Psychonomic Bulletin & Review
2008, 15 (5), 875-888
doi: 10.3758/PBR.15.5.875

Theoretical and Review Articles

The word-length effect provides no evidence for decay in short-term memory

Stephan Lewandowsky
University of Western Australia, Crawley, West Australia, Australia

and

Klaus Oberauer
University of Bristol, Bristol, England

Many researchers regard the word-length effect (WLE) as one of the strongest pieces of evidence for time-based decay in short-term memory. We argue that the WLE is, in fact, undiagnostic for the decay hypothesis for two reasons. First, the WLE represents a correlation across words between articulation duration and memory performance, and articulation duration is inevitably confounded with other word characteristics. Recent research has confirmed that such confounds are responsible for much, maybe all, of the WLE. Second, there is strong evidence for an attentional mechanism of refreshing memory traces that can operate concurrently with articulation. Any viable decay-based model must include such a mechanism, but such a model no longer necessarily predicts a WLE, because longer spoken duration does not imply longer postponement of refreshing. We conclude that the WLE is not diagnostic for decay in short-term memory.

There has been much renewed interest in the causes underlying forgetting in short-term memory, with some theorists proposing that it results from inexorable temporal decay of memory representations (e.g., Page & Norris, 1998) and with others negating any role of time, preferring instead to ascribe forgetting to nontemporal processes, such as interference (e.g., Farrell & Lewandowsky, 2002; Lewandowsky & Farrell, 2008). Although there are several sources of evidence that can constrain these competing theories, decay theorists inevitably cite the word-length effect (WLE) as a crucial piece of evidence (e.g., Cowan, 1995).

The WLE, first reported by Baddeley, Thomson, and Buchanan (1975), refers to the now well-established finding that lists composed of long words (e.g., *hippopotamus*, *retromingent*) are recalled less accurately than lists of short words (e.g., *rat*, *cut*, *hip*). Empirical interest in the effect has continued unabated to this date (e.g., Mueller, Seymour, Kieras, & Meyer, 2003). The poorer memory for long words than for short words has been taken to reflect the fact that verbal memory traces inexorably decay over time, and that fewer long words than short words can be recalled or refreshed by rehearsal (if that is possible) in the limited time before decay has rendered the trace irrecoverable. The WLE has often been discussed in the context of the phonological loop model (e.g., Baddeley, 1986), which explains the WLE as a race between decay and articulatory rehearsal. Other models (e.g., Schweickert & Boruff, 1986) have explained the WLE as resulting from a race between decay and overt recall.

A particularly elegant attribute of the original WLE was that articulation durations were typically determined outside the experimental context in which memory was tested. For example, people were timed either while articulating single words in isolation or while reading a list of words, and the times were used to predict memory performance of the same people on a different occasion (or the performance of a different set of participants; see, e.g., Lovatt, Avons, & Masterson, 2002, Experiments 1A and 1B). This counterintuitive ability to predict memory from afar, on the basis of a seemingly inconsequential property of linguistic material, has enhanced the impact of the effect and its underlying theorizing.

The idea of a race between decay and rehearsal or recall is elegant and parsimonious, and the WLE has been identified as "perhaps the best remaining solid evidence in favor of temporary memory storage" (Cowan, 1995, p. 42). Nevertheless, we argue in this article that although the WLE may be of considerable interest in its own right, it is of limited value in identifying the mechanisms underlying forgetting in short-term memory. To foreshadow our principal conclusion, we propose that the WLE ought to no longer be relied upon to constrain contemporary theorizing about the sources of short-term forgetting.

S. Lewandowsky, lewan@psy.uwa.edu.au

To formalize and structure our argument, we cast it in Bayesian terms. *Bayesian diagnosticity* is defined as the likelihood ratio of two mutually exclusive and jointly exhaustive hypotheses (we restrict consideration to binary pairs of hypotheses). *Likelihood* refers to the conditional probability of an observation, given a hypothesis. An observation is evidence in favor of a hypothesis to the degree that its likelihood under the hypothesis exceeds its likelihood under the alternative hypothesis. Applied to the WLE as potential evidence for the decay hypothesis,

$$\text{Diagnosticity} = P(\text{WLE} \mid \text{decay})/P(\text{WLE} \mid \text{no decay}).$$

It follows that for the WLE to strengthen the decay hypothesis, the probability of occurrence of the WLE under the decay hypothesis must be greater than the probability of the WLE under the alternative hypothesis that there is no decay:

$$P(\text{WLE} \mid \text{decay}) > P(\text{WLE} \mid \text{no decay}).$$

To date, proponents of the decay-based interpretation of the WLE have considered this crucial inequality to be firmly established on the basis of two considerations. First, there appears to be no doubt that a decay-based model necessarily gives rise to a WLE, and second, the WLE does not appear to fall out of non-decay-based models without additional assumptions.

We will argue that there are no good reasons to assume this inequality. Our argument is twofold. First, we show that the WLE is inherently correlational in nature and therefore inevitably confounded with other variables. In the few instances in which these confounds have been brought under control, articulation duration had no bearing on memory performance (see, e.g., Service, 1998). We show that these considerations apply even to recent attempts to refocus the WLE to a purportedly more appropriate measurement of articulation duration (e.g., Mueller et al., 2003). It follows that there are, and always will be, alternative explanations for the WLE besides decay. In more formal terms, $P(\text{WLE} \mid \text{no decay})$ is certainly greater than zero, and we do not know how much greater unless we understand the effect of all the factors that are confounded with word length. Figure 1 provides a road map for the first part of our argument.

Second, in light of recent evidence for the existence of an attentional mechanism that can refresh verbal short-term memory traces concurrently with overt articulation (e.g., Hudjetz & Oberauer, 2007; Raye, Johnson, Mitchell, Greene, & Johnson, 2007), the WLE is no longer a necessary prediction following from the decay hypothesis. Indeed, whether or not a decay-based model predicts the WLE depends entirely on exactly how articulatory rehearsal and attentional refreshing are coordinated in that model. In the absence of any certainty that a viable decay-based model would predict the WLE at all, there is no reason to postulate that $P(\text{WLE} \mid \text{decay}) > P(\text{WLE} \mid \text{no decay})$. Less technically, we acknowledge that the WLE exists, but we conclude that it tells us nothing about whether forgetting is due to decay or to some other factor, such as interference.

Because this conclusion may be theoretically controversial, it is also important to clarify what it does *not* imply. We do not imply that the WLE does not exist. We do not imply that the WLE is uninteresting or trivial. We do not imply that theories of short-term memory should

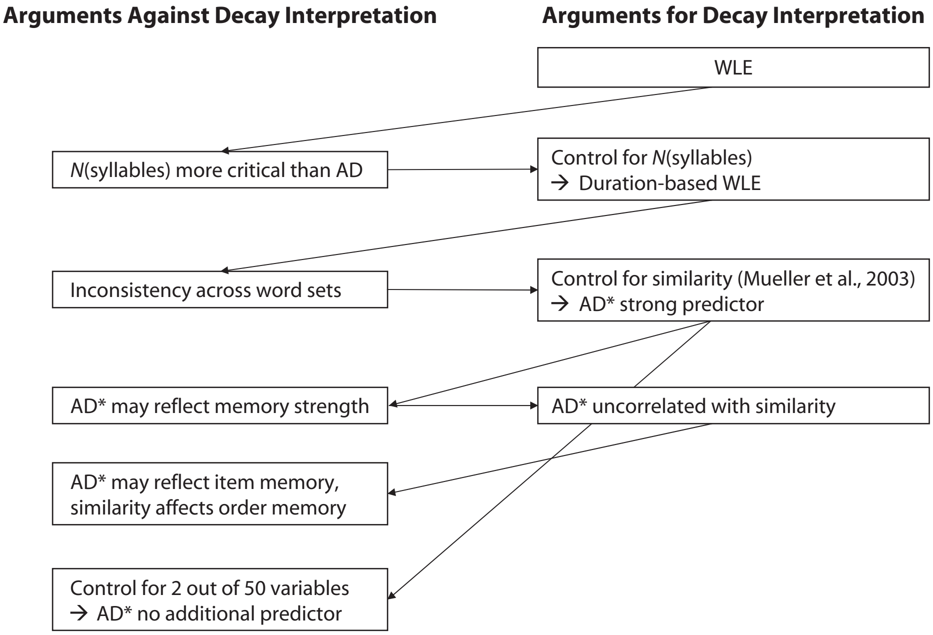
Arguments Against Decay Interpretation
Arguments for Decay Interpretation
WLE
N(syllables) more critical than AD
Control for N(syllables) → Duration-based WLE
Inconsistency across word sets
Control for similarity (Mueller et al., 2003) → AD* strong predictor
AD* may reflect memory strength
AD* uncorrelated with similarity
AD* may reflect item memory, similarity affects order memory
Control for 2 out of 50 variables → AD* no additional predictor

Figure 1. Road map of the argument that establishes the inescapability of confounds due to the correlational nature of the word-length effect. AD, articulation duration measured from reading speed; AD*, articulation duration as estimated from memorized speech by Mueller, Seymour, Kieras, and Meyer (2003). See the text for details.

ignore the WLE; on the contrary, a complete account of short-term memory must explain why, at least under certain circumstances, people's memory is worse for long words than it is for short words. Thus, we acknowledge that the WLE might tell us something about how words are encoded into working memory or about the role of language production for recall of verbal material. What we do conclude, however, is that the WLE is of no relevance to settling the issue of what processes explain forgetting from short-term memory. We therefore wrap up the article by pointing to existing methodological alternatives that present a more fruitful avenue for empirical exploration of the mechanisms underlying forgetting.

Before we present our arguments, we acknowledge that others have previously leveled criticisms against the WLE. For example, Nairne (2002) provided an incisive critique of much of the relevant research and its theoretical impact, and Brown and Hulme (1995), Lewandowsky and Farrell (2000), Neath and Brown (2006), and Neath and Nairne (1995) have provided computational implementations of the WLE in models that did not rely on decay. Our analysis, however, differs from those precedents because, although we rely on data or simulations where appropriate, our core arguments are conceptual, rather than based on data or predictions. Thus, instead of calling for resolution of outstanding empirical issues, we argue that there is no dearth of knowledge about the WLE and that no additional amount of knowledge can imbue the effect with the desired theoretical diagnosticity.

Correlational Effects Are Just That

There can be little doubt that the WLE is inherently correlational in nature: The experimenter selects items that differ in presumed articulation time; empirical measurements of the actual articulation times are obtained; and those are then used to predict memory performance, usually characterized by a *span* measure (e.g., the longest list that people can recall in perfect order half the time). Hence, one dependent variable is used to predict another, and the experimenter has no direct control over what causes the variability on one of the measures—namely, articulation time. Short and long words inevitably differ on many attributes other than their articulation duration, thus creating confounds with the variable of interest—namely, time in between rehearsals of a given item.[1]

Stated in this manner, the correlational nature of the WLE appears inescapable and hardly worthy of detailed examination. Indeed, one may wonder how a correlational effect could have attracted as much attention as the WLE has in the first place. However, as we show next, it is precisely its correlational nature that has—ironically—enabled proponents of the decay interpretation of the WLE to respond to repeated criticisms, notwithstanding the fact that those criticisms targeted the very same correlational problem. Indeed, we argue that some recent attempts to buttress the diagnosticity of the WLE have amplified, rather than resolved, the correlational problem. We therefore believe that a thorough conceptual analysis of the WLE is now required. To guide our analysis, we identify two cycles of criticisms that have been leveled against the decay interpretation of the WLE, each of which has engendered a rejoinder and a conceptual realignment by proponents of that interpretation. In Figure 1, these two cycles correspond to the first two boxes in the left-hand column, and the associated rejoinders by proponents of a decay interpretation are in the right-hand column.

Complexity versus duration. The first cycle of criticisms explored the confound between articulation duration and syllabic complexity. This criticism was anticipated and addressed by Baddeley et al. (1975) when they first established the WLE: Longer words typically have more syllables, and therefore, articulation duration is often confounded with phonological complexity (as indicated by syllabic length). Baddeley et al. therefore also contrasted two sets of bisyllabic words that differed in their spoken duration. Notwithstanding their identical syllabic complexity, words with shorter durations (such as *bishop* or *wicket*) were recalled more accurately than words with longer durations (such as *Friday* or *harpoon*). This finding indicated that there is not only a syllable-based WLE arising from words that differ in syllabic length (e.g., *hippopotamus* vs. *bat*), but also a purely duration-based WLE arising from words equated for syllabic length. We now examine in turn the criticisms that have been leveled at both manifestations of the WLE.

There is no doubt that the syllable-based WLE is robust and replicable (e.g., Caplan, Rochon, & Waters, 1992, Experiment 1), but it is now of reduced theoretical interest because the syllabic complexity of the material, rather than the passage of time, has been identified as the critical underlying variable. This was most convincingly demonstrated by Service (1998), who exploited the fact that in Finnish, words can contain single or double letters, and the change from single to double letters keeps the words identical in all respects other than their pronunciation duration. Service created a set of Finnish pseudowords that orthogonally varied in phonological complexity (i.e., number of syllables) and in pronunciation duration (i.e., single or double letters). Syllabic complexity was found to have a large effect on recall, whereas articulation duration had none.[2]

Tolan and Tehan (2005) provided a conceptual replication of Service's (1998) finding, using English words rather than Finnish pseudowords. In each of two experiments, Tolan and Tehan orthogonally compared words differing in complexity (one vs. two syllables) and pronunciation duration (short vs. long) and found that complexity, but not duration, affected serial recall (Experiment 1) and serial reconstruction (Experiment 2).

Similarly, Murray and Jones (2002) reexamined the hitherto widely accepted fact that memory span for digits spoken in Welsh is lower than the span for English digits (in the same bilingual subjects). This effect was first reported by Ellis and Hennelly (1980) and was ascribed to the longer articulation durations of Welsh digits in comparison with English digits. Murray and Jones showed that the effect is, instead, best understood as resulting from articulatory complexity at the boundary between consecutive digits on a list. Their analysis rested on two findings: First, Welsh digits were found to take longer to read than

English digits only when presented in a list, not when presented in isolation, and second, articulatory complexity at the boundary between English words was shown to affect memory span.

Finally, at a theoretical level, Brown and Hulme (1995), Lewandowsky and Farrell (2000), and Neath and Nairne (1995) showed that neither decay nor rehearsal is a theoretical construct necessary to account for the syllable-based WLE. These authors presented computational models that were based on a *segmented* representation of words and that reproduced the WLE without decay or rehearsal. The number of segments (or feature values, in the case of Lewandowsky & Farrell, 2000) was assumed to be proportional to the articulation duration of a word. These models might not be adequate explanations of the duration-based WLE, because one might question why words with an equal number of phonemes and syllables are represented by different numbers of segments. Words differing in syllabic length, however, can legitimately be represented by different numbers of segments, and, therefore, these three computational models can account for the syllable-based WLE without invoking the concept of decay. The existence of an alternative explanation establishes $P(\text{WLE} \mid \text{no decay})$ to be greater than zero, thus confirming the limited diagnosticity of the syllable-based WLE. (Neath & Brown, 2006, provided another computational instantiation of the WLE not involving a temporal explanation.)

In our view, this first cycle of criticism largely succeeded in identifying the syllable-based WLE as the result of phonological complexity, rather than time (see first argument on the left of Figure 1). Therefore, the linkage between the WLE and decay now hinges on the duration-based WLE (i.e., when words with the same number of syllables, but differing pronunciation durations, are compared; e.g., *platoon* vs. *racket*). Baddeley et al. (1975) were the first to report a memorial advantage for the shorter words, and their finding has been replicated repeatedly (Cowan et al., 1992; Longoni, Richardson, & Aiello, 1993; Lovatt, Avons, & Masterson, 2000; Nairne, Neath, & Serra, 1997). Nonetheless, the interpretation of the duration-based WLE has come under scrutiny during the second, more recent cycle of criticisms (see second argument on the left in Figure 1).

Duration versus stimulus selection. These criticisms have primarily focused on the sensitivity of the duration-based WLE to the selection of stimuli. Caplan et al. (1992) reported a reversal of the usual WLE, with long words being recalled better than short words from a set of disyllabic stimuli. Similarly, Lovatt et al. (2000) reported three experiments using disyllabic stimuli that variously found a reversed WLE (long words recalled better than short words), a null effect, or the expected advantage for short over long words. The expected duration-based WLE was found only for the exact set of materials used by Baddeley et al. (1975), whereas two additional sets of stimuli, selected according to the same criteria, gave rise to the opposing outcome or to a null effect. Lovatt et al. (2000) concluded that "there is no general effect of word duration on disyllabic-word recall, and . . . the differences originally observed arose as an accident of item selection" (p. 15).

In a follow-up study, Lovatt et al. (2002) replicated the sensitivity of the WLE to the particular set of stimuli (we consider other implications of this study later). Finally, Neath, Bireta, and Surprenant (2003) compared the stimuli used by Baddeley et al. (1975), Caplan et al. (1992), and Lovatt et al. (2000) within an identical methodology and replicated the precise pattern of variation of the WLE across stimuli. Like Lovatt et al. (2000), Neath et al. reported a duration-based WLE for Baddeley et al.'s stimuli, a reverse WLE with Caplan et al.'s stimuli, and a null effect for the words used by Lovatt et al. (2000), and also for a novel set of stimuli. The fact that the various outcomes for the different stimulus sets are all replicable is important, because it rules out the possibility that the occasional presence of the duration-based WLE is simply the result of a small effect suffering from a low signal-to-noise ratio. Instead, the replicability of the opposing outcomes strengthens the hypothesis that the duration-based WLE results from item selection artifacts.

Lest one think that item selection artifacts are rare, one must note that the conclusions by Lovatt et al. (2000, 2002) and Neath et al. (2003) do not represent an isolated occurrence: Within the word length arena, Bireta, Neath, and Surprenant (2006) showed that stimulus specificity also underlies certain manifestations of the syllable-based WLE. Specifically, Bireta et al. showed that differences in recall between short and long words that co-occur on mixed lists are tied entirely to the particular set of stimuli being used, thus impairing meaningful generalizations.

How might one respond to this rather alarming preponderance of the item specificity of the WLE? One possible response is to increase the samples of stimuli; rather than relying on small samples of short and long disyllabic words, researchers could draw large samples—in the limit, even highly representative ones—from all English words and let the law of large numbers average out all confounding factors. A relevant precedent can be found in the study by Tolan and Tehan (2005), which sampled study lists from an open pool of nearly 250 pairs of words differing in the duration of their vowel sound (e.g., *cut* and *cute*). As mentioned earlier, the study found no effect of articulation duration on memory. Although large-scale sampling reduces the impact of sampling error in compiling the material, it does not address the problem that articulation duration is likely to be correlated with a number of other linguistic attributes of the population of all possible words. That is, even with large, representative samples of words as stimuli, the WLE remains correlational, thereby precluding firm inferences about causality.

Another possible response is to try to identify the critical variable or ensemble of variables that distinguishes word sets that generate a duration-based WLE from those that generate no effect or even a negative WLE. This endeavor might point to confounding variables in some or all word sets, which could be controlled to achieve a purer measure of the duration-based WLE (e.g., Baddeley & Andrade, 1994). The latest and most elaborate effort along these lines is the work of Mueller et al. (2003). We therefore now present an analysis of their article.

Pushing Correlational Analysis to the Limits: A Case Study

In recognition of the pervasive effect of phonological similarity on memory (see, e.g., Baddeley, 1966; Conrad, 1964), Mueller et al. (2003) compared the phonological similarity among words within nine different sets, including some that had been used in previous studies to demonstrate the duration-based WLE (see the third box on the right-hand side of Figure 1). Mueller et al. developed a new, rather sophisticated algorithm to compare the phonological features of words, as well as a new procedure for measuring articulation duration. The latter is of particular interest here: Rather than read words aloud individually or in lists, participants were asked to commit lists of two to five words to memory and, when ready, to articulate them twice, rapidly and accurately. A nonlinear function was used to estimate mean articulation duration from these data on the basis of two parameters, thus replacing direct measurement of articulation speed by an estimation based on a measurement model.

Mueller et al. (2003) found that phonological similarity and estimated articulation duration predicted memory span across the nine word sets with great precision, whereas phonological complexity did not account for any additional variance (despite words varying in the number of syllables). Mueller et al. concluded that after phonological similarity was controlled, articulatory duration—but not phonological complexity—is a determinant of short-term memory performance, exactly as predicted by a decay-based theory such as the phonological loop model. The article by Mueller et al. represents a heroic effort to disentangle several features of words that potentially affect short-term memory. Indeed, it is impressive that their analysis was able to reconcile numerous previously conflicting outcomes by reexamining the various stimuli used by Caplan et al. (1992) and by comparing them with a number of novel sets of stimuli. Nonetheless, we believe that the efforts of Mueller et al. were insufficient to establish the theoretical diagnosticity of the WLE for three principal reasons.

Justification does not imply validity. First, their argument relies on the proposition that there is a single correct way of measuring phonological complexity, articulation duration, and phonological similarity. In the words of Mueller et al. (2003), "Experimenters have measured articulatory duration with five distinct methods whose rationale remains unclear. None of these measurements may be adequate for testing the model's predictions about serial recall accuracy" (p. 1357). Mueller et al. deserve credit for identifying the problem and for justifying their own measure explicitly; however, it does not necessarily follow that their measure is appropriate. This is best illustrated by supposing that articulation duration as measured by Mueller et al. had turned out not to be a good predictor of memory span. In that case, what would keep proponents of the decay interpretation of the WLE from arguing that yet another measure of articulation duration would have been a more appropriate measure of rehearsal time? For example, one could argue that overt articulation takes longer than covert articulation, and that therefore the best way to measure articulation duration would be to ask participants to silently mouth the words, rather than speak them aloud, or to speak them to themselves and stop the timer when they were finished. This is an in-principle problem: In the absence of any agreement on what constitutes articulation duration, no new measure can definitively be superior to any others.

The same argument holds for phonological similarity: Several alternative measures come to mind that seem as defensible as—or more so than—Mueller et al.'s (2003) algorithm, among them subjective ratings of similarity or the empirical probability of perceptual confusions between words (Conrad, 1964). Indeed, we submit that the latter option may be psychologically more attractive than the algorithmic method of Mueller et al.[3]

Predicting memory from memory. The second, more insidious problem is that Mueller et al. (2003) measured articulation duration in a way that moves the (intended) predictor variable dangerously close to the response variable. Recall that Mueller et al. obtained articulation times by asking participants repeatedly to recite words from memorized lists of varying lengths. In consequence, their principal analysis relies on predicting accuracy in immediate serial recall from speed in immediate serial recall. There is little doubt that speed and accuracy in a memory task are often strongly correlated across experimental conditions (Kahana & Loftus, 1999), suggesting that those two variables typically measure different facets of the same latent variable—namely, memory strength. It follows that the articulation duration of Mueller et al. likely includes variance in how well words can be recalled in correct order from short-term memory. This variance component, in turn, might have contributed to the articulation variable's predictive success.

This argument is buttressed by the fact that the articulation duration of a word, as measured by Mueller et al. (2003), increased with the length of the lists to be articulated from memory. The same observation has been made in studies measuring output timing in serial recall tasks (Cowan, 1992; Tehan & Lalor, 2000). Unlike Mueller et al., these studies distinguished between articulation time and pauses between successive words, and it was found that the duration of pauses, not articulation time, increased with list length. This empirical regularity suggests that the increase of overall articulation duration with list length observed by Mueller et al. likely reflected increasing pause durations, rather than the actual speaking time. Pauses, in turn, have been convincingly linked to retrieval efficiency (e.g., Cowan, 1992; Cowan et al., 1994). It follows that the articulation duration measure of Mueller et al. is likely to reflect, at least in part, variance due to retrieval efficiency (see the third argument on the left in Figure 1).

Thus, Mueller et al. (2003) relinquished the most interesting and valuable aspect of the correlation between articulation duration and memory span—namely, the prediction from afar. A guiding principle of research in individual differences is that correlations are the more surprising, and therefore the more theoretically interesting, the less similar the correlated variables are in terms of irrelevant sources of variance, such as method variance or variance in general cognitive ability. For example, nobody is impressed when one intelligence test predicts another, but the finding that simple reaction time tasks predict intelligence produced much debate (e.g., Jensen, 1982; Longstreth, 1984). In experimental research on working memory, the same logic has been accepted, at least implicitly. For example, in his seminal examination of the phonological similarity effect, Conrad (1964) demonstrated that letters found to be more confusable in an auditory identification task were also confused more frequently in immediate serial recall—including after visual presentation of the list. Had he found that letters that were more confusable in, say, serial recognition were also confused more often in serial recall, nobody would have cared much. Likewise, we venture that a correlation between memory span and the time that it takes to recall items from memory does not necessarily show much beyond the fact that two measures of memory efficiency correlate with each other.

Mueller et al. (2003) were aware of this problem and discussed the possibility that their measure of articulation duration, like memory span, simply reflected memorability of the words in the different sets. Mueller et al. argued against this possibility on the basis of three points, all of which we now rebut.

First, they suggested that a general memorability argument should be rejected because articulation duration was found to be uncorrelated with phonological similarity (see the last box on the right in Figure 1). Their argument was that if articulation duration reflects memorability, then phonological similarity must also correlate with articulation duration, because similarity, too, determines memorability. This argument is unconvincing, because it assumes a single dimension of memorability that mediates the effects of all variables that affect memory span. A more realistic possibility is that there are at least two latent variables of memorability, one reflecting how well a word can be recalled independently of the other words on the list and the other reflecting how likely a word is to be confused with a particular set of other words on the list (cf. the distinction of item memory and order memory; Healy, 1974; see penultimate argument on the left in Figure 1). Obviously, phonological similarity would affect the second memorability variable, and it is entirely plausible that articulation duration, when measured from memory, reflects the first memorability variable.

Second, Mueller et al. (2003) argued that because their measure of articulation duration included only trials on which participants did not hesitate, memorial difficulty did not contribute to their measure. In response, we note that eliminating trials with hesitations can only eliminate extreme cases in the distribution; it cannot alter what the variable reflects.

Mueller et al.'s (2003) third argument was that their measure of articulation duration correlated highly with reading time. A high correlation of these variables across words that contain between one and three syllables is not surprising; it might simply reflect the shared variance due to phonological complexity. The force of the argument for a purely duration-based WLE in Mueller et al.'s data rests on the finding that articulation duration predicts memory span over and above phonological complexity. This additional variance in articulation duration, however, might well be nothing but memorability.

The same conceptual shortcoming associated with predicting one indicator of memory by another applies to other work that is frequently cited in support of decay. For instance, Dosher and Ma (1998) observed that differences in immediate recall performance between materials can be entirely removed when accuracy is plotted as a function of recall time. Striking as it might seem, especially when demonstrated graphically, this finding merely reflects a correlation between recall accuracy and recall duration. The simplest explanation of this correlation is that the two variables are both indicators of the difficulty of recalling the different materials (a possibility acknowledged by Dosher & Ma, 1998, p. 329). Crowder (1976) recognized the correlational nature of this evidence in his discussion of a precursor of Dosher and Ma's finding (viz., Wingfield & Byrnes, 1972).

Confounding variables. The third, and most fundamental, problem illustrated by Mueller et al.'s (2003) work is that their efforts eliminated only two confounding linguistic variables—namely, phonological complexity and phonological similarity. Although those two variables are clearly among the most important ones that are known to determine memory, numerous other features of words play a similar role. For example, memory span for words is affected by their familiarity (Hulme, Maughan, & Brown, 1991), their imageability (Bourassa & Besner, 1994), their phonological neighborhood size and neighborhood frequency (Allen & Hulme, 2006; Roodenrys, Hulme, Lethbridge, Hinton, & Nimmo, 2002), their concreteness (Walker & Hulme, 1999), and the frequency of their constituent bigrams in the language (Thorn & Frankish, 2005). Accordingly, in their examination of the duration-based WLE, which replicated the stimulus-specific, capricious nature of the WLE, Neath et al. (2003) created a new stimulus set that controlled some seven linguistic variables, only two of which (number of phonemes and syllables) were controlled across all stimulus sets by Mueller et al.

Lest one think that control of seven linguistic variables is sufficient, Davis (2005) provided a program that computes nearly 50 word features (spanning about a dozen independent dimensions, such as frequency and neighborhood size, without any claim for that set of variables to be exhaustive) that are thought to be potentially important in the word-reading literature. Although Mueller et al. (2003) controlled two variables that are arguably among the most important ones in that set, this cannot eliminate the fact that several dozen other variables were left uncontrolled in their study.

To confirm this point, we obtained all of Davis's (2005) linguistic indicators that were available for the words used across both experiments by Mueller et al. (2003). We then regressed observed memory span on all possible pairs of those predictors and compared these two-predictor models with the regression model with phonological similarity and articulation duration favored by Mueller et al. The model favored by Mueller et al. accounted for only 1%

more variance (r^2 = .99) than did the next-best model formed by our alternative set of predictors (r^2 = .98). All predicted values of both models fell within the 95% confidence intervals of the data, confirming that the two models were statistically indistinguishable (see the final argument on the left in Figure 1).

Our alternative, nontemporal model included the (logarithm of the) number of phonemes and the neighborhood distribution as predictors. The *neighborhood distribution* refers to the number of letter positions at which orthographic neighbors can be formed (e.g., Mathey, Robert, & Zagar, 2004). For example, the neighbors of the word *probe* are all focused on one position (i.e., *prone*, *prose*, *prove*), whereas the neighbors of *impart* are spread across two positions (i.e., *impact* and *import*). In the word recognition literature, neighborhood distribution has been identified as an important determinant of performance, with lexical decision times generally being facilitated if neighbors are spread over more letter positions (Mathey et al., 2004; Mathey & Zagar, 2000; Robert, Mathey, & Zagar, 2007). Indeed, Pugh, Rexer, and Katz (1994; cited in Pugh, Rexer, Peter, & Katz, 1994) suggested that neighborhood distribution was a more important determinant of lexical decision times than was the number of orthographic neighbors. Similarly, the effects of neighborhood distribution have also been identified as critical in differentiating between rival models of word identification (Mathey et al., 2004).

In the short-term memory arena, examinations of neighborhood effects have so far been confined to comparing words with different numbers of phonological neighbors (Allen & Hulme, 2006; Roodenrys et al., 2002) without considering their distribution, but in light of evidence for a strong link between lexical access measures and memory performance (e.g., Tehan, Fogarty, & Ryan, 2004; Tehan & Lalor, 2000), neighborhood distribution might likewise turn out to be an important determinant of recall performance. Indeed, our reanalysis of the data of Mueller et al. (2003) could be taken to constitute a first step in that regard, should readers feel inclined to adopt that interpretation. (At a theoretical level, it is noteworthy that neighborhood distribution effects in word identification have been successfully modeled by the interactive activation model of McClelland & Rumelhart, 1981; a rough analogue to the processes embodied in that model can be found in dynamic redintegration approaches to short-term memory, such as those proposed by Farrell & Lewandowsky, 2002; Lewandowsky, 1999; and Lewandowsky & Farrell, 2000, 2008.)

In summary, our reanalysis of the data of Mueller et al. (2003) suggests several conclusions. First, in conjunction with phonemic similarity, the estimated articulation duration proposed by Mueller et al. is indeed an excellent predictor of memory performance (perhaps because it is also a measure of memory performance; see our earlier argument). Second, the data can be accommodated by an alternative set of nontemporal predictors—namely, the number of phonemes and the words' orthographic neighborhood distribution, without any statistically discernable loss of precision. Third, both models are theoretically plausible and supported by empirical and theoretical precedent, although Mueller et al.'s model may be considered more attractive because it is based on an a priori prediction of directly relevant decay models (e.g., Baddeley, 1986).

Concerning balance, we do not favor one model over the other: In our view, both are equally correlational and both underscore our principal concern—namely, that the lack of control over verbal material is a problem not of feasibility but of principle. There is no way to be certain at any point that all relevant variables have been controlled. Correlational data, can be only the second best way of evaluating causal hypotheses, and the path of our argument in Figure 1 suggests that all arguments for a decay interpretation of the WLE have, so far, been refuted. We therefore suggest that, wherever possible, we should aim to replace a correlational approach by experimental manipulation of the hypothesized cause. In the case of the decay hypothesis, this cause is the time during which items are held in short-term memory without being rehearsed.

Using Word Length to Manipulate the Retention Interval

Several researchers in the WLE arena have recognized the need for experimental manipulations of time, albeit within the methodological traditions of the WLE. Cowan et al. (1992) manipulated recall delay by independently varying the length of the first and the last three list words. The logic of this design was that beginning recall with three short words would engender less delay for the following three words, regardless of their length. For instance, a list with the structure LLLSSS (where L represents a long word and S a short word) would imply longer delay (in forward oral recall) before the final three words could be retrieved, compared with a list with the structure SSSSSS. Therefore, the same final three words should be recalled worse when preceded by long words (LLLSSS) than when preceded by short words (SSSSSS). Conversely, with backward recall, lists SSSLLL (recalled L→L→L→S→S→S) would result in more recall delay for the three list-initial short words, as compared with lists SSSSSS, and therefore the list-initial words should be recalled less well when the list-final words are long. This is exactly what Cowan et al. (1992) found.

At first glance, this method avoids the problems associated with the correlational nature of the WLE because performance was observable for the same target words after an experimental manipulation of recall delay. Alas, as in any other WLE methodology, the articulation duration of the words recalled first was inevitably confounded with other variables. Therefore, although Cowan et al. (1992) had full control over the characteristics of the last three words recalled, the experimenters could not control the characteristics of the three words recalled first. This is problematic if features of those words—for example, their capacity to generate output interference or their likelihood of being recalled correctly—have effects on the recall of the later words. There is evidence that this was indeed the case. Lovatt et al. (2002) replicated Cowan et al.'s (1992) design with several sets of words. They could replicate the original findings only with the material used by Cowan et al. (1992), not with their own sets of materials, pointing once again to the material-specific nature of WLE results. Moreover, even with the original material from Cowan et al. (1992), the effect of delay on the last-recalled words was eliminated when consideration was restricted to the trials in which the initial words were recalled without errors. Lovatt et al. (2002) concluded that the effect of delay arose not from the duration of speaking the initial words but, rather, from recall errors on the initial words, which had carryover effects on the later target items. Further doubt is cast on Cowan et al.'s (1992) conclusions by the findings of Bireta et al. (2006), who found that for all but one of their stimulus sets, both short and long words on mixed lists were recalled as accurately as words on lists consisting only of short words. This outcome is incommensurate with the notion of decay, because the delay arising from recall of long words on the mixed lists should have lowered performance on subsequent items, in comparison with performance on the short-only lists.

We focused on these studies not for their empirical outcome but to underscore our theoretical argument: At first glance, the preceding studies appear to have experimentally manipulated the retention interval of the later recalled target items. Upon closer inspection, however, the results again come down to a correlation between two list features—namely, the articulation duration of one half of the list and recall accuracy (or some other variable, such as differential interference) of the other half. The method of Cowan et al. (1992) is thus still correlational, and nothing has been gained.

To conclude, all evidence linking word length to recall accuracy is correlational and, therefore, open to alternative explanations that do not involve decay. In the case of the syllabic WLE, this alternative explanation may be phonological complexity. In the case of the duration-based WLE, it may be a common memorability variable that determines both recall and articulation speed, especially if the two are measured in similar circumstances. Alternatively, the duration-based WLE may be caused by an uncontrolled difference between the stimuli. From this existence of alternative explanations follows the first component of our argument: $P(\text{WLE} \mid \text{no decay}) > 0$. Because the extent to which $P(\text{WLE} \mid \text{no decay})$ exceeds zero is unknown, the crucial inequality stated at the outset—namely, that $P(\text{WLE} \mid \text{decay}) > P(\text{WLE} \mid \text{no decay})$—cannot be assumed to hold. This compromises the diagnosticity of the WLE in support of the decay hypothesis.

Should a Decay Model Predict the WLE?

We have shown that the WLE is correlational and therefore open to alternative explanations. Nevertheless, it could be that the likelihood of the WLE is still higher under the hypothesis of decay than under the alternative hypothesis that there is no decay, and in that case, by Bayesian logic, the observation of the WLE counts as support for the decay hypothesis. We now argue against this possibility because it is far from certain that decay models should predict a WLE in the first place. This second component of our argument will establish that $P(\text{WLE} \mid \text{decay})$ is not necessarily greater than zero.

Since Baddeley et al. (1975) discovered the WLE, there has been at least tacit consensus that every reasonable decay-based model of short-term memory must predict the WLE. This prediction, however, follows not from the decay assumption alone, but from a conjunction of two assumptions: (1) Memory traces decay rapidly, and (2) memory traces cannot be refreshed during overt or subvocal articulation. Without the second assumption, no WLE would be expected, because the decay arising from longer pronunciation durations could be compensated for by refreshing of memory traces that proceeds concurrently with articulation.

The second assumption can be satisfied in several ways. For example, one might postulate a simplistic decay model in which there is no rehearsal mechanism at all. Alternatively, one might postulate a more sophisticated variant, such as the phonological loop model, in which rehearsal of verbal material is possible but is exclusively mediated by subvocal articulation. Because rehearsal is linked to subvocal articulation, a given item cannot be rehearsed during rehearsal or overt spoken recall of another item.

Here, we argue that the second assumption is probably wrong, on the basis of recent findings that have identified a nonarticulatory memory restoration process that can operate concurrently with overt articulation. By implication, for a decay model to be in accord with those crucial recent data, it must discard the second assumption. However, any decay model that is reduced to only the first assumption does not necessarily predict a WLE, because decay might be counteracted by memory restoration or refreshing.

The case for attentional refreshing.[4] Several recent developments have questioned whether articulatory rehearsal is the only, or even the main, mechanism for revitalizing verbal representations in short-term memory.

The first development was initiated by Barrouillet and colleagues (e.g., Barrouillet, Bernardin, & Camos, 2004; Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007), who investigated the factors underlying forgetting in the complex-span paradigm. In the complex-span paradigm, encoding of list items for serial recall alternates with a short period of unrelated processing (e.g., solving arithmetic equations or reading aloud). Barrouillet and colleagues found that memory improved when a constant amount of processing inserted between memory items was stretched out over a longer period of time. For example, if the list items *A* and *B* were separated by five digits to be read aloud as distractors, memory was better if 5 sec intervened between *A* and *B*, rather than 3 sec.

This finding implies that memory improved as the retention interval was lengthened, contrary to what would be expected from uncompensated decay (i.e., Assumption 1 above). In order to reconcile their finding with the decay notion, Barrouillet et al. (2004; Barrouillet et al., 2007) assumed that participants used small pauses in between individual cognitive operations of the processing task to refresh the memory items. Stretching out the processing activity over more time implies longer pauses in between component operations and, thus, more opportunity for refreshing. Importantly, the time-based resource sharing

(TBRS) model of Barrouillet et al. (2004) assumes that refreshing is not based on subvocal articulation but rather on attention-based retrieval of the memory items. This refreshing mechanism requires an attentional bottleneck that is also engaged by the cognitive operations of the processing task, such that the bottleneck has to switch back and forth between processing operations and refreshing of memory traces. Therefore, memory depends on *cognitive load*, defined as the proportion of time that the bottleneck is engaged by the processing task. Stretching out the same number of operations over more time reduces the cognitive load and, therefore, improves memory by permitting more refreshing even during small pauses in the processing task (e.g., when one arithmetic operation has been completed and the next one has not yet commenced).

Barrouillet and colleagues have presented an impressive body of evidence in support of the effect of cognitive load (e.g., Barrouillet & Camos, 2001; Gavens & Barrouillet, 2004; Lépine, Barrouillet, & Camos, 2005; Lépine, Bernardin, & Barrouillet, 2005).[5] Particularly relevant here is an experiment by Barrouillet et al. (2007), in which the memory list consisted of verbal material (i.e., letters), whereas the processing task was a nonverbal two-choice task (i.e., discriminating high from low positions of a dot on the screen). Keeping processing time between encoding of successive letters constant, Barrouillet et al. (2007) varied the difficulty of the processing task by manipulating the spatial discriminability of the stimuli. More difficult choice trials were assumed to engage the bottleneck longer, and this was reflected in longer reaction times. As predicted, memory span was lower with the more difficult choice trials. Across several experiments, span declined as an approximately linear function of cognitive load, independently of the total amount of time available or the total number of processing operations required (see also Barrouillet et al., 2004). This finding is predicted by a theory that assumes competition between refreshing of verbal material and response selection in a nonverbal choice task, but it is harder to understand from the perspective of any model that assumes that rehearsal is articulatory, because articulatory maintenance rehearsal competes with nonverbal tasks only during a brief initial setup period (Naveh-Benjamin & Jonides, 1984).

The assumption of an attention-based refreshing mechanism was put to a direct test by Hudjetz and Oberauer (2007). They used a reading span task to test two possible versions of the TBRS, one assuming articulatory rehearsal and the other assuming attention-based refreshing as the mechanism counteracting decay during the processing phases (the latter is the version favored by Barrouillet et al., 2004). Participants read sets of sentences aloud and tried to remember the last word of each sentence for recall at the end of a set. Each sentence was presented in four segments of three words each, and cognitive load was manipulated by varying the presentation duration of each segment. A shorter presentation duration implies a higher load, because the attentional bottleneck is occupied with the reading task for a greater proportion of the available time. Two groups of participants worked with different reading instructions. Participants in one group read the sentences as they saw fit, as long as they kept up with the pace of presentation of each sentence's segments (*free reading*). The other group had to read continuously—that is, to articulate constantly without pauses—in synchrony with a metronome, pronouncing one word to each beat.

A control experiment confirmed that the free-reading condition permitted the overt articulation of an additional word several times during short reading pauses. The continuous, rhythmic reading condition, by contrast, reduced people's ability to articulate an additional word by a factor of 10. By implication, continuous reading virtually eliminated the opportunity for subvocal rehearsal during sentence reading. Notwithstanding, memory span increased with decreasing cognitive load (i.e., slower reading speeds) in both the free- and the continuous-reading conditions. In the free-reading condition, this effect of cognitive load could be explained by either version of TBRS, because people could have used small pauses in between reading words equally for articulatory rehearsal or for attentional refreshing. Of greater interest are the parallel results in the continuous-reading condition: Continuous reading demonstrably eliminated the opportunity for articulatory rehearsal. Yet memory increased with longer presentation duration of the sentence segments to the same degree as in the free-reading condition. If continuous reading had prevented all possible forms of rehearsal entirely, decay would have produced the opposite effect. Therefore, in order to accommodate the results of Hudjetz and Oberauer (2007), a decay-based model must assume another mechanism for compensating decay, one that is not impeded by continuous overt reading. The refreshing mechanism assumed in the TBRS could fill this role, but only if it is assumed to operate concurrently with overt articulation.

In conclusion, the findings of Barrouillet et al. (2004; Barrouillet et al., 2007) together with those of Hudjetz and Oberauer (2007) have several strong implications that are summarized graphically in Figure 2. (1) If verbal short-term memory traces decay, then any absence of forgetting implies that there is a mechanism to counteract decay, because otherwise stretching out the same processing episode over a longer period of time would lead to more, not less, forgetting. (2) This mechanism must operate concurrently with continuous overt articulation as effectively as in between gaps of articulation, because increasing processing time had the same beneficial effect with continuous reading as it did with free reading in the experiment of Hudjetz and Oberauer. (3) Therefore, the verbal short-term memory system must have a refreshing mechanism that is not impeded by concurrent articulation. This mechanism, by definition, cannot be articulatory rehearsal but must involve another, likely attentional, source of refreshing.[6]

Additional independent evidence for the existence of an attention-based, nonarticulatory refreshing mechanism comes from the work of Raye et al. (2007). They showed that attentional refreshing, which they operationalized as briefly thinking of a recently presented word or object, can be dissociated from articulatory rehearsal by its cortical activation pattern as reflected in fMRI. Moreover,

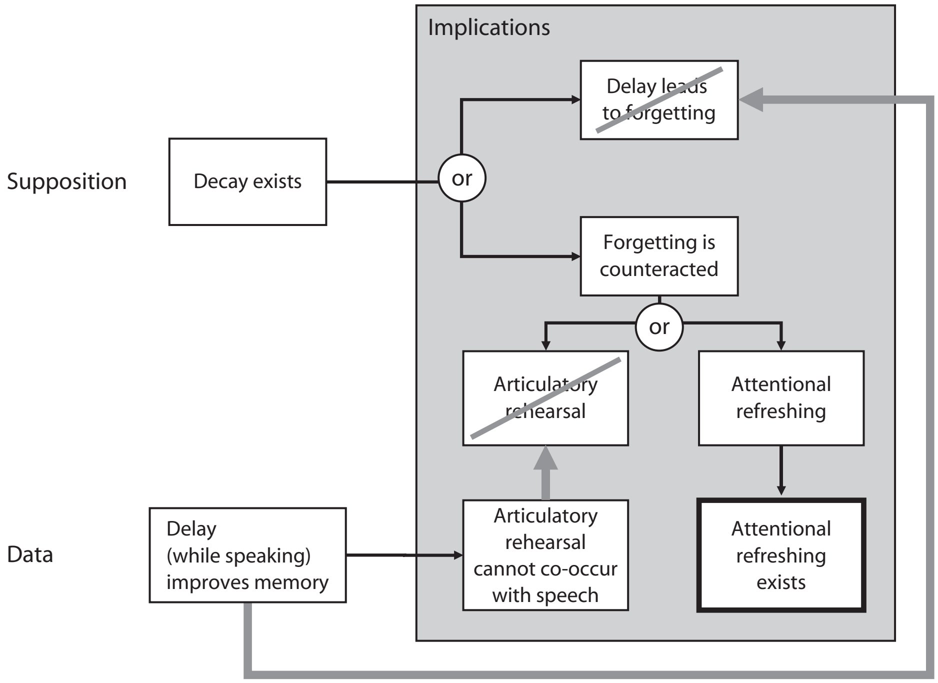
Implications
Delay ~~leads to forgetting~~
Supposition
Decay exists
or
Forgetting is counteracted
or
~~Articulatory rehearsal~~
Attentional refreshing
Data
Delay (while speaking) improves memory
Articulatory rehearsal cannot co-occur with speech
Attentional refreshing exists

Figure 2. Summary of the suppositional argument that establishes the existence of attentional refreshing. The large shaded box encloses all possible conclusions that are implied by the supposition that decay exists and by the data of Hudjetz and Oberauer (2007). The gray arrows represent implications that contradict other possible assertions. The box with thick borders corresponds to the conclusion suggested by our argument. See the text for further details.

refreshing has been found to enhance long-term recall (for a review, see Raye et al., 2007).

Decay, attentional refreshing, and the WLE. If we accept, then, that there is an attention-based refreshing mechanism that can operate concurrently with overt articulation, what are the implications of this process for the WLE and its theoretical interpretation? If a decay model included an attentional-refreshing mechanism in addition to articulatory rehearsal, would it still predict a WLE? It may or may not, but as we show next, the WLE can no longer be assumed to follow directly from the decay assumption.

We acknowledge that long words take longer to articulate than short words during encoding and during recall (especially when recall is spoken). We acknowledge that if people engaged in subvocal articulation, they would take longer to articulate a list of long words than they would a list of short words. Longer articulation durations, however, do not imply postponement of memory refreshing, because as we have shown in the preceding section, the attentional refreshing mechanism can operate concurrently with articulation. There is no reason to assume that the rate of attentional refreshing is tied to the rate of articulation in such a way that words with longer spoken duration take longer to refresh. It follows that word length need not affect recall performance in a decay model that includes attentional refreshing.

A decay model probably can be made to predict a WLE by the addition of certain assumptions. Subvocal articulation could be assumed to have an additional beneficial effect on memory over and above, and independent of, attention-based refreshing. Because the subvocal articulation rate would be determined by word length, under that assumption a WLE might still be predicted by a decay model. Nonetheless, whether a WLE is predicted by a decay-based model—or, for that matter, by any model that includes both articulatory rehearsal and attentional refreshing, regardless of the source of forgetting—depends on the detailed assumptions that the model makes about how rehearsal and refreshing interact.

In conclusion, the presence of attentional refreshing implies that the WLE cannot be assumed to follow directly from the decay assumption: Any model incorporating decay must include a sophisticated set of mechanisms for counteracting decay—both when articulation is possible and when it is not—to be in line with the data. Once these mechanisms are in place, it is far from obvious that a model would still predict the WLE, rendering it uncertain whether $P(\text{WLE} \mid \text{decay}) > 0$.

A decay model might predict the WLE if it included the assumption that articulatory rehearsal benefits memory independently of (and above and beyond) the effect of attentional refreshing. This assumption, however, can be made by any model, regardless of whether it assumes decay or another source of forgetting. Therefore, the assumption of decay does not increase the probability of the WLE; hence, $P(\text{WLE} \mid \text{decay}) = P(\text{WLE} \mid \text{no decay})$. It follows that the WLE is entirely undiagnostic for the question of whether memory representations decay over time.

That said, the question arises as to whether the decay and interference notions are at all empirically differentiable. Or are our arguments against the interpretation of the WLE tantamount to an acknowledgement that those two competing processes are not identifiable? Next, we show that, far from preventing empirical differentiation, our arguments against the decay-based interpretation of the WLE help identify a platform for a proper empirical test of alternative views of forgetting.

Experimental Alternatives to the Word Length Effect

There is pervasive agreement that an experimental test of a causal hypothesis is to be preferred over a correlational one. We wrap up this article by presenting techniques that can experimentally control the time during which representations in short-term memory are left to decay. The challenge is to manipulate time in a way that (1) avoids varying other variables besides time, among them the potential for interference with memory representation, and (2) prevents participants from compensating for the putative effect of decay by articulatory rehearsal or attentional refreshing. These two goals are in conflict with each other: To disable rehearsal and refreshing, researchers must fill the manipulated time interval with some cognitive activity that blocks them, but this activity could create interference. Several attempts have been undertaken to solve these problems—for example, by using a nonverbal signal-detection task to fill the retention interval (Reitman, 1974; Shiffrin, 1973) or by asking participants not to rehearse (Waugh & Norman, 1965). These early studies have remained inconclusive (see Crowder, 1976, for a thorough review of early experimental techniques). We now discuss several recent techniques, culminating with those that we believe hold the greatest promise.

One contemporary technique used to examine the role of time in forgetting was introduced by Cowan, Wood, Nugent, and Treisman (1997) and Cowan, Nugent, Elliott, and Geer (2000). Participants were instructed to recall lists of words either slowly or quickly, allowing a fixed window of 2.5 sec/word in both cases (i.e., when people took 1 sec to recall a word, a blank pause of 1.5 sec would follow, and if they took 2 sec for recall, the pause would be 0.5 sec). Cowan et al. (2000) found recall to be worse in the slow-recall condition, which they considered to refute the claims made by Service (1998) against the duration-based WLE. This technique is an improvement over correlational approaches, but we find it unsatisfactory, because it created silent periods during retrieval that were longer in the fast-recall condition than in the slow-recall condition and during which people's processing was entirely uncontrolled. Indeed, Cowan et al.'s (2000; Cowan et al., 1997) manipulation can be considered an (inadvertent) instantiation of a response-deadline methodology, with people in the fast-recall condition having more time to retrieve the next item in anticipation of the forthcoming retrieval opportunity than did people in the slow-recall condition. It is well known that providing more time for retrieval enhances memory performance (e.g., Reed, 1973); Cowan et al.'s (1997) results are, thus, not conclusive with respect to decay.

In confirmation of our doubts, Cowan et al. (2006) reported no effect of recall duration when retrieval was paced without varying the response deadline. In their first experiment, Cowan et al. (2006) varied the presentation pace of the items and left it to participants to recall at their own speed, which covaried with presentation rate. In the second experiment, participants were instructed to recall at two different speeds. Neither manipulation (which roughly doubled the recall rate from 1/sec to 2/sec) elicited a difference in recall accuracy, confirming our suspicion that the effects reported by Cowan et al. (2000; Cowan et al., 1997) arose not from recall duration per se, but from another factor introduced by the speed manipulation. We, therefore, do not agree that their results counter the finding by Service (1998) that the duration-based WLE fails to materialize with properly controlled material.

The method of Cowan et al. (2000; Cowan et al., 1997) was refined by Lewandowsky, Duncan, and Brown (2004). In their experiments, participants recalled a list of letters while repeating an irrelevant word aloud in between retrievals. There is broad consensus in the literature that this articulatory suppression blocks rehearsal (Baddeley, 1986, pp. 37, 86; Baddeley & Lewis, 1984; Page & Norris, 1998, pp. 764, 770). Lewandowsky et al. found that recall performance was unaffected by the number of times the irrelevant word was repeated in between memory retrievals, implying that time per se did not cause forgetting in serial recall. (Filling the retention interval with articulatory suppression potentially induces interference, but interference would, if anything, generate a negative effect of the number of repetitions of the irrelevant word. Thus, the finding that recall was invariant across the number of repetitions despite the potential for interference counts strongly against decay.)

The results of Lewandowsky et al. (2004) were replicated and extended by Oberauer and Lewandowsky (in press), who compared a baseline condition and a variety of conditions with filled delays during encoding or during retrieval, including some conditions in which the filler tasks blocked not only articulatory rehearsal, but also the attentional refreshing introduced earlier. Specifically, in those conditions, articulatory suppression was combined with a concurrent two-alternative speeded-choice task, and both tasks were performed for varying delays in between successive retrievals. The data showed that the length of the delay had at most a negligible effect on performance. Oberauer and Lewandowsky applied a number of competing computational models to their results and found that neither a decay model (the primacy model; Page & Norris, 1998) nor a temporal distinctiveness model (SIMPLE; Brown, Neath, & Chater, 2007) could handle the data. An interference-based model (SOB; Farrell & Lewandowsky, 2002; Lewandowsky & Farrell, 2008), by contrast, gave a good quantitative account of the differences between the various conditions.

It thus appears that the best available experimental manipulations of time—be it through manipulation of the articulation duration of items (Service, 1998), the time available for recall (Cowan et al., 2006), or delays filled with tasks that prevent articulatory rehearsal and also attentional refreshing (Oberauer & Lewandowsky, in press)—suggest that the time interval between encoding and retrieval of items in short-term memory has little, if any, effect on their memorability. Although any conclusions about the cause of forgetting from short-term memory are beyond the scope of this article, we submit that the experimental techniques illustrated in this section are a better vehicle for studying this issue than is the WLE.

Conclusions

The analyses presented in this article converge on three conclusions:

1. The WLE is correlational and, therefore, open to alternative explanations that do not assume time-based decay. Hence, $P(\text{WLE} \mid \text{no decay}) > 0$, and by implication, $P(\text{WLE} \mid \text{decay})$ is not necessarily greater than $P(\text{WLE} \mid \text{no decay})$. On that basis, it appears inadvisable to put much stock in the WLE as a key piece of evidence in favor of decay models.

2. Attentional refreshing is a necessary component of a viable decay model. Once this mechanism is included, a decay model may or may not predict the WLE, and whether it does depends on additional assumptions that are unrelated to the decay notion. Hence, $P(\text{WLE} \mid \text{decay})$ is not necessarily greater than zero and is certainly smaller than one. Together with Conclusion 1, this implies that $P(\text{WLE} \mid \text{decay})$ can be larger than, smaller than, or equal to $P(\text{WLE} \mid \text{no decay})$.

These two points are independent; each of them weakens the case for the WLE as evidence for decay. Together, they render the WLE undiagnostic for the decay hypothesis, according to the Bayesian logic of scientific reasoning: We have no reason to assume that the probability that there is a WLE is higher under the decay assumption than under the alternative hypothesis that there is no decay.

It may be helpful to consider the implications if one were to reject our second conclusion. The rejection of attentional refreshing implies that $P(\text{WLE} \mid \text{decay}) = 1$, thus restoring the tight and hitherto presumed link between decay models and the WLE. However, for the WLE to be a diagnostic of decay additionally requires that $P(\text{WLE} \mid \text{no decay}) < 1$. The latter constraint, in turn, implies that all plausible explanations for the WLE that are not based on decay must be convincingly ruled out. We do not believe that pursuit of this endeavor would be fruitful, for the reasons given in our third, and final, conclusion.

3. There are experimental alternatives to the WLE that afford experimental control over the time items reside in memory and that thereby permit a proper test of the decay hypothesis. So far, these tests have been largely negative, but we do not consider the case for or against decay closed. Future research should capitalize on the recent progress in developing experimental methods for studying the causes of short-term forgetting.

Our critical analysis of the interpretation of the WLE gives rise to the question of whether all research that relies on comparison of items is fraught with risk. The short answer is "yes," although there are instances in which this is inevitable. For example, researchers concerned with the properties of reading and word identification have no choice but to examine the effects of stimulus attributes on their favored response variables. Indeed, it is the very nature of the stimuli that is of conceptual and theoretical interest in those studies. Moreover, in the memory arena, it is of interest in and of itself to explore how various properties of the stimuli relate to performance. For example, the effects of phonological neighborhood size (see, e.g., Roodenrys et al., 2002) or the effects of phonotactic structure (see, e.g., Roodenrys & Hinton, 2002) may provide insights into likely coding mechanisms or the role of sequence probabilities on lists.

Our arguments against the interpretation of the WLE do not imply, therefore, that examining the effects of relevant stimulus dimensions on cognitive performance is inadvisable; when those dimensions are of interest or when there are no alternatives, comparisons across sets of material that differ in linguistic characteristics are essential. However, insofar as its diagnosticity for causes of forgetting is concerned, the WLE is qualitatively different from those other areas of inquiry, because there is an alternative—namely, the well-controlled studies that manipulated retention interval. In this context, it is important to bear in mind that word length has always been acknowledged to be a surrogate for the variable of actual theoretical interest—namely, the effect of time on forgetting. Given the problems associated with the WLE, more value must be placed on the results from other, more controlled alternatives that permit manipulation of the variable of interest.

Should we continue studying the WLE? There is a large body of empirical work on the WLE, creating a strong temptation to continue its investigation, because further experiments can be readily motivated by issues left unresolved by previous studies. Researchers who wish to pursue this path should be aware that the WLE was primarily of interest as a potential window into the architecture of immediate memory and the causes of forgetting. We have shown that this link does not hold. The WLE tells us nothing about whether time-based decay causes short-term forgetting. Other methods are available that hold more promise to resolve this issue. For those interested in finding out why short-term memory remembers so little and forgets so quickly, it is time to move on.

AUTHOR NOTE

Preparation of this article was facilitated by a Discovery Grant from the Australian Research Council to both authors and an Australian Professorial Fellowship from the Australian Research Council to the first author. We thank Gordon Brown for his comments on an earlier version of this article and Leo Roberts for assistance with the N-Watch analysis. We are grateful to Shane Mueller for providing us with the information required to compute confidence intervals for the data reported by Mueller et al. (2003). Correspondence concerning this article should be addressed to S. Lewandowsky, School of Psychology, University of Western Australia, Crawley, WA 6009, Australia (e-mail: lewan@psy.uwa.edu.au).

REFERENCES

Allen, R., & Hulme, C. (2006). Speech and language processing mechanisms in verbal serial recall. *Journal of Memory & Language*, **55**, 64-88.

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

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

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., & Lewis, V. J. (1984). When does rapid presentation enhance digit span? *Bulletin of the Psychonomic Society*, **22**, 403-405.

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

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, **133**, 83-100.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **33**, 570-585.

Barrouillet, P., & Camos, V. (2001). Developmental increase in working memory span: Resource-sharing or temporal decay? *Journal of Memory & Language*, **45**, 1-20.

Bireta, T. J., Neath, I., & Surprenant, A. M. (2006). The syllable-based word length effect and stimulus set specificity. *Psychonomic Bulletin & Review*, **13**, 434-438.

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

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

Brown, G. D. A., Neath, I., & Chater, N. (2007). A ratio model of scale-invariant memory and identification. *Psychological Review*, **114**, 539-576.

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.

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

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

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

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

Cowan, N., elliott, E. M., Saults, J. G., Nugent, L. D., Bomb, P., & Hismjatullina, A. (2006). Rethinking speed theories of cognitive development: Increasing the rate of reacll without affecting accuracy. *Psychological Science*, **17**, 67-73.

Cowan, N., Keller, T. A., Hulme, C., Roodenrys, S., McDougall, S., & Rack, J. (1994). Verbal memory span in children: Speech timing clues to the mechanisms underlying age and word length effects. *Journal of Memory & Language*, **33**, 234-250.

Cowan, N., Nugent, L. D., Elliott, E. M., and 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., 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.

Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, **37**, 65-70.

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

Ellis, N. C., & Hennelly, R. A. (1980). A bilingual word-length effect: Implications for intelligence testing and the relative ease of mental calculation in Welsh and English. *British Journal of Psychology*, **71**, 43-51.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review*, **9**, 59-79.

Gavens, N., & Barrouillet, P. (2004). Delays of retention, processing efficiency, and attentional resources in working memory span development. *Journal of Memory & Language*, **51**, 644-657.

Healy, A. F. (1974). Separating item from order information in short-term memory. *Journal of Verbal Learning & Verbal Behavior*, **13**, 644-655.

Hudjetz, A., & Oberauer, K. (2007). The effects of processing time and processing rate on forgetting in working memory: Testing four models of the complex span paradigm. *Memory & Cognition*, **35**, 1675-1684.

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

Jensen, A. R. (1982). Reaction time and psychometric *g*. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 93-132). New York: Springer.

Kahana, M., & Loftus, G. (1999). Response time versus accuracy in human memory. In R. J. Sternberg (Ed.), *The nature of cognition* (pp. 322-384). Cambridge, MA: MIT Press.

Lépine, R., Barrouillet, P., & Camos, V. (2005). What makes working memory spans so predictive of high-level cognition? *Psychonomic Bulletin & Review*, **12**, 165-170.

Lépine, R., Bernardin, S., & Barrouillet, P. (2005). Attention switching and working memory spans. *European Journal of Cognitive Psychology*, **17**, 329-346.

Lewandowsky, S. (1999). Redintegration and response suppression in serial recall: A dynamic network model. *International Journal of Psychology*, **34**, 434-446.

Lewandowsky, S., Duncan, M., & Brown, G. D. A. (2004). Time does not cause forgetting in short-term serial recall. *Psychonomic Bulletin & Review*, **11**, 771-790.

Lewandowsky, S., & Farrell, S. (2000). A redintegration account of the effects of speech rate, lexicality, and word frequency in immediate serial recall. *Psychological Research*, **63**, 163-173.

Lewandowsky, S., & Farrell, S. (2008). Short-term memory: New data and a model. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 49, pp. 1-48). San Diego: Academic Press.

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

Longstreth, L. E. (1984). Jensen's reaction time investigations of intelligence: A critique. *Intelligence*, **8**, 139-160.

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

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

Mathey, S., Robert, C., & Zagar, D. (2004). Neighbourhood distribution interacts with orthographic priming in the lexical decision task. *Language & Cognitive Processes*, **19**, 533-560.

Mathey, S., & Zagar, D. (2000). The neighborhood distribution effect in visual word recognition: Words with single and twin neighbors. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 184-205.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, **88**, 375-407.

Mueller, S. T., Seymour, T. L., Kieras, D. E., & Meyer, D. E. (2003). Theoretical implications of articulatory duration, phonological similarity, and phonological complexity in verbal working memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 1353-1380.

Murray, A., & Jones, D. M. (2002). Articulatory complexity at item boundaries in serial recall: The case of Welsh and English digit span. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 594-598.

Nairne, J. S. (2002). Remembering over the short-term: The case against the standard model. *Annual Review of Psychology*, **53**, 53-81.

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

Naveh-Benjamin, M., & Jonides, J. (1984). Maintenance rehearsal: A two-component analysis. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 369-385.

Neath, I., Bireta, T. J., & Surprenant, A. M. (2003). The time-based word length effect and stimulus set specificity. *Psychonomic Bulletin & Review*, **10**, 430-434.

Neath, I., & Brown, G. D. A. (2006). SIMPLE: Further applications of a local distinctiveness model of memory. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 46, pp. 201-243). San Diego: Academic Press.

Neath, I., & Nairne, J. S. (1995). Word-length effects in immediate memory: Overwriting trace decay theory. *Psychonomic Bulletin & Review*, **2**, 429-441.

Oberauer, K., & Lewandowsky, S. (in press). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review*.

Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review*, **105**, 761-781.

Pugh, K. R., Rexer, K., Peter, M., & Katz, L. (1994). Neighborhood effects in visual word recognition: Effects of letter delay and nonword context difficulty. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 639-648.

Raye, C. L., Johnson, M. K., Mitchell, K. J., Greene, E. J., & Johnson, M. R. (2007). Refreshing: A minimal executive function. *Cortex*, **43**, 135-145.

Reed, A. V. (1973). Speed–accuracy trade-off in recognition memory. *Science*, **181**, 574-576.

Reitman, J. S. (1974). Without surreptitious rehearsal, information in short-term memory decays. *Journal of Verbal Learning & Verbal Behavior*, **13**, 365-377.

Robert, C., Mathey, S., & Zagar, D. (2007). The effect of the balance of orthographic neighborhood distribution in visual word recognition. *Journal of Psycholinguistic Research*, **36**, 371-381.

Roodenrys, S., & Hinton, M. (2002). Sublexical or lexical effects on serial recall of nonwords? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 29-33.

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. M. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 1019-1034.

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

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

Shiffrin, R. M. (1973). Information persistence in short-term memory. *Journal of Experimental Psychology*, **100**, 39-49.

Tehan, G., Fogarty, G., & Ryan, K. (2004). The contribution to immediate serial recall of rehearsal, search speed, access to lexical memory, and phonological coding: An investigation at the construct level. *Memory & Cognition*, **32**, 711-721.

Tehan, G., & Lalor, D. M. (2000). Individual differences in memory span: The contribution of rehearsal, access to lexical memory, and output speed. *Quarterly Journal of Experimental Psychology*, **53A**, 1012-1038.

Thorn, A. S. C., & Frankish, C. R. (2005). Long-term knowledge effects on serial recall of nonwords are not exclusively lexical. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 729-735.

Tolan, G. A., & Tehan, G. (2005). Is spoken duration a sufficient explanation of the word length effect? *Memory*, **13**, 372-379.

Walker, I., & Hulme, C. (1999). Concrete words are easier to recall than abstract words: Evidence for a semantic contribution to short-term serial recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 1256-1271.

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

Wingfield, A., & Byrnes, D. L. (1972). Decay of information in short-term memory. *Science*, **176**, 690-692.

NOTES

[1] This argument is unaffected by whether articulation time is measured directly or inferred on the basis of the item's characteristics—for example, by counting the number of syllables. In both cases, no experimental variable is manipulated, but the characteristics of an existing corpus are observed. For the purposes of this article, we consider evidence to be *correlational* whenever it involves measurement of all variables under consideration without experimental intervention and control. This differs from *causal* evidence, in which one or more variables are manipulated, and the effect of those manipulations on another set of variables is observed. Our usage of the term *correlational* is therefore equivalent to alternative labels such as *pseudoindependent variable* or *quasi-experiment* that have been applied to similar designs.

[2] The article by Service (1998) engendered several empirical responses, which we consider in a later section.

[3] In this context, note that Mueller et al. (2003) failed to find a duration-based WLE. In their Experiment 2—the only one to include stimuli that were equal in phonological complexity but that differed in articulation duration—memory span for long words (5.05 sec) did not differ significantly from span for short words (5.21 sec). This failure to find a duration-based WLE was accommodated by their two-factor model, which predicted performance on those two sets of words to be equal because of a compensatory difference in phonological similarity; however, acceptance of this model relies on the questionable suppositions just outlined.

[4] We use the term *attentional refreshing* to maximize the contrast with articulatory forms or rehearsal. Elsewhere, attentional refreshing has also been referred to as attentional rehearsal (e.g., Hudjetz & Oberauer, 2007), but here we avoid use of the term *rehearsal* to maximize distinctiveness.

[5] Although Barrouillet and colleagues interpreted their results exclusively within a decay-based framework, the existence of nontemporal alternative accounts of their data (e.g., within an interference framework) cannot be ruled out (see, e.g., Barrouillet & Camos, 2001; Gavens & Barrouillet, 2004; Lépine, Barrouillet, & Camos, 2005; Lépine, Bernardin, & Barrouillet, 2005). This issue is of little relevance here, because irrespective of whether forgetting is due to decay or to interference, the data of Barrouillet and colleagues implicate an attentional refreshing process that is distinct from verbal rehearsal.

[6] A critic might argue that the evidence for attentional refreshing was derived using the complex-span paradigm, rather than the serial-recall paradigm that underlies most WLE effects. In response, we note that if the system has an attentional-refreshing mechanism available for use in the complex-span paradigm, there is no reason it would not use this mechanism in serial recall as well, in particular because the attentional refreshing mechanism is sufficiently flexible and fast to exploit small pauses in the processing episodes in a complex-span task.

(Manuscript received October 19, 2007;
revision accepted for publication February 9, 2008.)