Memory & Cognition
1996, 24 (3), 305-321

Group aggregates and individual reliability: The case of verbal short-term memory

ROBERT H. LOGIE and SERGIO DELLA SALA
University of Aberdeen, Aberdeen, Scotland

MARCELLA LAIACONA
Clinica Del Lavoro Foundation, Centro Medico, Veruno, Italy

and

PAT CHALMERS and VAL WYNN
University of Aberdeen, Aberdeen, Scotland

Two experiments examined the generalizability of the effects of word length and phonological similarity with visual and auditory presentation in immediate verbal serial ordered recall. In Experiment 1, data were collected from 251 adult volunteers drawn from a broad cross-section of the normal population. Word length and phonological similarity in both presentation modes significantly influenced the group means. However, 43% of the subjects failed to show at least one of the effects, and the likelihood that effects appeared was highly correlated with verbal memory span. In Experiment 2, 40 subjects of the original sample were retested, 20 of whom had failed to show one or more effects in Experiment 1. Whether or not an effect had appeared for individual subjects on the first test session was a poor predictor of whether the effect would appear on retest. Finally, an analysis of subject reports demonstrated that the patterns of experimental data could be accounted for in part by the strategies that subjects reported using, and the effect of strategy was independent of the effect of span. The implications of these findings for theories of verbal short-term memory are discussed.

Current views as to the characteristics of verbal short-term memory owe much to the discovery and exploration of the effects on immediate serial ordered recall of word length and phonological similarity. The *word-length effect* refers to the tendency for normal adult subjects to have more difficulty in immediate serial ordered recall of a sequence of long words (e.g., *university, aluminium, hippopotamus, refrigerator*) than in that of a sequence of short words (e.g., *scroll, switch, zinc, maths*). The *phonological similarity effect* arises from the relative difficulty in serial ordered recall of phonologically similar items (e.g., *man, mad, map, mat*) compared with recall of a sequence of items that are phonologically distinct (e.g., *day, boy, sup, few*). Word-length and phonological similarity effects appear whether the subjects read or listen to the word sequence for recall.

These are robust effects that have been replicated widely with a range of materials using groups of normal subjects (e.g., Baddeley, Thomson, & Buchanan, 1975; Caplan, Rochon, & Waters, 1992; Conrad, 1964; Levy, 1971; Murray, 1965, 1968; Vallar & Baddeley, 1984; for a review see Logie, 1995). Moreover, the study in neuropsychological patients of the effects (or lack of them) of phonological similarity and word length has provided insight into both the memory deficits of such patients and the nature of normal verbal temporary memory (e.g., Baddeley, Lewis, & Vallar, 1984; Baddeley & Wilson, 1985; Cubelli & Nichelli, 1992; Vallar & Baddeley, 1984; Vallar & Cappa, 1987; for a review, see Della Sala & Logie, 1993). In particular, the findings obtained from both normal subjects and neuropsychological patients have contributed to the development of a model of the system responsible for the appearance of these effects (Baddeley, 1992).

Table 1 shows the patterns typically obtained for groups of normal subjects and for patients with short-term memory deficits.

However, the robust nature of the effects of word length and phonological similarity results from the mean performance of groups of subjects. In contrast, the lack of these effects in brain-damaged patients generally is reported for individual cases rather than for groups. There are as yet no definitive data on the distribution of these effects in the normal population, and, in particular, it is not at all clear what proportion (if any) of the normal subject population fail to show the standard effects. If there are normal subjects who do not show one or other effect, how might we interpret their pattern of data? Do such subjects reliably fail to show the typical pattern? If so, then is this due to an

Part of the work on which this paper is based was supported by a travel grant given to the first author from the Trustees of the journal *Brain*. We are grateful to Alan Baddeley for very helpful discussions on this topic, and to Marc Marschark for useful discussions on the incidence of head injury in the college population. We are also grateful to Nelson Cowan and George Wolford for very useful comments on an earlier draft of this manuscript. Correspondence should be addressed to R. H. Logie, Department of Psychology, University of Aberdeen, Aberdeen AB9 2UB, U.K. (e-mail: r.logie@aberdeen.ac.uk).

Table 1
Summary of the Patterns of Word-Length (WL) and Phonological Similarity (PS) Effects With Visual and Auditory Presentation as Assumed for Normal Subjects, as Observed in Short-Term Memory Patients, and as Observed in the Sample of 251 Subjects Reported in This Paper

	WL		PS	
	Auditory	Visual	Auditory	Visual
Assumed normal pattern	+	+	+	+
Short-term memory patients	0	0?	+	0
No effects missing				
$n = 143$	+	+	+	+
One effect missing				
$n = 4$	+	+	0	+
$n = 13$	+	+	+	0
$n = 25$	0	+	+	+
$n = 36$	+	0	+	+
Two effects missing				
$n = 1$	+	+	0	0
$n = 3$	+	0	0	+
$n = 5$	+	0	+	0
$n = 19$	0	0	+	+
Three effects missing				
$n = 2$	0	0	+	0

Note—+ = Effect present. 0 = Effect absent. 0? = Absent for some patients or on some occasions.

alternative strategy for performing the task (e.g., Gilhooly, Logie, Wetherick, & Wynn, 1993; Jorm & Share, 1983; Siegler, 1987; Simon & Reed, 1976)? For example, subjects may attempt to use some form of mnemonic rather than subvocal rehearsal to retain the word list (Della Sala, Logie, Marchetti, & Wynn, 1991). One other possibility is that such "recalcitrant" subjects may have some form of mild, hitherto undetected, brain damage so that they are performing as if they have a short-term memory deficit. More crucially, if there are apparently normal subjects who show patterns that are similar to those obtained for individual patients with short-term memory deficits, this complicates the interpretation of the patient data.

To deal briefly with this last point, neuropsychological researchers routinely use a variety of converging tests with short-term memory patients, rather than relying solely on the effects of phonological similarity and word length. Nonetheless, the availability of information as to the distribution of these effects in the normal population would ease the interpretation of neuropsychological data and could provide further insights into verbal short-term memory function in normal adults.

In a study of our own (Della Sala et al., 1991), we investigated the pattern of phonological similarity and word-length effects in 15 normal subjects and in an anarthric patient. Anarthric patients have suffered brain damage resulting in an inability to control the speech output mechanisms, such as articulation, without impairment of central language processing. Despite being unable to produce overt speech, some anarthric patients still show word-length and phonological similarity effects (Baddeley & Wilson, 1985; Bishop & Robson, 1989; Logie, Cubelli, Della Sala, Alberoni, & Nichelli, 1989; Vallar & Cappa, 1987). Baddeley and Wilson interpreted this finding as suggesting that the ability to produce overt speech is not required for the adequate functioning of subvocal rehearsal or phonological recoding and storage. Like Baddeley and Wilson, we discovered that our own anarthric patient did indeed show the standard pattern of effects. However, to our considerable surprise, a small number of our normal subjects *did not* show these effects. Thus, for example, 1 subject remembered more visually presented long words than short words, and 2 subjects recalled more phonologically similar words than phonologically dissimilar words, when the words were presented auditorily.

To investigate the reliability of our findings, we retested these seemingly "aberrant" normal subjects and discovered that their data were indeed not wholly reliable. Much of this unreliability could be accounted for by the use of alternative strategies, such as a visual mnemonic that acted to undermine the use of subvocal rehearsal and phonological storage on which the effects of word length and phonological similarity depend.

This serendipitous finding led us to question the extent to which all normal subjects reliably show these effects, despite the presence of the effects in the pattern of performance for the group as a whole. The present paper reports a more comprehensive set of data on these effects in normal subjects. In doing so, we explore further factors that may lead to the appearance or absence of the effects. Also, we aim to provide a means of assessing the performance of short-term memory patients in relation to the reliability and distribution of these effects in the normal population. The paper comprises two experiments. Experiment 1 involved collecting data on effects of word length and phonological similarity from a sample of 251 normal adult subjects. In Experiment 2, we studied the reliability of the results from Experiment 1 by retesting a group of 40 subjects drawn from the original sample of 251. Finally, we report a more detailed study of the reported strategies of each of the subjects tested in Experiments 1 and 2.

In the final discussion of the paper, we shall return to the more general point concerning the use of findings based on groups of normal subjects.

EXPERIMENT 1

The purpose of Experiment 1 was twofold. First, we wished to examine what proportion of subjects from a large sample of adult volunteers show the widely replicated group effects of word length and phonological similarity in verbal serial ordered recall. Second, we aimed to investigate which of a number of variables might affect the size or presence of each of these effects.

Method

Subjects. We compiled a reasonably large sample of English language native speakers, with a roughly even gender balance, and covering a broad range of age, educational level, and social background. The subjects were drawn from the general population of the city of Aberdeen (total population approximately 250,000). They were recruited from the psychology department panel of volunteers, from clubs, societies, and church groups in the city, and from the local job center. There was a total of 251 subjects (117 male and 134 female),

whose mean age was 42.8 years (SD = 15.6), with a range of 18–70 years, and a mean educational level of 13.1 years (SD = 3.1), with a range of 8–22 years.

Stimuli. Four sets, each of nine stimulus words, were constructed. There was one set of phonologically similar words and one set of phonologically different words drawn from Baddeley (1966), and the words in each set were matched for frequency. The sets of long words and short words were selected from Baddeley et al. (1975), with words again matched for frequency.

Procedure. The subjects were tested individually with each of the word sets with both visual and auditory presentation. The order of each of the conditions was systematically varied from one subject to the next but blocked by the phenomena under study. For example, the set of words for testing phonological similarity with visual presentation always occurred immediately before or immediately after the set for testing phonological similarity with auditory presentation, with the order of word sets varied across subjects. Across the full subject sample, the order of presentation was counterbalanced.

Words were selected at random from the relevant set and presented at a rate of one per second using a span procedure and oral serial ordered recall. Presentation started with three trials of two words, moving on to three trials with three words, and so on, until the subject failed to recall the correct sequence on two successive occasions. Span was measured by taking the mean of the three longest sequences correctly recalled. This particular measure of span was chosen because it has been adopted widely in standard tests of intellectual ability (e.g., Weschler Adult Intelligence Scale, Weschler Memory Scale) and it is used widely in clinical neuropsychological settings. It has also been used in a variety of experimental settings both with normal subjects (e.g., Baddeley et al., 1984; Hulme, Maughan, & Brown, 1991) and with neuropsychological patients (e.g., Baddeley, Logie, Bressi, Della Sala, & Spinnler, 1986; De Renzi & Nichelli, 1975; Milner, 1971).

For auditory presentation, the experimenter read aloud the words. For visual presentation, the words were typed in black lowercase letters on individual white cards. The experimenter used a hand gesture to signal when the subject was to commence recall.

Immediately following the collection of span data, each subject was asked about the strategies they had adopted in attempting to retain and recall the items from each sequence. The questioning was entirely open ended, and no strategies were suggested to the subject. The subjects were then interviewed informally as to whether they had suffered any serious illness, neurological damage, or head injury.

These subject reports will be considered after the discussion of Experiment 2.

Results

Order of presentation did not have any effect on the presence or absence of each of the effects under study and was ignored for the main analyses.

To ensure that our procedure had indeed resulted in the typical group result, we carried out separate analyses of variance (ANOVAs) on each of the four conditions. Mean spans obtained for each of the different list types with auditory presentation and with visual presentation are shown in Figure 1.

As is evident from the figure, there were clear effects of word length with auditory presentation [$F(1,250) = 273$, $p < .001$] and with visual presentation [$F(1,250) = 194$, $p < .001$]. There were even stronger effects of phonological similarity with both auditory presentation [$F(1,250) = 891$, $p < .001$] and visual presentation [$F(1,250) = 528$, $p < .001$]. This demonstrates a clear replication of a widely reported set of findings, but with a much larger subject sample than has been used previously, and gives us considerable confidence in the procedures we adopted.

We next went on to explore whether there were individual differences in the presence or absence of each of the four main effects. Table 2 shows the number of subjects in the present experiment who showed or failed to show each of the expected effects as measured by our span procedure. From the table, it is clear that despite the highly significant group effects, a reasonable proportion of subjects

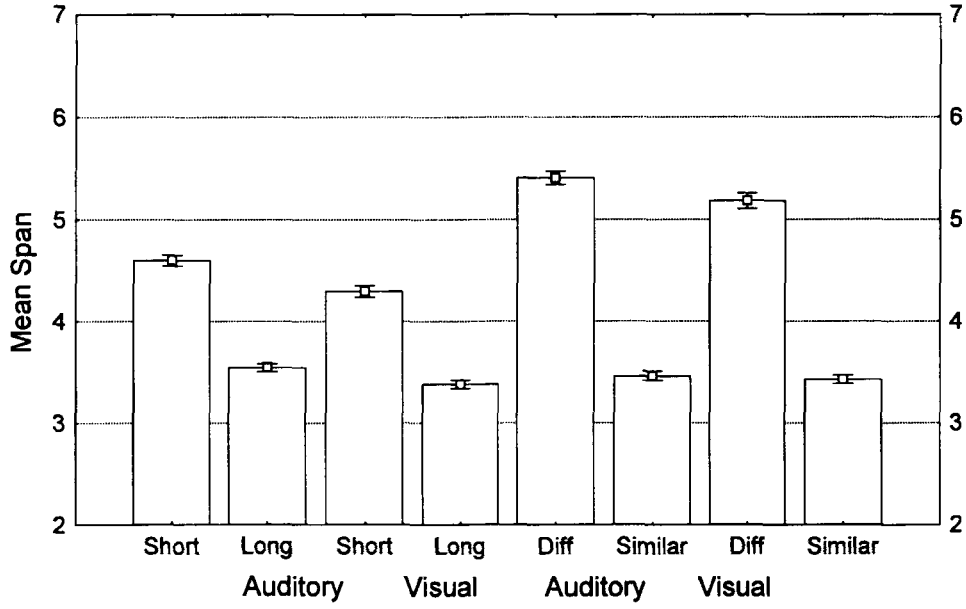

Figure 1. Mean span for immediate, oral, serial recall of short, long, phonologically different, and phonologically similar words with auditory and visual presentation for 251 normal adults.

Table 2
Number of Subjects (N = 251) Who Showed or Failed to Show Typical Group Effects for Word Length and Phonological Similarity With Auditory and Visual Presentation

	Zero or Negative Effect	Presence of Effect
Word Length		
Auditory	49	202
Visual	62	189
Phonological Similarity		
Auditory	5	246
Visual	24	227

failed to show word-length effects, and a smaller number failed to show phonological similarity effects. It appears that phonological similarity with auditory presentation is the most robust; however, even here, the effect was absent for 5 out of 251 subjects.

Evidently, a number of subjects failed to show our four standard effects. A possible account of the lack of effects for some subjects is that the effects themselves are quite small but that they appear for the majority of subjects. Thus, effect size may be clustered just above zero, with a tail end of the distribution just below zero. Therefore, our next step was to investigate the range in the observed magnitude of the standard effects.

Effect magnitude was measured by calculating the size of the difference in span between the two word sets for each condition and then expressing this as a percentage of the span for the word set that normally gives the better performance. So, for example, if a subject achieved a span of 4.6 with the phonologically similar material and a span of 5.7 with the phonologically different material, the percentage difference in span score would be calculated as follows: $5.7-4.6 = 1.1$; and $(1.1/5.7) \times 100 = 19.30\%$.

We chose this measure rather than a simple difference score to give us a measure of effect size that was conservative with respect to possible unreliability. That is, taking percentage difference between spans reduces the magnitude of the difference scores and thereby reduces the estimated size of the unreliability. This also gives us a more fine-grained measure.

One implication of this measure is that differences in span scores that go in the opposite direction from that obtained with the majority of subjects would yield a negative value. Figure 2 shows the frequency distribution of effects obtained displayed as a stem and leaf plot, with effect magnitude expressed as a percentage according to the formula given above.

From Figure 2, the effect magnitudes show a fairly wide spread, with a large number of subjects showing effects greater than 20% in the predicted direction. In contrast, a number of subjects showed effects in the opposite direction. A reasonable number of these were clustered at or just below zero, but there were a notable number of subjects who showed effects greater than 20% in the direction opposite from that of the majority. This occurred most frequently with the word-length effect, particularly with visual presentation.

These results illustrate the frequency of the presence or absence of the effects if we treat each of the four conditions as having been obtained from separate groups. However, with neuropsychological patients, there typically is more than one effect absent, and it is important to discover whether the obtained aberrant findings are being produced by the same subjects. A total of 108 subjects (43%) failed to show at least one of the effects, and, of these, 30 subjects (12%) failed to show two or more of the effects. Of these 30 subjects, two (0.8%) failed to show three out of the four effects. No subject failed to show all four effects. Therefore, the chances are quite high of finding a subject with one or even two effects missing. Subjects who show three effects missing are much rarer.

The next step was to investigate which factors determine whether or not an effect appears and the size of that effect.

Typically, mental performance test scores are influenced by several factors, such as age, education, occupation, or gender. In addition, in neuropsychological patients, the absence of the effects being studied here is commonly associated with very poor short-term memory span. We carried out a series of regression analyses to determine which of these subject variables best predicted the percentage magnitude of each of the effects.

We carried out one analysis for each of the four effects, in each case looking at the influence of age, education, sex, and span. Occupation was found to correlate highly with education (Spearman rho = .70); therefore, occupation was omitted from the regression analyses. Our measure of span was taken as the mean span score obtained for each subject on word sets that were not considered in a given analysis but that were in the same modality of presentation. For example, in the analysis of the phonological similarity effect (visual presentation), the measure of span was taken as the mean span score derived from combining the scores for the list of short words and the list of long words used for testing the word-length effect with visual presentation. Conversely, in the analyses of the word-length effect, we used a mean span score derived from the list of phonologically different words and the list of similar words, again using the same modality of presentation. This procedure was adopted to avoid using the same span for prediction as that used to calculate the magnitude of the various effects. We used a mean of the two available alternative spans in order to maximize the reliability of the span score employed in the regression analyses.

To verify the influence that each one of the variables had on each one of the word-length and phonological similarity effects, we analyzed our data with a covariance model. The effect of each variable was evaluated alone and within the complete model. That is, we partialed out the effect of each variable that was in common with each of the other variables present, but only when more than one variable was found to be significant.

The results of the analyses were very clear. There was no significant variance in the magnitude of the effects accounted for by age, sex, or education. The only significant contribution for all four effects was made by span. The re-

Table 3
Regression Analysis on the Relationship Between Span and Each of the Effects of Word Length and Phonological Similarity With Auditory and Visual Presentation

	df	*F*	*p*	Correlation	*p*
Word Length					
Auditory	1,249	42.042	< .0001	.380	< .001
Visual	1,248*	45.075	< .0001	.419	< .001
Phonological Similarity					
Auditory	1,249	7.910	< .005	.175	< .05
Visual	1,249	9.922	< .005	.196	< .01

*After partialing out the effect of a small significant contribution from educational level.

sults are summarized in Table 3. There was an apparent, small contribution from education for word length with visual presentation. However, this contribution from education disappeared after partialing out the effect of span. It therefore appeared that the size of any of the standard effects was directly related to the magnitude of span. That is, the lower the span, the smaller were the effects of word length or phonological similarity. These relationships are apparent from Figure 3.

Despite the strong relationship between span and the size of each of the four effects, the total amount of variance accounted for is at most 20%, indicating that there were a number of subjects with relatively high spans who also did not show these effects. Note also that our measure of effect size was based on a proportion of the span for that particular variable. Therefore, scores for effect sizes at low spans would have been inflated relative to effect sizes for high spans and relative to a simple difference measure for effect size. This decreases the chances of obtaining the correlation we obtained—namely, that lower span scores are associated with smaller effect sizes—and reinforces our earlier argument that the measures we have used here are conservative.

Discussion

It is clear from the results of Experiment 1 that when using standard procedures for testing short-term verbal memory in normal subjects, a substantial minority of our sample of 251 normal adult subjects apparently failed to show widely reported and replicated effects. This observation is in contrast to the highly significant replication of these effects in the group average data. Our sample of subjects represents a broader age range and educational level than is common in laboratory studies of normal subjects. As such, a larger number of subjects may have failed to show these effects than has been typical in previous studies on this topic. However, neuropsychological patients also are drawn from a population with a wider range of demographic variables than is typical in laboratory experiments. Moreover, our sample is substantially larger than that generally used in studies of normal cognition. As such, our data provide a better benchmark against which to assess the data from individual patients. These new data also provide an indication as to the generalizability to the wider population of findings that have been established in small, homogeneous groups.

What implications might our seemingly aberrant data from normal subjects have for the development of theories of verbal short-term memory? As we discussed in the introduction, such theories are derived on the whole from group aggregate data. The theories are also based on the assumption of a common cognitive architecture. It is further assumed that the group aggregate data pattern reflects the characteristics or operation of this architecture. If we find that a substantial minority of subjects fail to show the group pattern, this might well undermine the assumption of a common architecture. With respect to our own study, theories of verbal short-term memory are in part derived from the phenomena of phonological similarity and word length in serial ordered recall. How can we interpret the fact that a number of subjects do not show these effects?

With respect to the patterns shown in Table 1 for normal subjects and for short-term memory patients, it is interesting to reflect on the patterns of effects found in our own data shown in the same table. Most of our subjects show the pattern typically obtained with normal subjects. What we did not expect was that 2 of our subjects showed a pattern that is similar to that commonly obtained with short-term memory patients. Moreover, some of the patterns shown in Table 1 cannot easily be encompassed by this particular model. For example, it is difficult to account for the performance of 3 of our subjects who failed to show a word-length effect with auditory presentation coupled with the lack of an effect of phonological similarity with visual presentation. These patterns of data also suggest that there is a need to collect converging evidence and assess consistency of data patterns before using neuropsychological data in discussing a theoretical model of cognition. For example, the data in Table 1 suggest that the probability of obtaining a subject with only a word-length effect missing for visual presentation is 14% (36/251), whereas it is very rare to find subjects with three out of four effects missing (0.8%). Therefore, a patient who reliably fails to show three or four out of the four effects is likely to have a genuine deficit of verbal short-term memory, and their data may be informative for theories of normal memory function. In contrast, researchers should exercise caution when attempting to interpret data from a patient who fails to show just one of the four effects. For a detailed discussion of the clinical implications of these findings, see Della Sala and Logie (in press).

Next, we explored possible factors that could have led to the data pattern obtained, including demographic variables and verbal memory span. Dealing first with demographic variables, if subjects differ in the organization of their cognition, we might expect that these differences would be related to other differences between subjects. However, our data show clearly that differences among subjects in age, education, or sex are unrelated to the presence or magnitude of the effects of phonological similarity and word length.

On the other hand, the differences in memory span did seem to affect the pattern of data obtained. In particular, failure to show the effects was closely linked with a low memory span. This result fits quite well with the data from short-term memory patients who, by definition, have low

A)

Stem%Leaf	Word Length Effect: Auditory	Cases	Percentiles
-50%9		1	
-40%699		3	
-40%3		1	
-30%67799		5	
-30%		0	
-20%99999		5	
-20%23		2	
-10%11111122223		11	
-10%		0	
00%00000000000000000000		21	
00%5567777777777778888888888999999999		33	25%
10%113334444.		9	
10%55555566666888		14	
20%000011111123333333333333444444		31	median
20%555555666667777788888888889999		32	
30%00000133333333		14	
30%55555555555555667788888888		26	75%
40%00000011122333333333344		22	
40%6777777799		10	
50%00000234		8	
50%557		3	
60%		0	
	Total N:	251	

B)

Stem%Leaf	Word Length Effect: Visual	Cases	Percentiles
-80%5		1	
-70%		0	
-60%2		1	
-60%		0	
-50%5		1	
-50%		0	
-40%6666669		7	
-40%		0	
-30%55777789		8	
-30%222		3	
-20%7799999		7	
-20%04		2	
-10%11223		5	
-10%		0	
00%000000000000000000000000000		27	
00%778889999999999999999		23	25%
10%1111234444		10	
10%555555666666668888888999		24	
20%1111111233333333344444		22	median
20%55555555666677888888888889		26	
30%00000001133333333		17	
30%555555555566666777788888		23	75%
40%0000111222222222233344		22	
40%5666666777799999		16	
50%0233		4	
50%6		1	
60%		0	
60%6		1	
70%		0	
	Total N:	251	

Figure 2. Stem and leaf plots for each of the four effects studied: (A) word-length auditory presentation; (B) word-length visual presentation; (C) phonological similarity auditory presentation; (D) phonological similarity visual presentation. The diagrams display the stem as a decade percentage and the leaf as instances occurring within that decade.

C)

Stem%Leaf Phonological Similarity Effect: Auditory	Cases	Percentiles
-64%	1	
-60%	0	
-50%	0	
-40%	0	
-30%9	1	
-30%	0	
-20%	0	
-10%12	2	
-10%	0	
0%	1	
0%57777899	8	
10%1333444444	11	
10%56666678888	11	
20%00011111111223333333333	23	
20%55555566666666666677788888888999999	35	25%
30%00011111111333333333	20	
30%555555555555555556666667788888888888889	39	median
40%00000000001111111111222222223334444444	39	75%
40%55666667777777777799	20	
50%00000022222222222222334	22	
50%555666778899	12	
60%00234	5	
60%8	1	
70%	0	
Total N:	251	

D)

Stem%Leaf Phonological Similarity Effect:Visual	Cases	Percentiles
-60%2	1	
-60%	0	
-50%	0	
-40%	0	
-30%	0	
-20%999	3	
-20%3	1	
-10%111223	6	
0%0000000000000	13	
0%67777788999999	14	
10%11233444	8	
10%556666788888888999	18	25%
20%00111112333333333444	20	
20%556666677778888888999	21	
30%000011113333333333333333	23	median
30%555555555555566667777777888888889	32	
40%00000000111111122222222333344444	34	75%
40%666677777777777777778999	25	
50%000000000222222222223	20	
50%55557789	8	
60%	0	
60%13	2	
70%02	2	
Total N:	251	

spans and who also fail to show some of the standard effects in immediate verbal serial recall (see Table 1). This point is reinforced by the fact that the 2 subjects who failed to show three out of the four effects had low overall mean spans of 2.7 and 3.5.

Although we have investigated a number of possible causes of variability in the data, our discussion has been based on just one sample of each subject's performance on each set of experimental materials. A further aspect of variability is the extent to which a given measure is reliable from one occasion to another, and it is unusual for test–retest reliability to be reported in studies of groups of normal subjects. In this respect, it is notable that neuropsychological patients generally are tested extensively; however, even with studies of patients, this usually involves a range of tests and rarely is the test–retest reliability of any given test measured explicitly. Therefore, in Experiment 2, we attempted to determine whether those subjects who failed to show one or more of the effects in Experiment 1 would reliably fail to show these effects if retested. This was the primary purpose of Experiment 2.

EXPERIMENT 2

Method

Subjects. A total of 40 subjects took part in this experiment selected from our original group of 251. Twenty (12 female, 8 male) of these subjects were selected because they failed to show one or more of the standard effects on the first test session. They were also selected to represent as wide a range as possible of age (M = 46.2 years, SD = 13.9, range = 26–68) and of education (M = 12.2 years, SD = 3.4, range = 9–19). Furthermore, we attempted to cover the full range of span among those failing to show the effects. For the purpose of subject selection, we used the most common measure of verbal short-term memory span—namely, that for phonologically different words with auditory presentation. For this "effects-absent" group, the mean span was 4.45 (SD = 1.01, range = 2.33–6.33). The numbers of subjects in this group who failed to show each of the four effects on the first test are shown in Table 4. Note that some subjects may have failed to show more than one effect—hence, the numbers do not add up to 20.

The remaining 20 subjects (12 female, 8 male) for retest showed all four of the verbal memory effects on the first test session. These subjects were, as far as possible, individually matched with one of the subjects in the effects-absent group on span (phonologically different words, auditory presentation), age, education, and sex. Mean span for this "effects-present" group was 4.73 (SD = 0.71, range = 3.67–6.33), mean age was 46.85 years (SD = 14.4, range = 27–70), mean education was 11.75 years (SD = 2.6, range = 9–17).

Procedure. The procedure was identical to that used for Experiment 1, using the same materials, but the subjects were tested approximately 1 year later. Again, as for Experiment 1, each subject was asked to describe any strategies that they had used during the course of the experiment and was asked whether they had experienced any form of head injury or neurological damage.

Results and Discussion

Table 4 shows the numbers of subjects in each group for whom each of the four effects was absent or present on the first and second test session. It is immediately apparent that

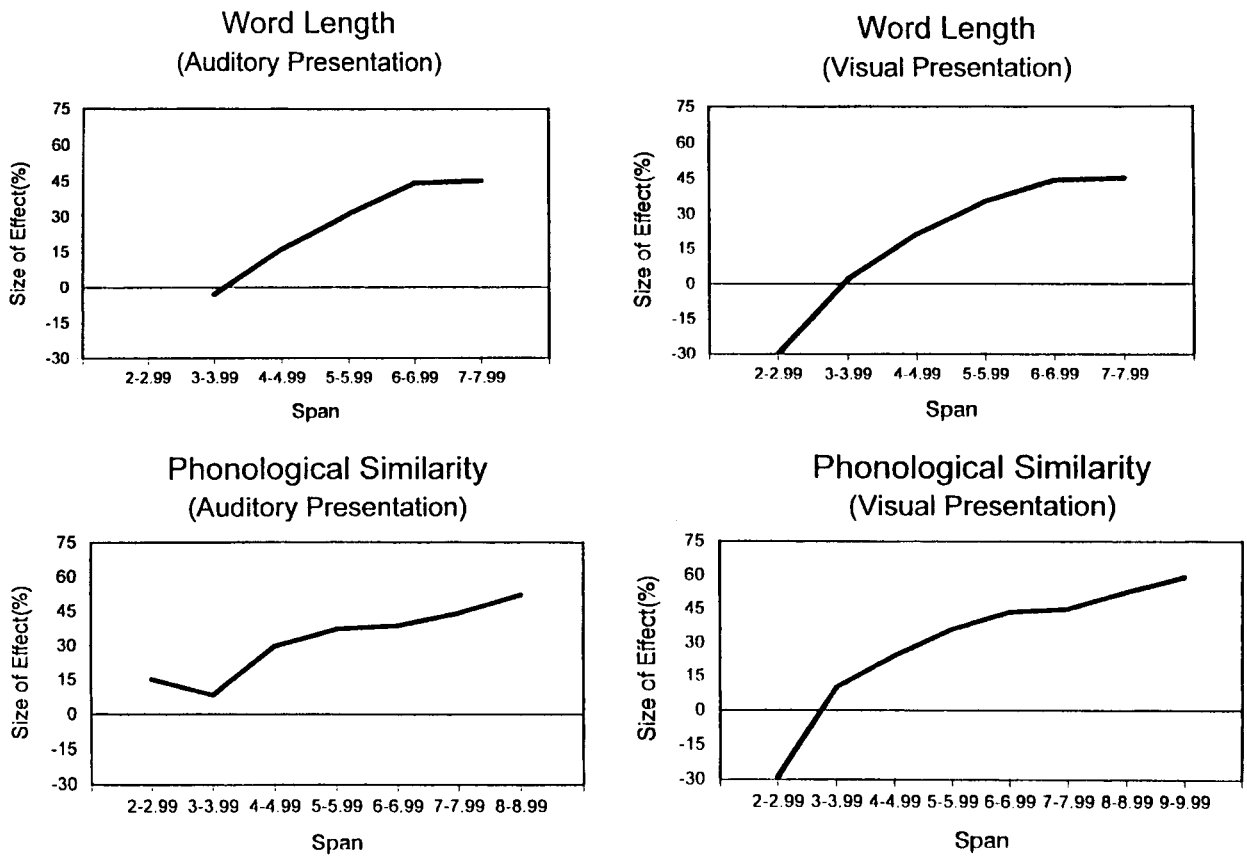

Figure 3. Magnitude of each of four effects in verbal short-term memory in relation to verbal memory span.

Table 4
Numbers of Retested Subjects Showing or Failing to Show Effects of Word Length and Phonological Similarity in Experiments 1 and 2

	Group A	Group B
Word Length With Auditory Presentation		
Missing first and second test	0	0
Missing first test only	0	11
Missing second test only	7	0
Word Length With Visual Presentation		
Missing first and second test	0	3
Missing first test only	0	11
Missing second test only	3	0
Phonological Similarity With Auditory Presentation		
Missing first and second test	0	0
Missing first test only	0	3
Missing second test only	0	1
Phonological Similarity With Visual Presentation		
Missing first and second test	0	0
Missing first test only	0	5
Missing second test only	2	2

Note—Group A = Group of 20 subjects who showed all four effects on the first test session. Group B = Group of 20 subjects who failed to show one or more effects on the first test session.

there was indeed considerable variability from the first test session to the second. For three of the four effects, none of the subjects in the effects-absent group failed to show the effects both on the first test session and on retest. Only in the case of word length with visual presentation did 3 subjects fail to show the effect on both occasions. From Table 4 it is also notable that several subjects in the effects-present group failed to show effects the second time around, and, generally, the word-length effect tended to be less reliable than the phonological similarity effect. We analyzed the consistency with which effects were present or absent using Cohen's k statistic (Cohen, 1960). All of the calculated k values were nonsignificant both when the data were treated as a whole and when they were treated individually for each of the four standard effects, suggesting a low level of consistency between the two sessions.

We also examined the correlation between the effect sizes obtained in each of the two test sessions across all 40 retested subjects. The correlations were −.31 for word length with auditory presentation, −.02 for word length with visual presentation, −.01 for phonological similarity with auditory presentation, and .09 for phonological similarity with visual presentation. These correlations indicate a complete lack of test–retest reliability for any of the effects. Since these four effects have been replicated very widely with groups of normal subjects, including the group tested in Experiment 1, this suggests that the effects are statistically reliable across a group tested on a single occasion. Thus, it would seem reasonable to expect that the effects also would be reliable from one occasion to another with the same individuals. However, when examined explicitly, the test– retest reliability did not live up to expectations. One caveat is that half of the subjects were selected for retest on the grounds that they failed to show one or more of the effects. In Experiment 1, we showed that the effect size was closely related to memory span, and, as such, our selected subjects tended to have memory spans from the lower half of the distribution. Since the other 20 subjects were matched on span, it follows that they too had lower spans. Thus, the poor reliability may have resulted in part from this process of selecting subjects and may not be wholly representative of test–retest reliability for the full sample. Nevertheless, the dramatically low or negative correlations obtained together with the lack of consistency for the presence of the effects can be taken as indicative that the poor reliability of the measure is genuine.

A further possibility is that the poor correlations may have arisen from the fact that we combined data from two rather different groups of subjects: One group showed effects on the first test session, and the second group did not. On retest, each of these groups may have a tendency to produce scores that regress toward the mean, but in different directions. It may then be misleading to report correlations based on data from the two groups combined.[1] One way to assess whether this might be the case is to examine the variability in the scores for each group and each effect across test sessions. The relevant means and standard deviations are shown in Table 5, from which it appears that, for the subjects who failed to show effects the first time around, the variability in the group was somewhat greater on the first test session than on the second. For the group who did show effects the first time around, the variability appeared to increase the second time around. Therefore, it appears that only one of the groups shows signs of regressing to the mean. We also calculated the correlations on the effect sizes across the first and second test sessions, separately for each of the two groups. For the effects-present group, the correlations were as follows: phonological similarity with auditory presentation, r = .08; phonological similarity with visual presentation, r = .14; word length with auditory presentation, r = −.10; word length with visual presentation, r = .11. For the effects-absent group, the correlations were as follows: phonological

Table 5
Means and Standard Deviations of Proportion Effect Sizes for Each of the Two Test Sessions for Subjects Who Failed to Show One or More Effects or Who Showed All of the Effects on the First Test Session

	Word Length				Phonological Similarity			
	Auditory		Visual		Auditory		Visual	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Effects Absent								
Test 1	0.04	0.22	−0.09	0.28	0.26	0.26	0.15	0.24
Test 2	0.25	0.14	0.24	0.22	0.32	0.15	0.25	0.18
Effects Present								
Test 1	0.26	0.11	0.27	0.08	0.28	0.12	0.37	0.12
Test 2	0.11	0.30	0.19	0.20	0.37	0.12	0.24	0.17

similarity with auditory presentation, $r = -.05$; phonological similarity with visual presentation, $r = .12$; word length with auditory presentation, $r = -.42$; word length with visual presentation, $r = .05$. In summary, none of the correlations are significant and the highest correlation obtained is negative. These results should, of course, be treated with some caution, since each is based on data from only 20 subjects. However, they indicate an impressive lack of test–retest reliability within each of the groups, thereby supporting our earlier conclusion about the reliability of these effects.

The primary purpose of Experiment 2 was to assess the test–retest reliability of the four effects under study. The low level of reliability found suggests that, in attempting to apply these findings to studies of neuropsychological patients, it is not legitimate to take the lack of effects on a single test occasion with a given test as necessarily reflecting an underlying cognitive deficit. By the same token, these findings suggest that it is difficult to draw firm conclusions for models of normal cognition from patient data patterns.

A further issue with respect to the reliability of each of the effects concerns the extent to which a failure to find an effect is due primarily to random error in the measurements being used. The lack of a consistency for individual subjects suggests that random error may indeed be a factor. We tested this more formally by calculating the confidence limits for each subject's effect magnitude for each of the four effects, taking into account their overall variability over the two test sessions.

The effect magnitude confidence limits (90%) for each condition were calculated using a formula derived by Huber (1973). This formula takes into account the test–retest reliability and the variability in the scores as follows:

$$\text{Observed effect magnitude} \pm [(1.645 \times SD) \times \sqrt{(1-\text{reliability})}],$$

where the value 1.645 allows inclusion of 90% of scores (excluding 5% at each end of the distribution), and reliability is essentially the correlation between the effect magnitudes on each of the two test sessions. The formula allows us to calculate the expected confidence limits for each effect size for each individual subject. In other words, given a particular observed effect magnitude in Experiment 1, we can derive with 95% confidence either the maximum or the minimum effect magnitude that each subject could have achieved. If the maximum expected effect magnitude is then equal to or less than zero, we can be confident that the effect was not present for that subject. As we mentioned above, it is possible that the method used to select subjects for retest might have led to lower reliability than if the retested sample had been more representative of the whole sample. However, a lower level of reliability will lead to wider confidence limits for each subject's score, making it less likely that we will find effect magnitudes of zero or less that fall below the upper confidence limit. Therefore, our derived scores will give us a conservative measure as to whether there are some subjects who still fail to show one or more effects.

From this analysis, 6 of our original 251 subjects failed to show one of the four effects. The spans for phonologically different words for these subjects were 3.0, 4.0, 4.7, 5.0, 5.7, and 6.7. For 3 of these subjects, the missing effect was word length with visual presentation. The other 3 subjects each failed to show one of each of the remaining effects. Thus, when taking into account confidence limits and test–retest reliability, the number of subjects failing to show effects dramatically decreases, and no subject fails to show more than one of the effects. Nevertheless, there were still 6 subjects from our sample of ostensibly neurologically intact individuals—some of whom had relatively high span scores—who failed to show at least one of the effects even when the reliability of the measures is taken into account.

These results from an exploration of the test–retest reliability for measures of short-term verbal memory performance suggest that it would be useful to assess the test– retest reliability of measures of normal cognitive performance before drawing conclusions from the group aggregate data. Nonetheless, the question remains as to why the measures examined here should be so unreliable. One possible response would be to attribute this to random error. However, this too begs the question as to whether the researcher has simply failed to take into account relevant factors that might have a systematic effect on the data. We pointed to one such factor in Experiment 1 when we noted that memory span accounted for at least some of the variability. Other possible factors are strategy choice and possible undetected brain damage. For example, in a previous paper (Della Sala et al., 1991), we have shown that strategy choice can have a substantial influence on word length and phonological similarity and that this influence is independent of the subject's span. The potential contribution of these factors to the variability in our own data is addressed next.

ANALYSIS OF SUBJECT REPORTS

One major factor that could contribute to variability in our data from Experiments 1 and 2 is the use of differing strategies across our subject sample. With respect to the effects under scrutiny, models of short-term verbal memory are based on the assumption that subjects use verbal rehearsal and phonological coding when attempting to retain the items. The appearance of phonological similarity

and word-length effects even with visually presented words is taken as evidence that the visually presented material is translated into a phonological or articulatory-based code (Conrad, 1964). However, if subjects attempted retention by means of a semantic strategy or a visual mnemonic, this would undermine all of the four standard verbal short-term memory effects. For example, subjects may use semantic associations between words in the list or visual imagery for linking the words together, rather than relying on verbal rehearsal. In this respect, it was interesting to note that despite the dramatic influence of span on the appearance or otherwise of the four effects in Experiment 1, a small number of our subjects with relatively normal spans of around 6 items also failed to show one or more of the standard effects. Of the 6 subjects failing to show effects following our consideration of confidence limits in Experiment 2, the spans ranged from 3.0 to 6.7. In our previous work (Della Sala et al., 1991), we reported one normal subject (W.D.) who persistently failed to show both phonological similarity and word-length effects with auditory presentation, despite having a very high span of 8.0, and despite extensive testing. When questioned about adopted strategies, he reported using mnemonic techniques. On a third test session, we specifically asked him to use subvocal rehearsal in the retention of the word sequences, and, under these conditions, the effects were clearly present. These previous observations together with the findings in Experiments 1 and 2 led us to suspect that the unreliability in these measures could well reflect the use of differing strategies by different subjects. They may also reflect differing strategies by the same subject on different occasions, or even for different trials within the same test session. We tackled this issue by considering the reports from subjects who took part in Experiments 1 and 2 about the strategies that they adopted in attempting to perform the task.

A second issue in this analysis was the possible contribution from hitherto undetected brain damage. In studies of normal subjects, it is generally assumed that subjects who take part in experiments are genuinely drawn from the normal, non–brain-damaged population. However, it is possible that some subjects who volunteer to take part in such experiments have in fact suffered some form of mild brain damage. Such subjects may not spontaneously report this, and not every experimenter specifically asks about medical history. In studies where subjects have been asked about head injury, approximately 20% of students in secondary or higher education reported having suffered an injury that was sufficiently severe to cause a loss of consciousness for a variable amount of time (Crovitz, Horn, & Daniel, 1983; Segalowitz & Brown, 1991). This is substantially higher than the number of people who spontaneously seek medical advice (Carlson, 1986) or require hospitalization (Richardson, 1990). These studies suggest that it would be worth exploring further the incidence and the effects of possible brain damage in our own data.

Reported Strategies

From our sample of 251 subjects in Experiment 1, 209 were interviewed about strategies they had used. Of these, 196 reported using one or more strategies. The remaining 13 subjects' reports included statements such as "I just concentrated" or "I found the visually presented words easier" without reporting a coherent strategy.

The classification scheme for the reported strategies and the number of subjects from Experiment 1 reporting each strategy are as follows: *verbal rehearsal* (55 subjects), subjects reported repeating the words inside their heads or repeated "parrot fashion"; *chunking* (30 subjects), subjects reported grouping items in twos or threes; *first letter* (11 subjects), subjects reported remembering just the first letter of each of the words rather than the whole word; *semantic mnemonic* (20 subjects), subjects reported forming semantic associations between the words for recall; *visual mnemonic* (14 subjects), subjects reported generating visual images of the meanings of the words; *mixed strategy* (66 subjects), subjects reported using varying combinations of the above strategies.

Influence of Strategy on Effect Magnitude

The effect magnitudes for each of the four effects were entered into an ANOVA to examine the influence of each of the reported strategies. The analysis revealed that there was an overall difference in effect magnitude according to the strategy adopted [$F(5,190) = 5.509$, $p < .0001$]. There was also an overall difference in the effect magnitudes for each of the four effects [$F(3,570) = 44.142$, $p < .0001$]. Reported strategy did not interact with effect type [$F(15,570) = 1.331$, $p > .1$]. Given the influence of span on effect magnitude shown in Experiment 1, it is possible that the influence of strategy could be entirely accounted for by span. To assess this, we ran another ANOVA where mean span over all list types was taken as a covariate. Even when span was taken into account, there was still a significant influence of reported strategy [$F(5,189) = 2.825$, $p < .02$]. The mean magnitude effect sizes obtained for each reported strategy are shown in Figure 4.

It is evident from Figure 4 that the largest effect sizes were obtained from subjects who used verbal rehearsal or rehearsal of "chunks." Newman-Keuls tests indicated that the means for rehearsal and for chunking did not differ but that both of these strategies differed ($p < .01$) from each of the other reported strategies (first letter, semantic, visual, or mixed). These other reported strategies did not significantly differ from one another.

These results indicate that the magnitude of both phonological similarity and word-length effects are heavily reliant on the consistent use of a verbal rehearsal strategy and that this reliance is independent of the size of span. In this respect, it is notable that only 85 of our subjects adopted a consistent strategy of using some form of verbal rehearsal. Thus, the majority of our sample reported using strategies that do not necessarily rely on the use of short-term verbal memory. This provides a coherent account as to why so many subjects failed to show one or more effects in Experiment 1.

It is also notable that around one fourth of our subject sample (66 subjects) reported changing strategies within the test session. For example, some subjects reported try-

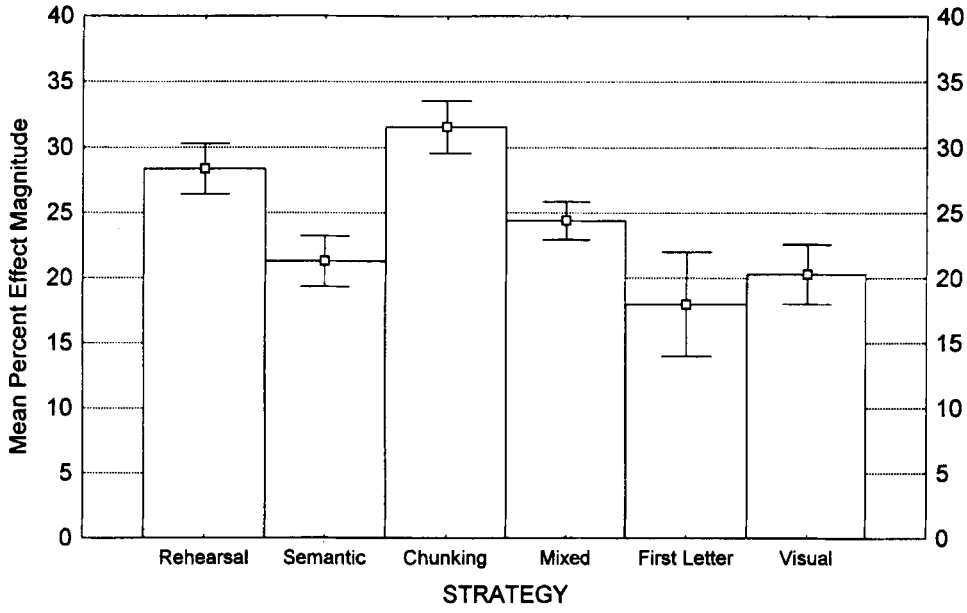

Figure 4. Magnitude of all verbal short-term memory effects as a function of subject-reported strategies.

ing to use some form of word association or visual mnemonic but then found that it did not work very well and switched to verbal rehearsal. A few subjects reported that they used verbal rehearsal for visually presented words but that they used visual imagery for the auditorily presented words. Several subjects reported switching from one strategy to another within trials and between trials as well as between types of list and presentation modalities. This tendency to swap strategies may provide an indication as to why the effects are so unreliable. To test this, we examined the consistency in reported strategies between the first and second test session.

Of the 40 subjects retested, 7 in the effects-absent group and 8 in the effects-present group reported using mixed strategies on the first occasion. Only 7 subjects in total reported using the same strategy on the second session as the one they had used on the first session, and 3 of these were from the effects-absent group; 2 of these reported using verbal rehearsal, and 1 reported using a semantic strategy. From the effects-present group, 3 reported using verbal rehearsal, and 1 reported using a semantic strategy. A further 2 subjects reported using a mixture of strategies on both occasions. The remaining 31 subjects reported using different strategies on the two test sessions. Among these, 8 did not report any strategy on one of the test sessions. It appears then that a lack of consistency in strategy choice could account in a large part for the lack of test–retest reliability in the four effects.

We also examined the strategies of the 6 subjects who failed to show effects even after test–retest reliability was taken into account. Three of these subjects reported a mixture of strategies involving visual mnemonics, first letter, or some form of word association. One subject reported consistently using visual mnemonics. The remaining 2 subjects reported using verbal rehearsal, and both failed to show phonological similarity effects. However, 1 of the subjects who failed to show the effect with auditory presentation had a span of 3.00 for auditorily presented, phonologically different words. The other subject failed to show the effect with visual presentation and had a span of 2.33 for visually presented, phonologically different words. To summarize, the subjects who reliably failed to show effects either were not using subvocal rehearsal or had very low spans, and this provides additional support for the conclusion that the effects depend on either memory span or choice of strategy.

Possible "Undetected" Brain Damage

Variability among our subject sample might also have arisen from possible brain damage in some of our ostensibly "normal" subject group. From the posttest interview, 7 of our 251 subjects reported aspects of their medical history that could be associated with possible brain damage. These included epilepsy, minor stroke, alcohol abuse, head injury, and possible sequelae of chicken pox encephalitis. One way in which brain damage could affect performance in the current context is in the form of poor verbal short-term memory capacity (Vallar & Shallice, 1990). On this basis, we looked at the verbal span performance of these 7 subjects in relation to the distribution of span for the remainder of the subject sample. For this purpose, we chose to examine span for auditory verbal short-term memory as measured by scores for the auditorily presented, phonologically different words. Mean span for the 244 subjects who did not report possible brain damage was 5.45 items, with a range of 3.00–8.67 (SD = 1.03). Span for our 7 possible brain-damaged subjects ranged from 2.33 to 5.67, with a mean of 4.00. Three of them had spans of 3.67 or less, which were in the bottom 5% of the distribution. Only 1

of these subjects (Subject 236) reported having a "lousy memory," and he was 1 of the 2 subjects who failed to show three out of the four verbal short-term memory effects in Experiment 1. As such, he appeared to present a pattern similar to that found for verbal short-term memory patients.

Three of the subjects who reported that they might have suffered from brain damage were included in the sample for retest. All 3 subjects, when retested, showed clear effects of phonological similarity and word length with both presentation modes. This result mirrors the low test–retest reliability found for the retested group as a whole and is in contrast to the absence of effects found for these 3 subjects in Experiment 1.

The extent to which mild head injuries or other unreported potential causes of brain damage may affect cognitive performance has still to be determined (Richardson, 1990), although our own data are at least suggestive that some of these subjects might exhibit patterns similar to those described for single neuropsychological cases.

GENERAL DISCUSSION

Our broad intention in this paper was to explore a potential, largely neglected problem in interpreting results derived from groups of normal subjects. Our arguments have been directed primarily toward the implications for developing models of normal cognition, using data derived from both normal subjects and neuropsychological patients. The arguments also have implications for clinical neuropsychology, but that has not been our focus here.

We have concentrated on the effects in immediate, serial, oral recall of word sequences, of phonological similarity, and of word length with auditory and visual presentation. At least in the case of verbal short-term memory, it appears that some of our concerns were justified. A substantial minority of apparently normal subjects failed to show one or more of these effects that have been widely reported and widely replicated with groups of normal subjects and that have been used to interpret data from patients with verbal short-term memory deficits. Moreover, individual normal subjects appear to show an element of unreliability in the magnitude of the effects when tested on separate occasions. Two factors that influence the size of these effects appear to be memory span and reported strategy.

These findings have implications for the development of at least one model of verbal short-term memory and for the study of patients with verbal short-term memory deficits. They also have wider implications for the interpretation of data from studies of other aspects of normal cognition. We shall discuss each of these topics in turn.

One widely adopted model of verbal short-term memory has been developed by Baddeley and his colleagues. They have interpreted the effects of phonological similarity and word length in terms of a model of verbal short-term memory referred to as the *phonological loop* (e.g., Baddeley, 1992). The model comprises a passive phonological store and an active subvocal rehearsal process, and these two components commonly act in concert. The phonological similarity effect is thought to reflect mutual interference among items in the passive phonological store, the contents of which are prone to decay over time and to disruption from concurrently or subsequently presented material (Baddeley et al., 1984; Salamé & Baddeley, 1982). The word-length effect is seen as the signature of subvocal rehearsal, in that longer words take longer to rehearse, and the system is limited by the amount of material that can be rehearsed in around 2 sec (Baddeley et al., 1975; Ellis & Hennelley, 1980). Subvocal rehearsal serves the additional functions of refreshing the contents of the phonological store and translating visually presented verbal material into a phonological code. There is now a large body of evidence in support of the model, some of which we discussed in the introduction (see Baddeley, 1986, 1992; Della Sala & Logie, 1993; Logie, 1995).

How might such a model account for the findings reported here? One approach is to consider the two major determining factors that arose from our data—namely, verbal memory span and strategy. The phonological loop is thought to play a key role in verbal memory span tasks, with subvocal rehearsal serving to enhance performance. One reason for believing this to be the case is that when presentation of the material for recall is accompanied by the subject's repeating aloud an irrelevant word (articulatory suppression), then memory span is severely impaired (Murray, 1968). Furthermore, articulatory suppression removes the effect of word length with auditory or visual presentation and removes the effect of phonological similarity when the material is presented visually (Baddeley et al., 1984). Articulatory suppression is commonly interpreted as a technique to suppress the use of subvocal rehearsal. Thus, when subjects are prevented from using subvocal rehearsal, their memory span is very low, and they fail to show the effects under scrutiny in our experiments. One very plausible interpretation of our own results would be that, should subjects simply fail to use subvocal rehearsal for oral serial recall tasks, this would have an effect similar to that found when subvocal rehearsal is suppressed. This interpretation fits well with our finding that those subjects who reported consistently using a verbal rehearsal strategy produced the largest effects. Subjects who reported other strategies or a mixture of strategies showed smaller effects. The interpretation also fits with our finding that subjects with very poor memory span were less likely to show each of the effects. A further implication of having a low span is that there is little variability in the data, with performance close to floor. This leaves little room for the appearance of effects that are, in any case, somewhat unreliable.

One possible criticism of our methodology is that we have relied on one particular measure of span, and it may be that this measure is peculiarly unreliable. There are indeed alternative measures of span (e.g., Gregg, Freedman, & Smith, 1989). However, as mentioned earlier, our chosen measure of span has been widely used in clinical and educational settings as a measure of verbal short-term memory capacity, as well as in experimental studies both with

normal subjects and with neuropsychological patients. Moreover, we have previously demonstrated similar findings with a smaller subject sample, but using a different measure of span (Della Sala et al., 1991).

A second possible criticism is that we have collected too few data points per subject, with one data point for each type of list for recall. However, it is worth noting that each span score is derived from performance on a series of trials. Thus, for example, a span of 4.33 would involve the subjects' being required to undertake 12 trials, with 3 trials at each list length from 2 items to 5 items. Thus, the span is a measure that summarizes a number of data points for each subject in each condition. In this sense, the span measure serves a function that is very similar to the widely adopted practice of using mean data to summarize subject performance. Of course, there is still no guarantee that increasing the number of data points per condition or per subject would increase the reliability in any useful way. This view stems from the observation that a large number of subjects reported changing strategy within the testing sessions (mixed-strategy group). If subjects change strategies from one occasion to another, they may well be using different aspects of their cognitive architecture on each of those occasions. Thus, a summary measure that incorporated a larger sample of data would most likely be generated from the use of a larger range of strategies. Therefore, it might provide a statistically more reliable measure, but it would be a very poor indicator of the cognitive functions underlying performance on the task.

Evidence for this argument comes from our previous work on this topic (Della Sala et al., 1991) described earlier, in which we gathered data on phonological similarity and word length from 15 subjects using both visual and auditory presentation. The appearance or absence of the effects of phonological similarity and word length was successfully manipulated by requiring the use of subvocal rehearsal or by allowing subjects to choose an alternative strategy. Moreover, as in the experiments reported here, without specific instructions on strategy, subjects spontaneously changed strategy from one occasion to another. Therefore, had we taken a simple summary measure from all of the testing sessions with these subjects, the measure would not have provided a very accurate reflection of the subjects' performance patterns.

A further issue arises from the observation in Experiment 1 that each of the four effects seemed to be differentially reliable. Nearly all of the subjects showed phonological similarity effects, especially with auditory presentation, whereas a large number of subjects failed to show word-length effects. One possible view is that since articulatory rehearsal is a control process that subjects may choose not to employ, then the word-length effect may be vulnerable to strategy choice. In contrast, the passive phonological store may be less prone to such strategic effects, particularly with auditory presentation, since, in this case, there is thought to be obligatory access to the store (Salamé & Baddeley, 1982). However, such an interpretation gains little support from our analysis of effect size as a function of reported strategy. There was no interaction between these variables; therefore, it is unclear whether the magnitude of the word-length effect benefited from subvocal rehearsal any more than did the magnitude of the phonological similarity effect.

A possible interpretation of our data is that the phonological loop model of short-term verbal memory is simply wrong. Another possibility is that different subjects may have different cognitive architectures, and these differences are reflected in the differing patterns of data obtained. The results from our analysis of reported strategies and from our previous work (Della Sala et al., 1991) suggest an alternative view that there is indeed a common organization for human cognition and the phonological loop system is just one of a number of possible cognitive mechanisms, one or more of which may be applied to oral verbal serial recall. In other words, different subjects adopt different strategies for performing verbal temporary retention tasks, and these strategies may or may not use the phonological loop. Instead, such subjects may perform the task by using, for example, a visual short-term memory system, a lexical system, or a semantic system.

A related argument has been put forward by Wetherick (1975, 1976; Wetherick & Alexander, 1977) who demonstrated that when items are drawn from a single semantic category, immediate verbal recall is better than when items are drawn from several different categories. That is, the semantic category could act as a cue for recall. An analogous result was observed by Hulme et al. (1991) who reported that memory span for familiar items is higher than it is for unfamiliar items. Both these and other studies (e.g., Bourassa & Besner, 1994; Cowan, Wood, & Borne, 1994) are persuasive that semantic and/or lexical information in long-term memory can contribute to performance on memory span tasks.

It is interesting to note that, in each of these lines of argument, verbal recall performance was measured on group aggregate data and semantic information contributed significantly to the variance across the group as a whole. Nevertheless, these data could be interpreted as suggesting that most (or all) subjects took some advantage of semantic information. Alternatively, they could suggest that some subjects adopted a semantic strategy while other subjects adopted other strategies, such as subvocal rehearsal. Which of these interpretations is closer to reality could have a profound effect on the kind of model proposed to account for task performance. For example, a model based on the former interpretation might argue that subvocal rehearsal has a very minor role in these tasks. A model based on the latter interpretation would view subvocal rehearsal as one component of a "constellation" of components (Morris, 1986) that can be employed strategically. Unfortunately, it is not possible to determine which interpretation is correct without taking into account the individual strategies adopted by subjects in these tasks. Wetherick (1976, 1978) recognized that strategies during retrieval affected whether semantic information influenced recall performance. Our own data support the second suggestion—namely, that some of the subjects in these studies were using a semantically based strategy some of the time but that other subjects were using

verbal rehearsal or other strategies. That is, long-term memory could make a contribution to memory span to a much greater extent in some subjects than in others, depending on which strategy they adopt or which components of their cognitive architecture are employed. Moreover, it is not possible to glean from memory span alone which aspects of cognitive architecture are being used. For example, it is possible for the span scores obtained to fall on a normal distribution, with no evidence of bimodality and yet for scores from different subjects to be derived from different underlying cognitive processes. We can glean more from the appearance or absence of specific phenomena associated with span such as the effects studied here. However, even here, it seems crucial to assess the reliability of the effects across individuals and within the same individuals across time. Thus, studies that fail to take account of individual differences in the strategies adopted may be presenting a misleading picture of the nature of the cognitive architecture responsible for task performance.

Of course, we have to rely on subject reports as to the strategy adopted. However, we obtained a systematic effect of the reported strategy, and our findings are unlikely to have reflected simply the ease with which the subjects could accurately report their strategies in retrospect (Ericsson & Simon, 1980, 1984).

We would argue that the phonological loop survives intact as a useful model of verbal short-term memory, although it does not necessarily provide an entirely adequate account of performance on oral verbal serial recall tasks. What about the status of neuropsychological data as a source of evidence for the characteristics of the model? If the use of the phonological loop is dependent on one particular strategy that normal subjects may or may not adopt, there is no guarantee that a failure to use subvocal rehearsal reflects impairment of the phonological loop. By the same token, there is no guarantee that a lack of the effects of phonological similarity and word length reflect damage to verbal short-term memory. It is true to say that patients with damage to their verbal short-term memory system might tend to avoid using such a system or be unable to use it, but we cannot tell if this is the case solely by looking at their facility with words differing in phonological similarity and word length. Neither would it be sufficient to demonstrate that patients with verbal short-term memory impairments reliably failed to show these effects, since it is possible that they may reliably use a strategy that does not rely on subvocal rehearsal.

Patients may, of course, choose to use such a strategy because the phonological loop system is damaged, or they may choose an alternative strategy regardless of any damage. Thus, we cannot tell if the data pattern reflects an attempt to use a strategy that relies on a damaged system, or if it reflects the operation of an alternative part of the cognitive system that is not optimal for oral serial verbal recall tasks.

One way forward in using the neuropsychological data to develop a theory of verbal short-term memory might be to ask patients to use subvocal rehearsal and then examine whether this instruction results in the appearance of the verbal memory effects. Vallar and Baddeley (1984) tested articulatory rehearsal in their short-term verbal memory patient PV and found that her articulation was relatively normal. They also found that articulatory suppression had no effect on PV's verbal memory span and concluded that subvocal rehearsal was possible for PV but that she gained little benefit from it in verbal short-term memory tasks. However, Vallar and Baddeley did not conduct the crucial test of asking PV to use subvocal rehearsal as a strategy on the grounds that it might affect her performance on subsequent testing sessions. Given the rarity of relatively pure verbal short-term memory patients, this is a sensible precaution. In other words, the patient might have adopted a strategy that could undermine the data pattern thought to reflect the damage to her cognitive system and, hence, make her data less useful for theory development. For this very reason, it might be highly beneficial to explore which strategies patients adopt spontaneously in the light of their damage and perhaps systematically to explore the extent to which instructions to use particular strategies affect the data pattern obtained. Any researcher clearly would have to bear in mind the possible impact of such procedures on future testing sessions with the same patients. However, it is also possible that the patients might spontaneously change their strategies from one occasion to another, and a systematic investigation of strategy use in patients could be highly informative for both theory development and rehabilitation.

The issue of strategy choice has wider implications for studies of human cognition other than those on verbal temporary memory. Much of experimental cognitive psychology rests on the assumption of a common cognitive architecture, and this is an assumption that we have adopted in our discussion thus far. Notably, research on human cognition also rests on a second, related assumption—that, given a common cognitive architecture, most, if not all, subjects will use the same components of that architecture to perform the same task. Given that researchers rarely report individual anomalies in their group results, we have no way of knowing how many other well-established and well-replicated phenomena have similar (or even more extreme) patterns of variability when tested on more than one occasion with the same subjects or with a wider range of subjects. Theory development that does not incorporate such tests of generality may lead to theories that have very limited utility. One paper that cogently illustrates this point was concerned with the topic of children's arithmetic (Siegler, 1987). Siegler points to the "perils of averaging over strategies." In his study of children's addition, he noted that when taking group average data, he successfully replicated results reported by other researchers. However, on closer examination, he discovered considerable individual variability in the strategies adopted by different children, suggesting that the group effect was largely a statistical artifact. Therefore, basing a theory of children's addition on the average data was very misleading.

A further, general point stems from the low test–retest reliability found in our own data. One response to this finding would be to suggest that most researchers use a range of tests to provide converging evidence for a particular interpretation. This is not always the case in studies of normal

subjects, although, as we have discussed, this is common practice with single case studies of patients. However, we have already pointed to one difficulty with this interpretation—namely, that subjects may adopt different strategies for a given test conducted on different occasions. By the same argument, subjects could adopt different strategies for each of a range of tests on different occasions. Thus, the test–retest reliability of a battery of tests may not be any better than the test–retest reliability of each of the tests individually. Unless the test–retest reliability of each test and of the test battery is measured explicitly, the argument that equivocal data converge remains an untested assumption. A record of reported strategies might add greatly to the confidence placed in the interpretation of the data obtained. Subjective reports of strategy choice may, of course, not always be informative, and their utility will depend on the nature of the task. It is a truism to say that subjects are not always aware of how they perform a task (Ericsson & Simon, 1984; Nisbett & Wilson, 1977; Pylyshyn, 1973). However, if the test–retest reliability of a test proves to be very low, this might indicate the differential use of strategies by subjects on different occasions, even if subjects are unable to report their strategies adequately. Thus, test– retest reliability measures may be highly informative but are rarely collected or reported. This is true in studies of normal subjects and in studies of neuropsychological patients. In the case of the latter, the argument applies to the collection of data for development of theories of cognition and to studies conducted for clinical purposes.

In sum, we have attempted to contribute to studies of verbal short-term memory by providing some new insight into the characteristics of some widely reported short-term memory phenomena. It appears that the test–retest reliability is rather poor for these phenomena, and their extent is determined both by memory span and by strategy choice. It appears also that several different components of cognitive architecture contribute to performance in verbal short-term storage tasks, not all of which could be considered as part of a verbal short-term memory system. Whether the researcher is interested in a model of short-term verbal memory or a model of how subjects perform verbal short-term storage tasks, it would appear essential to consider both the range of memory span and the range of strategy choice among the experimental participants at all stages of data collection. Also, we have argued that these findings have implications for the generality of conclusions drawn from studies of normal subject groups and for the use of neuropsychological data in developing theories of cognition (for related discussions, see Caramazza, 1986, 1990, and Cohen, 1994). In this respect, the rather neglected data sources of test–retest reliability, of intersubject reliability, and of reported strategy choice could be highly informative routine, rather than periodic additions to the toolkit wielded by researchers in human cognition.

REFERENCES

Baddeley, A. D. (1966). Short-term memory for word sequences as a function of acoustic, semantic, and formal similarity. *Quarterly Journal of Experimental Psychology*, **18**, 362-365.

Baddeley, A. D. (1986). *Working memory*. Oxford: Oxford University Press.

Baddeley, A. D. (1992). Is working memory working? The fifteenth Bartlett Lecture. *Quarterly Journal of Experimental Psychology*, **44**, 1-31.

Baddeley, A. D., Lewis, V. J., & Vallar, G. (1984). Exploring the articulatory loop. *Quarterly Journal of Experimental Psychology*, **36**, 233-252.

Baddeley, A. [D.], Logie, R. [H.], Bressi, S., Della Sala, S., & Spinnler, H. (1986). Senile dementia and working memory. *Quarterly Journal of Experimental Psychology*, **38A**, 603-618.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning & Verbal Behavior*, **14**, 575-589.

Baddeley, A. D., & Wilson, B. (1985). Phonological coding and short-term memory in patients without speech. *Journal of Verbal Learning & Verbal Behavior*, **24**, 490-502.

Bishop, D. V. M., & Robson, J. (1989). Unimpaired short-term memory and rhyme judgement in congenitally speechless individuals: Implications for the notion of "articulatory coding." *Quarterly Journal of Experimental Psychology*, **41A**, 123-141.

Bourassa, D. C., & Besner, D. (1994). Beyond the articulatory loop: A semantic contribution to serial order recall of subspan lists. *Psychonomic Bulletin & Review*, **1**, 122-125.

Caplan, D., Rochon, E., & Waters, G. S. (1992). Articulatory and phonological determinants of word length effects in span tasks. *Quarterly Journal of Experimental Psychology*, **45A**, 177-192.

Caramazza, A. (1986). On drawing inferences about the structure of the normal cognitive system from the analysis of patterns in impaired performances: The case for single-case study. *Brain & Cognition*, **5**, 41-66.

Caramazza, A. (Ed.) (1990). *Cognitive neuropsychology and neurolinguistics*. Hillsdale, NJ: Erlbaum.

Carlson, G. S. (1986). Head injury in a population study. *Acta Neurosurgica*, **36**, 13-15.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement*, **1**, 37-47.

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, **49**, 997-1003.

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology*, **55**, 75-84.

Cowan, N., Wood, N. L., & Borne, D. N. (1994). Reconfirmation of the short-term storage concept. *Psychological Science*, **5**, 103-106.

Crovitz, H. F., Horn, R. W., & Daniel, W. F. (1983). Inter-relationships among retrograde amnesia, post-traumatic amnesia, and time since head injury: A retrospective study. *Cortex*, **19**, 407-412.

Cubelli, R., & Nichelli, P. (1992). Inner speech in anarthria: Neuropsychological evidence of differential effects of cerebral lesions on subvocal articulation. *Journal of Clinical & Experimental Neuropsychology*, **14**, 499-517.

Della Sala, S., & Logie, R. [H.] (1993). When working memory does not work: The role of working memory in neuropsychology. In F. Boller & H. Spinnler (Eds.), *Handbook of neuropsychology* (Vol. 8, pp. 1-62). Amsterdam: Elsevier.

Della Sala, S., & Logie, R. H. (in press). Impairments of methodology and theory in cognitive neuropsychology: A case for rehabilitation? *Neuropsychological Rehabilitation*.

Della Sala, S., Logie, R. H., Marchetti, C., & Wynn, V. (1991). Case studies in working memory: A case for single cases? *Cortex*, **27**, 169-191.

De Renzi, E., & Nichelli, P. (1975). Verbal and nonverbal short term memory impairment following hemispheric damage. *Cortex*, **11**, 341-353.

Ellis, N. C., & Hennelley, R. A. (1980). A bilingual word length effect: Implications for intelligence testing and the relative ease of mental calculation in Welsh and English. *British Journal of Psychology*, **71**, 43-52.

Ericsson, K. A., & Simon, H. (1980). Verbal reports as data. *Psychological Review*, **87**, 215-251.

Ericsson, K. A., & Simon, H. (1984). *Protocol analysis*. Cambridge, MA: MIT Press.

GILHOOLY, K. J., LOGIE, R. H., WETHERICK, N. E., & WYNN, V. (1993). Working memory strategies in syllogistic reasoning tasks. *Memory & Cognition*, **21**, 115-124.

GREGG, V. H., FREEDMAN, C. M., & SMITH, D. K. (1989). Word frequency, articulatory suppression and memory span. *British Journal of Psychology*, **80**, 363-374.

HUBER, H. P. (1973). *Psychometrische Einzelfalldiagnostik* [Psychometric case studies]. Weinheim und Basel: Beltz Verlag.

HULME, C., MAUGHAN, S., & BROWN, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory & Language*, **30**, 685-701.

JORM, A. F., & SHARE, D. L. (1983). Phonological recoding and reading acquisition. *Applied Psycholinguistics*, **4**, 103-147.

LEVY, B. A. (1971). The role of articulation in auditory and visual short-term memory. *Journal of Verbal Learning & Verbal Behavior*, **10**, 123-132.

LOGIE, R. H. (1995). *Visuo-spatial working memory*. Hillsdale, NJ: Erlbaum.

LOGIE, R. H., CUBELLI, R., DELLA SALA, S., ALBERONI, M., & NICHELLI, P. (1989). Anarthria and verbal short-term memory. In J. Crawford & D. Parker (Eds.), *Developments in clinical and experimental neuropsychology* (pp. 203-211). New York: Plenum.

MILNER, B. (1971). Interhemispheric differences and psychological processes. *British Medical Bulletin*, **27**, 272-277.

MORRIS, N. (1986). *Working memory constellations*. Unpublished doctoral thesis, University of Durham.

MURRAY, D. (1965). Vocalization-at-presentation, with varying presentation rates. *Quarterly Journal of Experimental Psychology*, **17**, 47-56.

MURRAY, D. (1968). Articulation and acoustic confusability in short-term memory. *Journal of Experimental Psychology*, **78**, 679-684.

NISBETT, R. E., & WILSON, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, **84**, 231-259.

PYLYSHYN, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin*, **80**, 1-24.

RICHARDSON, J. T. E. (1990). *Clinical and neuropsychological aspects of closed head injury*. London: Taylor & Francis.

SALAMÉ, P., & BADDELEY, A. D. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning & Verbal Behavior*, **21**, 150-164.

SEGALOWITZ, S. J., & BROWN, D. (1991). Mild head injury as a source of developmental disabilities. *Journal of Learning Disabilities*, **24**, 551-559.

SIEGLER, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, **116**, 250-264.

SIMON, H. A., & REED, S. K. (1976). Modelling strategy shifts in a problem solving task. *Cognitive Psychology*, **8**, 86-97.

VALLAR, G., & BADDELEY, A. D. (1984). Fractionation of working memory: Neuropsychological evidence for a phonological short-term store. *Journal of Verbal Learning & Verbal Behavior*, **23**, 151-161.

VALLAR, G., & CAPPA, S. F. (1987). Articulation and verbal short-term memory: Evidence from anarthria. *Cognitive Neuropsychology*, **4**, 55-78.

VALLAR, G., & SHALLICE, T. (1990). *Neuropsychological impairments of short-term memory*. Cambridge: Cambridge University Press.

WETHERICK, N. E. (1975). The role of semantic information in short-term memory. *Journal of Verbal Learning & Verbal Behavior*, **14**, 471-480.

WETHERICK, N. E. (1976). Semantic information in short-term memory: Effects of presenting recall instructions after the list. *Bulletin of the Psychonomic Society*, **8**, 79-81.

WETHERICK, N. E. (1978). Strategies and memory. In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory* (pp. 644-651). London: Academic Press.

WETHERICK, N. [E.], & ALEXANDER, J. (1977). The role of semantic information in short-term memory in children aged 5 to 9 years. *British Journal of Psychology*, **68**, 71-75.

NOTE

1. We are grateful to Nelson Cowan for making us aware of this possibility.

(Manuscript received August 30, 1994;
revision accepted for publication July 12, 1995.)