may be seen as partial evidence for a change in the manner in which the material was recalled. For example, recall of categories is thought to index a person’s retrieval plan, whereas the number of words recalled per category taps the extent to which the items were organized at encoding (Tulving & Pearlstone, 1966). As has been suggested before (e.g., Burke & Light, 1981), the declines observed in the quantitative measure of performance may be linked more to inefficient retrieval from memory, rather than to initial encoding difficulties. Of course, this tentative conclusion is not based on systematic manipulation of encoding or retrieval conditions, but rather on performance-based evidence. It is interesting to note that the retrieval-oriented interpretation of qualitative word recall is consistent with patterns of individual differences in quantitative recall change reported by Hultsch and colleagues (1998). We found that individuals who decline most
on total words recalled from the lists also decline the most in performance on a fact retrieval test which measures access to information in semantic memory. These kinds of materials have been reported to produce greater frequency of retrieval blocks and tip-of-tongue states in older adults (e.g., Maylor, 1990). The strong association of individual differences in amount of cognitive decline on episodic (i.e., word recall) and semantic (i.e., fact recall) may be viewed as indirect evidence that retrieval problems play a role in declining free recall performance in old age.

Our results also suggest that with increasing age, some aspects of older adults’ ability to monitor their retrieval output may become impaired. Although the number of intrusions that were produced increased across the three occasions, the other variable purported to index the breakdown of monitoring, number of repeated words, showed no significant increases in frequency over time. The quite low base rates of these errors, however, should be noted. Again, these indicators are performance-based, and other monitoring variables have been used (Hertzog & Dixon, 1994), although not in such longitudinal designs.

The results pertaining to clustering were quite interesting. Overall, the values indicated that the recall output was quite well organized, equivalently so for young-old and old-old adults. This result is similar to that obtained by other researchers comparing similar age groups cross-sectionally (e.g., Bäckman & Larsson, 1992). Moreover, longitudinal analyses revealed not only that this high level of organizability was maintained across the three occasions by both age groups, but that there was actually a small, but significant improvement over time. Although this may not be attributable to differences in instructions (always free recall) or lists (equivalent within and across times of measurement), it is possible that adults may gain through practice some awareness that the lists are organizeable, and that this may be reflected in the structure of their recall protocols. Notably, however, the cross-sectional age similarities are maintained across three occasions.

Assuming that adults use the organization characteristics of the list items to support their memory performance, the ARC score should be related to overall recall scores. Based on previous cross-sectional work (Bäckman & Larsson, 1992), we had expected that the correlations might vary as a function of age. Results from this previous study showed apparent cross-sectional differences in ARC-recall performance correlations. Whereas young adults produced a relatively high correlation ($r = .5$), young-old ($r = .3$), old ($r = .1$) and old-old ($r = .2$) adults produced quite low correlations. However, our ARC-total recall correlations were relatively high (ranging from .36 to .55) and showed no apparent age- or occasion-related patterns. This suggests that both our young-old and old-old adults identified and used organizational characteristics of the word lists to support their overall memory performance. Despite reliable declines in quantitative performance, and corresponding declines in two aspects of qualitative performance (i.e., number of categories and intrusions), some characteristics of quality of performance were maintained into very late life.

Our text recall results both contrast and complement the word recall results. In contrast to the reliable age-related changes in quantitative word recall, corresponding declines in quantitative text recall were absent. In fact, the young-old age/period group showed significant, albeit rather small, increments in performance. By contrast, the old-old group exhibited slight, but not statistically reliable, decrements over six years. This dissociation between changes in text and word recall is consistent with previous studies that have examined longitudinal changes in multiple measures of episodic memory. For example, McCarty and colleagues (1982) reported significant declines in Visual Reproduction and difficult paired associates (Associate Learning subtest) from the Wechsler Memory Scale, but found little evidence for significant declines in easy paired associates or delayed recall of short stories (Logical Memory subtest). Although this result was qualified by the presence of selective attrition in the data that may have reduced the level of decline observed, it provides partial evidence for differential trajectories of change in word and text recall. Similarly, one possibility for the lack of change in story recall in the current study may be related to selective attrition of the longitudinal sample. However, whereas significant positive selective attrition effects were present for both word and text recall at the second time of testing, these effects were absent at Time 3 (Hultsch et al., 1998). Thus, the lack of significant negative changes in text recall performance observed here, especially those between the second and third times of measurement, cannot simply be ascribed to the effects of selective attrition. Finally, the stories that were used were relatively well structured and the estimated reading level is at the typical adult level. Had ill-structured expository tasks on difficult or arcane subjects been examined, significant negative changes in performance may have been observed.

The presence of significant six-year increments in text recall performance for the young-old group may be related to the presence of a practice effect. Indeed, longitudinal improvements in episodic memory have been reported before (e.g., McCarty et al., 1982; Zelinski et al., 1993), although not over such long time intervals as great as six years. To test for the presence of practice effects, two new samples were drawn from the same population using comparable recruitment procedures (Hultsch et al., 1998). Analyses revealed small, but reliable positive practice effects for the longitudinal sample in story recall between the first and second time of measurement. However, there was no evidence of further practice effects for story recall at Time 3, nor was there any evidence of significant practice effects in word recall at either time of measurement (Hultsch et al., 1998). Taken together, the results from these new samples suggest that the relative stability in mean levels of story recall cannot solely be ascribed to the presence of substantial practice effects. Further, because participants received different test materials at the three measurement points, any practice-related improvements that would occur could be attributed to individuals simply remembering stories they had previously read. The inclusion of multiple, parallel forms of the memory materials is unique among longitudinal studies of cognitive performance, and allows us to be confident that the relative stability we observed cannot be due to item-specific practice effects.

One additional reason for the lack of significant decline in text recall may be due to the length of the follow-up interval. Unlike word recall, much longer retest intervals may be required in order to observe reliable changes in performance. For example, Zelinski and Burnight (1997) observed significant decline in story recall, but this was only after a 16-year retest interval. Future longitudinal measurement points with our sample should help to clarify at what point longitudinal decrements in text recall begin to appear.
Turning to the results from the qualitative analysis of text recall, we examined whether this maintenance for young-old adults was associated with underlying stability or change in qualitative indicators of propositional level or relative meaning. In principle, quantitative maintenance could be a function of many different combinations of recall rates across propositional levels. That is, the same rate of overall recall could result from qualitative patterns emphasizing main ideas (more than details) or details (more than main ideas). We observed a robust levels main effect across all three occasions. This implies that both young-old and old-old adults continued to be able to identify and use the underlying organization of the stories. This replicates the results of previous cross-sectional studies (e.g., Dixon et al., 1982; Stine et al., 1996), all of which involved comparisons of young with older adults. Furthermore, our results extend those of previous studies, in that we observed no young-old and old-old age differences and no significant changes in recall structure across six years. In addition, there was no evidence for interactions between propositional level and time of measurement or age/cohort group. These results suggest that when performance differences do appear, these changes are distributed across all propositional levels. That is, they do not recall fewer details or fail to remember the general theme of the stories, rather, the decrements appear gradually at all levels of the propositional hierarchy. However, the lack of this interaction may also be related to the limited amount of longitudinal change in overall story recall.

Our gender analyses replicate in part those reported elsewhere (Herlitz et al., 1997; Schie, 1996). Specifically, our trend reflecting gender differences in word recall in favor of women coincides with the direction observed by these researchers. We were also able to test Herlitz and colleagues’ (1997) contention that the gender difference may be due to encoding deficits rather than retrieval problems. However, this hypothesis was not supported by our qualitative performance data. Whereas the encoding-related number of words per category variable showed no gender differences, the retrieval-related number of categories recalled indicator revealed significant differences in favor of women. Perhaps in the future, a combination of experimental- and performance-based indicators will resolve this discrepancy.

Although Herlitz and colleagues (1997) did not include a text recall measure in their battery, they proposed that gender differences would likely be found also for such tasks. They supported this speculation by citing previous data showing some gender differences in quantitative story recall (Hultsch et al., 1991; Zelinski et al., 1993). In addition, we have recently observed a significant standardized effect for gender on these episodic memory factors (Hultsch et al., 1998). In contrast, Zelinski and Burnight (1997) reported a lack of gender differences in three episodic memory measures, word recall, word recognition, and text recall. However, the limited statistical power of their study to detect small differences in performance may have mitigated against finding such effects. In the present study, we observed gender differences in quantitative text recall, although gender did not interact with other factors. In addition, we explored this effect further by examining possible post hoc explanations related to “familiarity” of protagonist or story theme (McKelvic, Standing, St. Jean, & Law, 1993). Specifically, we explored whether gender differences would appear differentially across the six stories according to a simple gender-matching principle. That is, would women perform better than men on stories featuring female protagonists, and vice versa? Consistent with the speculation of Herlitz and colleagues, no systematic gender effects were observed across multiple story comparisons. Thus, the overall gender effect is due to slight differences in story recall at the aggregate level. As Herlitz and colleagues acknowledge, the explanation for gender differences in episodic memory is not yet available.

Overall, our analyses of changes in qualitative indicators of word and text recall converge on an important theme. Some underlying structural characteristics of word and text recall are maintained into very late life, even when there is significant decline in overall total recall. For both word and story recall, young-old and old-old adults identified and used organizational characteristics of the to-be-remembered materials. The results from the story recall analyses indicated no significant changes in the manner in which the information was recalled over time. However, the word recall analyses suggest that the source of the statistically reliable longitudinal decrements may be related to problems associated with the retrieval of information from memory, as evidenced by significant decreases in categories recalled and increases in the frequency of intrusions. It is interesting to note that the potential retrieval-related changes in word recall closely match arguments we have made for declines in the semantic memory indicator, fact recall (Hultsch et al., 1998). In these analyses, we speculated that the source of the changes on a task typically very well preserved in old age may be related to difficulties in retrieving the information from stored knowledge. The results of the present analyses also suggest that the significant declines in word recall we observe may have a source in impaired retrieval.

In summary, the results of the present study are informative regarding changes in quantitative and qualitative measures of word and text recall. We repeat, however, our earlier acknowledgment that we have not sampled the entire range of potential qualitative indicators. Furthermore, it is important to note that the VLS sample is a positively selected one. Not only is the original sample relatively well-educated, but typical moderate selective attrition occurred from the first to second time of measurement (Hultsch et al., 1992). Whether the present change data underestimate episodic memory decline for these age groups is being addressed with ongoing research comparing the VLS with other less-select samples. Finally, we underscore several contributions the present study makes to the literature on aging and qualitative indicators of memory functioning. First, we provide novel information regarding aging-related changes in qualitative indicators of two episodic memory tasks. To date, all of the studies that have examined this issue have been cross-sectional in nature, comparing young and older adults. Moreover, most previous studies have included only one exemplar of episodic memory. The second strength is that because our quantitative and qualitative indicators of word and text recall are closely linked, we were able to speculate about the source of the quantitative deficits in memory by examining performance on the qualitative indicators. As we saw with text recall, the deficit for the old age/cohort group may reflect declines at all levels of text recall. On the other hand, there is some evidence to suggest that the changes in word recall have a source in impaired retrieval of information from memory.
ACKNOWLEDGMENTS

This research was supported by a grant from the National Institute on Aging (AG08355). We thank all staff members and volunteer participants of the Victoria Longitudinal Study for collaboration and data collection. We appreciate Scott Maitland’s efforts in organizing the present data set. This article was originally drafted while Roger Dixon and Brent Small were guest scientists at the Stockholm Gerontology Research Center. Helpful discussions in Stockholm with Lars Bäckman and Agneta Herlitz are appreciated.

Address correspondence to Brent J. Small, PhD, Department of Gerontology, SOC 107, University of South Florida, 4202 E. Fowler Ave., Tampa, FL 33620.

E-mail: bsmall@luna.cas.usf.edu

REFERENCES


Received January 19, 1998
Accepted October 19, 1998

---

**Nominations for Prestigious Awards**

**DEADLINE MAY 8, 1999**

**Robert W. Kleemeier Award** To a fellow of The Gerontological Society of America in recognition of outstanding research in the field of gerontology.

**Donald P. Kent Award** To a fellow of The Gerontological Society of America who best exemplifies the highest standards for professional leadership in gerontology through teaching, service, and interpretation of gerontology to the larger society.

**Glenn Foundation Award** Open to all scientists, regardless of field or nationality, for significant research contributions to the biology of aging.

**Nathan Shock New Investigator Award** For outstanding contributions to new knowledge about aging through basic biological research.

**PGC Polisher Research Institute Award** Open to all disciplines, for a significant contribution from applied research that has benefited older people and their care.

**Maxwell A. Pollack Award for Productive Living** Open to individuals or teams who have demonstrated excellence in bridging the worlds of research and practice.

**Doris Schwartz Gerontological Nursing Research Award** To a member of the Society in recognition of outstanding and sustained contribution to geriatric nursing research.

For further information about these and additional awards, contact:

**The Gerontological Society of America**
1030 15th Street NW, Suite 250
Washington, DC 20005-1503
Phone 202/842-1275 • FAX 202/842-1150
www.geron.org