Theoretical Implications of Articulatory Duration, Phonological Similarity, and Phonological Complexity in Verbal Working Memory

Shane T. Mueller, Travis L. Seymour, David E. Kieras, and David E. Meyer
University of Michigan

The phonological-loop model provides a prominent theoretical description of verbal working memory. According to it, serial memory span should be inversely related to the articulatory duration and phonological similarity of verbal items in memorized sequences. Initial tests of these predictions by Baddeley and colleagues appeared to support the phonological-loop model, but subsequent researchers have obtained conflicting data that putatively disconfirm its assumptions. Such conflicts may have stemmed from inadequate measurements of articulatory duration and phonological similarity. The present article discusses these concerns and proposes new theoretically-principled methods for properly measuring articulatory duration and phonological similarity. Two experiments that used these methods in the context of a verbal serial recall task are reported The results of these experiments confirm and extend the predictions of the phonological-loop model, while disarming previous criticisms of it.

Key words: Verbal working memory, phonological-loop model, memory span, serial recall, word length effect, articulatory duration, phonological similarity, phonological complexity

Since Miller's (1956) classic article on the magical number seven, plus or minus two, short-term verbal working memory (VWM) has become an increasingly important topic of investigation in experimental psychology and cognitive science. VWM plays a major role during performance of many basic mental tasks, ranging from serial recall to sentence comprehension, syllogistic reasoning, and arithmetic problem solving (Baddeley, 1986; Baddeley, 1992; Baddeley & Logie, 1999). Consequently, to account for such performance, investigators have proposed numerous hypotheses about the mechanisms that mediate VWM (e.g., Anderson & Matessa, 1997; Baddeley & Hitch, 1974; Brown & Hulme, 1995; Cantor & Engle, 1993; Cowan, 1999; Dosher & Ma, 1998; Drewnowski, 1980; Estes, 1972; Kieras, Meyer, Mueller, & Seymour, 1999; Laughery & Pinkus, 1970; Lewandowsky & Murdock, 1989; Murdock, 1993; Nairne, 1990; Page & Norris, 1998a; Sperling, 1967; Sternberg, Monsell, Knoll, & Wright, 1978; Waugh & Norman, 1965).

Following these proposals, significant controversies have arisen about which hypotheses best explain and predict various empirical phenomena related to VWM (Miyake & Shah, 1999). A major focus of controversy has been the influential phonological-loop model of VWM proposed by Baddeley (1986, 1992) and his colleagues (Baddeley & Hitch, 1974; Baddeley, Thomson, & Buchanan, 1975; Baddeley, Lewis, & Vallar, 1984). The basis for this controversy concerns the putative effects of phonological complexity (i.e., the numbers of phonemes and syllables in a word)[1] and articulatory duration (i.e., the time taken to pronounce a word) on performance of the verbal serial recall task. Critics of the phonological-loop model have argued that its accounts of these effects are either incomplete or incorrect (e.g., Caplan, Rochon, & Waters, 1992; Caplan & Waters, 1994; Cowan, Wood, Nugent, & Treisman, 1997; Cowan et al., 1998; Service, 1998; Cowan, Nugent, Elliot, & Geer, 2000; Service, 2000; Lovatt, Avons, & Masterson, 2000). In reply, proponents of the phonological-loop model have claimed that such arguments are dubious and do not require its assumptions to be substantially modified or discarded (e.g., Baddeley & Andrade, 1994; Baddeley & Logie, 1999).

We believe that this controversy has stemmed at least partly from inadequate measurement and control of factors such as articulatory duration, phonological similarity, and

Funding for this research was provided by Grant N00014-92-J-1173 from the U. S. Office of Naval Research to the University of Michigan. We thank members of the Brain, Cognition, and Action Laboratory at Michigan (David Fencsik, Darren Gergle, Jennifer Glass, Leon Gmeindl, Cerita Jones, Eric Schumacher, Adam Krawitz, and Mollie Schweppe) for helpful comments and Beth Emmons, Sean Ferguson, Michelle Gayewski, Ryan Kettler, Mollie Schweppe, and Michaela Ferguson for assistance in conducting the experiments. Editorial suggestions by Nelson Cowan and Gary Dell are gratefully acknowledged. Correspondence regarding this article should be addressed to Shane T. Mueller (smueller@umich.edu) or David E. Meyer (demeyer@umich.edu), Cognition and Perception Program, Dept. of Psychology, University of Michigan, 525 East University, Ann Arbor, MI, 48109-1109, USA.

[1] The term "phonological complexity" (as used by Service, 1998), which we have adopted here, should not be confused with the term "articulatory complexity" (as used by Caplan et al., 1992). Articulatory complexity can be thought of as indexing "ease of articulation", and so is more closely related to articulatory duration than is phonological complexity, which indexes the size of the sublexical phonological representation for a word.

phonological complexity. Thus, the purpose of the present article is to consider how these factors have been measured (and mismeasured) in the past, and to propose new methodological techniques for measuring them that may help resolve some of the persistent controversies about their roles in VWM tasks. Our approach here includes the following steps: First, we describe the phonological-loop model of VWM. Second, we discuss several recent empirical studies that have attempted to measure the effects of articulatory duration, phonological similarity, and phonological complexity on VWM. Third, we introduce more appropriate theoretically-principled methods that can be used for measuring phonological similarity and articulatory duration. Fourth, we present results of new experiments that confirm the utility and informativeness of our methods. Fifth, we evaluate the status of the phonological-loop model based on of our findings.

The Phonological-Loop Model

Several related versions of the phonological-loop model have been proposed (e.g., Baddeley & Hitch, 1974; Baddeley et al., 1984; Baddeley, 1986, 1992; Sperling, 1967). The one that we consider here comes mainly from Baddeley (1986). This model has three interconnected components: the phonological short-term store, articulatory motor-program processor, and central executive. The phonological short-term store is a repository of temporary phonological codes for verbal items. The articulatory motor-program processor enables the phonological codes to be refreshed through a cyclic rehearsal process. During its operations, codes in the phonological short-term store are converted sequentially to articulatory motor programs and vocalized overtly or covertly one after another. Finally, the central executive supervises interactions between the articulatory motor-program processor and phonological short-term store as well as other working-memory components.

Empirical Support

Support for the phonological-loop model has come from experiments with a wide range of tasks that stress VWM. Although many variables are known to affect VWM, we focus primarily on the effects of two factors: "word length", and phonological similarity.

The "word length" effect.[2] Some researchers have found that serial recall accuracy is better for sequences of words whose articulatory durations are shorter (e.g., Baddeley et al., 1975; Schweickert & Boruff, 1986; Longoni et al., 1993). Baddeley et al. (1975, Experiment IV) investigated this effect by comparing serial recall accuracy for two sets of words that had different mean articulatory durations but equal phonological complexity (i.e., numbers of phonemes and syllables). They found that sequences with shorter durations were remembered better than sequences with longer durations.

According to the phonological-loop model, articulatory-duration effects occur because words that are spoken more rapidly can be rehearsed more frequently. Supposedly, words that are rehearsed more frequently are less likely to decay before an entire sequence of them can be recalled. Thus, serial recall accuracy for sequences of shorter words should tend to be higher than for sequences of longer words.

Phonological-similarity effect. Another important factor that influences serial recall accuracy is phonological similarity. Typically, sequences of phonologically similar words are remembered less well than sequences of dissimilar words (e.g., Conrad & Hull, 1964; Baddeley, 1966, 1968; Schweickert et al., 1990). The phonological-similarity effect supports the phonological-loop model's assumption that verbal information is represented in a modality-specific phonological store, rather than another type of storage system such as a visual or semantic buffer.

Under the phonological-loop model, there are at least two explanations of the phonological-similarity effect. One possibility is that during rehearsal and recall, phonological codes decay more quickly when they are similar to each other (Posner & Konick, 1966). Another possibility is that the codes decay at the same rate regardless of their similarity, but at the time of recall, the partially-degraded codes of similar items are more difficult to reconstruct or "redintegrate" (Hulme et al., 1997; Nairne, 1990; Schweickert, 1993). Either of these possibilities could yield an inverse relationship between recall accuracy and phonological similarity in the serial recall task.

Other factor effects. Many other factors may also affect performance in VWM tasks. These factors include unattended irrelevant speech (e.g., Colle & Welsh, 1976; Salamé & Baddeley, 1982), rehearsal strategy (e.g. Logie et al., 1996; Standing, Bond, Smith, & Isely, 1980; Standing & Curtis, 1989; Cowan et al., 1997), articulatory suppression (e.g., Levy, 1971; Murray, 1967, 1968; Baddeley et al., 1975), and focal brain damage (e.g., Shallice & Warrington, 1970; Basso, Spinnler, Vallar, & Zanobio, 1982; Baddeley & Wilson, 1985; Della Sala, Logie, Marchetti, & Wynn, 1991; D'Esposito & Postle, 1999). Research on the effects of these factors has provided considerable evidence concerning the structure of VWM. This evidence has, for the most part, supported the assumptions of the phonological-loop model. However, for present purposes, the effects of word length and phonological similarity are most directly relevant to the prevailing criticisms of the phonological-loop model.

Criticisms of the Phonological-Loop Model

Despite the success of the phonological-loop model, it has become the target of many critiques by researchers who claim that its assumptions are incorrect. Consequently, there are now probably more experiments whose results appear

[2] Both articulatory duration and phonological complexity can be used to specify a word's "length", making the term "word-length effect" ambiguous. To avoid this ambiguity, we use either "articulatory duration" or "phonological complexity" to specify the particular aspect of "word length" being considered at the moment. However, when the source of the effect is unknown or controversial, we use the more general term "word length".

to contradict Baddeley's (1986, 1992) conclusions about the word-length effect than there are experiments that support them. Some of the first such critical experiments were reported by Caplan, Rochon, and Waters (1992).

Caplan et al. (1992). Based on studies of neuropsychological patients, it has been hypothesized by Caplan et al. (1992) that the word-length effect stems from speech planning rather than overt or covert articulation. Their hypothesis assumes that speech-planning times are influenced by the phonological complexity of words, as indexed by numbers of phonemes and syllables.[3] Caplan et al. (1992) noted that most of the evidence for the word-length effect has come from cases in which articulatory duration and phonological complexity were confounded, and they dismissed the few cases wherein these confounds did not occur (e.g., Experiment IV of Baddeley et al., 1975).

To support their hypothesis and assumptions, Caplan et al. (1992) performed three experiments with various sets of words. Each experiment involved versions of the immediate serial recall task where stimuli were presented either auditorily or visually, and responses were made with a "picture-pointing" procedure. They measured articulatory durations for each word set by recording words spoken in isolation by a confederate speaker. Computer software was used to determine the acoustic onset and offset of each word.

Caplan et al.'s (1992) first experiment yielded typical effects of both articulatory duration and phonological similarity: sequences of one-syllable words were recalled more accurately than sequences of three-syllable words, and sequences of dissimilar words were recalled more accurately than sequences of similar words. However, this pattern failed to occur in their second and third experiments. During Caplan et al.'s (1992) second experiment, to-be-recalled sequences were constructed from two new word sets that had either long or short articulatory durations but approximately equal phonological complexity. Contrary to Experiment 1, results of Experiment 2 revealed that sequences from the "long" set were remembered *better* than sequences from the "short" set. Furthermore, during Caplan et al.'s (1992) third experiment, sequences were constructed from two new word sets that were either "difficult" or "easy" (with long and short articulatory durations, respectively), but approximately equal phonological complexity. Results of Experiment 3 revealed that sequences from the long word set were remembered as well as sequences from the short set, contrary to both Experiments 1 and 2.

As a result, over these three experiments, the data appear to seriously challenge the phonological-loop model. In them, a word-length effect consistent with this model occurred only in Experiment 1, where articulatory duration was confounded with phonological complexity. In the other two experiments, phonological complexity was equated across word sets, but the model's predictions were not substantiated. Consequently, Caplan et al. (1992) proposed that the word-length effect does not stem from articulatory rehearsal, but rather from rehearsal through speech planning based on sublexical phonemic representations whose duration varies with the numbers of phonemes and syllables per word. Yet they did not explain why the results from their Experiments 2 and 3 differed.

Baddeley and Andrade (1994). In reply, Baddeley and Andrade (1994) criticized these experiments by claiming there were two problems with Caplan et al.'s (1992) word sets: their articulatory durations had not been measured properly, and they were not equal in phonological similarity.

To support their claims, Baddeley and Andrade (1994) measured the articulatory durations for the words of the two sets in Caplan et al.'s (1992) Experiment 2.[4] These additional measurements were obtained by having participants rapidly repeat pairs of words ten times, either overtly or covertly. This procedure was similar to one used by Baddeley et al. (1975), but differed from the isolated-word measurement procedure used by Caplan et al. (1992). As a result, there were only slight differences between the articulatory durations for the two sets of words (Table 1). This led Baddeley and Andrade (1994) to argue that the obtained duration differences were too small (i.e., less than 15 ms per word) to justify rejecting the phonological-loop model.

Baddeley and Andrade (1994) also obtained participants' judgments about the phonological similarity for the word sets used by Caplan et al. (1992, Exp. 2). Here the "short" words were rated as more phonologically similar than the "long" words. This suggests that Caplan et al.'s (1992) results stemmed, at least in part, from systematic confounded differences between phonological similarity and articulatory duration.

Caplan and Waters (1994). Caplan and Waters (1994) responded to Baddeley and Andrade's (1994) criticisms by repeating Experiments 2 and 3 from Caplan et al. (1992). In this replication, articulatory durations were measured by recording participants' speech while they read the actual word lists used in the memory-span task. Results like those of Caplan et al.'s (1992) original Experiment 3 were not obtained here: sequences of longer words from this experiment were remembered less well than sequences of shorter words, contrary to their speech-planning hypothesis. However, results like those of Caplan et al.'s (1992) original Experiment 2 *were* obtained, contradicting the results of Baddeley and Andrade (1994). Here, "long" words from Experiment 2 had longer durations than "short" words (Table 1),

[3] Although the speech-planning hypothesis may seem plausible, it should be taken with some reservations. Not all empirical evidence is consistent with it. For example, Sternberg et al. (1978) had participants utter memorized sequences of either one-syllable or two-syllable words upon command. For sequences that contained one to five words each, the latencies of these utterances in response to a "go" signal were not affected reliably by the phonological complexity (i.e., syllable numerosity) of the individual words. Given that the onset latency of an utterance may manifest speech-planning times (Rosenbaum et al., 1987; cf. Sternberg et al., 1978), these results are inconsistent with Caplan et al.'s (1992) hypothesis and assumptions.

[4] Only Caplan et al.'s (1992) Experiment 2 was addressed because the results of their Experiment 1 were not controversial, and the results of their Experiment 3 were not replicated subsequently by Caplan and Waters (1994).

Table 1
Articulatory durations of “short” and “long” words measured in previous studies of VWM.

		Mean Articulatory Duration per Word (ms)	
Source of Measurement	Articulated Items	“Long” Words	“Short” Words
Caplan et al. (1992, Exp. 2)	Individual words*	720	546
Baddeley & Andrade (1994, Exp. 1)	Overt word pairs	360	353
Baddeley & Andrade (1994, Exp. 2)	Overt word pairs*	364	353
Baddeley & Andrade (1994, Exp. 2)	Covert word pairs	331	320
Caplan & Waters (1994)	Overt word lists*	525	473

Note: Values in Baddeley and Andrade (1994) were obtained by dividing reported mean total articulation times by 20 (the number of words in ten repetitions of word pairs). Values in Caplan and Waters (1994) were obtained by dividing mean total articulation times by 5.05, the average list length. Asterisks denote differences that were statistically reliable ($p < .05$).

and “long” words were also recalled better. Caplan and Waters (1994) also had participants rate the phonological similarity of their words. They found no reliable difference between the rated similarity for the word sets from Caplan et al.’s (1992) Experiment 2, which contradicts Baddeley and Andrade’s (1994) results.

Consequently, the evidence remains rather equivocal about whether rehearsal in VWM is based on speech-planning or on actual articulatory execution.[5] Caplan and colleagues have shown twice that the words in a “long” set had longer articulatory durations than the words in a “short” set, but that serial recall accuracy was better for sequences of the “long” words. They also have shown that words in these two “short” and “long” sets were about equal in rated phonological similarity. In contrast, Baddeley and Andrade (1994) found that words in these sets did not differ significantly with respect to articulatory duration, but did differ with respect to rated phonological similarity. However, their data were produced by a different group of participants who spoke British English, which limits their relevance to the results produced by North American participants in Caplan and colleagues’ studies (Caplan et al., 1992; Caplan & Waters, 1994).

Cowan et al. (1997). Other objections have also been raised about the phonological-loop model’s account of the word-length effect. In one case, Cowan et al. (1997) explored the extent to which articulatory duration and phonological complexity have separable effects on verbal serial recall accuracy. They conducted an experiment in which participants performed backwards recall for sequences of “simple” one-syllable words or “complex” two-syllable words. Participants were instructed to recall the words at either a rapid or slow pace, so that the articulatory durations for the “simple” and “complex” words would be approximately equated during the recall phase.

As a result of these manipulations, two separate effects emerged. Overall, recall accuracy decreased as articulatory duration during recall increased, consistent with the phonological-loop model. However, when the durations of the “simple” and “complex” words were approximately equated during recall, participants recalled the “complex” words more accurately than the “simple” words. This latter result cannot be easily explained by either the phonological-loop model or the speech-planning hypothesis of Caplan et al. (1992). Nevertheless, Cowan et al. (1997) concluded that the phonological-loop was essentially correct, but should be supplemented with an assumption that phonological complexity enhances recall accuracy.

Service (1998). A study by Service (1998) also addressed issues regarding memory span, articulatory duration, and phonological complexity. She exploited the fact that in Finnish, some words contain double-vowels that are identical to single vowels except for their articulatory durations. This enabled her to construct two sets of “simple” pseudowords that were identical in terms of their phonological complexity and phonological similarity, yet differed in their articulatory durations. She also constructed a third set of “complex” pseudowords that had longer articulatory durations and more phonemes per pseudoword.

Service (1998) found that sequences of “simple” double-vowel pseudowords were recalled as well as sequences of “simple” single-vowel pseudowords. Both of these types of sequences were recalled better than sequences of longer, “complex” pseudowords. To check whether her “simple” single-vowel and double-vowel pseudowords actually had different articulatory durations, she measured the times taken by participants to read lists of them. Double-vowel pseudowords took 31% longer than lists of single-vowel pseudowords.

These results, which reveal a dissociation between recall accuracy and articulatory duration, raise further doubts about the phonological-loop model. Consequently, Service

[5] The distinction between speech planning and articulatory execution is important, especially with respect to understanding clinical disorders (e.g., dysarthria) that putatively affect vocal articulation. A primary motivation for Caplan et al.’s (1992) speech-planning hypothesis was to explain certain aspects of serial recall from verbal WM in patients with dysarthria of speech, who show normal “word length” effects even though they have deficient overt articulation. Veridical theories of speech-based rehearsal should account for such effects. Consequently, we define “articulatory execution” to mean the operation of the articulatory motor-program processor not at a peripheral effector level but rather at a central level, where it produces phonological codes that maintain and update the contents of the phonological buffer (cf. Kieras et al., 1999; Monsell, 1987). This type of covert operation does not require physical movement of the speech articulators, so it may remain intact in patients who have dysarthria of speech. Furthermore, it can account for why these patients show a ”word length” effect, because implicit (covert) speech occurs at the same rate as overt speech (Landauer, 1962).

argued that phonological complexity influences serial recall accuracy, but articulatory duration and time-based decay do not. She suggested that Cowan et al.'s (1997) findings (i.e., that participants recalled longer words less well than shorter words) were artifactual, and that they probably occurred because participants had to perform backward recall at a fixed, unnatural rate.

Cowan, Nugent, Elliot, and Geer (2000). In response to Service (1998), Cowan, Nugent, Elliot, and Geer (2000) conducted a new study that replicated several conditions of the experiments by Cowan et al. (1997). In this replication, participants were trained to recall sequences of "simple" one-syllable words and sequences of "complex" two-syllable words at both rapid and slow paces. Unlike in Cowan et al. (1997), forward recall was required here. As a result, recall was more accurate under the rapid recall condition, but when recall rate was approximately equal, recall was more accurate for the "simple" words than for the "complex" words. These results again suggest that articulatory duration *does* affect serial recall accuracy. However, Cowan et al.'s (1997) conclusion about phonological complexity is called into question, because the complexity effect in the present experiment was opposite to what occurred previously.

Service (2000). Furthermore, Service (2000) suggested that Cowan, Nugent, Elliot, and Geer's (2000) study still had defects. For example, its artificial manipulation of recall rate may have induced atypical effects because participants divided their attention between recalling words and controlling their recall rate, which constitutes an unusual dual-task situation. Consequently, Service (2000) argued that her original procedure and theoretical inferences (Service, 1998) remain on solid ground. Subsequently, however, Cowan, Nugent, and Elliot (2000) have disagreed with her argument and continued to advocate their position over Service's (1998, 2000). Insofar as we can tell, this debate has not been satisfactorily resolved to date.

Lovatt et al. (2000). Meanwhile, research by Lovatt et al. (2000) has created yet more uncertainty about the status of the phonological-loop model. They conducted three experiments with two sets of two-syllable words per experiment (including a total of six different word sets overall). The mean articulatory durations of the words in each set were measured, and subjective ratings of the words' phonological similarity were also obtained. These measurements showed that for each experiment, the articulatory durations of the two word sets differed reliably, and the phonological similarity of the words in the sets was approximately equal. However, Lovatt et al.'s three experiments yielded an inconsistent pattern of results, contrary to the phonological-loop model.

Specifically, in their Experiment 1, Lovatt et al. (2000) measured articulatory duration with a list-reading technique, and found that sequences with longer durations were recalled *better* than sequences with shorter durations. In Experiment 2, using two more word sets, Lovatt et al. (2000) obtained three measurements of "word length" (isolated word durations, list-reading times, and articulation times for rapidly-repeated word pairs). Here they found that although articulatory duration of the words from the one set were consistently shorter than those from the other set, there was almost no difference in recall accuracy for the two sets. Finally, in Experiment 3, using a third pair of word sets, Lovatt et al. (2000) measured articulatory durations with the same three procedures of Experiment 2. Here they found that sequences of shorter-duration words were recalled more accurately than sequences of longer-duration words. Thus, on balance, there was no consistent relationship between articulatory duration and serial recall accuracy across the word sets of these three experiments. Although Lovatt et al. (2000) did not provide any principled theory to explain their results, they imply that the phonological-loop model is probably incorrect.

Interim Assessment

Although some data from the studies reviewed here appear to offer evidence against the phonological-loop model, together they do not suggest a clear alternative theory, because of inconsistencies and contradictions among them. Perhaps these inconsistencies indicate that the phonological-loop model is difficult to test and needs to be specified more precisely. This difficulty is highlighted by the fact that across past tests of the model, experimenters have measured articulatory duration with four distinct methods whose rationale remains unclear. None of these measurements may be adequate for testing the model's predictions about serial-recall accuracy. Likewise, in these same experiments, phonological similarity was either disregarded or measured only through informal subjective ratings. Thus, the true effects of this factor may have been frequently confounded with those of articulatory duration and/or phonological complexity.

In light of these considerations, previous results that appear to contradict the phonological-loop model may just reflect the mismeasurement of articulatory duration or phonological similarity. Consequently, to pursue these matters further, we have developed more theoretically-principled methods for measuring phonological similarity and articulatory duration, whereby predictions of the phonological-loop model and other alternative hypotheses can be more validly and precisely tested. In the following sections, we describe our new methods for measuring phonological similarity and articulatory duration. Then we report two experiments with these methods that evaluate the predictions of the phonological-loop model and the criticisms of its opponents.

PSIMETRICA: A Formal Method for Measuring Phonological Dissimilarity

The effect of phonological similarity on serial recall implies that coded information in VWM has a phonological representation (e.g., Baddeley, 1966, 1968). Yet the phonological-loop model makes no explicit claims about the details of this representation. In order to test this and other models of VWM and serial recall, we have developed *PSIMETRICA* (**P**honological **Si**milarity **Metric A**nalysis), a formal method for quantifying the phonological dissimilarity of paired words.

PSIMETRICA assumes that the phonological dissimilarity between words is a multi-dimensional vector of psychologically-relevant aspects of dissimilarity, which we call a *dissimilarity profile*. The individual dimensions of this profile may include a variety of quantities, such as an index of the extent to which two words rhyme, the similarity of their stress patterns, and the degree to which the words' syllable onsets match or mismatch. The use of a dissimilarity profile enables us to quantify various distinct dimensions of dissimilarity and determine which ones matter more or less in any given situation. In the present article, we focus on three such dimensions, related to syllable onsets, nuclei, and codas (see below).[6]

Application of PSIMETRICA involves four main steps. First, given a set of words, we specify the contents of the phonological representation for each individual word. Second, for each dimension of the dissimilarity profile, we identify and align the phonemes in a pair of words, so that they may be compared to assess how much they match or mismatch. Third, we quantify the degrees of match and mismatch for each dimension of the dissimilarity profile. Fourth, we average the results of this quantification over all of the possible word pairs of the word set. This yields the mean phonological dissimilarity profile for the word set. In the following subsections, these steps are described more fully.

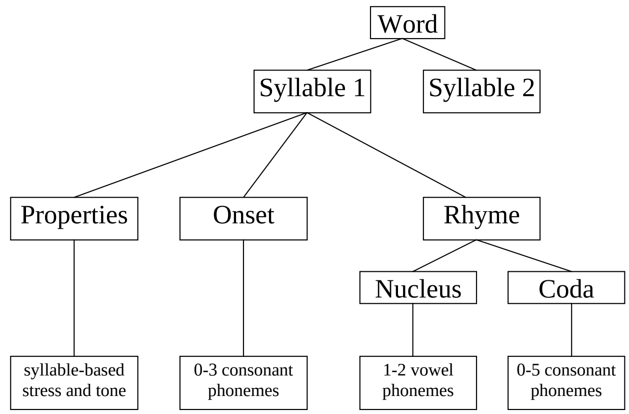
Phonemic Representation of a Word

Figure 1 shows the schematic hierarchical structure that we use for representing English words in PSIMETRICA. Here, a word is assumed to be composed of syllables. Each syllable has several properties (e.g., stress and tone) and contains three phoneme clusters: the *onset* (initial consonants), *nucleus* (vowels), and *coda* (final consonants). The nucleus and coda together are called the *rhyme*. The onset contains between zero and three consonant phonemes; the nucleus contains either one or two vowel phonemes; and the coda contains between zero and five consonant phonemes.

Decomposition of words into syllables and phoneme clusters. To decompose words into their constituent syllables and phoneme clusters, we use a standard linguistic procedure (O'Grady & Dobrovrolsky, 1992, pp. 76-82). In this procedure, the vowel nucleus of each syllable is identified first. Next, the longest string of consonants to the left of each syllable's nucleus that conforms to phonotactic and linguistic constraints is placed in the syllable's onset. Finally, all remaining consonants to the right of the syllable's nucleus form the coda. This decomposition is made for each syllable of a word, starting from the final syllable and working toward the first syllable.[7]

Decomposition of phonemes into phonological features. Next, for each phoneme in a word's representation, we identify its phonological features, using the system of Chomsky and Halle (1968). Under this system, each phoneme has thirteen or fewer binary features, which describe the states of certain vocal articulators when they produce the speech sound associated with a phoneme. A list of some common English phonemes and their constituent phonological features appear in Appendix A1. On this basis, the phonological representations for any word may be generated.

Figure 1. The hierarchical word structure used in PSIMETRICA for phonological dissimilarity measurement. Here a word is assumed to be composed of one or more syllables, which consist of properties and clusters of phonemes. Phonemes are associated with thirteen binary features that describe the articulatory states necessary to produce appropriate speech sounds (see Appendix A1).



Illustrative examples. The four steps that would be followed to obtain the phonological representations of the words "amount" and "placemats" appear in Table 2. For the word *placemats*, the syllable boundary is placed between the two constituent words, "place" and "mats", even though the phoneme cluster /sm/ is a legal onset (see Footnote 7). For "*amount*", the syllable boundary is placed between the phonemes /ə/ and /m/. When phonemes are assigned to their appropriate clusters, a null phoneme /Ø/ is used as a placeholder for clusters that contain no phonemes (such as the onset of the first syllable in *amount*).

Alignment Procedures For Individual Dissimilarity Dimensions

After the phonological representation of each word has been generated, PSIMETRICA's next step aligns these representations according to the constraints associated with each

[6] Other dimensions of dissimilarity might also be relevant, depending on the task that participants have to perform for the words to which PSIMETRICA is applied. For example, some of these other dimensions might include initial phoneme dissimilarity, stress dissimilarity, inter-syllable or inter-cluster dissimilarity, syllable or phoneme numerosity dissimilarity, phoneme distribution dissimilarity, and so forth. We have not yet exhaustively investigated the effects of all these dimensions in serial recall and other tasks.

[7] Multi-syllabic words are frequently composites of smaller words or morphemes (phoneme sequences that carry meaning), and the syllable-parsing technique must be sensitive to these semantic constraints, so that syllable boundaries are placed between morphemes. For example, the word *passport* should be divided into the syllables *pass* and *port* (maintaining each constituent word), rather than *pa* and *sport*. There may also be other cases in which alternate syllable boundaries might be preferred, such as those discussed by Treiman and Zukowski (1990). Nevertheless, a large majority of words can be decomposed and represented as described in the above text.

Table 2
Description and illustration of steps in PSIMETRICA for generating phonemic representations of the words "placemats" and "amount".

Step	Procedure	Representation: Placemats	Representation: Amount
1	Determine constituent phonemes	/plesmæ ts/	/əmaʊnt/
2	Decompose word into syllables	/(ples) (mæ ts)/	/(ə) (maʊnt)/
3	Decompose syllables into phoneme clusters	/((pl)(e)(s)) ((m)(æ)(ts))/	/((Ø)(ə)(Ø)) ((m)(aʊ)(nt))/
4	Decompose phonemes into features	Feature-level decomposition of phonemes is presented found in Appendix A1.	

dimension of the dissimilarity profile. These constraints ensure that corresponding phonemes of the paired words are properly compared with each other in terms of whatever psychological hypotheses are embodied by a given dimension. This step is essential for making the dissimilarity profiles be valid indicators of how two words are related phonologically. If the phonemes in their component syllables are not properly aligned, then relevant similarities and differences between them cannot be taken fully into account. Also, because the phonological structure of the words is complex, procedures for aligning the words properly must be somewhat intricate.

In the next sections, we describe the alignment procedures for measuring three dimensions of the dissimilarity profile: the onset, nucleus, and coda dissimilarity. These dimensions are not exhaustive—other interesting ones (e.g., Footnote 6) might also be considered. Furthermore, they are not unique—different choices could be made about many of their definitional details; whether these choices matter for our purposes is primarily an empirical question. Nevertheless, the three dimensions of onset, nucleus, and coda dissimilarity that we have quantified here provide a reasonably complete basis to measure phonological dissimilarity between syllables and to predict serial recall accuracy precisely across different word sets.

Onset alignment. Onset phoneme clusters may contain between zero and three consonant phonemes. To align the onset clusters for a pair of syllables, we use a template related to structures discussed by Fudge (1969) and Hartley and Houghton (1996). This template has three positions or "slots", each of which holds either an actual or null phoneme. If a syllable onset contains the phoneme /s/, /ʃ/ (pronounced *sh*), /z/, or /ʒ/ (pronounced *zh*), then it is put in the first template position. This is done because these phonemes are highly similar to each other and none of them can follow any other phonemes in the onsets of English syllables. Next, if the onset contains the phoneme /r/, /l/, /y/, or /w/, then it is placed in the third template position. This is done because each of these phonemes can follow phonemes other than those placed in the first template position, and must immediately precede the vowel of a syllable. Any other onset phonemes are put in the middle position of the onset template. Remaining empty positions of the onset template are filled with null consonants (/Ø/). This yields a uniform schema for representing the onset of any English syllable, and allows corresponding onset phonemes to be aligned and compared directly.

Nucleus alignment. Nucleus phoneme clusters may contain either one or two vowel phonemes. When two nuclei with the same numbers of vowels are compared, we align them vowel-by-vowel. When a nucleus with one vowel is compared to a nucleus with two vowels, we double the vowel of the smaller nucleus, so that the doublets are aligned with the vowels in the larger nucleus.

Coda alignment. Coda phoneme clusters may contain between zero and five consonant phonemes, including two /s/s. Because of this potential complexity, it is difficult to use a single template with a fixed number of positions for aligning the phonemes of two codas. If such a template were used, the most appropriate template position for an /s/ would depend upon the coda with which it is being compared.

For example, consider the codas of the following four words: *mast* (/mæst/), *mass* (/mæs/), *mats* (/mæts/), and *masts* (/mæsts/). A template must accommodate *masts*, and so must have two slots for /s/, before and after the slot for /t/. Clearly, the placement of the /s/ of *mass* depends upon whether *mass* is being compared to *mast* or *mats*. When *mass* is compared to *mast*, the /s/ should be placed in the first /s/ slot, but when *mass* is compared to *mats*, the /s/ should be placed in the second /s/ slot to embody the similarities between these words.

To deal with these complexities, our alignment procedure for codas takes a different approach than for onset clusters. When we align two coda clusters, we construct multiple candidate representations for each coda by adding null phonemes to the beginning and end of each cluster. We then find which two candidate representations are most similar, using a metric described later. This procedure yields representations that have the /s/ in *mass* aligned with both the /s/ in *mast* and the /s/ in *mats*.

Illustrative example. Our procedures for constructing the appropriate alignments between paired words for the onset, nucleus, and coda dimensions of the dissimilarity profile may be illustrated with the words *amount* and *placemats*. Table 3 shows the final alignments that would result for this pair of words. For the onset dimension, the onset phonemes are placed into templates. Here, the onset of the first syllable in *placemats* is aligned with three null phonemes in *amount*. The onsets of the second syllables are both represented as /Ø mØ/, and aligned accordingly. For the nuclei of the first pair of syllables, the phonemes are aligned directly because each word contains a single vowel, /ə/ and /e/. In contrast, for the nuclei of the second pair of syllables, the /æ/ of *placemats* is doubled, and then aligned with both the /a/ and the /ʊ/ of *amount*. For the codas, the most similar pair of all possible comparisons is chosen. Because the coda of the first syllable in *amount* has no phonemes, a null phoneme is aligned with the /s/ in *placemats*. Null phonemes are added to the codas of the second syllables in each word, so that the /t/s are aligned directly, while the /s/ of *placemats* and the /n/ of *amount* are each aligned with a null phoneme.

This process yields an alignment of the phonemes in *placemats* and *amount* for each dimension of our dissimilarity profile, across both syllables of the two words. None of the word sets used in our experiments contain words with differing numbers of syllables, and so this technique is sufficient to obtain dissimilarity profiles for these words. If paired words that have different numbers of syllables are compared, additional alignment procedures would be needed for each dimension of the dissimilarity profile.

Obtaining a Dissimilarity Profile

After the phonemic representations for a pair of words have been aligned, PSIMETRICA's next step is to measure the phonological dissimilarity on each dimension of the dissimilarity profile. For present purposes, the measurement of phonological dissimilarity proceeds in substeps. It begins with calculations at the level of corresponding phonemes in the aligned representations of two words. Next the mean dissimilarity between corresponding phoneme clusters and syllables is calculated. Then we calculate the overall mean dissimilarity between the words. Each of these substeps is described below.

Measurement of phonological dissimilarity between phonemes. In our phonemic representations of words, each phoneme has a unique combination of binary phonological features. On the basis of these features, we measure the phonological dissimilarity between two corresponding phonemes in a pair of words by counting the number of mismatching features that the phonemes have and dividing it by the total number of relevant features per phoneme.[8] This is analogous to measuring the "distance" between phonemes. Consequently, two identical phonemes would have a numerical dissimilarity of 0, whereas two extremely different phonemes would have a dissimilarity value closer to 1.

There are also other special cases that must be accommodated in measuring phonological dissimilarity at the phonemic level. As mentioned already, our procedures for aligning the phonemic representations of paired words frequently require consonant phonemes in syllable onset and coda clusters to be compared with corresponding null phonemes. Regarding these cases, we give them appropriate default dissimilarity values, depending on the phonemes' serial positions. For each distinct position in a syllable's onset, we calculate the mean dissimilarity between (a) the consonant phonemes that can legally fill this position and (b) the full set of consonant phonemes. This yields default dissimilarity values of 0.356, 0.37, and 0.53 for the three possible syllable-onset positions; comparisons between consonant and null phonemes in the first, second, and third serial positions of syllable onsets are all given these values, respectively. For comparisons between consonant and null phonemes in syllable codas, our dissimilarity-measurement method involves an analogous calculation; based on it, these comparisons are each given default dissimilarity values of 0.37. The dissimilarity between a pair of null phonemes is given a value of 0.

Calculation of the dissimilarity profile for a pair of words. After dissimilarity values have been calculated for each pair of phonemes identified by and aligned for every dimension of the dissimilarity profile, the mean values on these dimensions are obtained. This is first done for each pair of syllables, by finding the average dissimilarity of the phoneme pairs of each dissimilarity dimension. Then, the mean value across syllables for each dimension of the dissimilarity profile is calculated, producing the overall phonological dissimilarity profile for that word pair.

Calculation of the dissimilarity profile for a set of words. For an entire set of words, such as might be used to construct word sequences in the serial recall task, phonological dissimilarity is measured by averaging across the dissimilarity profiles of all possible word pairs from the set. In an n-word set, each word can be paired with $n-1$ other words, yielding $n(n-1)/2$ pairs. Averaging the dissimilarity values on each dimension for these pairs gives a mean dissimilarity profile for the word set as a whole.

Illustrative example. To illustrate how PSIMETRICA may be applied, let us again consider the words *placemats* and *amount* (Table 4). For these words, the syllables, phoneme clusters, and phonemes must first be aligned in the manner explained previously (cf. Table 3). Next we compare the aligned phonemes in each pair, and give them a dissimilarity value. When the phonemes of a pair are both non-null, this value equals the proportion of non-null features that differ between the phonemes. These proportions appear as fractions in Table 4. For example, the phonemes /e/ and /ə/ have nine relevant features, and only two of these features are different. So, when /e/ and /ə/ are compared in the first syllables of *placemats* and *amount*, they are given a dissimilarity value of $2/9$. Pairs of phonemes that contain one actual phoneme and one null phoneme are given a default dissimilarity value that depends on the position of the phoneme pair. These default values appear as decimal numbers in Table 4. For example, in the codas of the first syllables of *placemats* and *amount*, the phonemes /s/ and /Ø/ are paired with each other, and because one of them is null, this phoneme pair is given a dissimilarity of 0.37. Pairs of null phonemes have no phonological features, and dissimilarity values are not given to them.

After dissimilarity values have been given to all of the aligned phoneme pairs of *placemats* and *amount*, a single mean value for each phoneme cluster is calculated by averaging across the dissimilarity values of the phoneme pairs within the cluster. Dissimilarity values for pairs of null phonemes are not included in this average. These mean dis-

[8] Most phonemes do not have values for all thirteen features, because some features are specific to consonants and others to vowels. Consequently, the number of feature mismatches is divided by the number of features for which at least one phoneme has a value.

Table 3
Alignment used in PSIMETRICA for phonemic representations of the words "placemats" and "amount".

		First-Syllable Alignment			Second-Syllable Alignment		
Word	Phonemic Representation	Onset	Nucleus	Coda	Onset	Nucleus	Coda
placemats	/((pl)(e)(s)) ((m)(æ)(ts))/	/Ø p l/	/e/	/s/	/Ø mØ/	/æ æ/	/Ø t s/
amount	/((Ø)(ə)(Ø)) ((m)(aʊ)(nt))/	/Ø Ø Ø/	/ə/	/Ø/	/Ø mØ/	/a ʊ/	/n t Ø/

Table 4
Illustration of phonological dissimilarity measurement with PSIMETRICA for the words "placemats" and "amount".

	Word		Phonological Dissimilarity	Mean Dissimilarity
Phoneme Clusters	Placemats	Amount	of Phonemes	of Clusters
First Onset	/Ø p l/	/Ø Ø Ø/	n.a., 0.37, 0.53	0.45
First Nucleus	/e/	/ə/	2/9	0.22
First Coda	/s/	/Ø/	0.37	0.37
Second Onset	/Ø mØ/	/Ø mØ/	n.a., 0/11, n.a.	0.00
Second Nucleus	/æ æ/	/a ʊ/	1/9, 4/9	0.28
Second Coda	/Ø t s/	/n t Ø/	0.37, 0/11, 0.37	0.25

similarity values for the clusters of *placemats* and *amount* appear in the right-most column of Table 4. For example, the mean dissimilarity value of the second nuclei in *placemats* and *amount* is 0.28. Finally, a dissimilarity profile is obtained by averaging the dissimilarity values of each phoneme cluster across syllables. For *placemats* and *amount*, this profile is 0.225, 0.25, and 0.31 (respectively, the mean dissimilarity values of the onsets, nuclei, and codas). By averaging such profiles from all possible pairs of words in a set, we obtain the word sets' overall mean dissimilarity profile.

Relation to Previous Treatments of Phonological Coding and Similarity

PSIMETRICA has some significant advantages over previous treatments of phonological similarity in research on VWM and serial recall (cf. Baddeley, 1986; Baddeley & Andrade, 1994; Caplan & Waters, 1994; Lovatt et al., 2000). Unlike measures of phonological similarity or dissimilarity obtained through subjective ratings, the ones obtained through PSIMETRICA have a formal theoretical basis; they are precise, reproducible, and sensitive to fundamental aspects of representation that presumably mediate phonological-dissimilarity effects on performance of VWM tasks. PSIMETRICA assumes that phonological representations of words and their phonological features encode basic information relevant to the vocal production of speech sounds (Chomsky & Halle, 1968). This assumption may be especially apt because according to the memory literature (e.g., Gupta & MacWhinney, 1995; Murdock, 1974; Wickelgren, 1965, 1966), verbal working memory typically has an articulatory rather than auditory (e.g., acoustic waveform or formant) nature.

In other respects, PSIMETRICA resembles methods used by Vitz and Winkler (1973) for measuring the judged dissimilarity of words and by Nairne (1990) for modeling immediate memory. Like them, we obtain numerical values between 0 and 1 for each pair of elements being compared, with larger values corresponding to greater dissimilarity. Also related to PSIMETRICA is a method used by Frisch (1997), who measured similarity in terms of the number of "natural classes" that the phonological features of two phonemes share.

Nevertheless, PSIMETRICA makes some new contributions to the measurement and analysis of phonological dissimilarity that go significantly beyond previous attempts. To be specific, because different aspects of a word's phonetic structure may be important under different conditions, PSIMETRICA treats phonological dissimilarity as a multidimensional quantity. Also, PSIMETRICA's use of phonological features is more precise than in Vitz and Winkler (1973), but our hierarchical representation allows the dissimilarity of entire words to be assessed, and is consequently broader than Frisch's (1997) similarity metric, which compared only individual phonemes.

A Method for Measuring Articulatory Duration

Of course, just measuring phonological dissimilarity will not enable the phonological-loop model to be tested against other models of VWM. A valid method for measuring the articulatory durations of to-be-recalled items is also required. Such a method must be sensitive to the articulatory durations that are claimed by the phonological-loop model to cause the "word-length" effect. According to this model, the relevant durations are the actual times taken by participants to rapidly rehearse sequences of words. If the durations of articulatory rehearsal are not measured properly, then an experiment may yield misleading or irrelevant measurements for testing the phonological-loop model. Yet few (if any) previous experiments that attempted to test this model have measured articulatory durations in an adequate manner.

Given these considerations, in the next section, we review various methods that have been used previously for measuring articulatory durations, and we outline exactly what their deficiencies are. Next, we establish a set of relevant criteria that a valid method for measuring articulatory duration should satisfy. Then we describe such a method that we have developed and used in our present experiments.

Previous Methods for Measuring Articulatory Duration

In previous experiments on VWM and serial recall, articulatory durations have been measured with a variety of methods. These methods fall into four general categories: (1) measurement of durations for isolated words; (2) measurement of durations for words in short repeated constant-length sequences; (3) measurement of durations for words read from lists; and (4) measurement of durations for words produced during final serial recall. Each of these methods has its own strengths and weaknesses, but because of their weaknesses, none of them are sufficient for present purposes.

Measurement of articulatory durations for isolated words. Perhaps the simplest way to measure the articulatory duration for a word is to have a participant or confederate pronounce the word aloud in isolation and record the time taken to do so. This method was used in several of Baddeley et al.'s (1975) original experiments. Also, it was the sole method used by Caplan et al. (1992), and it was a method used by Lovatt et al. (2000) to facilitate the initial selection of stimuli.

Not all aspects of isolated-word duration measurement are irrelevant; results from this method do provide some indication of whether one word is longer than another. However, measuring articulatory durations in this way also has serious deficiencies for testing the phonological-loop model. Because this method does not measure the durations of rapidly articulated sequences of words, which are the *sine qua non* of overt and covert articulatory rehearsal, the obtained measurements may be inadequate to test this model's predictions about duration effects on serial recall accuracy and memory span. For example, they may fail to take coarticulation effects on the duration of word sequences properly into account.

Measurement of articulatory durations for words in short repeated constant-length sequences. Another common method for measuring articulatory duration involves having a participant rapidly repeat two or three words several (e.g., ten) times and recording the total time of their utterance. This method was used by Baddeley et al. (1975) in their Experiment V, by Baddeley and Andrade (1994) under both overt and covert articulation conditions, and by Lovatt et al. (2000). It is an improvement over isolated articulation, but it also has serious weaknesses, because it fails to measure aspects of articulatory duration relevant to the phonological-loop model.

For example, in the immediate serial recall task, participants often have to rehearse and recall sequences that contain between five and seven words. The mental processes involved in planning and executing the articulation of such sequences certainly differ from the processes involved in articulating word pairs repeatedly. Additionally, some factors that influence articulatory duration, such as interactions between "tongue twister" words, might only occur when specific combinations of words are articulated, and these combinations may not be present when limited sets of word pairs or triplets are involved. Consequently, experiments that use this method for measuring articulatory durations cannot adequately test the phonological-loop model.

Measurement of articulatory durations for words read from lists. A third method that has been used to measure articulatory duration is list reading. For this method, typical lists of words from the memory-span task are presented visually, and participants are instructed to read the lists aloud. The duration of their articulation is then measured. This method was used by Baddeley et al. (1975), Caplan and Waters (1994), Service (1998), and Lovatt et al. (2000).

One notable advantage of this method is that it involves word sequences like those used in the serial-recall task. Consequently, the obtained articulatory durations are more relevant than some other measures. Nevertheless, this method has some serious deficiencies too. For example, the lists of words are *read*, which is inadequate for testing the phonological-loop model because during the immediate serial-recall task, words must be recalled from memory. Additionally, list reading probably encourages participants to enunciate the words clearly and deliberately, rather than to articulate them rapidly (as is presumably done during rapid memory-based rehearsal). Consequently, experiments that use the list-reading method may not test the phonological-loop model adequately.

Measurement of articulatory durations for words during final serial recall. A fourth method for measuring articulatory duration has been to record how long a participant takes to produce each sequence of words during final serial recall. This method was used by Service (1998) and Cowan et al. (1997), as well as in other experiments that are less directly relevant here (e.g., Schweickert et al., 1990; Cowan, 1992; Dosher & Ma, 1998; Cowan et al., 1998). There are two major strengths of final-recall duration measurement: it measures articulatory duration *in vivo*, while the participant is actually performing the serial recall task, and it measures an aspect of articulatory duration that the phonological-loop deems important—recall duration.

However, this method also has some weaknesses. Most importantly, articulatory durations measured during final serial recall may not closely approximate those that occur during rehearsal, because recall might involve more memory search and reconstruction than rehearsal does. Given the phonological-loop model assumes that articulatory duration primarily affects rehearsal, this measurement method may not test the model adequately either.

A New Method for Measuring Articulatory Duration

The preceding critique suggests that a new method for measuring articulatory durations is needed in order to properly test the phonological-loop model. Specifically, to measure the mean duration per word for a set of words, multiple (e.g., 15 to 20) sequences of several lengths (e.g., 2 through 5 words per sequence) should be constructed from each word set. On each duration-measurement trial, one sequence of words should be presented, and the participant should be allowed to study it until he or she has committed it to memory. When the participant is ready, the sequence should be articulated from memory, as presumably happens during motivated rehearsal. The sequence should be articulated at least twice, as may happen during iterative rehearsal, and the total time taken by the participant to articulate the words overtly should be measured.[9] Participants should be encouraged to articulate the words rapidly and accurately. A duration-measurement trial should be repeated if a speech or memory error occurs during it. Finally, a mathematical analysis of total articulation times should be performed, and parameters of articulatory duration that are constant, linear, and curvilinear with respect to sequence length should be estimated.[10] As a result, mean articulatory durations that are most relevant for testing the phonological-loop model may be estimated. The procedure and formulae we use for this estimation appear in Appendix A2.

Overview of Experiments

The present article reports two experiments that used our new methods of measuring articulatory duration and phonological dissimilarity to test the phonological-loop model versus other models of VWM. During both experiments, participants performed a standard version of the verbal serial recall task (Baddeley, 1986). For each experiment, sequences of words were constructed from word sets that differed in terms of their mean articulatory durations, phonological dissimilarity, and phonological complexity.

In Experiment 1, we show that mean articulatory durations and phonological dissimilarity account for an extremely large proportion of variance in memory-span data across six word sets. The results of Experiment 1 also reveal that phonological complexity per se may have no reliable effects on memory spans over and above those attributable to mean articulatory durations and phonological dissimilarity. Experiment 2 extends these results to include three additional sets of words. We show that mean memory spans for these three word sets can be accurately predicted a priori based on parameter estimates from the results of Experiment 1. Taken overall, our findings are consistent with the phonological-loop model and constitute strong evidence for the importance of phonological coding, repetitive time-consuming articulatory rehearsal, and trace decay in the serial recall task.

Experiment 1

The purpose of Experiment 1 was to replicate and extend two experiments by Caplan et al. (1992, Exps. 2 and 3). We believe that their experiments warrant closer inspection for three reasons. First, in Caplan et al. (1992), Baddeley and Andrade (1994), and Caplan and Waters (1994), articulatory durations were measured five times (Table 1), yet none of these measurements were made under conditions that (according to the phonological-loop model) mediate the articulatory-duration effect on memory span. Second, in one experiment, Caplan et al. (1992, Exp. 2) found that the articulatory-duration effect was reversed; sequences of “long” words were recalled more accurately than sequences of “short” words. Although this reversal contradicts the phonological-loop model, Caplan et al.’s (1992) speech-planning hypothesis cannot explain it either, because the two word sets were equally complex and so should have yielded equal memory spans. Thus, this finding needs to be evaluated further. Third, Caplan and Waters (1994) attempted but failed to replicate other results from Caplan et al. (1992, Exp. 3) that had shown no difference between serial-recall accuracy when phonological complexity was equated but mean articulatory durations differed across word sets. Additional testing might provide more information about why this latter failure happened. To achieve these objectives, we measured phonological dissimilarities and articulatory durations of words with our new methods.

Method

Participants. The participants were six undergraduate students at the University of Michigan with normal perceptual, cognitive, and motor abilities. They were paid for their participation.

Apparatus. The experiment was conducted with a Pentium-class computer using special-purpose software. Auditory stimuli were presented via headphones, and visual stimuli were presented on the computer’s SVGA display. Performance was monitored by an experimenter who sat next to the participant and interacted with the computer in order to record the participant’s responses.

Stimuli. Six different sets of words were used for constructing word sequences to test verbal serial recall (Appendix A3). Sets 1 and 2 contained the “long” and “short” two-syllable words from Experiment 2 of Caplan et al. (1992). Sets 3 and 4 contained the “difficult” and “easy” one-syllable words from their Experiment 3. Sets 5 and 6 contained new “short” and “long” words that we selected to be low and approximately equal in concreteness and written frequency (Coltheart, 1981).[11] We also attempted to maximize intra-set dissimilarity in Sets 5 and 6 by choosing words that began with distinct letters and sounds. Table 5 shows relevant characteristics of the words in each of the six sets used here.

Measurement of phonological dissimilarity. For Experiment 1, preliminary data analysis revealed that only the dis-

[9] The use of overt articulation is justified for at least three reasons: First, maximum overt and covert articulation rates are essentially the same (Landauer, 1962, Baddeley & Andrade, 1994). Second, for overt articulation, the experimenter can monitor participants’ speech errors, thereby enabling durations from aberrant trials to be assessed and discarded if necessary. Third, the moments at which articulation starts and stops can be identified more accurately when participants speak overtly.

[10] Curvilinear parameters may be needed in measuring mean articulatory durations because Sternberg et al. (1978) found that production times for sequences of rapidly recalled words are a concave-upward function of sequence length.

[11] Both concreteness (ranging between 100 and 700) and written frequency (ranging between 0 and 70,000 based on the corpus of Kuçera and Francis, 1967) were obtained from the MRC Psycholinguistic Database Machine Usable Dictionary: Version 2.00 (Wilson, 1987).

Table 5 *Characteristics of the six word sets in Experiment 1.*

Variable	Statistic	Word Set 1	2	3	4	5	6
Syllable numerosity	mean	2	2	1	1	1	3
Phoneme numerosity	mean	6.00	5.13	3.00	3.00	3.25	7.50
	s.d.	0.7	0.3	0	0	0.6	1.3
Phonological dissimilarity	mean	0.401	0.344	0.266	0.336	0.399	0.363
	s.d.	0.14	0.08	0.14	0.18	0.11	0.08
Written frequency	mean	16	16	30	45	18	17
	s.d.	16	16	27	67	16	23
Concreteness	mean	-	-	-	-	304	296
	s.d.	-	-	-	-	40	26

Note: Caplan et al. (1992) reported the mean and standard deviations for the written frequency of Sets 1 through 4 as shown above. Corresponding statistics regarding the concreteness of these words were not reported.

similarity of syllable onsets within each word set appeared to affect performance—the dissimilarity of syllable rhymes and stress did not.[12] Because of this, we used PSIMETRICA's measure of onset dissimilarity as the present index of phonological dissimilarity for each pair of words. The phonological dissimilarity for an entire set of words was obtained by calculating the average pair-wise onset dissimilarity for the words in the set.

It should be emphasized that none of the six word sets in Experiment 1 were selected to contain highly similar words. Thus, our present situation differs from previous experiments that have investigated the effects of phonological similarity on serial recall accuracy (e.g., Conrad & Hull, 1964; Baddeley, 1964, 1966; Schweickert et al., 1990), where results were compared for sets of highly "similar" and "dissimilar" words. In those experiments, variations of phonological dissimilarity were generally large and obvious, whereas in our Experiment 1, variations of phonological dissimilarity were subtle and perhaps undetectable through casual inspection.

Design. The participants were tested individually with three procedures: verbal serial recall, articulatory-duration measurement for isolated words, and articulatory-duration measurement for words in memorized sequences. Testing occurred during two sessions on separate days. During the first session, articulatory durations of words in memorized sequences were measured for word Sets 1, 2, 5, and 6. During the second session, articulatory durations for isolated words in each of the six word sets were measured, followed by measurement of articulatory durations for words in memorized sequences from Sets 3 and 4. After the articulatory-duration measurements in the second session, participants performed a verbal serial recall task for each of the six sets of words. We randomized the order in which the word sets were used for each participant.

Measurement of articulatory durations for isolated words. To measure articulatory durations for isolated words, we had participants separately read each word from a set, speaking clearly and pausing before each word. The participants' utterances were recorded digitally. Their articulatory durations were measured later by identifying the beginning and end of the waveform for each word, using standard computer software.

Measurement of articulatory durations for words in memorized sequences. To measure the articulatory durations for words in memorized sequences, we used the method introduced earlier. With this method, a sequence of words was displayed on the video screen in a horizontal row at the start of each trial. The word sequence remained on the screen until the participant verbally indicated that he or she was ready to begin. Then, three 100-ms tones were presented at 500 ms intervals. Immediately after the third tone, the word sequence disappeared from the video screen and a timer began. As instructed, the participant then clearly articulated the sequence of words twice from memory at a moderately rapid pace, as if he or she was rehearsing the sequence for later serial recall. Trials on which the participant committed a speech error or failed to recall all of the words in a sequence were discarded and repeated for this sequence. When the participant completed the final word of a sequence, the experimenter pressed a key that stopped the digital timer. Then a new sequence of words was presented and the process was repeated.[13]

Articulatory-duration measurements for each set were made in a single trial block. Before each block began, 13 three-word sequences, 10 four-word sequences, and 8 five-word sequences were constructed randomly with uniform probability so that no word occurred more than once per sequence, and all words occurred with approximately equal frequency throughout the block. Because Set 6 contained words that were considerably longer than those in the other sets, and because previous pilot participants had difficulty accurately repeating sequences of five words constructed from Set 6,

[12] There are various possible reasons why the contributions of syllable nuclei and codas to phonological-dissimilarity effects on performance were relatively unimportant here. For example, during rapid rehearsal, participants may have articulated mainly the onsets of syllables, which could lead them to be most important. However, different experimental procedures might induce other parts of the syllable to have more influence, so our present emphasis of syllable-onset dissimilarity might not generalize to all situations.

[13] Subsequent supplementary data analysis revealed that the measures of mean articulatory duration obtained through this timing procedure were virtually as accurate as ones based on evaluation of digitized speech waveforms (see Appendix A2 and Footnote 17).

the blocks in which articulatory durations were measured for Sets 5 and 6 included 20 three-word sequences and 15 four-word sequences but no five-word sequences.

Measurement of serial memory spans. To measure participants' memory spans for each word set, we presented sequences of words constructed from each set in six serial-recall trial blocks (one block per word set). The words of each sequence were constructed by randomly choosing words from a set without replacement.

At the beginning of each serial recall trial, the participant was informed about the length of the impending to-be-recalled word sequence. The participant then listened to a sequence of words presented over headphones, with 1.5-second intervals between onsets. A brief tone was presented 1.5 seconds after the final word, signalling that serial recall should begin. Participants recalled the words aloud, and received credit for being correct only if an entire sequence was reproduced in its original serial order. After recall was complete, the computer provided feedback about whether the sequence was recalled correctly.

A staircase testing procedure, similar to the one of Schweickert et al. (1990), was used to select the length of each sequence in our task. Each block of trials began with a sequence of four words. If a sequence was recalled correctly, then the next presented sequence was one word longer; if a sequence was not recalled correctly, the next sequence was one word shorter. For each word set, this continued for a total of 16 sequences.

Preliminary Data Analysis

Before evaluating the results of Experiment 1 in detail, we conducted preliminary data analyses to obtain estimates of memory spans and articulatory durations for words in memorized sequences. These analyses yielded an estimate of memory span and mean articulatory duration for each participant for each of the experiment's six word sets.

Estimation of memory spans. Memory span was estimated to be the fractional sequence length at which a participant had a 50% chance of recalling a sequence of words correctly. This estimation involved fitting a generalized binomial linear-regression model with a logistic link function (McCullagh & Nelder, 1989) for each participant/word-set combination. Memory spans—analogous to LD50, the "lethal dose for 50% of the cases"—were calculated from the best-fitting parameters of these functions.

Estimation of mean articulatory durations for words in memorized sequences. We estimated the mean articulatory durations for words in memorized sequences by analyzing the total times taken to repeat particular sequences twice during the duration-measurement trials. Details of this analysis appear in Appendix A2. For each word set, the analysis involved fitting concave-upward functions of sequence length (Appendix A2, Equation 2) to the mean total articulation times produced by each participant. As discussed subsequently, the fits of the time functions were typically excellent (see *Results*). A quantitative combination of parameters associated with these functions (Appendix A2, Equation 4) provided an estimate of the overall mean articulatory duration per word for each word set (Table 6).

Also, in conjunction with this analysis, outliers among the total articulation times from the duration-measurement trials were deleted. We did this by excluding any such time that differed by more than 2.5 standard-deviation units from what the fitted concave-upward functions of sequence length (Appendix A2, Equation 2) would have predicted. The outliers occurred about equally infrequently across the different participants, word sets, and sequence lengths; they comprised less than 1.6% of the total data set.

Results

Table 6 summarizes the mean values of articulatory duration and memory span from the six sets of words used in Experiment 1. Table 5 shows the corresponding mean values of phonological dissimilarity and phonological complexity for these word sets. We next discuss each of these variables and how they are related.

Phonological dissimilarity. According to our PSIMETRICA method for measuring phonological dissimilarity, the words in Set 1 were considerably more dissimilar from each other than were the words in Set 2 (Table 5). This difference agrees with the subjective ratings of Baddeley and Andrade's (1994, Exp. 3) participants. However, the present dissimilarity measurements for word Sets 1 and 2 disagree with Caplan and Waters' (1994) participants, who rated the two sets as having nearly identical levels of phonological dissimilarity.

Such disagreement also occurred for Sets 3 and 4. Measurements based on PSIMETRICA suggest that the words in Set 4 were considerably more dissimilar from each other than were the words in Set 3. However, participants in Caplan and Waters' (1994) study rated these two sets as having virtually identical levels of dissimilarity. Consequently, it appears that Caplan and Waters' (1994) procedures were not sensitive enough to detect measurable differences in phonological dissimilarity. Yet it remains to be seen whether these differences affected performance in the serial recall task.

Articulatory durations for isolated words. Across the six sets of words in Experiment 1, there were large differences between the mean durations of words when participants articulated each word in isolation. A within-participants analysis of variance revealed that differences between these durations were highly reliable; $F(5,25) = 55.0$, $p < .001$. As shown in Table 6, we found that Set 1 ("long" words) yielded reliably longer mean articulatory durations than did Set 2 ("short" words); mean difference $= 130 \pm 20$ ms, $t(25) = 6.50$, $p < .001$. This is analogous to what Caplan et al. (1992, Exp. 2) found with these two word sets (mean difference $= 174$ ms). Furthermore, we found that Set 3 ("difficult" words) yielded reliably longer mean articulatory durations than did Set 4 ("easy" words); mean difference $= 106 \pm 20$ ms, $t(25) = 5.30$, $p < .001$. This is analogous to what Caplan et al. (1992, Exp. 3) found with these two word sets (mean difference = 96 ms).

Articulatory durations for words in memorized sequences. The total articulation times that participants took on aver-

Table 6 *Results from Experiment 1.*

Variable	Word Set 1	2	3	4	5	6
Duration of isolated words (ms)	562	432	509	403	418	672
Duration of words in memorized sequences (ms)	254	275	257	201	245	405
Observed Memory Span (number of words)	6.22	5.78	5.50	6.05	6.27	5.09
Predicted Memory Span (number of words)	6.19	5.77	5.45	6.14	6.23	5.12
Residual Memory Span (number of words)	0.027	0.012	0.049	-0.094	0.038	-0.033
Standard Error of Residual Memory Span	0.186	0.186	0.186	0.186	0.186	0.186

Note: Predicted memory spans are based on parameter values in Table 7. Residual memory spans are differences between observed and predicted memory spans.

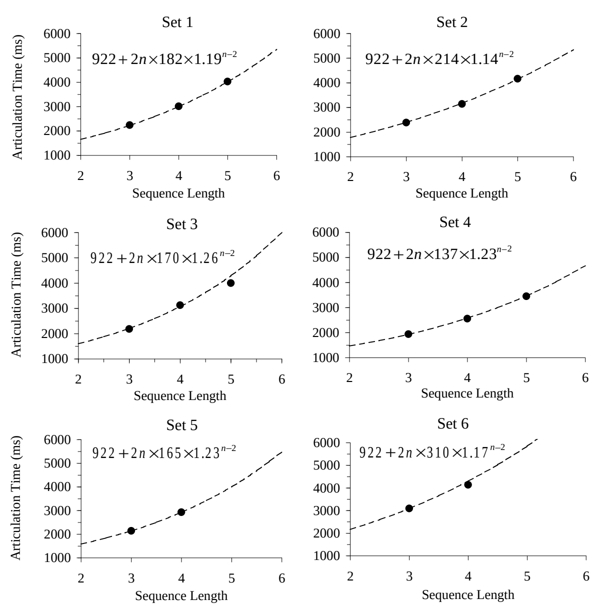
age to produce memorized sequences of words during the duration-measurement trials are shown in Figure 2 (solid points) as a function of sequence length. Also shown here are the theoretical concave-upward functions whose parameters yielded the mean articulatory duration per word for each of the six word sets in Experiment 1, as described by Appendix A2. The fit between these functions and the observed mean total articulation times was reasonably good (root mean-squared error $= 87$ ms, relative to total articulation times that typically exceeded 2 sec).[14]

On the basis of results from this analysis, Table 6 shows the estimated mean articulatory duration per word for each word set. We found that the words articulated by participants in memorized sequences had durations whose means differed reliably across these sets; $F(5,25) = 66.1$, $p < .001$. For example, under these conditions, the difference between the mean articulatory durations of Set 5's and Set 6's words was 160 ± 12 ms; $t(25) = 13.3$, $p < .001$.

However, when participants articulated words in memorized sequences, their mean durations were considerably less than those obtained for isolated words; mean difference $= 226 \pm 37$ ms, $t(5) = 6.10$, $p < .005$ (Table 6). Also, the rank orders of the mean articulatory durations across word sets differed reliably, depending on how the durations were measured; $F(5,25) = 13.26$, $p < .001$. For example, when we measured the durations of Set 2's nominally "short" words by having participants articulate them in memorized sequences, their mean duration was actually a bit longer than the corresponding mean for Set 1's nominally "long" words (275 vs. 254 ms), whereas Set 2's words had a much shorter mean duration than did Set 1's words when they were articulated in isolation (432 vs. 562 ms). The interaction contrast between these mean duration differences (151 ± 23 ms) was highly reliable; $t(25) = 6.57$, $p < .001$.

As a result, our mean articulatory durations for words in memorized sequences differed reliably from the articulatory durations reported by Caplan and Waters (1994), who measured them by having participants read words in lists that were visible throughout the articulation period. With their procedure, Caplan and Waters (1994) found that Set 1's words were considerably longer than Set 2's words (531 vs. 477 ms). In contrast, with our sequence-recall method, we found that the mean articulatory durations of these sets reversed their ordering (254 vs. 275 ms). Such a reversal suggests that list reading may not yield adequate measure-

Figure 2. Observed mean total articulation times (solid points) and theoretical total articulation-time functions (dashed lines) for each word set and sequence length in Experiment 1. Theoretical time functions are based on Equation 2 in Appendix A2.



ments of articulatory durations for testing the phonological-loop model.

Memory spans. Across the six word sets in Experiment 1, participants' memory spans depended reliably on which set was used for constructing the to-be-recalled word sequences; $F(5,25) = 6.09$, $p < .005$. For example, on average, the

[14] Figure 2 shows both observed mean data and total articulation-time functions based on the mean parameter values of Equation 2 (Appendix A2) for each word set and participant. To calculate the observed mean data and time functions, the obtained function intercepts were first subtracted from each observed mean total articulation time. The geometric mean (across participants) of the residual values was then calculated, and these values were then added to the arithmetic mean of the previously-subtracted intercepts to produce the observed mean data points. For each word set, parameters of the displayed time function were obtained by calculating the arithmetic mean of the *I* values and the geometric mean of the *a* and *d* values across participants. This type of averaging maintains the form of Equation 2 for the mean data.

words of Set 4 yielded a reliably greater memory span than did the words of Set 3; mean difference $= 0.55 \pm 0.26$ words, $t(25) = 2.12$, $p < .05$. This difference is similar to what Caplan and Waters (1994) found for these word sets; an analogous difference was also found by Caplan et al. (1992). On average, our participants also had marginally greater memory spans for the words of Set 1 than for the words of Set 2; mean difference $= 0.44 \pm 0.26$ words, $t(25) = 1.69$, $p < .10$. Again this difference is analogous to what Caplan and Waters (1994) found.

It therefore appears that considerable agreement exists between our memory-span data and those of some previous investigators who have criticized the phonological-loop model. What we and they disagree about is the extent to which the systematic variance of mean memory spans across word sets can be explained by their correlations with mean articulatory durations and phonological dissimilarity, which are two principal predictor variables that should account well for memory spans if the phonological-loop model is veridical. Thus, to help resolve this disagreement, the following additional analyses were performed.

Correlations between memory spans, articulatory durations, phonological dissimilarity, and phonological complexity. Using the results from the six word sets of Experiment 1 (Table 6), we performed a multiple-regression analysis with the mean articulatory durations and phonological dissimilarities of words in memorized sequences as the predictor variables, and the mean memory spans as the predicted variable. Across Sets 1 through 6, this analysis accounted for an extremely high and reliable percentage of variance in the memory-span data; multiple $R^2 = .986$, adjusted $R^2 = .977$, $F(2,3) = 106.0$, $p < .002$. The root-mean-squared-error between the observed and predicted mean memory spans in this analysis was extremely small; $RMSE = 0.05$ words. The estimated regression coefficients that yielded this excellent fit appear in Table 7. Both of the predictor variables contributed reliably to the overall goodness-of-fit: for the mean articulatory durations, partial $r = -.99$, $t(3) = -12.60$, $p < .001$; for the mean phonological dissimilarities, partial $r = .98$, $t(3) = 8.50$, $p < .005$. Thus, if either articulatory duration or phonological dissimilarity had been omitted as a predictor variable in this regression analysis, then the fit to the observed memory spans would have been significantly poorer (Figure 3).[15]

After the estimated contributions of these predictor variables to memory spans were removed, there was no reliable correlation between memory-span residuals and phonological complexity ($r = -.064$, $t(4) = -0.13$, $p > .05$). Furthermore, the residual span values were all very small and fell within confidence intervals that surrounded a null span value (Table 6). These results are what would be expected if the phonological-loop model is basically correct about the nature of the mechanisms that mediate memory spans, whereas a model based on the speech-planning hypothesis of Caplan et al. (1992) cannot account for what we found.

To further test whether phonological complexity might be a significant predictor of memory spans, as Caplan et al.'s (1992) hypothesis implies, we conducted two more multiple-regression analyses. In one of these, mean memory spans were the predicted variable, and there were three predictor variables: mean phonological complexity, phonological dissimilarity, and articulatory durations of words in memorized sequences. Across Sets 1 through 6 of Experiment 1, this analysis accounted for virtually the same percentage of variance in the memory-span data as did the preceding analysis that was based on only articulatory duration and phonological dissimilarity; multiple $R^2 = .986$, adjusted $R^2 = .966$, $F(3,2) = 48.1$, $p < .05$. Including phonological complexity as a predictor variable did not reliably improve the goodness-of-fit between the observed and predicted mean memory spans; $F(1,2) = 0.04$, $p > .50$. The correlation between the mean memory spans and phonological complexity was low and unreliable; partial $r = -.14$, $t(2) = -0.20$, $p > .50$.

Table 7
Coefficients of multiple linear-regression analysis for predicting memory span based on articulatory duration and phonological dissimilarity in Experiment 1.

Predictor	Coefficient	Standard Error
Intercept (number of words)	5.50	0.244
Articulatory Duration (ms)	-0.00573	0.00045
Phonological Dissimilarity	5.35	0.629

We also conducted a complementary multiple-regression analysis that included phonological complexity and phonological dissimilarity as predictor variables, but excluded articulatory duration. This analysis, compared to the preceding ones, accounted for substantially less systematic variance in the memory-span data across the six word sets; multiple $R^2 = .82$, adjusted $R^2 = .69$. The fit between the observed and predicted memory spans obtained from this latter analysis was reliably worse than the fit obtained when articulatory duration was included as a predictor variable; $F(1,2) = 24.9$, $p < .05$. These results further suggest that articulatory duration is crucial in accounting for memory spans, and if the contributions of this predictor variable are taken properly into account, then phonological complexity may be irrelevant.

Nevertheless, there is still another conceivable hypothesis that needs to be considered here (N. Cowan, personal

[15] For example, Figure 3 (top panel) reveals that the mean memory span for word Set 3 was markedly over-predicted by the linear predictor coefficient associated with articulatory duration. Such over-prediction occurred because this predictor coefficient, by itself, neglects to take account of phonological dissimilarity, which was especially low for Set 3. This is why phonological dissimilarity as well as articulatory duration must be taken into account. Conversely, the mean memory span for word Set 6 was markedly over-predicted by the predictor coefficient associated with phonological dissimilarity (Figure 3, bottom panel). Such over-prediction occurred because this predictor coefficient, by itself, neglects to take account of articulatory duration, which was especially long for Set 6. This is why articulatory duration as well as phonological dissimilarity must be taken into account. Only by taking into account the contributions by both articulatory duration and phonological dissimilarity can mean memory spans be predicted with uniformly high accuracy.

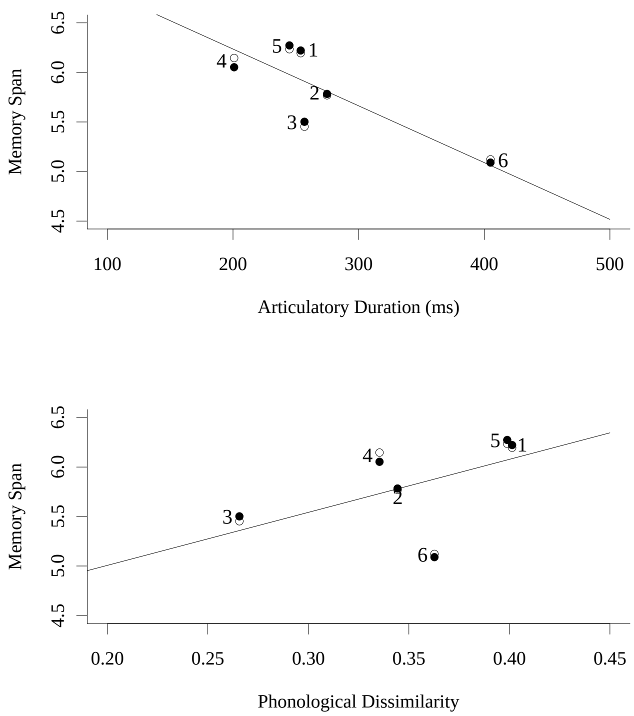


Figure 3. Mean memory span for each word set of Experiment 1 plotted against (top panel) mean articulatory duration for words in memorized sequences and (bottom panel) phonological dissimilarity. The lines in the panels show the estimated effect of each respective predictor variable on memory span, as determined from the linear regression coefficients in Table 7. The dark points in each panel represent observed mean values and the adjacent light points represent corresponding predicted mean values when both predictor variables are taken into account. The numbers beside these values indicate the word sets from which they came.

communication). According to it, both memory span and mean articulatory duration of words in memorized sequences are by-products of a single more basic underlying factor: *item memorability*. Perhaps such memorability determines the rate at which items are reproduced from sub-span sequences as well as the probability of correct recall from supra-span sequences. If so, then our results would not support the phonological-loop model per se. Rather than showing that mean articulatory duration is a causal determinant of memory span as this model implies, the correlation that we found between these two variables might manifest some other mechanism whose operation is really responsible for “memorability”.

Yet, on balance, our results provide evidence against the latter memorability hypothesis. For example, it predicts that mean articulatory durations should be shorter in sequences of words with high phonological dissimilarity, because high phonological dissimilarity presumably increases their memorability. However, we found no support for this prediction; the correlation between phonological dissimilarity and articulatory duration was only .09 in Experiment 1.

Furthermore, this evidence against the memorability hypothesis is bolstered by two other facts. First, during our articulatory-duration measurement procedure, trials on which participants hesitated or made recall errors were discarded. Consequently, all measured articulation times came from cases in which the participants recalled the word sequences with relative ease. Second, the correlation between our two types of articulatory-duration measurement (memorized lists vs. isolated speech) was .84. This high correlation indicates that both of these measurements manifested similar processes, even though one imposed essentially no memory load whereas the other required some memory.

Discussion

Taken all together, the results of Experiment 1 strongly support the phonological-loop model and its predictions about performance of the serial-recall task. We found that articulatory duration and phonological dissimilarity can account for nearly all of the variance in mean memory spans across the present six word sets, including several ones used by previous investigators (Baddeley & Andrade, 1994; Caplan et al., 1992; Caplan & Waters, 1994). In contrast, phonological complexity offers relatively little predictive power here. Our results yielded no evidence that the “word-length” effect stems from phonological complexity per se, so there is no reason to believe that rehearsal during the serial-recall task is based solely on speech planning, as hypothesized by Caplan et al. (1992). Instead, the original assumptions of the phonological-loop model still appear to be veridical, and we conclude that rehearsal requires overt or covert articulation of word sequences whose duration affects memory spans significantly.

Experiment 1 also demonstrates that different procedures for measuring articulatory durations can yield values that differ in both absolute and relative magnitudes. When we measured the articulatory durations for words in memorized sequences constructed from Sets 1 and 2, Caplan et al.’s (1992) set of nominal “long” words yielded durations that were actually shorter than those for the matched set of nominal “short” words. There are several reasons why this may have occurred. First, our measurement method involves articulation of words from memory, whereas previous methods have not (cf. Caplan et al., 1992; Caplan & Waters, 1994). Second, our instructions encouraged participants to articulate the words as if they were rehearsing, and required repeating each memorized sequence of words twice. Consequently, this method creates a situation that closely resembles rehearsal in the verbal serial-recall task, so participants may be less prone to unnecessarily extend their pronunciations of words. Third, we measured articulatory durations for several different sequence lengths. These measurements reveal that articulatory durations of words may differ in terms of both a baseline duration and a sequence-length amplification factor. As a result, an apparent duration difference between words for one sequence length may not generalize to differences for other sequence lengths. Thus, it is essential that these differences be taken into account when attempting to predict memory-span data accurately.

Given the results of Experiment 1, it can be seen likewise that PSIMETRICA is more valid and informative than some previous methods of measuring phonological similarity, such as those based on subjective ratings (cf. Baddeley & Andrade, 1994; Caplan & Waters, 1994). With phonological-dissimilarity measurements from PSIMETRICA, we successfully predicted reliable differences between the mean memory spans for word Set 1 versus word Set 2 and for word Set 3 versus word Set 4 (Figure 3, bottom panel). These predictions succeeded even though Caplan and Waters (1994) claimed, on the basis of subjective ratings, that Sets 1 and 2 had about equal degrees of phonological dissimilarity, and that Sets 3 and 4 did as well.

The present success of PSIMETRICA presumably stems from the fact that it takes account of detailed matches and mismatches at the level of individual phonological features, whereas subjective ratings of phonological dissimilarity perhaps do not (cf. Vitz & Winkler, 1973). We have been able to predict consistent effects of phonological dissimilarity on memory spans for sets of words that do not obviously differ in phonological dissimilarity, which is possible only because we characterize phonological dissimilarity at the phonological-feature level. Indeed, it appears that combinations of individual phonological features may be crucial for coding and storing phonemic information about words during serial recall and other VWM tasks (cf. Baddeley, 1986, 1992).

In summary, we conclude that the data of Caplan et al. (1992), who mounted one of the most influential critiques against the phonological-loop model, do not really justify rejecting this model's assumptions. The apparent inconsistencies between the predictions of the phonological-loop model and the data of Caplan et al. (1992) were probably obtained only because they measured phonological dissimilarity and articulatory duration in less than ideal ways. If, instead, these variables are measured more appropriately, the strong correlations obtained between them and memory spans support the phonological-loop model.

Experiment 2

A principal objective of Experiment 2 was to generalize the results of Experiment 1. Our approach involved measuring participants' memory spans for word sequences constructed from three new sets of words across which mean articulatory durations and phonological complexity varied in a quasi-orthogonal manner. For two of these word sets, their phonological complexity was held constant, but the mean articulatory durations of their words differed significantly between sets. This aspect of Experiment 2 is analogous to the experimental designs of Baddeley et al. (1975, Exp. 4), Caplan et al. (1992), Caplan and Waters (1994), Service (1998), and Lovatt et al. (2000). However, unlike some of these investigators, we show that differences in mean articulatory durations account for large reliable amounts of variance in memory spans even when phonological complexity is constant across word sets. Furthermore, for two of the word sets in Experiment 2, their phonological complexity differed significantly, but the mean articulatory durations of their words were approximately equal. This aspect of Experiment 2 is analogous to the experimental designs of Cowan et al. (1997) and Service (1998). Unlike these investigators, however, we show that differences in phonological complexity fail to account for large reliable amounts of variance in memory spans when mean articulatory durations are equated across word sets. Taken together, such results are like what Experiment 1 yielded and what the phonological-loop model predicts, whereas they conflict with the predictions of some alternative models (e.g., the speech-planning hypothesis; Caplan et al., 1992).

Two crucial methodological innovations, which have been advocated and adopted already in this article, enabled Experiment 2 to generalize our results from Experiment 1. As during Experiment 1, we measured articulatory durations for words in recently memorized sequences with variable lengths. Also, to quantify the effects of phonological dissimilarity, we measured it on the basis of PSIMETRICA.

In addition, there was another important aspect of Experiment 2. As part of analyzing its results, the effects of both phonological dissimilarity and articulatory duration were quantified by the values of the multiple-regression predictor coefficients that fit the memory spans from Experiment 1. Thus, with Experiment 2, we show that the effects of these factors on memory spans may be essentially invariant across different groups of participants and experimental stimuli.

Method

Participants. The participants were twelve undergraduate students from the University of Michigan. None had participated previously.

Apparatus. The apparatus was the same as in Experiment 1.

Stimuli. To construct sequences of words for the serial recall task, we used three new word sets: Set 7, which contained "short" two-syllable words; Set 8, which contained "long" two-syllable words; and Set 9, which contained "complex" three-syllable words (see Appendix A3). The words of Sets 8 and 9 were selected to have approximately equal mean articulatory durations even though Set 8 had fewer syllables and phonemes per word than did Set 9. The words of Set 7 were selected to have significantly shorter mean articulatory durations than did those of Sets 8 and 9. However, the words of Sets 7 and 8 were equal in phonological complexity (i.e., two syllables, and seven phonemes per word), whereas the words of Set 9 were significantly more complex (i.e., three syllables, and eight phonemes per word). All three sets of words were approximately equal in familiarity and concreteness.[16] Table 8 shows the mean values (and standard deviations) that each word set had for concreteness, familiarity,

[16] Both concreteness and familiarity ranged between 100 and 700 as reported by the MRC database (Wilson, 1987).

Table 8

Mean values of variables for word sets in Experiment 2

	Word Set		
Variable	7	8	9
Syllable Numerosity	2 (0)	2 (0)	3 (0)
Phoneme Numerosity	7 (0)	7 (0)	8 (0)
Phonological Dissimilarity	0.344 (0.087)	0.389 (0.077)	0.324 (0.096)
Familiarity	489 (99)	464 (62)	485 (73)
Concreteness	387 (93)	453 (112)	397 (129)

Note: Standard deviations are shown in parentheses.

phonological complexity (indexed by syllable and phoneme numerosity), and phonological dissimilarity.

Given the several constraints imposed on these word sets, we could not perfectly equate the mean phonological dissimilarity of the words in each set with the mean phonological dissimilarity of the words in the other sets (see Table 8). Consequently, there was a partial confounding among phonological dissimilarity, phonological complexity, and articulatory durations across Sets 7 through 9. Nevertheless, by taking previous results from Experiment 1 into account, we separated the contributions of these predictor variables to the memory spans in Experiment 2.

Design. Each participant was tested individually in a single 1.5 hour session. After an introductory instruction period, each session included two phases of testing. In the first phase, the participant's mean articulatory durations were measured for the words in Sets 7, 8, and 9, respectively. In the second phase, the participant's memory spans were measured for sequences of words constructed from each set. The order in which the word sets were used during each phase was counterbalanced across participants, but each participant received the sets in the same order during both phases. Participants were paid $8 plus bonuses for good performance on the serial recall task.

Articulatory-duration measurement. The articulatory durations for words in memorized sequences were measured as in Experiment 1. Each participant received sequences ranging from two to five words, and performed eight measurement trials per length for each word set.

Memory-span measurement. For each word set, memory spans were measured as in Experiment 1. Each memory-span trial block included twenty trials. To encourage good performance, whenever the memory span for a word set exceeded four words, we paid the participant a bonus of $S-4$ dollars, where S was the magnitude of the memory span.

Preliminary data analysis. One atypical participant correctly recalled sequences that included as many as ten words. Although an articulatory-duration effect occurred for him, his memory spans far exceeded those of the other eleven participants, and he reported using a unique mnemonic strategy to achieve this performance. We therefore omitted his data from subsequent analyses, because they were clearly unusual and would have distorted the remaining group means. For all other participants, mean articulatory durations and memory spans were calculated and analyzed as in Experiment 1.

Results

Table 9 summarizes the mean values of articulatory duration and memory span that resulted from the three sets of words in Experiment 2 . Here we discuss our results with respect to each of these variables and the relationships among them compared to those from Experiment 1 (cf. Table 6).

Articulatory durations. As before, when participants articulated recently memorized sequences of words, their total articulation times increased in a concave upward trend with sequence length. Figure 4 shows how well Equation 2 of Appendix A2 fit these times on average for each word set of Experiment 2 ($RMSE = 19$ ms). On the basis of estimates obtained from parameters of this equation, the means of the articulatory durations per word differed reliably across word sets; $F(2,20) = 28.6$, $p < .001$. In particular, the difference between the mean articulatory durations for the "long" two-syllable words of Set 8 and the "short" two-syllable words of Set 7 was large and reliable (Table 9); mean difference $= 58 \pm 8$ ms, $t(20) = 7.25$, $p < .001$. This result confirms that we succeeded in constructing two sets of words across which the mean articulatory duration per word differed even though their degrees of phonological complexity (i.e., numbers of syllables and phonemes per word) were equated. Moreover, the difference between the mean articulatory durations for the "long" two-syllable words of Set 8 and the "complex" three-syllable words of Set 9 was small and unreliable (Table 9); mean difference $= 7 \pm 8$ ms, $t(20) = 0.87$, $p > .25$. This result confirms that we succeeded in constructing two sets of words across which the mean articulatory durations per word were nearly equal despite there being a marked difference in their phonological complexity. Thus, in Experiment 2, mean articulatory durations and phonological complexity varied quasi-orthogonally, which enables us to separate the effects of these predictor variables on memory spans.[17]

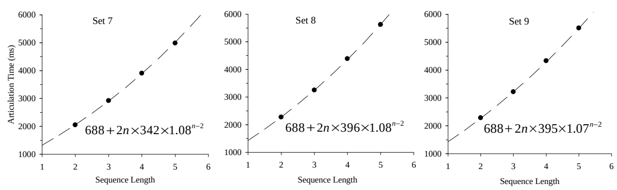
[17] Our mathematical analysis was designed to remove any biases contributed by the experimenter to our measurement of participants' mean word durations in articulating sequences from memory. To verify that this analysis succeeded, we have re-examined the data from four participants whose performance during Experiment 2 was tape recorded. Their utterances from articulating words in memorized sequences were evaluated with digital-waveform analysis software to obtain new estimates of the mean articulatory duration for each word set and each participant. Results showed that mean estimated articulatory durations from the original measurements differed on average by 10 ms (ranging between 0 and 20 ms) compared to those from the waveform analysis. Furthermore, these esti-

Table 9 *Results from Experiment 2.*

Variable	Word Set 7	8	9
Mean Articulatory Duration (ms)	393	451	444
Observed Memory Span (number of words)	5.21	5.05	4.71
Predicted Memory Span (number of words)	5.09	5.00	4.69
Residual Memory Span (number of words)	0.121	0.053	0.021
Standard Error of Residual Memory Span	0.099	0.099	0.099

Note: Predicted memory spans are based on the same predictor coefficients for articulatory duration and phonological dissimilarity as in Experiment 1.

Figure 4. Observed (solid points) mean total articulation times and theoretical total articulation-time functions (dashed lines) for each word set and sequence length tested in Experiment 2. Theoretical time functions are based on Equation 2.



Memory spans. Participants' memory spans (Table 9) depended reliably on which word set was used for constructing the to-be-recalled sequences of Experiment 2; $F(2,20) = 6.56$, $p < .01$. For example, on average, the "short" two-syllable words of Set 7 yielded a reliably larger memory span than did the "complex" three-syllable words of Set 9; mean difference$= 0.50 \pm 0.14$ words, $t(20) = 3.52$, $p < .001$. The mean memory span for the "long" two-syllable words of Set 8 was intermediate.

Correlations between memory spans, articulatory durations, phonological dissimilarity, and phonological complexity. To discover which predictor variables accounted best for the mean memory spans across the three word sets of Experiment 2, we used the same regression equation that had accounted well for memory spans from Experiment 1. According to this equation,

$$S = 5.50 - 0.00573 \cdot D + 5.35 \cdot P, \quad (1)$$

where S is the predicted mean memory span for sequences constructed from a particular word set, D is the mean articulatory duration (in milliseconds) of the words in the set, P is their mean phonological dissimilarity, and the numerical coefficients have been estimated with Experiment 1's results (cf. Table 6). Substituting the values of D and P for the three new word sets of Experiment 2 in Equation 1 yielded the predicted mean memory spans in Table 9, which may be compared to the observed mean memory spans for these sets.

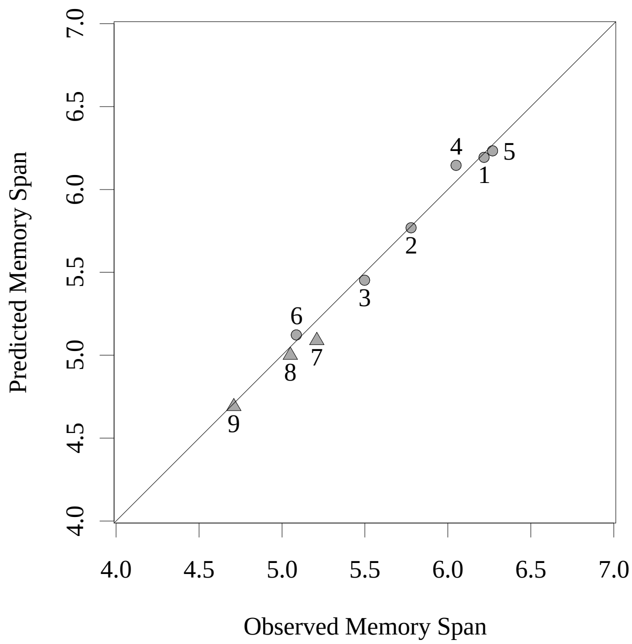
From this comparison, we see that the predicted mean memory spans based on Equation 1 fit the observed mean memory spans closely; $R^2 = .991$; $RMSE = 0.076$ words. The residual differences between the observed and predicted mean memory spans are not reliable, $F(2,20) = 0.28$, $p > .05$, indicating that Equation 1 accounted for essentially all of the systematic variance in the observed mean memory spans across the three word sets of Experiment 2. The difference between the grand means of the observed and predicted memory spans is not reliable either; $F(1,10) = 0.10$, $p > .5$, indicating that Equation 1 accounted not only for the word-set effects on observed memory spans but also for the memory spans' overall absolute magnitude.

This excellent fit was achieved even though Experiment 2 involved new sets of words and participants, but no new parameter values were estimated from its results (i.e., the fit provided by Equation 1 is based solely on predictor coefficients estimated from Experiment 1's results). The pairs of observed and predicted mean memory spans in Experiment 2 fit almost seamlessly with those that Experiment 1 yielded (Figure 5). Thus, because Equation 1 is based on articulatory durations and phonological dissimilarity, rather than phonological complexity, its successful a priori prediction of the mean memory spans in Experiment 2 further supports the phonological-loop model of verbal serial recall.

On the other hand, the results of Experiment 2—again like the results of Experiment 1—provide no support for alternative models under which phonological complexity is claimed to be a powerful predictor of memory span (e.g., Caplan et al., 1992; Caplan & Waters, 1994; Cowan et al., 1997). After the expected contributions of the mean articulatory durations and phonological dissimilarity were removed from the observed mean memory spans through Equation 1, the residual memory spans (Table 9) revealed no reliable effect of phonological complexity; $F(2,20) = 0.28$, $p > .5$. The residual memory spans were all very small and fell within confidence intervals that surrounded a null value. In particular, the residual memory span for the "complex" three-syllable words of Set 9 was neither less than zero nor reliably less than the residual memory span for the "long" two-syllable words of Set 8; $t(10) = 0.25$, $p > .50$, contrary to what Caplan et al. (1992; Caplan & Waters, 1994) would have expected. That the "complex" words of Set 9 yielded the smallest observed mean memory span (Figure 5, left-most triangle) appears to have resulted simply because their articulatory durations were longer than those in Set 7 and their phonological dissimilarities were less than those in Set 8.

mates were essentially unbiased (the mean difference between original and verification measurements equaled 0.4 ms across words sets and participants).

Figure 5. Observed and mean memory spans based on Equation 1 from Experiment 1. Shaded circles come from word sets in Experiment 1, and shaded triangles come from word sets in Experiment 2. Numerals indicate which word sets yielded individual pairs of observed and predicted memory spans.



Discussion

Experiment 2 bolsters the results of Experiment 1, confirming that phonological complexity is not a significant predictor of verbal serial memory span when the effects of articulatory duration and phonological dissimilarity have been taken properly into account. If complexity reduced (or enhanced) memory span independent of these other factor effects, then memory span for the complex three-syllable words of Set 9 in Experiment 2 should have been considerably less (or more) than we found. Although memory spans for sequences of words may depend on many factors, phonological complexity does not seem to matter when phonological dissimilarity and articulatory duration are measured appropriately, and then used in predicting performance. This is exactly what the phonological-loop model implies.

Experiment 2 also provides further empirical justification for the methods that we have introduced to measure articulatory duration and phonological dissimilarity. When measured as we have described, these variables enable highly accurate accounts of memory span that generalize readily to new participants and different sets of words. If our measurement methods had been seriously flawed, such generalization would not have been possible here, and the results of Experiment 2 could not have been predicted so accurately from the results of Experiment 1.

General Discussion

In two experiments on verbal working memory, we have obtained instructive new evidence about the extent to which three theoretically relevant factors affect serial recall accuracy and memory span. Our results show that when articulatory duration and phonological dissimilarity are measured with the methods introduced here, these factors may be extremely accurate predictors of memory spans across word sets. However, contrary to other alternative hypotheses, phonological complexity is a substantially less accurate predictor; it may contribute essentially nothing to serial recall accuracy after articulatory duration and phonological dissimilarity have been taken properly into account. The present evidence has important implications for the status of the phonological-loop model, especially with respect to what some previous studies have concluded about it.

Reinterpretation of Results from Past VWM Studies

Although the phonological-loop model provides a good explanation for the results of our two experiments, it is not entirely clear why the results from several past studies on VWM (e.g., Caplan et al., 1992; Caplan & Waters, 1994; Cowan et al., 1997, 2000a; Lovatt et al., 2000; Service, 1998, 2000) appeared to contradict this model. If these studies had measured phonological dissimilarity and articulatory duration more appropriately, perhaps many of their results would have supported this model rather than disconfirming it. To determine whether this may in fact be the case, we next examine each of these studies more carefully.

Caplan et al. (1992). In Experiment 2 of Caplan et al. (1992), performance was compared for two sets of words that were equated on phonological complexity but had either "long" or "short" articulatory durations when measured in isolation. Participants' serial recall accuracy was higher for the "long" words than for the "short" words. Caplan et al. (1992) interpreted these results as evidence against the phonological-loop model, because it predicts that ceteris paribus, longer words should yield lower rather than higher serial recall accuracy.

Nevertheless, upon closer inspection, these results appear to be consistent with the phonological-loop model and the findings from our Experiment 1. We found that when participants repetitively articulated Caplan et al.'s (1992) "short" and "long" words in variable-length memorized sequences (rather than in isolation), the mean duration of the "short" words (i.e., Set 2) actually exceeded the mean duration of the "long" words (i.e., Set 1). Also, as measured by PSIMETRICA, the phonological dissimilarity of the "short" words was significantly less than the phonological dissimilarity of the "long" words. Together, these factors account almost perfectly for the difference in mean memory spans between Caplan et al.'s (1992, Experiment 2) "short" and "long" word sets, as the phonological-loop model would predict (Figure 5). Our findings therefore lead us to conclude that Experiment 2 of Caplan et al. (1992) suffered from two problems: articulatory duration was measured in a less than ideal way, and the effects of phonological dissimilarity were not taken properly into account.

Still it remains unclear how such problems could underlie some other results of Caplan et al. (1992). In their Experiment 3, they compared participants' performance for two more sets of words that were equated on phonological complexity but nominally either "difficult" or "easy" to articulate. Confirming this latter difference, the "difficult" words yielded a longer mean duration than the "easy" words did when a confederate speaker articulated them in isolation. However, participants' serial recall accuracy was virtually equal (71.3% and 71.4%, respectively) for the "difficult" and "easy" words, seemingly contrary to the phonological-loop model, which predicts that in this case, serial recall accuracy should have been lower for the "difficult" words.

These latter results cannot be explained simply by a mismeasurement of articulatory duration or failure to take account of countervailing phonological-dissimilarity effects. During our Experiment 1, when we measured the articulatory durations for the "difficult" and "easy" words in variable-length memorized sequences, the "difficult" (Set 3) words yielded somewhat longer durations than did the "easy" (Set 4) words, consistent with what Caplan et al. (1992, Exp. 3) found. Also, PSIMETRICA revealed that the "difficult" words of Set 3 were less phonologically dissimilar from each other than were the "easy" words of Set 4. Thus, according to the phonological-loop model, serial recall accuracy should have been lower for the "difficult" words than for the "easy" words. Indeed, the results from our Experiment 1 confirmed this prediction. On the other hand, however, the results of Caplan et al. (1992, Exp. 3) differ with ours, because they found that the "difficult" and "easy" words yielded essentially equal serial recall accuracy.

Although it may appear at first that this discrepancy is simply a statistical fluke, we believe these different results can be explained without rejecting the phonological-loop model. This is possible if we assume that some atypical aspects of Caplan et al.'s (1992, Exp. 3) procedure discouraged participants from using their phonological loops to perform the serial recall task, and instead induced participants either to adopt other idiosyncratic (non-verbal) rehearsal strategies or to entirely forego rehearsal. These assumptions seem plausible because Caplan et al. (1992, Exp. 3) required serial recall to be made through picture-pointing (rather than vocal responses), and their participants were not instructed to rehearse articulatorily. Thus, the procedure used by Caplan et al. (1992, Exp. 3) lacked prerequisite features that may be essential to testing the phonological-loop model fairly, so it is not surprising that they found relatively little difference between serial recall accuracy for their "difficult" and "easy" words.

Caplan and Waters (1994). This reinterpretation is supported by subsequent findings that Caplan and Waters (1994) obtained. They repeated Experiment 3 of Caplan et al. (1992) with the same "difficult" and "easy" words, but measured the accuracy of immediate vocal serial recall for these words, instead of requiring recall to be made through picture-pointing responses. In this modified replication, whose procedure probably encouraged articulatory rehearsal more than before, serial recall accuracy was significantly lower for the "difficult" words than for the "easy" words (64.1% vs. 72.8%). This finding is like what we obtained for these words (Table 6), supporting the phonological-loop model and our reinterpretation of the previous atypical results reported by Caplan et al. (1992, Exp. 3).

Cowan et al. (1997). Cowan et al. (1997) obtained results that suggest both phonological complexity and articulatory duration affect serial recall accuracy. In their experiment, participants were presented sequences of either "simple" (one-syllable) or "complex" (two-syllable) printed words. During both presentation and later (backward) recall, there were visual timing cues with either "short" (about 500 ms) or "long" (about 900 ms) durations, and the participant was instructed to speak each word of a sequence aloud so that it matched the durations of these signals. Consequently, quasi-orthogonal combinations of both articulatory duration and phonological complexity occurred here. Cowan et al. (1997, Exp. 2) found that serial recall accuracy was significantly lower when articulatory durations were "long" rather than "short", and that the "complex" words yielded higher serial recall accuracy than did the "simple" words.

Without further elaboration, the phonological-loop model can not explain this beneficial positive effect of phonological complexity on serial recall accuracy. Consequently, Cowan et al. (1997) suggested that this model is either incomplete or incorrect. They proposed that other theoretical constructs—perhaps involving principles of interference rather than time-based decay—are needed to account for serial recall accuracy in their experiments.

However, the apparent positive effect of phonological complexity found by Cowan et al. (1997) may have stemmed from differences in phonological dissimilarity that were confounded with the observed differences in phonological complexity. Evidence for this possibility can be found in Table 10, where we have used PSIMETRICA to quantify the phonological dissimilarity of both the "simple" and "complex" words of Cowan et al. (1997). According to this quantification, it appears that the syllable onsets of these "complex" words were significantly more dissimilar from each other than were those of the "simple" words (0.419 vs. 0.365, respectively). Furthermore, the "complex" words had syllable nuclei that were significantly more dissimilar from each other than were those of the "simple" words (0.261 vs. 0.204, respectively). In contrast, the syllable codas of the "complex" words were just slightly less dissimilar from each other than were the codas of the "simple" words (0.252 vs. 0.272, respectively). Thus, on average, the "complex" words seem to have been more phonologically dissimilar from each other than were the "simple" words. Given this systematic variation across these two types of word, our previous regression analysis (Equation 1) implies that phonological-dissimilarity effects may have contributed at least a 0.3 memory-span unit advantage for the "complex" words over "simple" words. Such contributions are sufficiently large to suspect that the nominal phonological-complexity effect found by Cowan et al. (1997) stemmed mostly, perhaps even entirely, from

Table 10
Phonological dissimilarity of each syllable constituent for the word sets used by Cowan et al. (1997) and Lovatt et al. (2000).

		Phonological Dissimilarity		
Source	Word Set	Onset	Nucleus	Coda
Cowan et al., 1997	"Simple"	0.365	0.204	0.272
	"Complex"	0.419	0.261	0.252
Lovatt et al., 2000, Exp. 1	"Short"	0.318	0.108	0.200
	"Long"	0.337	0.169	0.235
Lovatt et al., 2000, Exp. 2	"Short"	0.348	0.148	0.164
	"Long"	0.297	0.240	0.278
Lovatt et al., 2000, Exp. 3	"Short"	0.334	0.159	0.213
	"Long"	0.397	0.193	0.178

a confounding with phonological dissimilarity.

Cowan et al. (2000a). Confirming our suspicions, Cowan et al. (2000a) conducted a subsequent study that failed to replicate some of their previous results. In this study, they used the same procedure with paced articulation as before (Cowan et al., 1997, Exp. 2), but required forward instead of backward serial recall. Under these modified conditions, serial recall accuracies conformed more closely to predictions of the phonological-loop model, which is what should happen if the requirement of forward serial recall especially encourages participants to use incremental cyclic articulatory rehearsal.

Specifically, Cowan et al. (2000a) found that with forward serial recall, there were two changes in their results compared to what they reported previously (cf. Cowan et al., 1997, Exp. 2). Longer articulatory durations during the presentation and recall phases of each trial yielded lower recall accuracies at all of the serial positions of the word sequences, as the phonological-loop model would ordinarily predict. Also, unlike before, when participants' articulatory durations during sequence presentation and forward serial recall were approximately equated for "simple" and "complex" words, the "complex" words did not yield higher serial recall accuracies. Thus, Cowan et al.'s (2000a) results provide additional strong support for the phonological-loop model.

Service (1998). Like Cowan et al. (1997, 2000a), Service (1998) tried to separate the effects of articulatory duration and phonological complexity on serial recall accuracy. In her experiments, participants recalled sequences of Finnish pseudowords that were either (1) "simple" disyllables with nominally "short" articulatory durations, (2) "simple" disyllables with nominally "long" articulatory durations, or (3) "complex" trisyllables with "long" articulatory durations. Articulatory durations of the "simple" pseudowords were manipulated by exploiting a special property of Finnish, which has phonemes whose durations are either "short" or "long" but whose other phonological features are supposedly identical. Consequently, the two "simple" pseudoword sets were identical except for the difference in the lengths of their vowels. However, despite this difference, they produced nearly equal levels of serial recall accuracy, whereas the "complex" trisyllabic pseudowords yielded much lower recall accuracy. Service (1998) therefore concluded that phonological complexity (but not articulatory duration) influences performance in VWM tasks. She also concluded that investigators should abandon the phonological-loop model in favor of some other theoretical account (e.g., one involving interference and redintegrative mechanisms instead of articulatory rehearsal and time-based decay).

Yet this conclusion may be unwarranted, because in Service's (1998) experiments, articulatory durations were not measured under conditions relevant to the phonological-loop model. She measured the articulatory durations of Finnish pseudowords by having participants read visually presented lists, and found that the simple "long" pseudowords yielded much (31%) longer articulatory durations than did the simple "short" pseudowords. However, it is likely that this list-reading task imposed certain demand characteristics on the participants, compelling them to speak the two sets of "simple" pseudowords at distinctively different rates, even if they could rehearse these pseudowords at the same rate during the serial-recall task.

To substantiate this possibility, we note that Service (1998) also measured participants' recall durations for the pseudoword sequences in the serial recall task. Based on this measurement, which may approximate the ideal method more closely than list reading does, there was only a 7.7% difference between the durations of the simple "short" and "long" pseudowords. This difference was small, even though participants were required to extend their overt articulation of the pseudowords with longer vowels in order to be scored correctly. So quite possibly, when they were not compelled by experimental demands to speak these words at different rates, the participants may have rehearsed the "long" and "short" pseudowords at nearly the same rate.

As a consequence, Service's (1998) claims about the importance of phonological-complexity effects on serial recall accuracy require qualification. Because her experiments used unfamiliar pseudowords instead of familiar words for constructing to-be-memorized sequences, they may have induced participants to adopt atypical encoding, rehearsal, or recall strategies that magnify the effect of phonological complexity far beyond what occurs for real words under more typical conditions. If so, then the phonological-loop model may still usually provide a veridical account of serial recall accuracy, and further research will be needed to assess phonological complexity's relevance for VWM.

Lovatt et al. (2000). Following Service (1998), these investigators conducted three experiments with the verbal serial recall task to test the phonological-loop model. Each experiment involved two sets of disyllabic words that differed in their nominal mean articulatory durations ("short" versus "long") but were putatively matched with respect to several other linguistic variables (e.g., frequency, familiarity, phonological dissimilarity, number of phonemes, and semantic associations). Although the word sets changed from one experiment to the next, other aspects of Lovatt et al.'s (2000) designs and procedures were essentially identical across their experiments. In particular, articulatory durations were always measured for isolated words and for words read from lists.[18] Phonological dissimilarities between the words of each set were measured by having participants rate them on a five-point scale, as in some other studies (e.g., Baddeley & Andrade, 1994; Caplan & Waters, 1994). These ratings suggested that the mean phonological dissimilarity between the "long" words almost exactly equaled the mean phonological dissimilarity between the corresponding "short" words.

Nevertheless, Lovatt et al. (2000) obtained extremely inconsistent articulatory-duration effects on serial recall accuracies. During their first experiment, the "long" words yielded a reliably higher recall accuracy than did the "short" words (65.1% vs. 60.7% with auditory stimuli and vocal responses), whereas the "long" and "short" words of the second experiment yielded virtually equal recall accuracies (58.5% vs. 58.7%, respectively). Only during Lovatt et al.'s (2000) third experiment did the "long" words yield a reliably lower recall accuracy than the "short" words (65.5% vs. 70.7%, respectively) as the phonological-loop model would ordinarily predict. Given such apparent empirical inconsistency, Lovatt et al. (2000) strongly questioned the general veracity of this model's assumptions.

However, the experiments of Lovatt et al. (2000) suffer from many of the same problems as the previous experiments that we have discussed. For example, these investigators did not measure the articulatory durations of words in variable-length memorized sequences. Consequently, it is unclear whether the articulatory durations of these words differed in ways that are relevant to the phonological-loop model. Of even greater concern, phonological dissimilarity was assessed by Lovatt et al. (2000) through a subjective rating procedure. Because subjective ratings may not be sensitive enough to detect relevant differences in phonological dissimilarity that can affect serial recall accuracy, perhaps the word sets of Lovatt et al. (2000) were not actually equated with respect to phonological dissimilarity.

To examine this further, we have measured the phonological dissimilarity of Lovatt et al.'s (2000) word sets (Table 10) using PSIMETRICA. For the "short" words and "long" words of their first experiment, the dissimilarities between the respective onsets, nuclei, and codas of the "long" words were uniformly greater than the dissimilarities between those of the "short" words. From the perspective of the phonological-loop model and our previous regression analysis (Equation 1), this greater dissimilarity between the "long" words could have counteracted the effect of articulatory duration and produced a higher serial recall accuracy than did the "short" words.

In Lovatt et al.'s (2000) second experiment, the phonological dissimilarities between the nuclei and codas of the "long" words were again greater than the corresponding dissimilarities for the "short" words. However, these "long" words had onsets whose phonological dissimilarity was considerably less than those of the "short" words' (Table 10). Such opposing degrees of phonological dissimilarity might explain why the "long" and "short" words of Lovatt et al.'s (2000) second experiment yielded almost equal serial recall accuracies.

In their third experiment, the "long" words yielded lower serial recall accuracy than did the "short" words. This happened even though the "long" words again tended to be more dissimilar than were the "short" words (Table 10). So in this last case, the difference between their mean articulatory durations was apparently great enough to counteract the effect of phonological dissimilarity.

Summary of reinterpreted experiments. When initially examined, these previous experiments (i.e., Caplan et al., 1992; Caplan & Waters, 1994; Cowan et al., 1997, 2000a; Lovatt et al., 2000; Service, 1998, 2000) each appeared to offer evidence against the phonological-loop model. However, taken as a whole, they do not support any clear alternative theory. Furthermore, when we reexamine their results through the lenses of PSIMETRICA and our articulatory measurement method, it appears that the phonological-loop model is consistent with all of them. This model can even explain why some of the attempts to replicate these past results failed.

Status of The Phonological-Loop Model

On the basis of our results from Experiments 1 and 2, as well as our reinterpretation of previous experiments, several core assumptions of the phonological-loop model have received additional strong empirical support.

Phonological coding of stored word sequences. The phonological-loop model assumes that during VWM tasks such as serial recall, the words of memorized sequences are coded and stored in the form of temporary phonological representations (Baddeley, 1986). Supporting this assumption, we have found that precise quantitative measurement of phonological dissimilarity through PSIMETRICA can help predict serial recall accuracy. Such predictability is what would be expected if the temporary stored representations of words are closely tied to information used by the vocal articulators for producing covert and overt speech during rehearsal, as in the phonological-loop model.

Information loss through time-based decay. A second related assumption is that the loss of information from VWM occurs through time-based decay. Because of such decay, the principal limit on the functional capacity of VWM may be the time required to refresh the memory traces of stored

[18] During Lovatt et al.'s (2000) second and third experiments, articulatory durations were also measured for words in short constant-length (two-word) repeated sequences. These durations approximately equaled those obtained through list reading.

items. If so, VWM would not have a fixed small number of "slots" for storing subunits of information (cf. Kieras et al., 1999). Supporting this assumption, we have found that memory spans depend reliably on the articulatory durations of words in memorized sequences but not on their phonological complexity per se. Consistent with the phonological-loop model, there is no apparent constant upper bound on the numbers of phonemes or syllables that can be retained in VWM.

Memory-trace retention by strategic articulatory rehearsal. The phonological-loop model also assumes that the retention of memory traces entails strategic articulatory rehearsal. This assumption, like the preceding one, is supported by our finding that articulatory durations measured for words in memorized sequences are reliable predictors of memory spans, whereas articulatory durations measured for isolated words are not. On the other hand, if articulatory rehearsal were irrelevant to memory-trace retention in VWM, then how we measured the articulatory durations of words should have been irrelevant, and memory spans should have been uncorrelated with all of these duration measures. The absence of duration effects on serial recall accuracy when recall involves picture-pointing rather than vocal responses (Caplan et al., 1992, Exp. 3) likewise demonstrates the optional strategic nature of articulatory rehearsal.

New Insights About Articulatory Rehearsal and Phonological-Dissimilarity Effects

While the results of Experiments 1 and 2 support several core assumptions of the phonological-loop model, they also provide additional insights about ways in which this model should be refined and elaborated further. For example, we found that under the present conditions, memory spans could be predicted reliably by the phonological dissimilarity between syllable onsets, but the dissimilarity between syllable rhymes was an unreliable predictor. This may have occurred because participants' articulation during rehearsal truncated the syllables in order to cycle through the memorized word sequences more rapidly. Such truncation may have "nullified" the phonological features in the syllable rhymes, making the different features of the onsets relatively more important for the effect of phonological dissimilarity on serial recall accuracy. If so, then the possible contributions of these variations in rehearsal strategies must be factored into future formulations of the phonological-loop model.

A second insight from our Experiment 1 about articulatory rehearsal is that the measured durations of words depend on how participants have to articulate them. On average, words articulated in memorized sequences have much shorter durations than do words articulated in isolation. Consequently, for certain paired word sets, the difference between the mean articulatory durations of their words may be reliably positive or negative, depending on whether sequential or isolated articulation is involved. This dependence can occur because articulatory rehearsal entails relatively rapid utterances in which syllable rhymes are presumably truncated more than syllable onsets, thereby changing which parts of the syllables contribute most to the measured durations of words.

A third insight, obtained through our reinterpretation of previous studies, is that procedures for testing VWM can influence what strategies participants use for performing the memory tasks. For example, some testing procedures may discourage the use of articulatory rehearsal (e.g., those in Experiment 3 of Caplan et al., 1992, and in Cowan et al., 1997). When this happens, it will yield a diminished effect of articulatory duration on recall accuracy compared to experiments that encourage participants' use of such rehearsal (e.g., those in our experiments, as well as in Caplan and Waters, 1994, and Cowan et al., 2000a).

Taken together, these new insights indicate that the effects of articulatory duration and phonological dissimilarity, often viewed as bellwethers of the phonological-loop model, will not necessarily occur whenever word sets differ in terms of these factors. In order for these factors to affect performance in a memory task, the task must encourage the participant to use a phonological code, and the participant must engage in articulatory rehearsal. Additionally, in order for the effects of these factors to be detected, appropriate measurement methods must be used. If these requirements are not met, then the phonological-loop model does not predict that articulatory duration or phonological dissimilarity will have reliable effects. The full range of conditions under which this model's predictions are applicable remains to be determined.

References

Anderson, J. R., & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review, 104*, 728-748.
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. (1968). How does acoustic similarity influence short-term memory? *Quarterly Journal of Experimental Psychology, 20*, 249-264.
Baddeley, A. D. (1986). *Working memory*. Oxford, UK: Oxford University Press.
Baddeley, A. D. (1992). Working memory. *Science, 255*, 566-569.
Baddeley, A. D., & Andrade, J. (1994). Reversing the word-length effect: A comment on Caplan, Rochon, and Waters. *Quarterly Journal of Experimental Psychology, 47A*, 1047-1054.
Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *Recent advances in learning and motivation, Vol. VIII* (pp. 47-90). New York: Academic Press.
Baddeley, A. D., Lewis, V., & Vallar, G. (1984). Exploring the articulatory loop. *Quarterly Journal of Experimental Psychology, 36A*, 233-252.
Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28-61). New York: Cambridge University Press.
Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14*, 575-589.
Baddeley, A. D., & Wilson, B. (1985). Phonological coding and short-term memory in patients without speech. *Journal of Memory and Language, 14*, 575-589.

Basso, A., Spinnler, H., Vallar, G., & Zanobio, E. (1982). Left hemisphere damage and selective impairment of auditory verbal short-term memory: A case study. *Neuropsychologia, 20*, 263-274.

Brown, G. D., & Hulme, C. (1995). Modeling item length effects in memory span: No rehearsal needed? *Journal of Memory and Language, 34*, 594-624.

Burgess, N., & Hitch, G. (1996). A connectionist model of STM for serial order. In S. Gathercole (Ed.) *Models of short-term memory* (pp. 51-72). Hove, U. K.: Psychology Press.

Cantor, J., & Engle, R. W. (1993). Working-memory capacity as long-term memory activation: An individual-differences approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1101-1114.

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.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper Row.

Colle, H. A., & Welsh, A. (1976). Acoustic masking in primary memory. *Journal of Verbal Learning and Verbal Behavior, 15*, 17-32.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language*. New York: Harcourt Brace Jovanavich.

Coltheart, M. (1981). The MRC Psycholingistic Database. *Quarterly Journal of Experimental Psychology, 33a*, 497-505.

Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion, and memory span. *British Journal of Psychology, 55*, 429-432.

Cowan, N. (1992). Verbal memory span and the timing of spoken recall. *Journal of Memory and Language, 31*, 668-684.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62-101). New York: Cambridge University Press.

Cowan, N., Day, L., Saults, J. S., Keller, 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 and Language, 31*, 1-17.

Cowan, N., Nugent, L. D., & Elliot, E. M. (2000). Memory-search and rehearsal processes and the word length effect in immediate serial recall: A synthesis in reply to Service. *Quarterly Journal of Experimental Psychology, 53A*, 666-670.

Cowan, N., Nugent, L. D., Elliot, 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.

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

Cowan, N., Wood, N. L., Nugent, L. D., & Treisman, M. (1997). There are two word-length effects in verbal short-term memory: Opposed effects of duration and complexity. *Psychological Science, 8*, 290-295.

Cowan, N., Wood, N. L., Wood, P. K., Keller, T. A., Nugent, L. D., & Keller, C. V. (1998). Two separate verbal processing rates contributing to short-term memory span. *Journal of Experimental Psychology: General, 127*, 141-160.

D'Esposito, M., & Postle, B. (1999). Neural correlates of component processes of working memory: Evidence from neuropsychological and pharmacological studies. In S. Monsell & J. Driver (Eds.), *Attention and performance XVIII*. Cambridge, MA: M. I. T. Press.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283-321.

Dell, G. S., Juliano, C., Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science, 17*, 149-195.

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.

Dosher, B. A., & Ma, 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, and Cognition, 24*, 316-335.

Drewnowski, A. (1980). Attributes and priorities in short-term recall: A new model of memory span. *Journal of Experimental Psychology: General, 109*, 208-250.

Drewnowski, A., & Murdock, B. B. (1980). The role of auditory features in memory span for words. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 319-332.

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

Estes, W. K. (1972). An associative basis for coding and organization in memory. In A.W.Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 161-190). Washington, DC: Winston.

Frisch, S. A. (1997). Against underspecification in speech errors. *Studies in the Linguistic Sciences, 27*, 79-97.

Fudge, E. C. (1969). Syllables. *Journal of Linguistics, 5*, 253-286.

Glasspool, D. W. (1995). Competitive queuing and the articulatory loop: An extended network model. In J. Levy, D. Bairaktaris, J Bullinaria, & D. Cairns (Eds.), *Connectionist models of memory and language*. London: UCL Press.

Gordon, P. C., & Meyer, D. E. (1987). Control of serial order in rapidly spoken syllable sequences. *Journal of Memory and Language, 26*, 300-321.

Gupta, P., & MacWhinney, B. (1995). Is the articulatory loop articulatory or auditory? Reexamining the effects of concurrent articulation on immediate serial recall. *Journal of Memory and Language, 34*, 63-88.

Hartley, T., & Houghton, G. (1996). A linguistically constrained model of short-term memory for non-words. *Journal of Memory and Language, 35*, 1-31.

Hodgeman, C. D. (1941). *Mathematical tables from Handbook of chemistry and physics, Seventh Edition*. Cleveland: Chemical Rubber Publishing Company.

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1217-1232.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122-149.

Kieras, D. E., Meyer, D. E., Mueller, S., & Seymour, T. (1999). Insights into working memory from the perspective of the EPIC architecture for modeling skilled perceptual-motor and cognitive human performance. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 183-223). New York: Cambridge University Press.

Kuçera, H. & Francis, W. N. (1967). Computational analysis of present-day American English. Providence, RI: Brown University Press.

Landauer, T. K. (1962). The rate of implicit speech. *Perceptual and Motor Skills, 15*, 646.

Laughery, K. R., & Pinkus, A. L. (1970). A simulation model of short-term memory: Parameter sensitivity studies and implications for two current issues. *Journal of Mathematical Psychology, 7*, 554-571.

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

Lewandowsky, S., & Murdock, B. B. (1989). Memory for serial order. *Psychological Review, 96*, 25-57.

Logie, R. H., Della Sala, S., Laiacona, M., Chalmers, P. (1996). Group aggregates and individual reliability: The case of verbal short-term memory. *Memory & Cognition, 24*, 305-321.

Longoni, A. M., Richardson, A. 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)*, 1-22.

McClellend, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1-86.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models, 2nd edition*. London: Chapman & Hall.

Meyer, D. E., & Gordon, P. C. (1985). Speech production: Motor programming of phonetic features. *Journal of Memory and Language, 24*, 3-26.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits to our capacity for processing information. *Psychological Review, 63*, 81-97.

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

Murdock, B. B. (1974). *Human memory: Theory and data*. Potomac, MD: Lawrence Earlbaum Associates.

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review, 100*, 183-203.

Murray, D. J. (1967). The role of speech responses in short-term memory. *Canadian Journal of Psychology, 21*, 263-276.

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

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition, 18*, 251-269.

Norris, D. (1994). SHORTLIST: A connectionist model of continuous speech recognition. *Cognition, 52*, 189-234.

O'Grady, W., & Dobrovrolsky, M. (1992). *Contemporary linguistic analysis*, 2nd *ed.* Toronto: Copp Clark Pittman Ltd.

Page, M. P. A., & Norris, D. (1998a). The primacy model: A new model of immediate serial recall. *Psychological Review, 105*, 761-768.

Page, M., & Norris, D. (1998b). Modeling immediate serial recall with a localist implementation of the primacy model. In J. Grainger & A. M. Jacobs (Eds.) *Localist connectionist approaches to human cognition*, (pp. 227-255), Mahwah, New Jersey: Lawrence Erlbaum Associates.

Posner, M. I., & Konick, A. F. (1966). On the role of interference in short-term memory. Journal of Experimental Psychology, 72, 221-231.

Rosenbaum, D. A., Gordon, A. M., Stillings, N. A., & Feinstein, M. H. (1987). Stimulus-response compatibility in the programming of speech. *Memory & Cognition, 15*, 217-224.

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 and Verbal Behavior, 21*, 150-164.

Schweickert, R., & Boruff, B. (1986). Short-term memory capacity: Magic number or magic spell? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 419-425.

Schweickert, R., Guentert, L., & Hersberger, L. (1990). Phonological similarity, pronunciation rate, and memory span. *Psychological Science, 1*, 74-77.

Service, E. (1998). The effect of word length on immediate serial recall depends on phonological complexity, not articulatory duration. *Quarterly Journal of Experimental Psychology: Human 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, Elliot, and Geer. *Quarterly Journal of Experimental Psychology, 53A*, 661-665.

Shallice, T., & Warrington, E. K. (1970). Independent functioning of verbal memory stores: A neuropsychological study. *Quarterly Journal of Experimental Psychology, 22*, 261-273.

Sperling, G. (1967). Successive approximations to a model for short-term memory. *Acta Psychologica, 27*, 285-292.

Standing, L., Bond, B., Smith, P., & Isely, C. (1980). Is the immediate memory span determined by subvocalization rate? *British Journal of Psychology, 71*, 525-539.

Standing, L., & Curtis, L. (1989). Subvocalization rate versus other predictors of the memory span. *Psychological Reports, 65*, 487-495.

Sternberg, S., Monsell, S., Knoll, R. L., & Wright, C. E. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In G. E. Stelmach (Ed.), *Information processing in motor control* (pp. 117-152). New York: Academic Press.

Treiman, R., & Zukowski, A. (1990). Toward an understanding of English syllabification. *Journal of Memory and Language, 29*, 66-85.

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

Vitz, P. C., & Winkler, B. S. (1973). Predicting the judged "Similarity of sound" of English words. *Journal of Verbal Learning and Verbal Behavior, 12*, 373-388.

Waters, G. S., Rochon, E., & Caplan, D. (1992). The role of high-level speech planning in rehearsal: Evidence from patients with apraxia of speech. *Journal of Memory and Language, 31*, 54-73.

Waugh, N. C., & Norman, D. A. (1965). Primary memory. *Psychological Review, 72*, 89-104.

Wickelgren, W. A. (1965). Distinctive features and errors in short-term memory for English vowels. *Journal of the Acoustical Society of America, 38*, 583-588.

Wickelgren, W. A. (1966). Distinctive features and errors in short-term memory for English consonants. *Journal of the Acoustical Society of America, 39*, 388-398.

Wilson, M. (1987). *The MRC psycholinguistic database machine usable dictionary: Version 2.00.*

Appendix

A1: Representation of phonemes

To measure phonological dissimilarity for a pair of words, PSIMETRICA represents each word in terms of its phonemes. Each phoneme is decomposed into values that it has for the following features: (a) vocalic, (b) consonantal, (c) high, (d) back, (e) low, (f) anterior, (g) coronal, (h) round, (i) tense, (j) voice, (k) continuant, (l) nasal, (m) strident. Vowels do not have values for the voice, continuant, nasal, and strident features; some consonants do not have values for the round and tense features. The following table shows the phonological-feature values for a standard set of phonemes as described by Clark and Clark (1977, p. 187) and Chomsky and Halle (1968).

phoneme	a	b	c	d	e	f	g	h	i	j	k	l	m
/i/	+	-	+	-	-	-	-	-	+	x	x	x	x
/ɪ/	+	-	+	-	-	-	-	-	-	x	x	x	x
/e/	+	-	-	-	-	-	-	-	+	x	x	x	x
/ɛ/	+	-	-	-	-	-	-	-	-	x	x	x	x
/æ/	+	-	-	-	+	-	-	-	+	x	x	x	x
/ɨ/	+	-	+	+	-	-	-	-	-	x	x	x	x
/ə/	+	-	-	+	-	-	-	-	-	x	x	x	x
/ʌ/	+	-	-	+	+	-	-	-	-	x	x	x	x
/a/	+	-	-	+	+	-	-	-	+	x	x	x	x
/u/	+	-	+	+	-	-	-	+	+	x	x	x	x
/ʊ/	+	-	+	+	-	-	-	+	-	x	x	x	x
/o/	+	-	-	+	-	-	-	+	+	x	x	x	x
/ɔ/	+	-	-	+	+	-	-	+	-	x	x	x	x
/y/	-	-	+	-	-	-	-	-	-	x	x	x	x
/w/	-	-	+	+	-	-	-	+	-	x	x	x	x
/r/	+	+	-	-	-	-	+	x	x	+	+	-	-
/l/	+	+	-	-	-	+	+	x	x	+	+	-	-
/h/	-	-	-	-	+	-	-	x	x	-	+	-	-
/p/	-	+	-	-	-	+	-	x	x	-	-	-	-
/b/	-	+	-	-	-	+	-	x	x	+	-	-	-
/t/	-	+	-	-	-	+	+	x	x	-	-	-	-
/d/	-	+	-	-	-	+	+	x	x	+	-	-	-
/č/	-	+	+	-	-	-	+	x	x	-	-	-	+
/ǰ/	-	+	+	-	-	-	+	x	x	+	-	-	+
/k/	-	+	+	+	-	-	-	-	x	-	-	-	-
/g/	-	+	+	+	-	-	-	-	x	+	-	-	-
/f/	-	+	-	-	-	+	-	x	x	-	+	-	+
/v/	-	+	-	-	-	+	-	x	x	+	+	-	+
/θ/	-	+	-	-	-	+	+	x	x	-	+	-	-
/ð/	-	+	-	-	-	+	+	x	x	+	+	-	-
/s/	-	+	-	-	-	+	+	x	x	-	+	-	+
/z/	-	+	-	-	-	+	+	x	x	+	+	-	+
/ʃ/	-	+	+	-	-	-	+	x	x	-	+	-	+
/ʒ/	-	+	+	-	-	-	+	x	x	+	+	-	+
/m/	-	+	-	-	-	+	-	x	x	+	-	+	-
/n/	-	+	-	-	-	+	+	x	x	+	-	+	-
/ŋ/	-	+	+	+	-	-	-	x	x	+	-	+	-

A2: Analysis for estimation of mean articulatory durations

The mean articulatory duration of word sequences tends to be a concave-upward function of sequence length (cf. Sternberg et al., 1978). It may also contain contributions of speech preparation and recording biases that are independent of word set and sequence length. Consequently, raw sequence articulation times should be analysed with a mathematical model that takes these potential contributions into account.

In the present experiments, sequences of between 2 and 5 words were repeated twice from memory for multiple word sets (indexed by j) and multiple participants (indexed by k). From the results of this procedure, we approximate the total articulation time on each trial by

$$T = I_k + 2 \cdot n \cdot d_{jk} \cdot a_{jk}^{n-2}. \tag{2}$$

Here I is an intercept parameter, n is the sequence length, d is a base articulatory duration per word, and a is a duration-amplification factor. For each combination of word set and participant, our analysis yields an estimate of the mean sequence duration based on I, d, a, and n.

In this analysis, we assume that I depends on neither n nor the word set from which a sequence has been constructed. However, I may vary across participants and includes temporal overhead due to sequence repetition as well as any biases contributed by the apparatus or experimenter to the measurement procedure. The base articulatory duration per word, d, depends on the words from which a sequence has been constructed. Also, on the right side of Equation 2, the product $2 \cdot n \cdot d_{jk} \cdot a_{jk}^{n-2}$ embodies contributions due to the n-word sequence being repeated twice and the duration per word increasing with n. As part of this product, the amplification factor a_{jk} is raised to the $n-2$ power because two-word sequences are the shortest ones for which Equation 2 may hold.[19]

Given the observed total articulation times, the accuracy of Equation 2 can be assessed by calculating a deviation score, δ_{ijk}, for each articulation trial i performed by participant j with a sequence of words constructed from word set k. In the calculation,

$$\delta_{ijk} = \frac{(t_{ijk,observed} - t_{ijk,predicted})}{n_{ijk}} \tag{3}$$

where n_{ijk} is the sequence length, and $t_{ijk,predicted}$ is based on Equation 2 . Values of $t_{ijk,predicted}$ can be derived by minimizing $\sum_{i,j,k}(\delta_{ijk})^2$, which yields a single intercept parameter I_j for each participant j, as well as an articulatory duration parameter d_{jk} and an amplification parameter a_{jk} for each word set k encountered by participant j. At this point, outliers among the values of $t_{ijk,observed}$ can be removed by excluding trials whose values of δ_{ijk} differ by more than some number of (e.g., 2.5) standard deviations from participant j's mean δ_{ijk}. On the basis of the parameters a_{jk} and d_{jk}, the mean articulatory duration per word in memorized sequences can be estimated for each word set. This involves two further steps. First, for each word set k with which participant j performs, the mean articulatory duration per word, D_{jk}, is estimated as

$$D_{jk} = \sum_{n=2}^{5} (d_{jk} \cdot a_{jk}^{n-2})/4, \tag{4}$$

where d_{jk} is the base word duration, and a_{jk} is the duration-amplification parameter in Equation 2 when participant j and

[19] In contrast to our current analysis, Sternberg et al. (1978) concluded that total articulation times for memorized word sequences could be fit well by a quadratic function of sequence length. However, we have found that an exponential function (i.e., Equation 2) provides more reasonable and interpretable parameter values under the conditions of our experiments, and it fits our data slightly better than does a quadratic function. Yet under other circumstances, given that $1 + n \cdot a \approx (1+a)^n$ when a is small (Hodgeman, 1941), it would be difficult to distinguish Equation 2 empirically from a quadratic function.

word set k are involved.

Second, for each word set k, the values of D_{jk} from Equation 4 are averaged across participants, yielding D_k, the mean articulatory duration per word in sequences constructed with this word set. These final D_k values are then taken into account as part of assessing articulatory-duration effects on the memory-span data.

Equation 4 is used to estimate D_{jk} because during presentation of to-be-recalled words in the serial recall task, rehearsal of short subsequences presumably precedes rehearsal of longer subsequences (Kieras et al., 1999). Also, by using Equation 4, we take into account that the articulatory duration per word increases with the length of the subsequences being rehearsed successively. Equation 4 omits the intercept of Equation 2 because I is presumably independent of the individual word durations and only embodies contributions from ancillary sources that do not vary systematically as a function of sequence length.

For purposes of interpreting the results from this subsequent assessment, it should be stressed that using Equation 2 to help measure the mean articulatory durations of words in memorized sequences has some especially significant benefits. Doing so enables us remove apparatus and experimenter biases that may contribute to the observed total articulation times but that should be excluded from estimates of the mean articulatory durations for words in memorized sequences. Consequently, regardless of whether these durations come from sources such as manual stopwatch timing or examination of acoustic waveform records, they may be reasonably veridical insofar as they conform to Equation 2 and are quantified appropriately in terms of its parameters, (e.g., see Footnote 17).

A3: Word sets used in Experiments 1 and 2

Set 1 (“long”)		Set 2 (“short”)		Set 3 (“difficult”)	
Word	Phonemic Transcription	Word	Phonemic Transcription	Word	Phonemic Transcription
crayon	/ˈkre yən/	carrot	/ˈke rət/	crow	/ˈkro/
vacuum	/ˈvæ kium/	bullet	/ˈbʊ lət/	stew	/ˈstu/
sirloin	/ˈsər loin/	ladder	/ˈlæ dər/	glue	/ˈglu/
spider	/ˈspai dər/	devil	/ˈdɛ vəl/	tree	/ˈtri/
balloon	/bəˈlun/	picnic	/ˈpɪk nɪk/	tray	/ˈtre/
baby	/ˈbei bi/	ticket	/ˈtɪ kɪt/	draw	/ˈdra/
tower	/ˈtau wər/	zipper	/ˈzɪ pər/	snow	/ˈsno/
orange	/ˈoʊ rənj/	cabin	/ˈkæ bən/		

Set 4 (“easy”)		Set 5 (“short”)		Set 6 (“long”)	
Word	Phonemic Transcription	Word	Phonemic Transcription	Word	Phonemic Transcription
hat	/ˈhæ t/	cult	/ˈkəlt/	advantage	/əd ˈvæ n təj/
book	/ˈbʊk/	dare	/ˈder/	behavior	/biˈ he vyər/
pen	/ˈpɛn/	fate	/ˈfet/	circumstance	/ˈsər kəm stæ ns/
rat	/ˈræ t/	guess	/ˈgɛs/	defiance	/dəˈfai əns/
rug	/ˈrʌg/	hint	/ˈhɪnt/	fantasy	/ˈfæ n tə si/
pig	/ˈpɪg/	mood	/ˈmud/	interim	/ˈɪn tə rəm/
hen	/ˈhɛn/	oath	/ˈoθ/	misery	/ˈmɪ zə ri/
		plea	/ˈpli/	narrowness	/ˈnæ ro nɛs/
		rush	/ˈrʌʃ/	occasion	/ə ˈke ʒən/
		truce	/ˈtrus/	protocol	/ˈpro tə kal/
		verb	/ˈvərb/	ridicule	/ˈri də kyul/
		zeal	/ˈzil/	upheaval	/əp ˈhi vəl/

Set 7 (“simple short”)		Set 8 (“simple long”)		Set 9 (“complex long”)	
Word	Phonemic Transcription	Word	Phonemic Transcription	Word	Phonemic Transcription
basement	ˈbes mənt	control	kən ˈtrol	accident	ˈæ k sə dənt
conscience	ˈkan ʃəns	dispute	dɪ ˈspiut	clarinet	kle rəˈ nɛt
discharge	ˈdɪs čarg	himself	him ˈsɛlf	discipline	ˈdɪ sə plɪn
entrance	ˈɛn trəns	imprint	ˈɪm prɪnt	exception	ək ˈsɛp ʃən
falsehood	ˈfals hʊd	junction	ˈjʌŋk ʃən	hexagon	ˈhɛk sə gan
household	ˈhaus hold	mixture	ˈmɪks čər	industry	ˈɪn də stri
grievance	ˈgri vəns	prefix	ˈpri fɪks	protocol	ˈpro tə kal
mistress	ˈmɪs trəs	respect	rəˈspɛkt	ridicule	ˈrɪ də kyul
sheepskin	ˈʃip skɪn	stipend	ˈstai pɛnd	specimen	ˈspɛ sə mən
traction	ˈtræ k ʃən	trumpet	ˈtrʌm pət	tradition	trəˈ dɪ ʃən