

The relationship between absolute and proportion scores of serial order memory: Simulation predictions and empirical data

C. PHILIP BEAMAN

University of Reading, Reading, England

The capability of a feature model of immediate memory (Nairne, 1990; Neath, 2000) to predict and account for a relationship between absolute and proportion scoring of immediate serial recall when memory load is varied (the list-length effect, LLE) is examined. The model correctly predicts the novel finding of an LLE in immediate serial order memory similar to that observed with free recall and previously assumed to be attributable to the long-term memory component of that procedure (Glanzer, 1972). The usefulness of formal models as predictive tools and the continuity between short-term serial order and longer term item memory are considered.

One of the more interesting things about immediate recall studies is the limitations they reveal in the ability to recall accurately, in the correct order, a short sequence of items presented only a matter of seconds before. A number of different functional accounts of why performance is restricted are available (see Miyake & Shah, 1999, p. 421, for eight different possible sources of limitation). However, the majority of studies on immediate ordered recall vary some aspect of the stimuli (e.g., word length) or the situation (e.g., the requirement to carry out another task simultaneously) rather than look directly at the relationship between the amount the participant is asked to recall and the amount actually recalled. This is not the case with free recall, for which the *list-length effect* (LLE; Murdock, 1962; Roberts, 1972) is well established. This “effect” is a simple observation of two things: First, the proportion of words recalled is an inverse function of the number of to-be-recalled words presented (list length), and second, simultaneously, the total number of words recalled *increases* with list length.

The first of these facts is also true of immediate serial recall and is responsible for the short-term memory (STM) span measure, the cutoff point at which proportion correct drops below 1.0 as a function of list length. It is perhaps surprising that, to the author’s knowledge, the possibility that the second observation might also be true for serial recall has never been investigated, and the existence of an LLE in short-term serial order memory has not been considered.

Within the free recall procedure, the LLE is often considered a long-term memory (LTM) phenomenon

(Glanzer, 1972) because it primarily has an impact on the preresency portion of the free recall curve. The list lengths tested when the effect is observed in free recall are considerably longer (10–30 items) than would generally be considered in a serial recall situation (usually 5–9 items) and even more so in the observation of an LLE in picture-recognition tasks (up to 150 pictures; Strong, 1912). Because many STM theorists assume that different mnemonic principles apply to the particularly long list lengths used in free recall as opposed to those involved in holding smaller sequence lengths within serial STM (e.g., Baddeley, 2003), there may be no LLE within immediate serial recall. Other researchers, however, have argued that because many of the stimuli and testing procedures are identical, serial and free recall involve similar principles (Anderson, Bothell, Lebiere, & Matessa, 1998; Brown, Neath, & Chater, 2004; Ward, 2001). If this second group of theorists is correct, there is no a priori reason why LLEs should not occur in serial recall.

To examine the relationships between the two measures of immediate recall (absolute number and proportion correct) involved in the LLE, it is desirable to have a fully worked-out theory of span limitations in serial recall. It is also desirable that any such model fit the same data set when the data are expressed in different ways (Laming, 1999). The model considered here is the feature model (FM) of Nairne (1990). For the present purposes, it has a number of interesting properties. It gives an account of span limitations in immediate memory performance (Nairne, 1990, pp. 263–264) and contains accounts of a variety of other mechanisms (e.g., to account for irrelevant speech effects; Neath, 2000) that can be investigated in conjunction with span but have not, as yet, been implemented in many of the competing models of immediate serial recall. It is a model of immediate ordered recall that relies on the relative “fit” of cues to a search set of target items to produce accurate recall, rather than the relative activation levels or “strength” of items in many free recall

Thanks to Philip Smith for help with the mathematical aspects of the feature model and to Ian Neath for explanation of some of the workings of the model at mathematical and conceptual levels. Correspondence regarding this article may be sent to C. P. Beaman, School of Psychology, University of Reading, Earley Gate, Whiteknights, Reading RG6 6AL, England (e-mail: c.p.beaman@reading.ac.uk).

models (Ward, 2002). It emphasizes continuity between forms of memory (Crowder & Neath, 1991; Nairne & Neath, 2001). An untested claim is that it will produce LLEs, although the form of these effects is not described (Nairne, 2001, p. 292). It follows that if an LLE is expected with immediate serial recall, it should be apparent within the output of the FM.

Background to the Feature Model

Nairne (2001) presents a general introduction to the ideas behind the FM, and technical summaries appear in Nairne (1990) and Neath (2000). A simplified version is presented here. The basic idea is that recall is guided by a set of cues (*primary memory*, PM), which may be more or less effective in identifying the target item or event from a search set defined within *secondary memory* (SM), or memory proper. Unlike in other models of immediate recall, there is no dedicated STM store as such; the same set of PM cues are assumed to suffice for both shorter- and longer term recall. Concepts often associated with dedicated short-term, serial-order storage, such as direct access, activation, and decay, are absent from the model (see Nairne, 2002, for a review). Cues do not decay but are subject to a process of interference. Anything that makes the cues less effective (e.g., similarity between the representations of to-be-recalled targets) or degrades the cues so that two or more targets could fit the profile specified by the cues (e.g., the presence of irrelevant, interfering material at encoding) renders recall less accurate. Increasingly ineffective cuing techniques make up the basic mechanism for memory limitations. Thus, the model's account of span is a combination of item-by-item interference degrading the cues and an increasing variety of possible targets fitting the specification given by the cues.

Formally, items in SM and cues in PM are presumed to be made up of internally generated (modality-independent) and externally generated (modality-dependent) features organized as row vectors. For the purposes of simulation, feature values are randomly generated and take binary values. The main source of forgetting in the model is retroactive interference: If feature x of item $n + 1$ is the same as feature x of item n , then feature x of item n is lost, or overwritten, and returns a value of 0. A simplifying assumption is that only immediately adjacent items overwrite. The final item of a to-be-recalled list is, by definition, not followed by a further list item of the same form, and so is normally overwritten only by internally generated (modality-independent) activity. Earlier items, however, are susceptible to overwriting by modality-independent and modality-dependent features of the succeeding item.

The relative number of accurate features available to cue the item in SM dictates recall performance. The probability that a particular SM trace, SM_j , will be retrieved from a given set of traces is calculated for a particular PM cue, PM_i . The distance between the target SM item and its PM cue in terms of their shared features is calculated according to Equation 1. The value M_k is equal to 1 if the

feature at position k of PM cue i does not match the feature at the corresponding position of SM representation j , and is equal to 0 otherwise. The number of mismatches across the features is summed in the numerator of Equation 1. The value N is the number of features in each of the vectors, and a is a scaling parameter representing overall level of attention.

$$d_{ij} = a \sum \frac{M_k}{N} \quad (1)$$

Next, the distance between the PM and SM items is transformed to provide a similarity metric (Equation 2).

$$s(i, j) = e^{-d_{ij}} \quad (2)$$

The similarity between PM cues and SM representations is used to calculate the probability that a particular SM trace, SM_j , will be "sampled" given a particular PM cue, PM_i . The probability of sampling a particular item is given by a similarity-based choice rule (Equation 3), and a simplified example of this process, calculating Equations 1–3, is shown in Figure 1.

$$P(SM_j | PM_i) = \frac{s(i, j)}{\sum_{l=1}^{\text{no. vectors}} s(i, l)} \quad (3)$$

By design, sampling probabilities for all the items under consideration in SM sum to unity. The choice of which of these SM items to sample is made by randomly generating a number from 1 to 100 and identifying the item corresponding to the region within which that random number falls when the sampling probabilities are put together as cumulative probabilities (see Figure 1).

Next, the probability of recovering a sampled item is given by Equation 4, where c is a constant and r is the number of times the sampled item has already been recalled on this trial. This equation, and the r parameter, are used to reduce the likelihood of recalling the same item on multiple occasions, which participants avoid doing even when the same item is repeated within the to-be-recalled list (the "Ranschburg effect"; Jahnke, 1969). If the item has not been recalled previously, Equation 4 gives e^{-0} —that is, a probability of 1.0 that the sampled item will be recovered and recalled. For each time the item has been previously recalled, however, the probability of recovery calculated by Equation 4 is reduced. If two attempts at recovery are unsuccessful, an omission error is recorded.

$$P_r = e^{-cr} \quad (4)$$

The intent of the FM is to provide qualitative, rather than quantitative fits to the data, showing the correct pattern of results by changing only the parameters that are identifiable as reflecting particular psychological processes. In the present study, this philosophy is taken one step further, to ask whether, without altering the default set of parameters, the FM can be used to predict the existence of an LLE in immediate serial recall and the par-

Primary memory cue degraded by retroactive interference
(Zeros represent feature values lost through overwriting.)

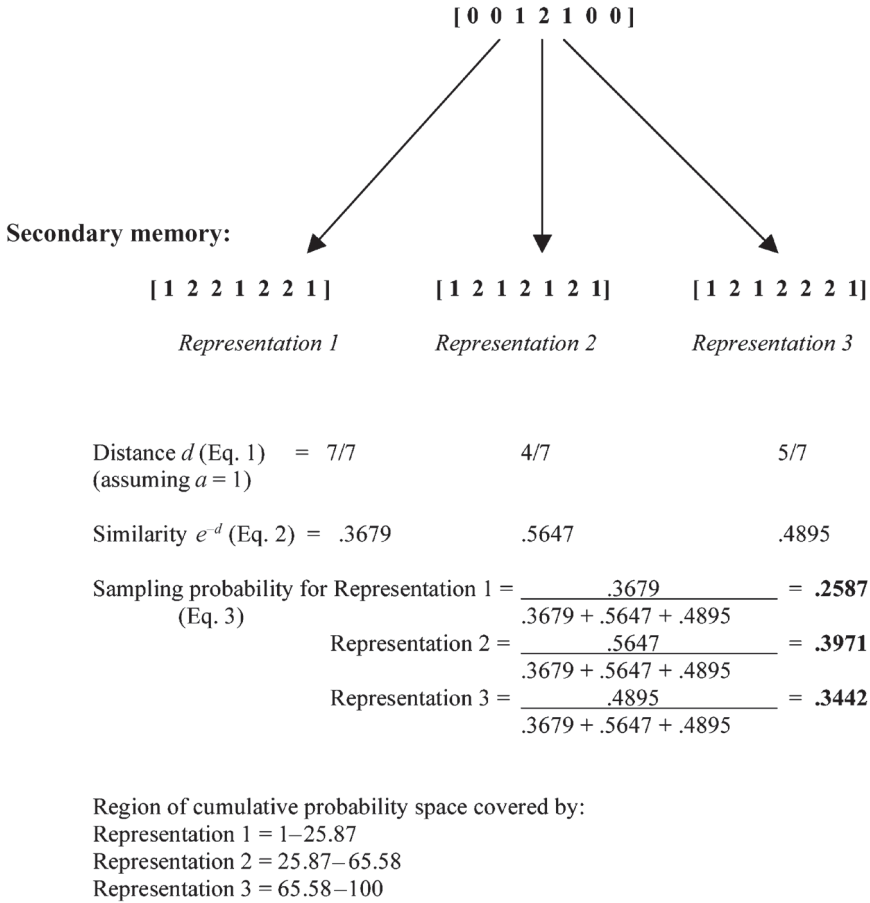


Figure 1. Example computation of distance and sampling probabilities. In this example, there are seven features, of which the values for all but three have been lost from primary memory. For clarity, binary feature values of 1 and 2 have been used rather than the -1 and +1 values used by Nairne (1990) and Neath (2000). Distance between the degraded representation in primary memory and three secondary memory representations is calculated as the number of feature mismatches divided by the number of feature comparisons. Similarity is the exponential of distance, and sampling probability is the ratio of these similarities. Random number generation determines which region (item) is sampled. The probability of recalling that item (correctly or not) is then determined by Equation 4 (not shown), which takes into account the number of times the item has already been recalled on the present trial.

ticular form it takes. The following describes a simulation study that was run using the default parameters described by Neath (2000; see Appendix); the results were plotted before any data were scored.

Initial Data Analysis and Simulation

A review of the literature found that almost all studies varying list length failed to report scores beyond span. Studies that examined supraspan lists did not vary list length independently of other variables. One study was identified as presenting data in a form appropriate for rescoreing to determine whether LLEs exist in short-term serial order memory (Drewnowski, 1980). Rescoreing proceeded by taking the average recall score across serial

positions to give an average proportion score and multiplying this by list length to give a mean number of items score. The results of this procedure are, of necessity, inexact, and give no indication that any LLE observed within this single data set will be statistically reliable, so a further initial experiment was run to confirm observations from these older data.

In the initial experiment, participants were presented with five-, seven-, or nine-item lists of digits (sampled randomly and without replacement from the set of nine digits), at a rate of one digit per second. Half the trials were accompanied by irrelevant speech, which the participants were asked to ignore. This procedure was followed to bring performance in the short-list conditions

down from ceiling and also to confirm, if possible, that any relationships between the variables to be considered and the effects of irrelevant speech followed the same qualitative pattern in the human data and the FM. For example, one plausible outcome could have been that if irrelevant speech degraded PM cues, such degradation might counteract the LLE and keep the absolute number of items correctly recalled fairly constant, increasing the rate at which the proportion of correct recalls dropped. Participants were presented with blocks of five-, seven-, and nine-item lists, in that order. The FM simulation was run before the rescoring of Drewnowski (1980) and before the results of this experiment were known. Thus, the simulation results are blind predictions, not the result of post hoc parameter fitting.

Predictions from the FM are shown in the upper panel of Figure 2, and the experimental data, in the lower. The left y-axis scores performance according to proportion correct, and the right y-axis scores it according to the mean number correct. Tests show main effects of list length and speech on the proportion correct scores [$F(2,74) = 80.84$, $p < .001$, and $F(1,37) = 77.72$, $p < .001$, respectively] and a speech \times length interaction [$F(2,74) = 5.21$, $p = .008$]. There was a linear effect of list length [$F(1,37) = 127.93$] and there were significant differences between the shortest and the longest lists, both with and without

speech [$t(37) = 10.21$ and $t(37) = 10.01$, respectively; $p < .001$ in both cases]. The same pattern was also shown when absolute number correct was used as the measure, rather than proportion. Both main effects were again significant [$F(2,74) = 10.27$, $p < .001$, for length, and $F(1,37) = 79.97$, $p < .001$, for speech], as was the interaction [$F(2,74) = 6.7$, $p = .004$]. Once again there was a significant linear effect of list length [$F(1,37) = 13.66$, $p = .001$]. This time, however, it was an upward rather than a downward trend (see Figure 2), and significantly more items were recalled from the longer than from the shorter lists, both in quiet and in speech [$t(37) = 3.92$, $p < .001$, and $t(37) = 2.84$, $p = .007$, respectively].

As is shown in Figure 2, these effects were predicted a priori by the FM. A better fit to the data can easily be obtained post hoc—for example, by manipulating the attention parameter to reduce the size of the irrelevant speech effect (which the model overestimates) and to improve overall performance levels (which it underestimates). However, such a manipulation would be contrary to the stated intent of this study to examine the basic predictions of the model without altering parameter values needlessly. The results of this simulation are satisfactory for the FM because it correctly predicts (1) the appearance of an LLE and (2) the observation that irrelevant background speech depresses both absolute number recalled and proportion

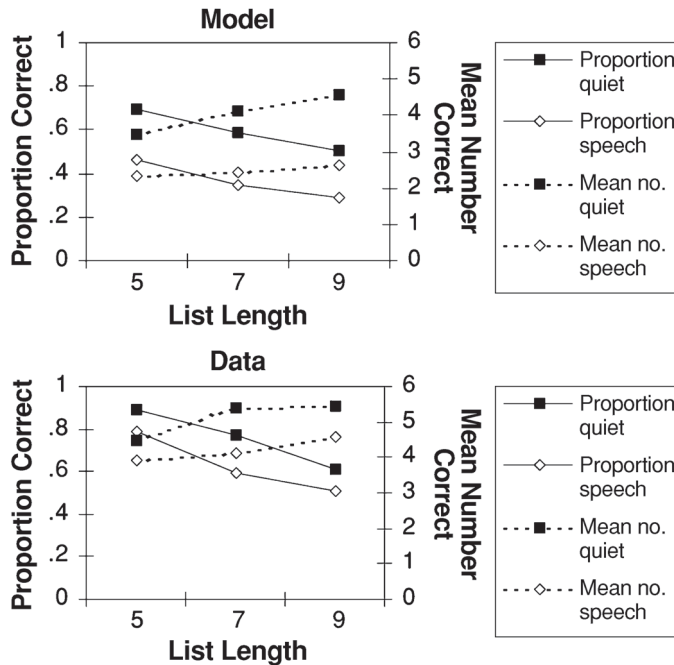


Figure 2. The decrease in proportion correct and the increase in mean number correct in immediate ordered recall as a function of list length. Stimuli were presented either in quiet or in the presence of irrelevant background speech. The upper panel shows the predictions of the feature model (FM) and the lower panel shows the experimental data. A version of the FM sufficient to demonstrate the list-length effect prediction can be found at memory.psych.purdue.edu/models/feature/.

correctly recalled but does not affect the appearance or form of the LLE. This successful set of predictions is based on only the standard set of parameters of the model.

The empirical findings of this study are also supported by the rescoring of data from Drewnowski's (1980) Experiment 1, shown in Figure 3. Although the list lengths used differed slightly, and different stimuli were employed in this earlier study (letters or consonant–vowel–consonant [CVC] syllables rather than digits), the results are consistent across studies.

One possible criticism of the experimental data reported here is that, since digits were the stimuli employed, the results could be conflated by guessing. With digits as the stimuli, there is, as has been mentioned, a possibility that mean performance will increase with list length (and proportion decrease) simply as a result of guesswork. Figure 4 shows how the expected number and proportion correct for five-, seven-, and nine-item lists are affected if correct performance is guaranteed for four items (cf. Cowan, 2001) and the remaining five items are randomly chosen either with or without replacement. The figure looks very similar to the predictions of the model, but for entirely different reasons.

A second logical possibility is that the order of presentation of list lengths (starting with short lists and increasing list length across the testing session) could affect the results. For example, proportion correct might stay constant across the list lengths and, by necessity of the design, absolute number correct will therefore increase as list length increases. However, if fatigue sets in at the later stages of the experiment, proportion correct might fall yet leave the absolute number correct at relatively high levels.

The results obtained by Drewnowski (1980) are inconsistent with both these objections. Drewnowski's data show the same pattern, although his experiments employed large numbers of stimuli (18 consonants for the letter condition and 28 unfamiliar trigrams for the CVC condition), making a guessing explanation of the data unlikely. The order of presentation of the lists was also randomized, in contrast to the blocked presentation employed here, reducing the potential for fatigue or practice effects. However, it is worthwhile to attempt to replicate

this general pattern of results in order to formally address these concerns. The Drewnowski values were estimated from the published paper and were not subject to formal analysis. It is therefore reasonable to ask that further data be produced that meet these two objections and are available for more detailed study.

METHOD

Participants

Twelve undergraduate students from the University of Reading participated in this experiment in return for course credit.

Materials and Design

All materials were presented using a Macintosh Performa 5400/160 PowerPC running HyperCard software.

To-be-recalled stimuli comprised 30 lists of monosyllabic words each with a Kučera–Francis (1967) written frequency of >50 per million, 10 each of five-, seven-, and nine-word lists. The presentation of different list lengths was blocked and counterbalanced. Words were presented on the computer screen, center justified in 48-point Geneva font, at a rate of 1 every 2.5 sec (2 sec on, 500 msec off). Following each list there was a 2-sec pause before the cue "Recall" was visually presented in the same manner. The recall cue remained on for 2 sec.

Procedure

The participants were told that they would see lists of words appearing one at a time on the computer screen and that, after a short pause, a recall cue would appear. When they saw the recall cue, they were required to write down, on the response sheet provided, the words they had seen, in the order in which they appeared.

RESULTS

The results are shown in Figure 5. ANOVA showed a main effect of list length on the number of correct scores [$F(2,22) = 6.32, p = .007$]. The linear contrast was also significant [$F(1,11) = 8.94, p = .01$], and a paired-sample t test showed significantly more items recalled in the longest than in the shortest lists [$t(11) = 2.99, p = .01$]. When proportion correct was the dependent variable, the main effect was again significant [$F(2,22) = 41.74, p < .001$], and once again there was a significant linear effect of list length [$F(1,11) = 89.8, p < .001$]. This time, how-

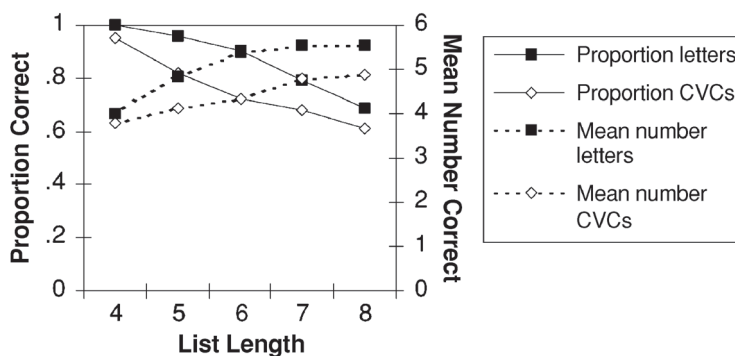


Figure 3. Experimental data from Drewnowski (1980), Experiment 1. These data confirm the list-length effect in immediate serial recall when using letters and CVC syllables as stimuli.

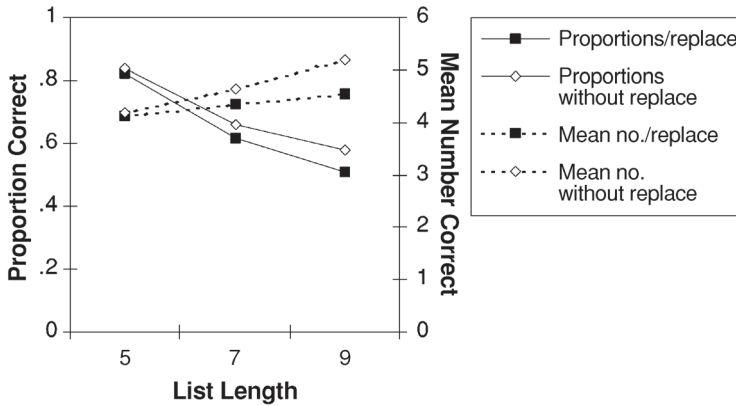


Figure 4. The expected list-length effect given recall of four digits and subsequent guessing. Expected values are presented for mean and proportion correct when sampling from the remaining five digits with and without replacement.

ever, it was a downward rather than an upward trend (see Figure 5) and a significantly lower proportion of items was recalled from the longer than from the shorter lists [$t(11) = 9.48, p < .001$].

DISCUSSION

The results of this experiment confirm the pattern of data observed in the initial study and in the rescoring of Drewnowski (1980). These experimental data also provide a useful constraint for any putative theory of immediate serial recall, because they point to a balance to be maintained between the interference caused by overloading and the extent to which the amount of information available for recall nevertheless increases. However, the data are very difficult to deal with without some form of formal model. The experimental results are best conceived not as an “effect” per se but as a relationship between absolute capacity (mean number correctly recalled) and relative recall (proportion correct) that is interpretable within some structural accounts of immediate memory but not within others. For

example, a simplistic “fixed slot” model of n chunks ($\pm x$) cannot account for the effect, because the capacity of a slot model is unaffected by the number of items in the recall set, even though the identity of the items actually recalled may change as the list length increases. Similarly, it is also difficult to accommodate this relationship in a straightforward fixed-duration rehearsal loop (see, e.g., Baddeley, 2003) or a chunk-loop combination (Zhang & Simon, 1985). Augmented versions of the loop, incorporating concepts such as activation and competition (Burgess & Hitch, 1999; Page & Norris, 1998), may fare better. A structural limit on the pool of activation available (see, e.g., Anderson et al., 1998) might, as suggested by Ward (2002), account for findings such as these. However, the spread of activation across the items in the recall set must be appropriate; hence, such a model need not predict the finding a priori and may require post hoc parameter fitting.

Finally, the results can be set alongside recent experimental work by Ward (2002) demonstrating that the LLE in free recall is related to the recency of rehearsals of pre-recency items. Although the account of LLEs provided by

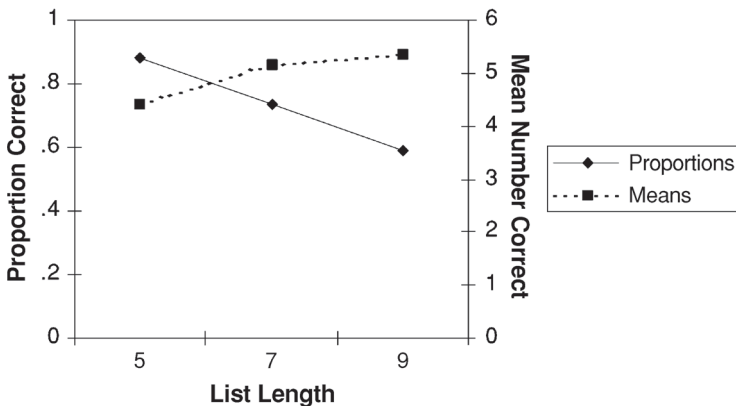


Figure 5. The list-length effect in lists of words presented in the experiment when guesswork is not a viable option and presentation order of each list length is fully counterbalanced.

the FM differs from that of Ward (2002), both accounts and both sets of data converge on the presumption that many free and serial recall phenomena may in fact be indicative of similar mnemonic processes operating over different time scales.

REFERENCES

- ANDERSON, J. R., BOTHELL, D., LEBIERE, C., & MATESSA, M. (1998). An integrated theory of list memory. *Journal of Memory & Language*, **38**, 341-380.
- BADDELEY, A. [D.] (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, **4**, 829-839.
- BROWN, G. D. A., NEATH, I., & CHATER, N. (2004). *A ratio model of scale-invariant memory and identification*. Manuscript submitted for publication.
- BURGESS, N., & HITCH, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, **106**, 551-581.
- COWAN, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral & Brain Sciences*, **24**, 87-185.
- CROWDER, R. G., & NEATH, I. (1991). The microscope metaphor in human memory. In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 111-125). Hillsdale, NJ: Erlbaum.
- DREWNOWSKI, A. (1980). Attributes and priorities in short-term recall: A new model of memory span. *Journal of Experimental Psychology: General*, **109**, 208-250.
- GLANZER, M. (1972). Storage mechanisms in recall. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 5, pp. 129-193). New York: Academic Press.
- JAHNKE, J. C. (1969). Output interference and the Ranschburg effect. *Journal of Verbal Learning & Verbal Behavior*, **8**, 614-621.
- KUČERA, H., & FRANCIS, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- LAMING, D. (1999). Testing the idea of distinct storage mechanisms in memory. *International Journal of Psychology*, **34**, 419-426.
- MIYAKE, A., & SHAH, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge: Cambridge University Press.
- MURDOCK, B. B., JR. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, **64**, 482-488.
- NAIRNE, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, **18**, 251-269.
- NAIRNE, J. S. (2001). A functional analysis of primary memory. In H. L. Roediger III, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 283-296). Washington, DC: American Psychological Association.
- 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. (2001). Long-term memory span. *Behavioral & Brain Sciences*, **24**, 134-135.
- NEATH, I. (2000). Modeling the effects of irrelevant speech on memory. *Psychonomic Bulletin & Review*, **7**, 403-423.
- PAGE, M. P. A., & NORRIS, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review*, **105**, 761-781.
- ROBERTS, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology*, **92**, 365-372.
- STRONG, E. K., JR. (1912). The effect of length of series upon recognition memory. *Psychological Review*, **19**, 447-462.
- WARD, G. (2001). A critique of the working memory model. In J. Andrade (Ed.), *Working memory in perspective* (pp. 219-239). Hove, U.K.: Psychology Press.
- WARD, G. (2002). A recency-based account of the list length effect in free recall. *Memory & Cognition*, **30**, 885-892.
- ZHANG, G., & SIMON, H. A. (1985). STM capacity for Chinese words and idioms: Chunking and acoustical loop hypotheses. *Memory & Cognition*, **13**, 193-201.

APPENDIX

Simulation Details

Table A1 shows the parameter values used in the reported simulation. These same values were used as default options by Neath (2000). Probability of overwriting (F) and other optional weights not reported here were all set to 1.0. The irrelevant speech simulation employed feature adoption and an alteration to the attention parameter as detailed by Neath (2000, Simulations 1-3).

Table A1
Parameter Values Used for Feature Model Simulation

Parameter	Value
Number of simulations	2,000
Overall attention (a)	10 (quiet), 12 (speech)
Number of modality-independent features	20
Number of modality-dependent features	2
Recovery constant (c)	2
Number of recovery attempts	2