The phonological similarity effect in immediate recall: Positions of shared phonemes

XIAOJIAN LI, RICHARD SCHWEICKERT, and JACK GANDOUR
Purdue University, West Lafayette, Indiana

Earlier literature proposes two ways phonological similarity could harm immediate recall: (1) It could increase the degradation of the representations of items in memory, or (2) it could decrease the probability that a degraded representation is correctly reconstructed. A multinomial processing tree model for each hypothesis was used to analyze an immediate recall experiment. Both gave a good account of the data, but, of the two, results favor the hypothesis that the effect of phonological similarity is to impair reconstruction of degraded representations. A second issue is whether positions of repeated phonemes in phonologically similar items matter. We found that mere repetition of phonemes produced a phonological similarity effect. Repeated phonemes in the same positions appeared to produce a greater effect. A final finding is that when reading rate was preequated, phonological similarity affected memory span by changing the time taken to recall a list of span length.

Memory span is the number of items in a list that can be immediately recalled in order, on half the presentations. One of the major variables influencing memory span is the phonological similarity of the items (Baddeley, 1966; Conrad, 1964; Hintzman, 1965). In an often-cited paper, Posner and Konick (1966) described two mechanisms by which phonological similarity could decrease correct recall. With one, the acid bath, "interfering items interact with the stored trace spontaneously during the retention interval to weaken its strength" (p. 222). With the other, trace comparison, "the role of interference is to vary the competition between traces at the moment of recall" (p. 222). We report an experiment comparing these two mechanisms, using a multinomial processing tree model proposed by Schweickert (1993; see also Batchelder & Riefer, 1999; Gathercole, Frankish, Pickering, & Peaker, 1999).

According to the model, at the time of recall, the representation of an item is intact with probability I and leads to correct recall. If the representation is not intact but degraded, the item can be correctly redintegrated (reconstructed) with probability R. The overall probability of correctly recalling the item, then, is $I + (1 - I)R$.

If phonological similarity has an effect via the acid bath, it will increase degradation of the memory representation. On the other hand, if phonological similarity has an effect via trace competition, it will decrease the probability of redintegration of a degraded representation.

In other words, when the model is fit to data, according to the acid bath view, a change of phonological similarity should change the parameter I, and according to the trace competition view, a change of phonological similarity should change the parameter R.

Positions of Shared Phonemes

We investigated another question as well. Typically, experiments varying phonological similarity use a pool of similar items having common phonemes in the same positions (e.g., the names of the letters {B, C, D, P, T, V}) and a control pool of dissimilar items having few phonemes in common (e.g., the names of the letters {F, H, K, N, R, Z}). An increase in the number of repeated phonemes is confounded with an increase in the probability of correctly reconstructing an item if part of it is lost.

In the above control pool, for example, if all that remained of the phonemes of an item were /_e/ ("_ay"), the participant would be able to reconstruct the item uniquely as "kay." In the phonologically similar pool, if the remnant were /_i/ ("_ee"), the participant would have little luck in reconstructing the item. Constraints—such as rhyming, in this case—can either help or hinder recall, depending on the task (Watkins, Watkins, & Crowder, 1974). Bower and Bolton (1969), for example, demonstrated that such constraints can sometimes help; they showed that the hypothesis of restriction of possible responses is sufficient to explain the ease with which rhymes are learned in paired-associate tasks. For the ordered recall required in memory span tasks, Sperling and Speelman (1970) argued that the constraints introduced by rhyming decrease the probability of correct guessing, thus hindering recall. Are the constraints of shared phoneme positions the only reason for the phonological similarity effect? If items have phonemes in common but in different positions, would the resulting phonological similarity affect immediate ordered recall?

Many models take no position on these questions. There are at least two that do—the neural network model of Burgess and Hitch (1992; Burgess, 1995; Burgess & Hitch, 1996), a similarity-based choice model, and the phomemic model of Sperling and Speelman (1970), a sophisticated-guessing model. The models are formulated carefully enough that one can see that they cannot both be true.

A similarity-based choice model. The model developed by Burgess and Hitch (1992; Burgess, 1995; Burgess & Hitch, 1996), has nodes for phonemes and nodes for words (and for other components not relevant here). For a word with n phonemes, each phoneme node contributes activation proportional to $1/\sqrt{n}$. At time of recall, the word with the highest activation is selected (Burgess & Hitch, 1992, pp. 437, 440). The upshot is that if a phoneme in the to-be-recalled item also appears in some other item, the other item will have an increased chance of being output, in error. As Houghton, Hartely, and Glasspool (1996) point out, in the model, "the sound of a word is represented as a set of phonemes with no order defined over them, so that, for instance, /tap/, /pat/, and /apt/ would be represented identically" (p. 109). Consequently, a repeated phoneme has an effect, but its position has no additional effect.

A sophisticated-guessing model. Sperling and Speelman (1970) used letters of the alphabet as stimuli. They assumed that the phonemes of items are lost individually in short-term memory. They further assumed that during recall, "if only one of the two phonemes of a letter is available, a guess is made from among those letters of the alphabet that contain the retained phoneme in the same position" (p. 155). The phonological similarity effect is explained by this sophisticated guessing, with order being critical. The effect is bigger when there are more candidates for guessing.

We note, incidentally, that phonological similarity operates in the Burgess and Hitch (1992; Burgess, 1995; Burgess & Hitch, 1996) model via acid bath, because increased similarity increases noise and decreases the strength of the representation. On the other hand, phonological similarity operates in the Sperling and Speelman (1970) model via trace comparision, because an increase in phonological similarity increases the number of competing alternatives for the reconstruction of a degraded representation. It happens that each model posits a different mechanism for the phonological similarity effect, and each makes a different assumption about the role of phoneme position. But in general, assumptions about acid bath versus trace comparison are logically independent of assumptions about the role of phoneme position.

Recall Duration

Our experiment also investigated the effect of phonological similarity on recall duration, a variable important because of its connection with an important factor influencing memory span—word length. Items requiring more

time to pronounce—that is, having greater word length—yield shorter memory spans (Baddeley, Thomson, & Buchanan, 1975; Hoosain & Salili, 1988; Mackworth, 1963; Schweickert & Boruff, 1986). In an experiment by Chase (1977), the phonologically similar items took more time to pronounce than did the dissimilar items, suggesting the obvious hypothesis that the phonological similarity effect is a by-product of the word-length effect. This hypothesis was rejected by Hulme and Tordoff (1989) and by Schweickert, Guentert, and Hersberger (1990), who found a phonological similarity effect on span in the absence of a difference in mean pronunciation time per item, for the items they used.

A loophole remains, because the time to utter the phonologically similar and dissimilar items in these studies was not measured during recall, when it is arguably most important (Dosher & Ma, 1998). For this reason, recall durations are reported here. (Another study reporting recall durations is Schweickert, McDaniel, & Riegler, 1994.)

Two experimental word pools were constructed to represent two different ways in which words can be phonologically similar. Two control pools contained dissimilar words constructed, primarily, to control for word length and, secondarily, to control for word frequency. Each experimental pool was given its own control pool, because it was not possible to make the two experimental pools match in terms of word frequency.

METHOD

Stimuli

The four word pools are shown in Tables 1A and 1B. For words in Experimental Pool 1, the constituent consonants are repeated often, but almost always in fixed positions. Control Pool 1 contains phonologically dissimilar words, chosen with the goal of matching the mean reading time (see below) of Experimental Pool 1. A secondary goal was to approximately equate the mean word frequencies for the two pools.

Each word in Experimental Pool 2 was obtained from a word in Experimental Pool 1 by a permutation of its phonemes. For example, SPIT in Experimental Pool 1 became TIPS in Experimental Pool 2, and SPAT became SAPPED. Therefore, words in Experimental Pool 2 have exactly the same phonemes as those in Experimental Pool 1. Words in Experimental Pool 2 have repeated phonemes, but the repeated phonemes almost always occur in different positions in different items, unlike the repeated phonemes in Experimental Pool 1. Control Pool 2, corresponding to Experimental Pool 2, was selected with the same goals as Control Pool 1. (Inadvertently, the item SEEKED was included in Experimental Pool 1, although it is not a word. No participant commented on it.)

Calibration

The mean reading times in Tables 1A and 1B are from a calibration study. Twenty-four undergraduates, native speakers of English, participated in this study to satisfy a course requirement. The method was the same as that for the recall experiment to be reported, except that, on each trial, a list of six items appeared on the screen and the participant simply read the list aloud. An analysis of variance (ANOVA) found no significant effect on reading time of similarity, type of repeated phoneme, or their interaction. (The ANOVA design was the same as that for the recall experiment to be reported.)

Table 1A
Stimuli

Experimental Pool 1 Similar (Fixed Positions)			Control Pool 1 Dissimilar (For Fixed Positions)		
spit	/spɪt/	(11)	wings	/wɪŋz/	(24)
spot	/spat/	(57)	climb	/klaɪm/	(12)
spout	/spaᵘt/	(1)	fuzzed	/fəzd/	(1)
spat	/spæt/	(8)	mined	/maɪnd/	(3)
skeet	/skit/	(2)	globe	/glob/	(12)
Scot	/skat/	(1)	foams	/fomz/	(22)
skate	/sket/	(1)	hinge	/hɪnʤ/	(1)
stick	/stɪk/	(39)	pride	/praɪd/	(40)
stoke	/stok/		frog	/frag/	(1)
scat	/skæt/		dorm	/dorm/	
scut	/skət/		lisp	/lɪsp/	
Mean frequency 15.0			Mean frequency 12.9		
Mean reading time 0.470 sec			Mean reading time 0.472 sec		

Note—The numbers in parentheses are frequencies from Francis and Kučera (1982). Words without frequency specification are not in the book and were not counted in the mean frequency. Mean reading time is mean reading time per item from the calibration study.

Table 1B
Stimuli

Experimental Pool 2 Similar (Variable Positions)			Control Pool 2 Dissimilar (For Variable Positions)		
tips	/tɪps/	(13)	curb	/kərb/	(13)
stop	/stap/	(120)	sound	/saᵘnd/	(202)
pouts	/paᵘts/		wired	/waɪrd/	(11)
sapped	/sæpt/	(3)	bilge	/bɪlʤ/	(2)
seeked	/sikt/		drug	/drəg/	(20)
cots	/kats/		blip	/blɪp/	
takes	/teks/	(86)	learn	/lərn/	(83)
kits	/kɪts/	(1)	gowns	/gaᵘnz/	(2)
coast	/kost/	(34)	claim	/klem/	(97)
acts	/ækts/	(37)	wiped	/waɪpt/	(19)
stuck	/stək/	(23)	coins	/kɔɪnz/	(9)
Mean frequency 39.6			Mean frequency 45.80		
Mean reading time 0.465 sec			Mean reading time 0.465 sec		

Note—The numbers in parentheses are frequencies from Francis and Kučera (1982). Words without frequency specification are not in the book and were not counted in the mean frequency. Mean reading time is mean reading time per item from the calibration study.

Participants

Twenty-four undergraduates, native speakers of English, participated in the recall experiment to satisfy a course requirement.

Apparatus

An IBM-compatible computer presented the stimuli. The computer was connected to two Lafayette Instrument voice activated relays, one used during reading aloud, the other during recall. When the participant began to speak, the appropriate voice key was activated and remained activated until 1.3 sec after the participant finished speaking. (The delay was needed to keep the voice key activated during pauses.) The duration of an utterance was the difference between offset and onset time, minus 1.3 sec. The experimenter used a footswitch to send the signal from the microphone to the appropriate voice key. A tape recorder connected to the microphone recorded the participants' reading and recall.

Design

For each of the four word pools, a block of word lists was generated. In each block, list lengths were determined trial by trial, using the staircase method explained below. Each list was obtained by randomly sampling the word pool without replacement. (A list with more words than the word pool contained was never needed.)

A practice block of digits was used at the start of the session, so the participant could learn to activate the voice key properly. After the practice block, each participant was tested on the four word pools.

There were 12 orders of presentation for the two experimental (phonologically similar) and two control (phonologically dissimilar) blocks. To avoid carryover effects from the repeated phonemes, the two experimental blocks were never adjacent to each other. Otherwise, all possible orders of the four blocks were used. A drawback of this design is that experimental lists appeared most often in the beginning and end of the sessions, whereas control lists appeared most often in the middle. The participants were randomly assigned to the 12 orders, with 2 participants in each order.

Procedure

Before each test block, the participant read aloud the words that would be presented in the block. Pronunciation was corrected if necessary.

Each test block started with six words in the first list. With the staircase method, if recall of the list was scored as correct by the experimenter, the list length was increased by one for the next trial. If recall was scored as wrong, the list length was decreased by one.

Each list was presented on the computer screen with the words in one horizontal row. One and a half seconds before the display of the list, a + sign was presented in the future position of the first item for the participant to focus on.

The participants rehearsed the list once by reading it aloud at any speed they liked, as long as no long pause occurred. A mask, composed of two lines of the symbol #, was turned on by the experimenter right after the participant finished reading the last item. The participant then attempted to recall the list in order immediately. Recall was scored as correct if it contained all the items in order with nothing extra.

For each trial, list length, correctness, rehearsal time, and recall time were collected. Trials with misreadings or other disturbances were invalidated by the experimenter and replaced with new trials. Each block contained 35 valid trials. The first 3 valid trials in each block functioned as practice, and only the last 32 valid trials were used for analysis.

RESULTS

For each participant and each word pool, five dependent variables were computed from the trials. *Memory span*, s, was the average list length. *Rehearsal time*, t_{rh}, was the average rehearsal time. *Rehearsal rate* was $r_{rh} = s/t_{rh}$. *Recall time*, t, was estimated with Equation A1 in Appendix A. (Simply averaging the recall times is not useful because, on half the trials, the list was not recalled correctly.) *Recall rate* was $r = s/t$. Notation is given in Appendix B. Mean memory spans, recall times, recall rates, rehearsal times, and rehearsal rates of the four word pools are shown in Table 2.

For each of the five dependent variables, a univariate ANOVA was conducted, with the mean value of the dependent variable for each block for each participant as the input. Participant (24 levels) was random and nested in order of presentation (12 levels), which was fixed as a

Table 2
Span, Recall Time and Rate, and Rehearsal Time and Rate

Dependent Variable	Similarity	
	Similar	Dissimilar
Memory span (items)		
Fixed positions	4.421	5.742
Variable positions	5.014	5.926
Recall time (sec)		
Fixed positions	2.044	2.810
Variable positions	2.343	2.756
Recall rate (items/sec)		
Fixed positions	2.230	2.093
Variable positions	2.185	2.198
Rehearsal time (sec)		
Fixed positions	2.165	2.964
Variable positions	2.507	2.957
Rehearsal rate (items/sec)		
Fixed positions	2.109	1.984
Variable positions	2.047	2.054

between-subjects factor. Similarity (phonologically similar vs. dissimilar) and type of repeated phoneme (fixed positions and associated control vs. variable positions and associated control) were treated as fixed within-subjects factors.

Main effects, when significant, are reported through planned analyses of simple main effects. Interactions, when significant, are reported with the omnibus F. The reader uninterested in such details can skip to the summary.

Memory Span

A planned analysis of simple main effects showed that whether the positions of the repeated phonemes were fixed or variable, the effect of similarity was significant [fixed positions, $F(1,12) = 177.32, MS_e = 0.12, p < .001$; variable positions, $F(1,12) = 102.5, MS_e = 0.10, p < .001$]. For the two similar pools, span was smaller when the repeated phonemes were in fixed positions than when they were in variable positions [$F(1,12) = 36.06, MS_e = 0.12, p < .001$]. For the two dissimilar (control) word pools, span was smaller for the control pool for repeated phonemes in fixed positions [$F(1,12) = 12.65, MS_e = 0.03, p < .01$]. The similarity $\times$ type of repeated phoneme interaction was significant [$F(1,12) = 13.63, MS_e = 0.07, p < .01$].

The small significant difference in span of 0.184 items between the two control pools may be due to their difference in word frequency. That is, the higher word frequency of the one pool leads to better reconstruction of degraded traces and, hence, to slightly better span (Hulme et al., 1997; Poirier & Saint-Aubin, 1996). Is the difference of 0.593 items between the two similar word pools also attributable to their difference in word frequency? Each similar word pool is close in mean frequency to its corresponding control word pool. Poirier and Saint-Aubin (1996) found that word frequency and phonological similarity have additive effects—that is, the effect of word fre-

quency is the same for phonologically similar words and phonologically dissimilar words. If word frequency were the source of the difference we found for the phonologically similar pools, their study would lead us to expect additivity also. Instead, we found a significant interaction; the difference between the two similar word pools is larger than can be accounted for solely by word frequency.

Recall Time

Time spent reading aloud the list of items from the screen, prior to recalling them, is called rehearsal time, and time spent recalling aloud is called recall time. In the planned analysis of simple main effects, phonological similarity led to shorter recall time, whether the positions of the repeated phonemes were fixed [$F(1,12) = 81.70, MS_e = 0.09, p < .001$] or variable [$F(1,12) = 44.25, MS_e = 0.05, p < .001$]. For the two similar pools, recall time was shorter when repeated phonemes were in fixed positions than when they were in variable positions [$F(1,12) = 15.49, MS_e = 0.07, p < .01$]. The similarity $\times$ type of repeated phoneme interaction was significant [$F(1,12) = 9.44, MS_e = 0.08, p < .01$].

Recall Rate

In the planned analysis of simple main effects, when the positions of the repeated phoneme were fixed, recall rate was faster (by 0.137 items per second) for the phonologically similar pool than for its control [$F(1,12) = 5.79, MS_e = 0.04, p < .05$]. We do not know why these rates differ, given that in the calibration experiment, the reading rates for these two pools did not differ. Note that the phonological similarity effect on span cannot be due to this difference, because the direction of the difference is opposite to what would be needed.

For the two control pools, recall rate was faster (by 0.105 items per second) for the control pool for repeated phonemes in variable positions [$F(1,12) = 5.61, MS_e = 0.02, p < .05$].

It is possible that the difference in span between the two control pools is due to a difference in word length (as indexed by the difference in recall rates). But our earlier conclusion that the difference in span between the phonologically similar lists exceeds the difference between the control lists would only be strengthened if the difference between the control lists were artificially high owing to a difference in word length.

Rehearsal Time

In the planned analysis of simple main effects, rehearsal time was shorter for each phonologically similar pool, as compared with its control pool, whether the positions of the repeated phonemes were fixed [$F(1,12) = 103.99, MS_e = 0.07, p < .001$] or variable [$F(1,12) = 39.52, MS_e = 0.06, p < .001$]. For the two phonologically similar pools, rehearsal time was shorter when repeated phonemes were in fixed positions [$F(1,12) = 24.46, MS_e = $

0.06, $p < .001$]. The similarity × type of repeated phoneme interaction was significant [$F(1,12) = 10.57$, $MS_e = 0.07, p < .01$].

Rehearsal Rate

In the planned analysis of simple main effects, rehearsal rate was faster (by 0.125 items per second) for the phonologically similar pool, as compared with its control, when repeated phonemes were in fixed positions [$F(1,12) = 8.11, MS_e = 0.02, p < .05$]. This difference in rehearsal rates parallels the difference for recall rates. Again, we do not know why the rehearsal rates differ, given that the reading rates did not differ in the calibration study. Again, we note that the difference in span for these two pools cannot be due to the difference in rehearsal rates, because of the direction of the difference.

Summary

First, mere repetition of phonemes in different positions produced a decrease in span. The effect of phonological similarity, of either type, on span is not due to word length. We conclude this because (1) reading rate in the calibration study is the same for the pools, (2) for variable positions, the recall rates and rehearsal rates were nearly equal for the similar pool and its control, and (3) for fixed positions, these rates were significantly different, but in the direction opposite to that needed to explain the effect. This replicates the results of Hulme and Tordoff (1989) and Schweickert et al. (1990).

Second, repetition of phonemes in the same positions produced a larger decrease in span. The phonologically similar word pools differ in word frequency, so one could argue that this causes the difference in spans for the two pools. However, the results of Poirier and Saint-Aubin (1996) suggest that the difference in spans for the two corresponding control pools would then be the same size, but it was not.

Finally, phonological similarity, of either type, evidently decreased memory span by decreasing the useful lifetime of the memory trace. This is indicated by the decrease in the time required to rehearse or recall a list of span length, for the phonologically similar items.

Rehearsal and Recall Times as Predictors of Span

Rehearsal time, t_{rh}, and recall time, t, both correlate highly with memory span, s. The correlation between t_{rh} and s is .993; that between t and s is .986.

The regression equations have the form span equals rate times time (i.e., $s = rt$). The slopes (rates) were significantly different from 0 [for rehearsal time, $t(2) = 12.11$, $p < .01$; for recall time, $t(2) = 8.25, p < .05$]; the intercepts were not [for rehearsal time, $t(2) = 1.45$, n.s.; for recall time, $t(2) = 1.03$, n.s.]. Intuitively, recall time is a better estimate than rehearsal time of the useful lifetime of the representation (Dosher & Ma, 1998).

Time is the predictor here. Sometimes, when rate is the predictor, for reasons that are not understood, a positive intercept is reported (e.g., Baddeley et al., 1975; Hulme, Maughan, & Brown, 1991).

THE MULTINOMIAL TREE MODEL

We use a multinomial tree model to compare two mechanisms proposed for the phonological similarity effect. In the model, an item is recalled correctly if it is intact or if it is not intact but is successfully redintegrated. For a given item, let I be the probability that the representation is intact at the time of recall. Then, the probability that the item is not intact (i.e., is degraded) is $1 - I$. If an item is degraded, let R be the probability it is redintegrated and recalled correctly. Then, the probability of correct recall of the item is $P = I + (1 - I)R$.

One hypothesis, the acid bath, is that phonological similarity increases the probability, $1 - I$, that the memory representation is degraded. Another hypothesis is that phonological similarity has its effect on trace comparison; this is interpreted as the assumption that phonological similarity decreases the probability, R, that a degraded representation is redintegrated.

In the following, subscripts s and d indicate the phonologically similar (experimental) and dissimilar (control) pools, respectively. Subscript v indicates the phonologically similar pool in which repeated phonemes are in variable positions and its corresponding dissimilar (control) pool. Subscript f, analogously, indicates the phonologically similar pool in which repeated phonemes are in fixed positions and its corresponding dissimilar (control) pool. Subscript i indicates the serial position of an item. For example, the probability of correctly recalling the item in serial position i, for the pool of phonologically similar items, with variable positions of repeated phonemes, is P_{svi}.

For both versions of the model, serial position is assumed to change the probability, I, that an item is intact. Simply put, items at the beginning and end of the list are more likely to be intact than those in the middle. Also, for both versions of the model, the probability, R, of redintegrating a degraded item is assumed to depend on word frequency, following Hulme et al. (1997). Given a degraded representation, the higher the frequency of the word, the more likely the participant is to reconstruct it correctly. For our word pools, this means that R depends on the type of phonological similarity (fixed or varied positions of repeated phonemes), because these types differ in word frequency.

Here are the two versions of the model. Consider the item in serial position i, for the pool of phonologically similar items, with variable positions of repeated phonemes. If phonological similarity changes the probability of redintegration, the model becomes

$$P_{svi} = I_i + (1 - I_i)R_{sv}. \qquad (1)$$

(Serial position affects I, similarity affects R, word frequency affects R.) If phonological similarity changes the probability an item is intact, the model becomes

$$P_{svi} = I_{is} + (1 - I_{is})R_v. \qquad (2)$$

(Serial position affects I, similarity affects I, word frequency affects R.) For lists made from a pool of dissim

ilar (control) items, the subscript d is substituted for the subscript s in Equations 1 and 2. For the pool in which phonologically similar items have repeated phonemes in fixed positions or its control pool, the subscript f is substituted for the subscript v in Equations 1 and 2.

Model Fits

To test the models quantitatively, parameter estimates and goodness-of-fit statistics were calculated, using the General Processing Tree program by Hu (1998). With the staircase method, the number of trials for a particular list length may differ from condition to condition. List lengths 5 and 6 were used for model fitting because they have the largest number of observations in each of the four conditions. The observed probability of correct recall is shown in Table 3, and the total number of trials, correct and incorrect, in each condition are shown in Table 4.

Parameter estimates for the version in Equation 1 are shown in Table 3. As an example, with these parameters, the model predicts $.93 + .76 - .93 \times .76 = .98$ for probability of correct recall, list length 5, serial position 1, dissimilar pool, variable positions. The observed value is .98.

Parameter estimates for the version in Equation 2 are as follows. For list length 5, dissimilar pools, $I_{d1}, \ldots, I_{d5}$ are, respectively, .97, .88, .70, .54, .62. For similar pools,

$I_{s1}, \ldots, I_{s5}$ are, respectively, .92, .73, .34, .02, .04. For both, R_v is .61, and R_f is .41. List length 6, dissimilar pools, $I_{d1}, \ldots, I_{d6}$ are, respectively, .98, .94, .75, .52, .35, .41. For similar pools, $I_{s1}, \ldots, I_{s6}$ are, respectively, .92, .81, .48, .17, .01, .00. For both, R_v is .37, and R_f is .23. It is hard to interpret parameter estimates across list lengths, because the staircase method leads to different sets of observations for each.

Both versions of the model fit quite well, as is shown by the correlations between predicted and observed values of the frequency of correct responses. For list length 5, for the version in Equation 1, $r^2 = .992$, and for the version in Equation 2, $r^2 = .984$. For list length 6, for the version in Equation 1, $r^2 = .997$, and for the version in Equation 2, $r^2 = .999$. Each version accounts for a large proportion of the variance.

A goodness-of-fit measure used with multinomial tree models is G^2, which is -2 times the likelihood ratio and has asymptotically a chi-square distribution (see, e.g., Bishop, Fienberg, & Holland, 1975). The larger the value of G^2, the worse the fit. Parameters were estimated to minimize G^2 (on the basis of frequencies of correct and incorrect responses, not of their probabilities). Degrees of freedom are different for the two versions of the model, because the acid bath version in Equation 2 has more pa-

Table 3
Observed Probability of Correct Recall and Interaction Contrasts and Parameter Estimates When Phonological Similarity Influences Redintegration

Condition	Serial Position						Parameter	Estimate
	1	2	3	5	4			
			List Length 5					
Probabilities								
Dissimilar variable	.98	.95	.87	.80	.78		R_{dv}	.76
Dissimilar fixed	.99	.94	.84	.82	.76		R_{df}	.75
Similar variable	.98	.92	.75	.63	.65		R_{sv}	.59
Similar fixed	.94	.82	.61	.44	.40		R_{sf}	.31
Parameter	I_1	I_2	I_3	I_5	I_4			
Estimate	.93	.76	.41	.17	.12			
Interaction contrasts								
Dissimilar variable	–	–	–	–	–			
Dissimilar fixed	–	$-.02$	$-.04$	.01	$-.02$			
Similar variable	–	$-.03$	$-.12$	$-.18$	$-.12$			
Similar fixed	–	$-.09$	$-.22$	$-.32$	$-.34$			

	Serial Position					
	1	2	3	4	5	6
			List Length 6			
Probabilities						
Dissimilar variable	.98	.96	.84	.69	.63	.57	R_{dv}	.59
Dissimilar fixed	.99	.95	.81	.63	.55	.52	R_{df}	.51
Similar variable	.96	.89	.69	.52	.35	.36	R_{sv}	.32
Similar fixed	.93	.82	.55	.30	.25	.25	R_{sf}	.16
Parameter	I_1	I_2	I_3	I_4	I_6	I_5		
Estimate	.95	.86	.56	.24	.08	.05		
Interaction contrasts								
Dissimilar variable	–	–	–	–	–	–		
Dissimilar fixed	–	$-.02$	$-.04$	$-.07$	$-.09$	$-.06$		
Similar variable	–	$-.05$	$-.12$	$-.15$	$-.25$	$-.18$		
Similar Fixed	–	$-.09$	$-.23$	$-.34$	$-.32$	$-.26$		

Note—Operations were performed before rounding.

Table 4
Proportions of Various Recall Errors

Similarity	Type	Wrong Position	Omission With Intrusion	Omission Without Intrusion	Incorrect Trials	Number of Trials
			List Length 5			
Similar	fixed	.09	.12	.27	189	256
Dissimilar	fixed	.14	.09	.19	72	233
Similar	variable	.07	.11	.23	141	284
Dissimilar	variable	.07	.06	.28	59	201
			List Length 6			
Similar	fixed	.07	.11	.37	70	83
Dissimilar	fixed	.11	.06	.25	158	267
Similar	variable	.11	.07	.31	138	184
Dissimilar	variable	.12	.04	.26	136	272

Note—Each proportion is the number of occurrences of the type of error divided by the product of list length and the number of incorrect trials. More than one type of error may occur on a given trial. Incorrect trials are those in which an error of any kind occurred in any position.

rameters (twice as many I parameters, half as many R parameters). For list length 5, for the version in Equation 1, in which phonological similarity influences redintegration, $G^2 = 9.28, df = 11$, and for the version in Equation 2, in which phonological similarity influences degradation, $G^2 = 20.90, df = 8, (p < .01)$. For list length 6, for the version in Equation 1, $G^2 = 21.38, df = 14$, and for the version in Equation 2, $G^2 = 7.29, df = 10$. The p value of .01 for list length 5 for the version of the model in Equation 2 must be taken as an approximation, because the repeated observations on the same participants are probably not independent. Nonetheless, the low p value for G^2 argues against the acid bath version.

The two versions of the model have different numbers of parameters; furthermore, neither is a special case of the other. To compare models in this situation, Akaike's information criterion (AIC) was developed (Akaike, 1973). It is a goodness-of-fit measure, which incorporates in a principled way the number of parameters in a model. If two models differ in the value of AIC, the model with the smaller value is better, according to this criterion (e.g., Takana & Shibayama, 1992). For a particular version of the model,

$$AIC = \Sigma G^2 + 2s,$$

where the summation is over list lengths 5 and 6 and s is the number of parameters in the version of the model.

For the version of the model in Equation 1, in which phonological similarity influences redintegration, $AIC = 68.66$. For the version in Equation 2, in which phonological similarity influences degradation, $AIC = 80.19$. The version in Equation 1 does better.

In brief, although both versions of the model give a good account of the data, the goodness-of-fit values favor the trace comparison version in Equation 1. (A model in which phonological similarity affects both I and R is a logical possibility, but the number of parameters is so large that it will not be considered here.)

Qualitative Predictions

Both versions of the model make qualitative predictions. Consider the trace comparison version in Equation 1, in which phonological similarity affects R, serial position affects I, and word frequency affects R. According to the model, the probabilities of correct recall can be arranged in a matrix, and the columns can be ordered so that I decreases across the columns (see Table 3). Then, (1) the probability of correct recall will decrease monotonically across the columns. Moreover, the rows can be ordered so that R decreases down the rows. Then, (2) probability will decrease monotonically down the rows. (The predictions are derived in Schweickert, 1993.)

Predictions (1) and (2) of monotonic probability are satisfied almost completely in Table 3. The exceptions are mostly found in the last serial position, where a recency effect not included in the model may be operating.

Further qualitative predictions are made (Schweickert, 1993). For a cell outside the first row and first column, an interaction contrast can be calculated, using cells in the first row and first column as baselines. For example, for list length 5, for the cell in row 2, column 2, the interaction contrast is $.94 - .95 - .99 + .98 = -.02$. The interaction contrast for each cell is shown in Table 3. The version of the model in which phonological similarity affects redintegration, Equation 1, makes two more qualitative predictions, these with regard to interaction contrasts. (3) The interaction contrasts all have the same sign—negative. (4) When the rows and columns are ordered to produce monotonic probability of correct recall as described in Predictions 1 and 2, the interaction contrasts will be monotonically decreasing across the columns and down the rows.

There is agreement between the above predictions and the observations in Table 3. A few interaction contrasts are not in the predicted order, but almost all are.

The version of the model in which phonological similarity affects degradation, Equation 2, also makes the qualitative predictions above, but only for cells within a

given level of phonological similarity. The reason is that the dissimilar word pools have one set of parameters, whereas the similar word pools have another set. The qualitative predictions are made for word pools having the same set of parameters. No predictions are made comparing two pools having different sets of parameters. For example, in Table 3, probability of correct recall is greater for the dissimilar fixed pool than for the similar variable pool in every serial position, for both list lengths 5 and 6. For the version of the model in Equation 2, this ordering is a chance occurrence, with no further explanation. But the model in Equation 1 predicts that performance should be better for one of these pools than for the other (without specifying which one). Qualitative patterns appear in the data that are predicted by the trace comparison version in Equation 1 but are ignored by the acid bath version in Equation 2, an advantage for the version in Equation 1.

Item and Order Information

Both versions of the model focus on errors made for individual items, yet phonological similarity is sometimes said to have its major effect on memory span by causing items to be recalled out of order (e.g., Burgess & Hitch, 1996). The classic source is Conrad (1965), who said, "items which acoustically confuse with each other are likely to transpose in recall" (p. 161). We can consider the relative contribution of order errors in our data, although not in fine detail since it was not designed for that purpose.

Experimenters listened to the tape recordings and noted, for each list, which words were spoken during recall and the positions they were recalled in. For list lengths 5 and 6, Table 4 shows what proportion of time a presented item was spoken during recall, but in the wrong position. The table also shows what proportion of the time a presented item was not recalled at all.

It is clear that for our stimuli and procedures, (1) the major source of errors is omissions, (2) the major difference between similar and dissimilar items is omissions, and (3) there is no tendency for items to be recalled more often in the wrong position for the phonologically similar lists than for the dissimilar lists. If the proportion of position errors is adjusted by dividing by the proportion of items not omitted (Murdock, 1976), the conclusions are unchanged, as the reader can check. Modeling the individual item errors is appropriate for these data.

Our procedures differ in several ways from those of the classic studies of Conrad (1964, 1965), probably leading to the difference in results. With our scoring, the participant knows that a list will be scored as incorrect once an error is made, so there is no incentive to continue recalling beyond the first error noticed. Omissions tended to increase as serial position increases. In Conrad's studies, a written copy of the stimulus pool was provided to the participant during recall, to avoid intrusion errors. The participants recalled by filling in boxes, one for each item presented; although the boxes were ordered, the participants were not required to fill them in order from first to last. Finally, in Conrad (1964), the only responses analyzed

were those having exactly one letter wrong, and in Conrad (1965), "sequences containing omissions or more than two wrong letters were ignored" (p. 162). The procedures followed by Conrad were quite different from ours.

DISCUSSION

Several models propose that representations of items are sometimes noisy but can be reconstructed (e.g., Burgess & Hitch, 1992; Cowan, 1992; Hulme et al., 1991; Lewandowsky & Murdock, 1989; Murdock, 1983; Nairne, 1990; Neath & Nairne, 1995; Sperling & Speelman, 1970). According to the acid bath hypothesis, phonological similarity increases the noise, whereas according to the trace comparison hypothesis, phonological similarity makes items harder to reconstruct. We compared the two hypotheses, using a multinomial tree. Each hypothesis could be interpreted as a special version of the model. The version incorporating the acid bath hypothesis actually accounts for a large proportion of the variance in probability of correct recall. Given this good fit, it would be hard for an alternative to do immensely better. But we found a distinctly better fit for the trace comparison version.

Our conclusion is consistent with earlier literature. Investigators have argued against the acid bath hypothesis in recognition memory (Ingleby, 1972), the Brown–Peterson paradigm (Butler & Chechile, 1976; Chechile, 1977), and immediate probed recall (Sanders & Willemsen, 1978).

We also find that memory span is different for our two pools of phonologically similar words. These two pools differ primarily in whether the positions in which the repeated phonemes occur are fixed or variable. Fallon and Groves (1998) recently reported a similar result. Our finding is consistent with an intriguing study on proactive interference by Tehan and Humphreys (1998). Their Experiment 4 was on immediate recall of lists. Consider a typical item, HOSE. Performance on HOSE was worse if (1) the list contained a word rhyming with HOSE and (2) the preceding list contained the words MUCK, ROOM, and DAY, which have the phonemes required to form an interfering foil, RAKE. What is of interest here is that the effect occurred even if the phonemes in the preceding list were not presented in the order they appear in RAKE. Note especially that individual phonemes and phoneme order (the rhyme) have separate roles in the effect.

Our results indicate that neither the memory mechanism proposed by Sperling and Speelman (1970) nor the one proposed by Burgess and Hitch (1992) can be the sole mechanism producing the phonological similarity effect. Recall that the phoneme model of Sperling and Speelman predicts no confusion between STOP (/stap/) and TIPS (/tIps/), because the three phonemes, /s/, /t/, and /p/, that STOP and TIPS have in common occur in different positions. On the other hand, the connectionist model of Burgess and Hitch (1992; Burgess, 1995; Burgess & Hitch, 1996) predicts that common phonemes have the same effect whether they are repeated in the same positions (e.g., SPIT [/spit/]

and SPOT [/spat/], or in different positions (e.g., STOP [/stap/] and TIPS [/tIps/]). Neither prediction is supported by our results. In the future, it may be fruitful to explore models combining both mechanisms.

Finally, when the probability of redintegration is higher, the representation can be used for a longer time (Schweickert, Chen, & Poirier, 1999; Schweickert et al., 1990). The result is the increase in recall time found for phonologically dissimilar items.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267-281). Budapest: Akademiai Kiado.

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., THOMSON, N., & BUCHANAN, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning & Verbal Behavior*, **14**, 575-589.

BATCHELDER, W. H., & RIEFER, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, **6**, 57-86.

BISHOP, Y. M. M., FIENBERG, S. E., & HOLLAND, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

BOWER, G. H., & BOLTON, L. S. (1969). Why are rhymes easy to learn? *Journal of Experimental Psychology*, **26**, 453-461.

BURGESS, N. (1995). A solvable connectionist model of immediate recall of ordered lists. In G. Tesauro, D. Touretzky, & T. K. Leen (Eds.), *Neural information processing systems* (Vol. 7, pp. 51-58). Cambridge, MA: MIT Press.

BURGESS, N., & HITCH, G. J. (1992). Towards a network model of the articulatory loop. *Journal of Memory & Language*, **31**, 429-460.

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

BUTLER, K., & CHECHILE, R. (1976). "Acid bath" effects on storage and retrieval PI. *Bulletin of the Psychonomic Society*, **8**, 349-352.

CHASE, W. G. (1977). Does memory scanning involve implicit speech? In S. Dornic (Ed.), *Attention and performance VI* (pp. 607-628). Hillsdale, NJ: Erlbaum.

CHECHILE, R. (1977). Storage-retrieval analysis of acoustic similarity. *Memory & Cognition*, **5**, 535-540.

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

CONRAD, R. (1965). Order errors in immediate recall of sequences. *Journal of Verbal Learning & Verbal Behavior*, **4**, 161-169.

COWAN, N. (1992). Verbal memory span and the timing of spoken recall. *Journal of Memory & Language*, **31**, 269-295.

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

FALLON, A. B., & GROVES, K. (1998, June). *Phonemic similarity and immediate serial recall: Different materials produce different effects*. Paper presented at the Quebec 98 Conference on Short Term Memory, Quebec.

FALMAGNE, J. C. (1986). Psychophysical measurement and theory. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 1, pp. 1-66). New York: Wiley.

FRANCIS, W. N., & KUČERA, H. (1982). *Frequency analysis of English usage*. Boston: Houghton Mifflin.

GATHERCOLE, S., FRANKISH, C. R., PICKERING, S. J., & PEAKER, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 84-95.

HINTZMAN, D. L. (1965). Classification and aural coding in short-term memory. *Psychonomic Science*, **3**, 161-162.

HOOSAIN, R., & SALILI, F. (1988). Language differences, working memory, and mathematical ability. In M. M. Greunberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 2, pp. 512-517). New York: Wiley.

HOUGHTON, G., HARTELY, T., & GLASSPOOL, D. W. (1996). The representation of words and nonwords in short-term memory: Serial order and syllable structure. In S. E. Gathercole (Ed.), *Models of short-term memory* (pp. 101-128). Hove, U.K.: Psychology Press.

HU, X. (1998). GPT.EXE [Computer program]. Memphis: University of Memphis.

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.

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, & Cognition*, **23**, 1217-1232.

HULME, C., & TORDOFF, V. (1989). Working memory development: The effects of speech rate, word length, and acoustic similarity on serial recall. *Journal of Experimental Child Psychology*, **47**, 72-87.

INGLEBY, J. D. (1972). Acoustic similarity and interference in short-term recognition memory: Some results of a decision theory analysis. *Quarterly Journal of Experimental Psychology*, **24**, 505-514.

LEWANDOWSKY, S., & MURDOCK, B. B., JR. (1989). Memory for serial order. *Psychological Review*, **96**, 25-57.

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

MURDOCK, B. B., JR. (1976). Item and order information in short-term memory. *Journal of Experimental Psychology: General*, **105**, 191-216.

MURDOCK, B. B., JR. (1983). A distributed memory model for serial-order information. *Psychological Review*, **90**, 316-338.

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

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

POIRIER, M., & SAINT-AUBIN, J. (1996). Immediate serial recall, word-frequency, item identity and item position. *Canadian Journal of Experimental Psychology*, **50**, 408-412.

POSNER, M. I., & KONICK, A. F. (1966). On the role of interference in short-term retention. *Journal of Experimental Psychology*, **72**, 221-231.

SANDERS, A. F., & WILLEMSEN, E. M. (1978). The course of proactive interference in immediate probed recall. *Acta Psychologica*, **42**, 133-144.

SCHWEICKERT, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition*, **21**, 168-175.

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

SCHWEICKERT, R., CHEN, S., & POIRIER, M. (1999). Redintegration and the useful lifetime of the verbal memory representation. *International Journal of Psychology*, **34**, 447-453.

SCHWEICKERT, R., GUENTERT, L., & HERSBERGER, L. (1990). Phonological similarity, pronunciation rate, and memory span. *Psychological Science*, **1**, 74-77.

SCHWEICKERT, R., McDANIEL, M., & RIEGLER, G. (1994). Effects of generation on immediate memory span and delayed unexpected free recall. *Quarterly Journal of Experimental Psychology*, **47A**, 781-804.

SPERLING, G., & SPEELMAN, R. G. (1970). Acoustic similarity and auditory short-term memory: Experiments and a model. In D. A. Norman (Ed.), *Models of human memory* (pp. 151-202). New York: Academic Press.

TAKANA, Y., & SHIBAYAMA, T. (1992). Structures in stimulus identification data. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 335-362). Hillsdale, NJ: Erlbaum.

TEHAN, G., & HUMPHREYS, M. S. (1998). Creating proactive interference in immediate recall: Building a DOG from a DART, a MOP, and a FIG. *Memory & Cognition*, **26**, 477-489.

WATKINS, M. J., WATKINS, O. C., & CROWDER, R. G. (1974). The modality effect in free and serial recall as a function of phonological similarity. *Journal of Verbal Learning & Verbal Behavior*, **13**, 430-447.

APPENDIX A
Recall Time

When the staircase method is used, the length of the list to be rehearsed and recalled increases or decreases, depending on whether recall was correct or incorrect on the previous trial. The average list length s is a good estimate of the memory span (see, e.g., Falmagne, 1986).

Let t denote the recall time for a list of length s. There is a difficulty in obtaining an estimate of t. Because recall time is measured on every trial, it might seem at first that recall time averaged over all the trials would be a good estimate of t. But on half the trials, the list is incorrectly recalled, often because only part of the stimulus list is uttered. Recall time from an incomplete recall is less than the time it would take to recall the entire stimulus list. Hence, recall time averaged over correctly and incorrectly recalled lists underestimates the time to recall a list of span length.

An estimate of the time to recall a list of length s can be found as follows. With the staircase method, a trial ends with correct recall or incorrect recall. Let the average list length across the correct trials be denoted by s_c, which is less than the memory span s. Let the recall time averaged across only the correct trials be denoted by t_c. Then, the time to recall a single item, when recall is correct, is t_c/s_c.

Our estimation procedure uses t_c/s_c as an estimate of the time to recall a single item in a list of span length. Then, we estimate the time t to recall a list of span length, s items, as s times t_c/s_c—that is,

$$t = st_c/s_c. \tag{A1}$$

The recall rate across correct trials is $r_c = s_c/t_c$. We use this rate as an estimate of the rate, r, at which a list of span length would be recalled. That is, we assume $r = r_c$, so $s = rt$.

APPENDIX B
Notation

s	Memory span, average list length for one block of trials
s_c	Average list length across the correct trials for a block
t_c	Average recall time across correct trials for a block
$r_c = s_c/t_c$	Recall rate for correct trials in a block
$t = s/r_c$	Estimated recall time for a list of length s
$r = s/t = r_c$	Estimated recall rate for a block of trials
t_{rh}	Rehearsal time for a block of trials
r_{rh}	Rehearsal rate for a block of trials
P	Probability of correct recall
I	Probability an item is intact
R	Probability a nonintact item is redintegrated

Subscripts

i	Serial position
s	Phonologically similar
d	Phonologically dissimilar
f	Phonologically similar with repeated phonemes in fixed positions or corresponding control
v	Phonologically similar with repeated phonemes in variable positions or corresponding control