Psychonomic Bulletin & Review
2003, 10 (2), 430–434

The time-based word length effect and stimulus set specificity

IAN NEATH, TAMRA J. BIRETA, and AIMÉE M. SURPRENANT
Purdue University, West Lafayette, Indiana

The *word length effect* is the finding that short items are remembered better than long items on immediate serial recall tests. The *time-based* word length effect refers to this finding when the lists comprise items that vary only in pronunciation time. Three experiments compared recall of three different sets of disyllabic words that differed systematically only in spoken duration. One set showed a word length effect, one set showed no effect of word length, and the third showed a reverse word length effect, with long words recalled better than short. A new fourth set of words was created, and it also failed to yield a time-based word length effect. Because all four experiments used the same methodology and varied only the stimulus sets, it is argued that the time-based word length effect is not robust and as such poses problems for models based on the phonological loop.

The most widely accepted view of immediate memory is a framework known as working memory (see, e.g., Baddeley, 1986; Cowan, 1995; Miyake & Shah, 1999). These models all share the assumption that "memory traces decay over a period of a few seconds, unless revived by articulatory rehearsal" (Baddeley, 2000, p. 419). Recall of items in situations thought to depend on working memory is thus a joint function of the decay rate of the items and the time needed for one to rehearse them. Given the assumption that there is a positive correlation between the rate of subvocal rehearsal and overt pronunciation time, working memory models must predict worse recall of items that take longer to pronounce than items that take less time to pronounce.

Cowan (1995, p. 42) has stated that the word length effect, the finding that people do indeed recall temporally shorter words better than temporally longer words, is the "best remaining solid evidence" in favor of this view. However, Lovatt, Avons, and Masterson (2000) recently presented data suggesting that the time-based word length effect is observable only with one set of stimuli; other stimulus sets produce no such effect, or even a reverse effect. Thus, the purpose of the present work is to examine, under standardized conditions, four different sets of stimuli that vary systematically only in pronunciation time and to determine how many produce a time-based word length effect.

The empirical demonstration of failing to observe a time-based word length effect has tremendous theoretical significance. The architectural core of the working memory framework (e.g., Baddeley, 1986, 2000) includes the phonological store and articulatory control process, which together constitute the articulatory loop. Items decay in the phonological store unless refreshed by rehearsal via the articulatory control process. Because shorter items take less time to rehearse, more decaying traces of short items can be refreshed than decaying traces of long items. By definition, differences will be apparent only in supra-span lists. If the time-based word length effect is not robust with respect to stimulus sets, then the foundation of the working memory view is greatly compromised.

Baddeley, Thomson, and Buchanan (1975) reported the first detailed examination of the time-based word length effect, although the high correlation between reading rate and memory span had been noted earlier (e.g., Mackworth, 1963). They found that a set of disyllabic words that took less time to pronounce was recalled better than an otherwise equivalent set of words that took more time to pronounce. There exist numerous replications of this finding using either the full set of stimuli or a subset: Cowan et al. (1992); Longoni, Richardson, and Aiello (1993); Lovatt et al. (2000); and Nairne, Neath, and Serra (1997).

Studies that have used different stimulus sets, however, have consistently failed to replicate the time-based word length effect. Caplan, Rochon, and Waters (1992) created a set of two-syllable words that differed in spoken duration but were matched for number of syllables and number of phonemes. They observed better recall of the temporally long words. Lovatt et al. (2000) used two different sets of items; in one instance, they found better recall of longer words, and in the other, no difference in recall of short and long words.

Similar results have been reported in other languages. For example, Service (1998) compared recall of three types of items by taking advantage of the structure of Finnish to vary the duration and complexity of the list items inde-

We thank Amanda Alexander, Mavis Fuller, Angela Pasyk, Jennifer Smith, Mariah Wells, and Kimberly Woods for assistance in collecting the data. Correspondence may be addressed to Ian Neath, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907-2004 (e-mail: ineath@purdue.edu).

pendently. She found that long two-syllable Finnish pseudowords were recalled about as accurately as short two-syllable pseudowords. Zhang and Feng (1990, reported in Lovatt et al., 2000) also found no difference in the level of recall of short and long Chinese disyllables.

The studies above differ in more ways than just composition of the stimulus set. There are variations in list length, presentation time, presentation modality, response modality, number of trials, type of test, response mode, subject population, and so on. Given the possibility that stimulus set characteristics may be critical in determining whether one sees a time-based word length effect, Experiments 1, 2, and 3 were designed to replicate the findings of Baddeley et al. (1975), Caplan et al. (1992), and Lovatt et al. (2000), respectively, using a consistent methodology.[1] In Experiment 4, we used the same methodology but used a new set of words to increase the number of stimulus sets tested.

GENERAL METHOD

In Experiments 1–4, we used the same methodology, with the experiments differing only in the stimulus items.

Subjects

Sixteen different Purdue University undergraduates participated in each experiment in exchange for credit in introductory psychology courses. All were native speakers of American English.

Design

Word length was a within-subjects variable. The subjects saw the to-be-remembered items presented on a computer screen and were asked to read each word silently. Occurrence of lists of long or short items was random, with the constraint that each occurred 10 times per block of 20 lists. The words were displayed in the center of the screen for 1 sec, and the next word was presented immediately after the offset of the previous word. List length was set at eight because three of the four stimulus sets had eight short and eight long words. Because all eight short or all eight long items were included in each list, each list of short stimuli remained equated on all factors other than pronunciation time with each list of long stimuli.

Procedure

The subjects were informed that we were interested in how accurately they could remember the order in which a series of words had been presented. They were asked to indicate the presentation order by clicking on appropriately labeled buttons on the screen using the mouse. For example, if they thought the first word was VACUUM, they should click on the button labeled VACUUM first. This method of testing has yielded robust time-based word length effects in the past (e.g., Nairne et al., 1997). All subjects reported having used a mouse previously, and no one reported any difficulty in using the mouse to respond. Each subject received 40 trials, 20 with long words and 20 with short words, randomly intermixed. Recall was self-paced. The subjects were tested individually, and the experimenter remained in the room to ensure that the instructions were followed appropriately.

Additional measurements

For each of the four sets of stimuli, 16 additional subjects, all of whom were native speakers of American English, provided measures of the temporal duration of all of the stimuli. Two such measures were obtained. For all the stimulus sets, the difference between short and long pronunciation times was significantly different by a two-tailed t test with an alpha level of .05 (see Table 1 for means).

Normal pronunciation time. The first measure was the time necessary for subjects to pronounce each word aloud at a "normal" speaking rate (e.g., Baddeley et al., 1975, Experiment 3; Lovatt et al., 2000, Experiment 1). These subjects were informed that we needed speakers to produce lists of items that we would use as stimuli in future experiments. The items were read as a list, with each subject receiving a different random ordering of the items, but there was a pause between each item and the item that followed. Each list was read through several times prior to recording. Each list was recorded onto cassette tape and then digitized at 44 kHz. The duration of each token was then estimated using digital sound editing software.

Speeded list pronunciation time. The second measure was the time needed for subjects to pronounce an entire list of 8 items as quickly as possible. The idea was that if subjects do indeed rehearse subvocally, as the working memory models posit, they are likely to do so as fast as possible. These subjects were informed that they should pronounce the entire list of items as quickly and as clearly as they could. Each subject received a different random order, and each list was read by the subject several times prior to recording. Each list was recorded onto cassette and then digitized at 44 kHz. The time to pronounce each list was then estimated using digital sound editing software.

Table 1
Proportion of Short and Long Items Recalled in Order in Experiments 1–4 and Pronunciation Time (in Milliseconds)

	Proportion Recalled	Normal Pronunciation Time (msec)	Speeded Pronunciation Time (msec)
Experiment 1			
Short	.403	484	390
Long	.351	587	456
Experiment 2			
Short	.400	470	364
Long	.513	570	416
Experiment 3			
Short	.411	490	353
Long	.418	572	423
Experiment 4			
Short	.470	626	475
Long	.472	744	543

EXPERIMENT 1

Method

The stimuli were 8 of the 10 long and 8 of the 10 short words from Experiment 3 of Baddeley et al. (1975). The short words were BISHOP, DECOR, EMBER, PECTIN, PEWTER, TIPPLE, WICKET, WIGGLE. The long words were COERCE, FRIDAY, HARPOON, HUMANE, MORPHINE, TYCOON, VOODOO, ZYGOTE. The words CYCLONE and NITRATE were dropped from the long pool, and the words HACKLE and PHALLIC were dropped from the short pool. The mean normal pronunciation times were 484 msec for the short words and 587 msec for the long words, a difference of 103 msec. The mean speeded pronunciation times were 390 msec for the short words and 456 msec for the long words, a difference of 66 msec.

Results

The proportion correctly recalled in order was .403 for the short items and .351 for the long items [$F(1,15) = 11.03$, $MS_e = 0.016$, $p < .01$]. There was a reliable effect of serial position [$F(7,105) = 47.71$, $MS_e = 0.018$, $p < .01$], and there was no interaction [$F(7,105) < 1$]. Experiment 1 thus replicated the many previous reports of a time-based word length effect using the original stimuli of Baddeley et al. (1975).

EXPERIMENT 2

Method

The stimuli were the eight long and eight short words from Experiment 2 of Caplan et al. (1992). The short words were BULLET, CABIN, CARROT, DEVIL, LADDER, PICNIC, TICKET, ZIPPER. The long words were BABY, BALLOON, CRAYON, ORANGE, SIRLOIN, SPIDER, TOWER, VACUUM. The mean normal pronunciation times were 470 msec for the short words and 570 msec for the long words, a difference of 100 msec. The mean speeded pronunciation times were 364 msec for the short words and 416 msec for the long words, a difference of 52 msec.

Results

Unlike the stimulus set used in Experiment 1, the stimulus set used in Experiment 2 yielded a *reverse* word length effect. The proportion of long items recalled (.513) was greater than the proportion of short items recalled (.400), replicating Caplan et al. (1992; Caplan & Waters, 1994). The main effect of word length was reliable [$F(1,15) = 30.79$, $MS_e = 0.027$, $p < .01$], as was the main effect of position [$F(7,105) = 65.63$, $MS_e = 0.020$, $p < .01$]. The interaction was also reliable [$F(7,105) = 2.42$, $MS_e = 0.008$, $p < .05$], reflecting a larger difference for later serial positions than for early serial positions.

EXPERIMENT 3

Method

The stimuli were the eight long and eight short words from Experiment 1B of Lovatt et al. (2000). The short words were BUTTON, CANDLE, PENCIL, POCKET, SHOVEL, SPIDER, TRACTOR, WHISTLE. The long words were BRANCHES, CANOES, CURTAINS, NECKLACE, NEEDLE, PEBBLES, ROBOT, STATION. The mean normal pronunciation times were 490 msec for the short words and 572 msec for the long words, a difference of 82 msec. The mean speeded pronunciation times were 353 msec for the short words and 423 msec for the long words, a difference of 70 msec.

Results

The stimulus set used in Experiment 3 yielded no difference between recall of short and long items, .411 and .418, respectively. The main effect of length was not reliable [$F(1,15) < 1$], whereas the main effect of serial position was [$F(7,105) = 60.90$, $MS_e = 0.017$, $p < .01$]. The interaction was not reliable [$F(7,105) < 1$]. Lovatt et al. (2000) found a reverse word length effect with these stimuli with pointing as the response mode; we used a conceptually similar response mode (clicking on buttons with a mouse) but observed no difference. Most importantly, no time-based word length effect was observable with these stimuli and the direction of the nonsignificant difference was in the same direction as in Lovatt et al.'s study.

EXPERIMENT 4

Method

In Experiment 4, we used a new set of stimuli, described fully in the Appendix. The eight short words did not differ significantly from the eight long words on any measure except pronunciation time. Any slight advantages favored the short items. The mean pronunciation times were 626 msec for the short words and 744 msec for the long words, a difference of 118 msec. The mean speeded pronunciation times were 475 msec for the short words and 543 msec for the long words, a difference of 68 msec.

Results

The proportion of short items recalled (.470) was the same as the proportion of long items recalled (.472). The main effect of word length was not reliable [$F(1,15) < 1$], whereas the main effect of serial position was [$F(7,105) = 56.10$, $MS_e = 0.019$, $p < .01$]. The interaction was not reliable [$F(7,105) < 1$].

GENERAL DISCUSSION

The time-based word length effect, better recall of words that take less time to pronounce than of equivalent words that take longer to pronounce, was observed for only one set of stimuli (from Baddeley et al., 1975). A reverse effect was observed for the stimuli created by Caplan et al. (1992), whereas there was no difference in recall for the stimuli from Lovatt et al. (2000) or the new set we created. If one includes the other pool of words used by Lovatt et al. that were not tested here, there exists one set of English words that show a time-based word length effect and there are four sets of words that do not show such an effect. In addition, one set of Finnish pseudowords (Service, 1998) and one set of Chinese words (Zhang & Feng, 1990, reported in Lovatt et al., 2000) also do not show a time-based word length effect. Our results confirm and extend the suggestion of Lovatt et al. (2000) that the time-based word length effect might be an artifact of the particular set of stimuli used.

One cannot plausibly argue that the differences in pronunciation time were too small to yield a difference in recall performance, because the difference in measured pronunciation time was comparable in all four experiments and a reliable advantage for short words was found in Experi-

ment 1 and a reliable advantage for long words was found in Experiment 2. One also cannot plausibly argue that aspects of our methodology compromised whether any effects would be observed, because all experiments were identical except for the stimuli and because the conditions in Experiment 1 were sufficient for one to observe a time-based word length effect.

One might argue that the sets of stimuli do not adequately control for all factors other than pronunciation time. This is likely, given the difficulty in matching a small set of words on all possible measures. It was almost impossible to find sufficient overlap among published norms to ensure that our stimulus items differed in pronunciation time but were equated for recent measures of frequency, familiarity, concreteness, imageability, number of phonemes, and number of syllables. Thus, the items may differ in terms of phonotactic probability, neighborhood density, or neighborhood frequency. If there are systematic differences between the short and long items in some of the stimulus sets tested here, these effects exert a stronger influence on recall than does pronunciation time. This in itself is a problem for phonological loop models which give temporal duration, effects of phonological confusability, and rehearsal rate the primary role. It would be highly problematic if a linguistic factor such as phonotactic probability was more important on serial recall tasks than temporal duration.

Two other sources of evidence for the idea of time-based decay offset by rehearsal come from two different studies by Cowan and his colleagues. Cowan et al. (1992), using a subset of the Baddeley et al. (1975) stimuli, examined recall when half of the list was made up of short items and half was made up of long items. They found better recall when the short items occurred in the first half of the list than when the long items occurred first. Lovatt, Avons, and Masterson (2002) used the same procedure, but with a different pool of words, and found no such advantage. In accord with the results of Lovatt et al. (2000) and those we report, effects of stimulus duration appear to vary with different stimulus sets.

In a second paradigm, Cowan, Nugent, Elliott, and Geer (2000) asked subjects to read the same set of words aloud either quickly or slowly. They found that performance was better when the lists were read quickly than when the lists were read slowly, a result that is consistent with the idea of time-based decay offset by rehearsal (but see also Service, 2000). Such a procedure might introduce a confound, however. It is possible that speaking quickly versus speaking slowly differs not only in the resulting temporal duration of the spoken item, but also in terms of the type of information emphasized. For example, according to the item-order hypothesis (Nairne, Riegler, & Serra, 1991), speaking quickly is likely to lead to more order information than speaking slowly, and order information is particularly useful on serial recall tasks.

These findings and conclusions do not apply to word length effects observed when factors other than pronunciation time (e.g., number of syllables, number of phonemes) are allowed to vary. However, they do suggest that variables other than temporal duration are likely to be important given that the effect of pronunciation time cannot be widely observed when it is the only stimulus characteristic that is systematically varied (e.g., Brown & Hulme, 1995; Dosher & Ma, 1998; Neath & Nairne, 1995).

Any model of memory based on the phonological loop predicts that lists of short words will be recalled better than lists of comparable long words. Only one set of words produces a time-based word length effect, whereas four sets of English stimuli and one set each of Finnish and Chinese stimuli yield no difference or a reverse effect. The challenge for proponents of working memory is to find other sets of stimuli that do produce a word length effect.

REFERENCES

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

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, **4**, 417-423.

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.

Brown, G. D. A., & Hulme, C. (1995). Modeling item length effects in memory span: No rehearsal needed? *Journal of Memory & Language*, **34**, 594-621.

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.

Caplan, D., & Waters, G. S. (1994). Articulatory length and phonological similarity in span tasks: A reply to Baddeley and Andrade. *Quarterly Journal of Experimental Psychology*, **47A**, 1055-1062.

Carroll, J. B., Davies, P., & Richman, A. (1971). *Word frequency book*. Boston: Houghton Miflin.

Cowan, N. (1995). *Attention and memory: An integrated framework*. New York: Oxford University Press.

Cowan, N., Day, L., Saults, J. S., Kellar, T. A., Johnson, T., & Flores, L. (1992). The role of verbal output time in the effects of word-length on immediate memory. *Journal of Memory & Language*, **31**, 1-17.

Cowan, N., Nugent, L. D., Elliott, E. M., & Geer, T. (2000). Is there a temporal basis of the word length effect? A response to Service (1998). *Quarterly Journal of Experimental Psychology*, **53A**, 647-660.

Dosher, B. A., & Ma, J.-J. (1998). Output loss or rehearsal loop? Output-time versus pronunciation-time limits in immediate recall for forgetting-matched materials. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 316-335.

Longoni, A. M., Richardson, J. T. E., & Aiello, A. (1993). Articulatory rehearsal and phonological storage in working memory. *Memory & Cognition*, **21**, 11-22.

Lovatt, P., Avons, S. E., Masterson, J. (2000). The word-length effect and disyllabic words. *Quarterly Journal of Experimental Psychology*, **53A**, 1-22.

Lovatt, P., Avons, S. E., & Masterson, J. (2002). Output decay in immediate serial recall: Speech time revisited. *Journal of Memory & Language*, **46**, 227-243.

Mackworth, J. F. (1963). The duration of the visual image. *Canadian Journal of Psychology*, **17**, 62-81.

Miyake, A., & Shah, P. (Eds.) (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. New York: Cambridge University Press.

Nairne, J. S., Neath, I., & Serra, M. (1997). Proactive interference plays a role in the word-length effect. *Psychonomic Bulletin & Review*, **4**, 541-545.

Nairne, J. S., Riegler, G. L., & Serra, M. (1991). Dissociative effects of generation on item and order retention. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 702-709.

Neath, I., & Nairne, J. S. (1995). Word-length effects in immediate

memory: Overwriting trace decay theory. *Psychonomic Bulletin & Review*, **2**, 429-441.

PAIVIO, A., YUILLE, J. C., & MADIGAN, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monographs*, **76** (1, Pt. 2), 1-25.

SERVICE, E. (1998). The effect of word length on immediate serial recall depends on phonological complexity, not articulatory duration. *Quarterly Journal of Experimental Psychology*, **51A**, 283-304.

SERVICE, E. (2000). Phonological complexity and word duration in immediate recall: Different paradigms answer different questions—A comment on Cowan, Nugent, Elliott, and Geer. *Quarterly Journal of Experimental Psychology*, **53A**, 661-665.

NOTE

1. Because our subjects were all native speakers of American English, we did not try to replicate the lack of a time-based word length effect with Finnish pseudowords or Chinese disyllables.

APPENDIX
Stimuli Used in Experiment 4

Note: The short and long items differ significantly (by a two-tailed *t* test) only in pronunciation time.

NPT: mean normal pronunciation time in milliseconds.‡
CNC: concreteness.*
FAM: printed familiarity.*
IMG: imageability.*
MEAN: meaningfulness.*
NP: number of phonemes.*
NS: number of syllables.*
PF: Paivio frequency.§
BNC: British National Corpus frequency.†
SF: standard frequency index.#

*Measures obtained from the MRC Psycholinguistic Database (http://www.psy.uwa.edu.au/MRCDataBase/mrc2.html). †A measure obtained from the British National Corpus (http://info.ox.ac.uk/bnc/). ‡A measure collected as part of the experiment. §A measure obtained from Paivio, Yuille, & Madigan (1968). #A measure obtained from Carroll, Davies, & Richman (1971).

	Measure									
	NPT	CNC	FAM	IMG	Mean	NP	NS	PF	BNC	SFI
Short										
acrobat	636	536	431	583	567	7	3	1	24	35.6
animal	521	587	620	575	700	6	3	68	6,629	62.7
daffodil	633	595	404	611	696	7	3	3	43	36.7
gentleman	621	516	537	559	580	8	3	28	5,036	52.9
medallion	699	577	338	565	632	8	3	1	97	31.0
physician	615	573	472	572	592	7	3	43	492	47.0
umbrella	628	606	511	592	676	7	3	13	809	48.6
vegetable	654	602	591	598	692	8	4	10	963	51.0
Mean	626	578	488	582	642	7.25	3.13	20.88	1,762	45.69
Long										
automobile	776	607	456	628	616	8	4	50	217	56.2
infirmary	758	557	437	546	665	8	4	1	187	26.6
macaroni	759	631	498	608	600	8	4	2	39	46.7
newspaper	776	576	641	616	612	8	3	65	5,017	57.1
performer	680	529	479	530	648	6	3	4	473	44.8
prosecutor	794	520	454	497	632	10	4	2	357	30.5
volcano	683	591	461	627	760	7	3	14	363	51.9
wholesaler	725	539	403	409	488	7	3	1	113	31.0
Mean	744	569	479	558	628	7.75	3.50	17.38	846	43.10

(Manuscript received December 10, 2001;
revision accepted for publication April 30, 2002.)