Visual similarity effects in immediate serial recall and (sometimes) in immediate serial recognition

Chrissy M. Chubala[1] · Dominic Guitard[2] · Ian Neath[1] · Jean Saint-Aubin[2] · Aimée M. Surprenant[1]

Abstract

Words that sound dissimilar are recalled better than otherwise comparable words that sound similar on both immediate serial recall and immediate serial recognition tests, the so-called acoustic similarity effect. Although studies using immediate serial recall have shown an analogous visual similarity effect, in which words that look dissimilar are recalled better than words that look similar, this effect has not been examined in immediate serial recognition. We derived a prediction from the Feature Model that a visual similarity effect will be observed in immediate serial recognition only when the items are acoustically dissimilar; the model predicts no effect when the items are acoustically similar. Experiments 1 and 2 used visually dissimilar and visually similar stimuli that were all acoustically similar and replicated the visual similarity effect in serial recall but revealed no effect in serial recognition. Experiments 3 and 4 used a second set of stimuli that were acoustically dissimilar and found a visual similarity effect in both serial recall and serial recognition. The experiments confirm the Feature Model's predictions and add to earlier findings that the two tests, serial recall and serial recognition, may show quite different results because the two tests are not as similar as previously thought.

Keywords Visual similarity · Serial recall · Serial recognition · Short-term memory · Memory models

Introduction

Conrad (1964) reported that lists of letters that sound alike were not recalled as well as those that sound different, even when the stimuli were presented in written form. He noted that the data suggested that "the majority of subjects verbalize the [visually presented] stimuli rather than attempting to store them in visual form" (Conrad, 1964, p. 75). A consensus quickly emerged that letters and words, whether seen or heard, are represented by an acoustic or phonological code (e.g., Baddeley, 1966; Wickelgren, 1965). The acoustic similarity effect noted by Conrad is a well-known and extensively studied phenomenon, but at the time it was quite a revelation and arguably initiated an era of systematic explorations of the ways in which information is coded in memory. One other consequence is that relatively little work has examined whether visual similarity effects obtain with verbal stimuli. In this paper, we briefly review some of the studies that have examined visual similarity with verbal material, and then assess a novel prediction from the Feature Model concerning whether such effects will be observed in serial recognition despite being observed in serial recall.

Serial recall versus serial recognition

In a typical immediate serial recall test, subjects see or hear a relatively short list of words (usually around six items or so) and are then asked to reproduce them in the exact same order. Jacobs (1887, p. 75) termed this ability *span*, a term still used today; for example, when the stimuli are digits the task is frequently referred to as digit span. The task has been studied extensively, so much so that there exist multiple different models of immediate serial recall (for a review, see Nairne & Neath, 2013). Serial recognition, on the other hand, has a slightly different methodology. Both tasks begin similarly, with presentation of a short list of items one at a time. In a serial recognition task, however, the same items are presented a second time at test. On half of the trials, the items are presented in the same order, but on the other half of the trials, two adjacent items are transposed. For example, if the original list

✉ Ian Neath
ineath@mun.ca

1 Department of Psychology, Memorial University of Newfoundland, St John's, Newfoundland A1B 3X9, Canada

2 Department of Psychology, Université de Moncton, Moncton, New Brunswick, Canada

was "ant, bee, cat, dog, elk, fox," the second list on a different trial might be "ant, bee, dog, cat, elk, fox," with "dog" and "cat" switching positions. The subject's task is to indicate whether the order is the same or different. Originally designed to assess short-term memory in neurologically impaired patients (e.g., Kinsbourne, 1972; Shallice & Warrington, 1977), the test is far less studied than its serial recall analog, and to our knowledge is addressed by only one model (Farrell & McLaughlin, 2007).

Most researchers generally view the two tasks as comparable, the primary difference being that immediate serial recognition is seen as a purer task in two ways. First, serial recognition is often described as a purer measure of order information because the items are presented at test (e.g., Gathercole, Pickering, Hall, & Peaker, 2001; Gisselgård, Uddén, Ingvar, & Petersson, 2007; Jefferies, Frankish, & Lambon Ralph, 2006; Lian, Karlsen, & Winsvold, 2001; Romani, McAlpine, & Martin, 2008). Second, it is often described as a purer measure of short-term or working memory because serial recall requires contributions from long-term memory whereas serial recognition does not (e.g., Gathercole et al., 2001; Gisselgård et al., 2007; Thorn & Gathercole, 1999; Thorn, Gathercole, & Frankish, 2002).

Whether these characterizations are accurate is difficult to ascertain because there are relatively few studies in the literature that have directly compared the two tasks within the same experiment using the same subject pool and the same stimuli. Most studies examine only one type of test, making it difficult to interpret the results. One notable exception is an experiment reported by Tse, Li, and Altarriba (2011), who found a semantic relatedness effect in serial recall, with worse order recall with lists comprised of words from the same category than lists comprised of words from different categories, and lower performance for those lists in serial recognition. A further complication is that many researchers report only "proportion correct" for serial recognition, despite the fact that it is a same/different task and there are established ways of taking into account both sensitivity and bias (Macmillan & Creelman, 2005). Furthermore, it is not always clear whether "proportion correct" refers to just hits, a combination of hits and correct rejections, or to some other measure.

In one of the few studies to directly compare these two tasks using a detection-type analysis, Chubala, Surprenant, Neath, and Quinlan (2018) found that dynamic visual noise, which involves an array of squares that change randomly between black and white several times a second, had a detrimental effect on memory for concrete words when assessed by d' in serial recognition, but had no effect on memory for concrete words when assessed by proportion correct in serial recall. The lack of an effect of dynamic visual noise on immediate serial recall has been replicated a number of times (e.g., Castellà & Campoy, 2018; Ueno & Saito, 2013). In another study, Chubala, Neath, and Surprenant (2019) found that

whereas serial recall shows frequency effects, with better recall of high compared to low frequency words, serial recognition shows no effect of frequency. Similarly, whereas serial recall shows semantic relatedness effects, with better recall of lists comprised of words from the same category, serial recognition does not, replicating the results of Tse et al. (2011). However, both tests show effects of concreteness, with better performance for concrete than abstract words, and both show effects of acoustic similarity, with better performance for dissimilar-sounding words than similar-sounding words.

Chubala et al. (2019) suggested that this pattern of similarities and differences could be explained by incorporating the main idea underlying Farrell and McLaughlin's (2007) serial recognition model into the Feature Model (Nairne, 1990; Neath, 2000), although they did not implement this suggestion. An item is represented in the Feature Model as a vector of features, with a distinction made between modality-independent and modality-dependent features. The latter represent aspects of the stimulus that are unique to the presentation modality (e.g., font information for visually presented words or regional accent for auditorily presented words) whereas the former represent those aspects that are the same regardless of presentation modality (e.g., a word's meaning). These features are subject to interference such that at the time of recall, the cues for the current list of items have a combination of missing, changed, and intact features. In the original version of the model, only serial recall was implemented. On a serial recall test, a degraded cue is compared to the intact items stored in secondary memory. The cue associated most strongly with the first position is selected first, and then the most similar intact item is selected as a response.

In the Farrell and McLauglin (2007) serial recognition model, an overall measure of similarity is computed between the original items, which are assumed to be noisy, and the test items, which are assumed to be noise free. This measure is compared to a criterion, and if the value exceeds the criterion, a response of "different" is given, otherwise a response of "same" is given. This conception of serial recognition can be incorporated into the Feature Model, albeit with a different implementation. In the Feature Model, the cues are already noisy due to feature overwriting and interference. Similarity can be readily computed by examining the proportion of mismatching features between the study list and the probe list. To anticipate, we will not be developing a complete model at this time, primarily because there exist too few studies to constrain options. Rather, we show how the proportion of mismatching features measure allows the model to predict quite different results for two types of similarity (acoustic and visual) in serial recognition regardless of the other aspects of the model.

Acoustic and visual similarity effects in memory

As noted previously, the acoustic similarity effect refers to the finding that lists of words that sound dissimilar are better

recalled than otherwise comparable lists of words that sound similar (Conrad, 1964). When subjects engage in concurrent articulation, repeatedly saying an irrelevant word or series of words out loud, the acoustic similarity effect is removed for visual items but not for auditory items (Murray, 1968; Peterson & Johnson, 1971). Nairne (1990) showed how the Feature Model accounts for this pattern.

The visual similarity effect has not been studied nearly as extensively. Logie, Della Sala, Wynn, and Baddeley (2000) presented subjects with lists of words that either looked similar or looked dissimilar, but all of which sounded similar. The dissimilar-looking words were GUY, LIE, PI, RYE, SIGH, and THAI and the similar-looking words were CRY, DRY, FLY, PLY, SHY, and TRY. At the end of the list, the subjects were asked to immediately write down the words in order. Logie et al. also varied list length and whether subjects engaged in concurrent articulation. Overall, they observed a visual similarity effect, with better recall of the dissimilar-looking items than the similar-looking items, 0.58 versus 0.52 proportion correct, respectively. This advantage for visually dissimilar words held regardless of whether subjects engaged in concurrent articulation. This result was then replicated using a variety of different stimulus sets. For example, Logie, Saito, Morita, Varma, and Norris (2016) have demonstrated the same effect with visually similar and dissimilar Japanese Kanji characters, and Lin, Chen, Lai, and Wu (2015) have replicated this with Chinese characters in probed serial recall.

Similarity effects and the feature model

The Feature Model can readily account for both acoustic and visual similarity effects in serial recall. Because most of the model remains the same as in past work, we present only a brief overview; for more details and discussion about the various assumptions, see Nairne (1990), Neath and Nairne (1995), and Neath (2000).[1] Visually presented items are usually represented by 20 modality-independent features and two modality-dependent features. Each feature is randomly set to +1 or −1. Subsequent items can interfere with earlier items, thus producing noisy cues. Similarly, concurrent articulation is seen as adding noise to the modality-independent features rather than as preventing rehearsal; there is no decay in the model. At test, the noisy cues in primary memory are compared to intact traces in secondary memory.

The probability that a particular secondary memory item, SM_j, will be sampled as a potential recall response for a particular primary memory cue, PM_i, is given by Equation 1:

$$P_s\left(SM_j | PM_i\right) = \frac{s(i,j)}{\sum\limits_{k=1}^{n} s(i,k)} \qquad (1)$$

The similarity between a cue and a target, $s(i,j)$, is given by Equation 2:

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

Distance, d in Equation 2, is calculated from the number of mismatching features, as shown in Equation 3, where N is the number of compared features and a is the main scaling parameter:

$$d_{ij} = \frac{a \sum M_k}{N} \qquad (3)$$

As in some other models, sampled items need to be recovered prior to output. As an analogy, to find information in a library, one first has to find the book (sampling) and then the book has to contain the desired information (recovery). P_r, the probability of recovering a sampled item, is a function of the number of times the item has previously been sampled and recalled, r, as given by Equation 4, where c is a scale constant.

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

The acoustic similarity effect occurs with both visual and auditory presentation because modality-independent features can represent how a word sounds. The effect is abolished by concurrent articulation for visual presentation because these features are overwritten. The effect remains for auditory presentation because the modality-dependent features retain information about how the auditory items sound (see Nairne, 1990, for more details).

The original version of the Feature Model used 20 modality-independent features and two modality-dependent features to model visually presented items, due in part to the hardware limits of computers in the 1980s. One potential problem for the current work is that two modality-dependent features may not provide sufficient sensitivity to model visual similarity. For example, when modelling acoustic similarity, modality-independent and modality-dependent features are randomly set to +1 or −1 in the dissimilar condition. In the similar condition, a certain proportion of the features are guaranteed to have the same value, reflecting, for example, that the letters BCDGPTV all have an "ee" sound, despite varying in other ways. With only two modality-dependent features, the probability that Item j will have the same value for modality-dependent feature X as Item k by chance alone is 0.5 because there are only two possible values. This means that there is a 25% chance that both modality-dependent features of Item j will be same as those of Item k in the control condition. If the number of modality-dependent features is increased, the probability decreases. With four modality-dependent features, the

---

[1] In particular, we omit discussion of perturbation (see Neath, 2000) for simplicity and because it does not affect the qualitative predictions made.

probability that all modality-dependent features of two control items will be the same by chance is far smaller at 0.0625, and with six, the probability is 0.015625.

We first show that increasing the number of modality-dependent features has little effect on modeling the acoustic similarity effect in serial recall. We set the number of modality-independent features to 40 and the number of modality-dependent features to 10 as a balance between mitigating the aforementioned problem but still keeping the model as similar to previous versions as possible. The top two rows of Table 1 show the results when we use the parameters from Simulation 4V of Neath (2000), with 20 modality-independent and two modality-dependent features, of which eight and one, respectively, were set to the same value in the similar condition. The bottom two rows show the results when we re-ran the simulation with 40 modality-independent and 10 modality-dependent features, of which 16 and three, respectively, were set to the same value in the similar condition. The column labelled "Serial Recall" shows the mean proportion of items correctly recalled in order, and there is little difference as a function of the number of features: both show an acoustic similarity effect: 0.503 versus 0.334 for the conditions with fewer features, and 0.529 versus 0.364 for the conditions with more features.

Table 1 also shows the predictions for serial recognition, which come from the same simulations that produced the predictions for serial recall. When modelling serial recall, one component is the number of mismatched features, which is used in Equation 3. These values are further processed (e.g., see Equation 2) to produce a similarity metric. Although the number of mismatched features is also the basis for serial recognition, the processing is different. For "same" trials, the presented list and the test list are in the same order. Therefore, we compare the noisy cue for each item with its intact representation and calculate the mean number of mismatched features. This can be thought of as a global measure of similarity. We then divide this by the number of features to produce the

proportion of mismatched features. For "different" trials, we take into account the transposition. In the experiments we will run, Item 1 was always presented first at test, but all other adjacent items were transposed equally often. Consider when Items 3 and 4 are transposed: The noisy cues for Items 1, 2, 5, and 6 are still compared to their intact representations, but the noisy cue for Item 4 is compared to the intact representation of Item 3, and the noisy cue for Item 3 is compared to the intact representation of Item 4. On average, there will be more mismatches than when compared to the appropriate intact item. We compute the mean for all possible transpositions, and again divide by the number of features. These values are also shown in Table 1. The penultimate column shows the difference in the proportion of mismatched features between same and different trials, and this can be considered an index of performance on the recognition task: The larger the difference, the easier it is to distinguish same from different trials. The current focus is on the final column, which can be considered an index of the magnitude of the acoustic similarity effect in serial recognition. Once again, the addition of more features makes little difference overall: In both cases, an acoustic similarity effect in serial recognition is predicted because there are more feature mismatches for similar than for dissimilar items due to feature overwriting (see Nairne, 1990).

The visual similarity effect is modeled in almost the same way, the only difference being that more modality-dependent features are set to the same value to reflect that the letters look similar. The visually dissimilar and similar English stimuli used by Logie et al. (2000, 2016) were all acoustically similar. Therefore, when modeling this instance of a visual similarity effect, the dissimilar condition is identical to the similar condition from the acoustic similarity effect simulation. The similar condition just has more modality-dependent features set to the same value. For the dissimilar condition, then, there were again 40 modality-independent and 10 modality-dependent features, of which 16 and three, respectively, were set to the same value. For the similar condition, the only difference was

Table 1 Predictions of the Feature Model for the acoustic similarity effect in serial recall (proportion correct) and serial recognition (proportion of mismatching features) when the number of modality-independent and modality-dependent features are set to 20 and 2, respectively (top two rows), and 40 and 10, respectively (bottom two rows)

Condition	Serial recall	Serial recognition			
	Proportion correct	Proportion of mismatching features			
		Same trial	Different trial	Diff – Same	Dis – Sim
Dissimilar (20,2)	0.504	0.493	0.621	0.128	0.052
Similar (20,2)	0.334	0.693	0.769	0.075	
Dissimilar (40,10)	0.529	0.484	0.612	0.129	0.049
Similar (40,10)	0.364	0.669	0.748	0.080	

Note: Dissimilar and similar refer to acoustic similarity; all stimuli are visually dissimilar. Diff – Same indexes discriminability between same and different trials whereas Dis – Sim indexes the magnitude of the acoustic similarity effect

that the number of modality-dependent features set to the same value was increased to eight. Table 2 shows the results in the top two rows.

As can be seen by comparing the bottom two rows of Table 1 and top two rows of Table 2, the Feature Model predicts a larger acoustic similarity effect than visual similarity effect in serial recall. The reason has to do with the smaller proportion of features used to code for visual similarity. The table also shows that the Feature Model predicts an acoustic similarity effect in serial recognition but essentially no visual similarity effect in serial recognition when the items are acoustically similar: The difference between the two is only 0.013. The Feature Model predicts the absence of a visual similarity effect in serial recognition when the items are acoustically similar because the proportion of features coding for acoustic similarity is far greater than the proportion coding for visual similarity. In essence, the contribution of visual similarity to the global measure is overwhelmed by the contribution of acoustic similarity.

The Feature Model makes another two predictions, which can be seen by comparing the top two rows to the bottom two rows in Table 2: First, there will be a visual similarity effect in serial recall when the items are acoustically dissimilar. This has been reported by Saito, Logie, Morita, and Law (2008) and Logie et al. (2016) with Japanese stimuli but has not to our knowledge been demonstrated with English stimuli. Second, it predicts a visual similarity effect in serial recognition when the items are acoustically dissimilar. For this second prediction, note that the difference between the proportion of mismatching features is numerically larger when the items are acoustically dissimilar than when they are acoustically similar. This larger difference holds for all sensible parameter choices, although of course the magnitude of the difference varies. The Feature Model predicts the presence of a visual similarity effect in serial recognition when the items are acoustically dissimilar because the contribution of acoustic

similarity to the global measure is reduced, potentially allowing the contribution of visual similarity to be detected.

The Feature Model then predicts acoustic similarity effects in both serial recall and serial recognition, and the latter finding has been confirmed (Chubala et al., 2019). It also predicts smaller visual similarity than acoustic similarity effects in serial recall, and no effect in serial recognition when the items are also acoustically similar. Finally, it predicts a small effect in serial recognition when the items are acoustically dissimilar. The purpose of the current experiments is to test these predictions. Experiments 1 and 2 both used Logie et al.'s (2016) stimuli, but the first experiment used immediate serial recall and the second used immediate serial recognition. Experiments 3 and 4 both use a new set of acoustically dissimilar words, and used immediate serial recall and immediate serial recognition, respectively.

Experiment 1

Experiment 1 was designed as a replication of Experiment 3 of Logie et al. (2016). Subjects saw six-item lists of either visually similar or visually dissimilar words and were asked to recall the words in strict serial order. There are two major differences between this experiment and that reported by Logie et al. First, we did not ask subjects to engage in concurrent articulation because Logie et al. (2000) had shown no interaction between concurrent articulation and the visual similarity effect. Second, we used only pure lists (lists contained either all dissimilar or all similar items but not both), whereas Logie et al. also used mixed lists in which dissimilar and similar items alternated. This second change was made because the focus of the current work is on serial recognition and a transposition of two adjacent items in an alternating pattern could be easy to detect.

Table 2 Predictions of the Feature Model for the visual similarity effect in serial recall (proportion correct) and serial recognition (proportion of mismatching features)

Condition	Serial recall	Serial recognition			
	Proportion correct	Proportion of mismatching features			
		Same trial	Different trial	Diff – Same	Dis – Sim
Ac Sim, Vis Dis	0.364	0.669	0.749	0.080	0.013
Ac Sim, Vis Sim	0.318	0.710	0.777	0.067	
Ac Dis, Vis Dis	0.535	0.483	0.612	0.129	0.022
Ac Dis, Vis Sim	0.461	0.550	0.657	0.107	

The first two rows are for acoustically similar items that are either visually dissimilar or similar. The last two rows are for acoustically dissimilar items that are either visually dissimilar or similar

Note: Diff – Same indexes discriminability between same and different trials whereas Dis – Sim indexes the magnitude of the visual similarity effect

Subjects

Thirty volunteers from Prolific Academic (ProlificAC) participated and were paid £9 per hour (pro-rated). Inclusion criteria for this and all other experiments was: (1) native speaker of English; (2) nationality must be from the UK, USA, or Canada; (3) approval rating of at least 90% on prior submissions at ProlificAC; (4) normal or corrected-to-normal vision; (5) no cognitive impairment or dementia; (6) no language-related disorders; (7) age between 19 and 30 years. The mean age was 25.73 years (SD = 3.42, range 19–30), and 20 participants self-identified as female and 10 as male.

The sample size for both serial recall experiments was determined as follows. The effect size of the pure lists in Experiment 3 of Logie et al. (2016) was estimated to be d = 0.62. Using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009), we determined a sample size of 30 would have power of 0.90 to detect such an effect.

Design

There were two independent variables, both manipulated within subjects, to form a two similarity (dissimilar vs. similar) × six serial-position repeated measures design.

Stimuli

The stimuli were the same as in Experiment 3 of Logie et al. (2016): The visually dissimilar words were GUY, HI, LIE, RYE, SIGH, and THAI, and the visually similar words were CRY, DRY, FLY, PLY, SHY, and TRY.

Procedure

After basic demographic information was collected, subjects were reminded of the instructions. They were informed that they would see lists of six words presented one at a time and were asked to read each word silently. Each word was shown for 1 s in upper case 24-point Courier, a monospace font. All instructions were shown in 28-point Helvetica. At the end of the list, a message appeared saying "Please type in the first word you saw." They were informed that they needed to type in the first word first, the second word second, and so on. If they could not remember a word, they were asked either to guess or to click on a button labelled "Skip." Six responses were required before the next list began.

Sixteen trials used the visually similar words and sixteen used the visually dissimilar words. The order of the words was randomized on each trial, and the order of the trials was also randomized for each subject. Subjects could take a break whenever they wished by simply waiting to click on the "Start next trial" button.

Results and discussion

For all experiments, the data were analysed using both frequentist and Bayesian techniques using JASP (JASP Team, 2018), reporting both a p value and effect size for the former, and a Bayes factor (BF) for the latter. BF_{10} between 3 and 20 indicates positive evidence for the alternate hypothesis (and therefore evidence against the null hypothesis); BF_{10} between 20 and 150 indicates strong evidence; and BF_{10} greater than 150 indicates very strong evidence (Kass & Raftery, 1995). BF_{01} indicates evidence in favor of the null hypothesis. Non-integer degrees of freedom indicate the Greenhouse-Geisser correction was applied because the assumption of sphericity was violated.

The typed responses were checked for obvious errors such as omitting an initial letter (IGH rather than SIGH), omitting a final letter (SIG rather than SIGH), typing two identical letters (TTIE instead of TIE), transpositions (TEI rather than TIE), and so on. A total of 56 out of 5760 responses (0.97%) were changed, 37 in the dissimilar condition and 19 in the similar condition. Because this did not affect the results of the statistical analyses, the uncorrected data were used for the analyses reported here.

As is evident in the left panel of Fig. 1, we replicated the finding of better serial recall of visually dissimilar than visually similar words when all words sound similar. This observation was supported by the results of two-levels of similarity (dissimilar vs. similar) × six serial-position repeated-measures ANOVA on the proportion of words correctly recalled in order. The main effect of similarity was significant, $F(1,29) = 22.338$, $MSE = 0.054$, $\eta_p^2 = 0.435$, $p < 0.001$, $BF_{10} = 3656.50$, with a larger proportion of words correctly recalled in order in the dissimilar condition ($M = 0.615$, $SD = 0.154$) than in the similar condition ($M = 0.499$, $SD = 0.172$), Cohen's $d = 0.863$. The main effect of position was also significant, $F(2.42,70.50) = 120.628$, $MSE = 0.041$, $\eta_p^2 = 0.806$, $p < 0.001$, $BF_{10} = 2.934 \times 10^{63}$. The interaction was not significant, $F(3.99,115.91) = 1.80$, $MSE = 0.013$, $\eta_p^2 = 0.059$, $p = 0.133$, $BF_{01} = 15.23$.

Experiment 1 replicated the results previously reported by Logie et al. (2000, 2016): When the stimuli were all acoustically similar, subjects recalled visually dissimilar items in order more accurately than visually similar items, and similarity and position did not interact. Having demonstrated a replication of the basic finding with Logie et al.'s stimuli, we can now test whether the effect is observable in serial recognition.

Experiment 2

The purpose of Experiment 2 was to test the prediction of the Feature Model that the visual similarity effect will not be observed in serial recognition when the items all sound
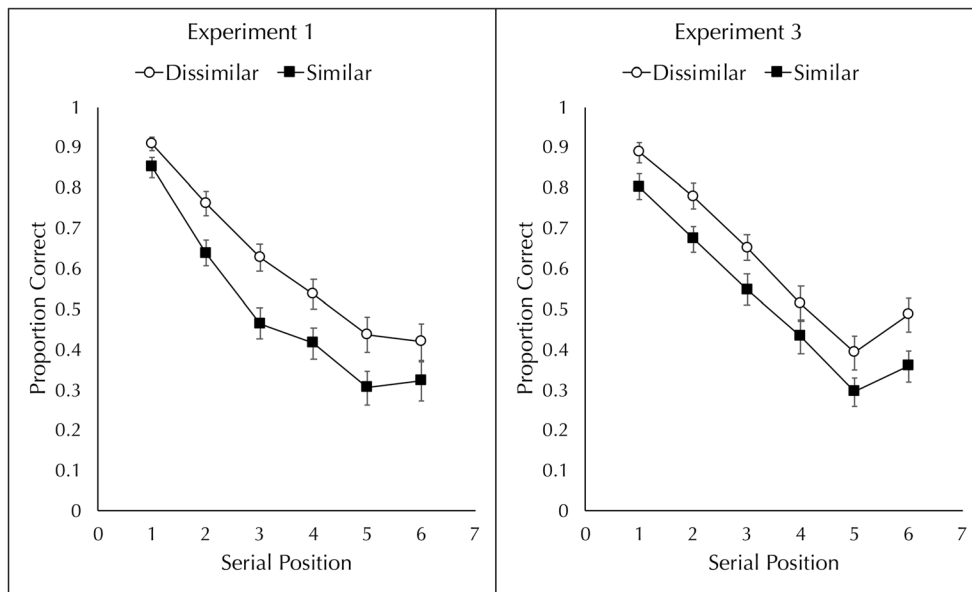
Fig. 1 The proportion of visually dissimilar and visually similar words recalled in order in Experiment 1 (left panel) and Experiment 3 (right panel). Experiment 1 used stimuli that were acoustically similar whereas Experiment 3 used stimuli that were acoustically dissimilar. Error bars show the standard error of the mean

similar. Chubala et al. (2019) reported two manipulations that produced an effect in serial recall but not serial recognition, word frequency (Exp. 2) and semantic relatedness (Exp. 4). The current experiment is identical to those except for the stimuli used.

Subjects

Sixty different people volunteered from Prolific.AC. The mean age was 25.17 years (SD = 3.52, range 19–30); 39 self-identified as female and 21 subjects self-identified as male.

The sample size for both serial recognition studies was determined as follows. Chubala et al. (2019) reported two significant effects in serial recognition, concreteness (d = 0.719) and acoustic similarity (d = 0.578) and two non-significant effects in serial recognition, word frequency (d = 0.008) and semantic relatedness (d = 0.178). Using G*Power (Faul et al., 2009), we determined a sample size of 60 would have power of 0.99 to detect an effect as large as the acoustic
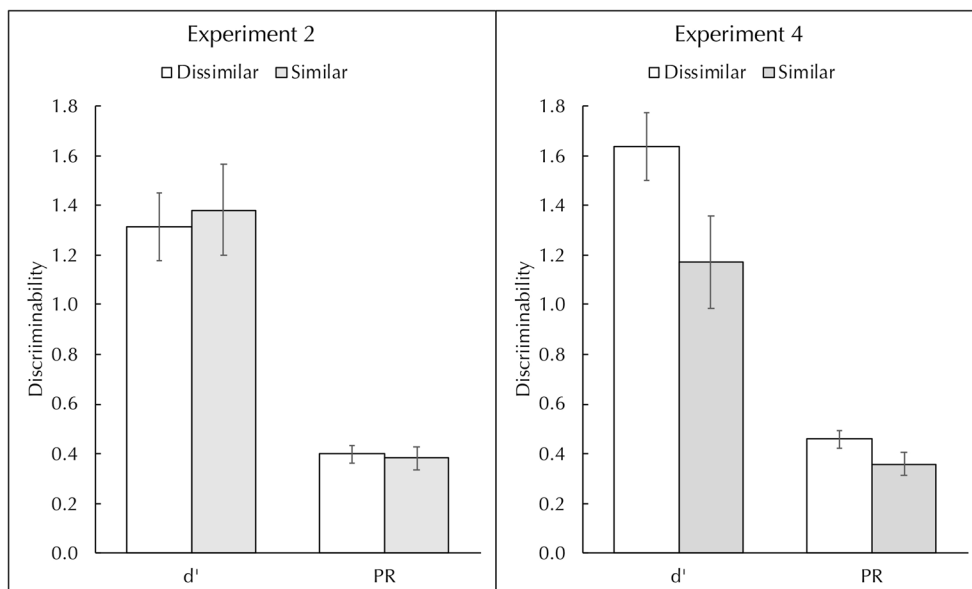

Fig. 2 Performance on the immediate serial recognition test, d' and Pr, for visually dissimilar and visually similar lists. In Experiment 2 (**left panel**), the words were acoustically similar whereas in Experiment 4 (**right panel**), the words were acoustically dissimilar. Error bars show the standard error of the mean

similarity effect, and power of 0.80 to detect an effect midway in size between the significant and non-significant effects.

Design

There was only one independent variable, visual similarity (dissimilar vs. similar), which was manipulated within subjects.

Stimuli

The stimuli were the same as in Experiment 1.

Procedure

After collecting basic demographic information, subjects were reminded of the instructions. They were informed that they would see a list of six words presented one at a time, and then they would see a second list of six words. They were further informed that on half the trials, the second list would be identical to the first. On the remaining trials, two adjacent words would be transposed. Their task was to indicate whether the lists were the same or different.

As in Experiment 1, each word was shown for 1 s and the words were shown in upper case in 24-point Courier. At the end of the first list, a message appeared saying "Is this the same order or a different order?" and the second list was presented. Two buttons, one labelled "Same" and one labelled "Different," then became active and a response was made by clicking on the appropriate button.

There were 32 trials, half of which were "same" and half of which were "different" trials. In addition, half of each type of trial used the similar words and half used the dissimilar words. For the different trials, Items 1 and 2 were never transposed. All other adjacent transpositions (e.g., transposing Items 2 and 3, transposing Items 3 and 4, etc.) were done twice. The order of the words was randomized on each trial, and the order of the trials was also randomized for each subject. Subjects could take a break whenever they wished by simply waiting to click on the "Start next trial" button.

Results and discussion

We followed the procedure described by Macmillan and Creelman (2005) for calculating hits, false alarms, and d'. When subjects responded "different" on "different" trials it was classified as a hit and when they responded "different" on "same" trials it was classified as a false alarm. The same hit and false-alarm rates were also used to calculate Pr. As in Chubala et al. (2019), a score of 0.99 or 0.01 replaced a hit rate of 1 or false-alarm rate of 0, respectively.

Hits and false alarms. There was no difference in the hit rate between the dissimilar ($M = 0.719$, $SD = 0.196$) and similar ($M = 0.706$, $SD = 0.232$) conditions, $t(59) = 0.514$, $p = 0.608$, $d = 0.066$, $BF_{01} = 6.237$. There was also no difference in the false-alarm rate between the dissimilar ($M = 0.321$, $SD = 0.177$) and the similar ($M = 0.324$, $SD = 0.177$) conditions, $t(59) = 0.008$, $p = 0.933$, $d = 0.011$, $BF_{01} = 7.055$.

d' and C. There was no difference in d' between the dissimilar ($M = 1.314$, $SD = 1.067$) and similar ($M = 1.382$, $SD = 1.408$) conditions, $t(59) = 0.424$, $p = 0.673$, $d = 0.055$, $BF_{01} = 6.496$. There was also no difference in C between the dissimilar ($M = -0.112$, $SD = 0.474$) and similar ($M = -0.102$, $SD = 0.505$) conditions, $t(59) = 0.145$, $p = 0.885$, $d = 0.019$, $BF_{01} = 7.008$.

Pr and Br. There was no difference in Pr between the dissimilar ($M = 0.398$, $SD = 0.349$) and similar ($M = 0.382$, $SD = 0.349$) conditions, $t(59) = 0.394$, $p = 0.695$, $d = 0.051$, $BF_{01} = 6.572$. There was also no difference in Br between the dissimilar ($M = 0.551$, $SD = 0.237$) and similar ($M = 0.559$, $SD = 0.251$) conditions, $t(59) = 0.228$, $p = 0.820$, $d = 0.029$, $BF_{01} = 6.905$.

The Feature Model predicted no visual similarity effect in immediate serial recognition when the words sound similar, and none was seen either in d' or Pr. The reason is that serial recognition is based on a global measure of similarity, and because the items all sound similar, there is very little difference between the similarity of the noisy cues and the test order when they are the same versus when two items are transposed.

Experiment 3

The stimuli that produced a visual similarity effect in serial recall in Experiment 1 failed to produce a visual similarity effect in serial recognition in Experiment 2, as predicted by the Feature Model. The purpose of Experiments 3 and 4 was to assess another prediction: if acoustically dissimilar words are used, a visual similarity effect will be observed in both serial recall (Experiment 3) and serial recognition (Experiment 4). Logie et al. (2016; see also Saito et al., 2008) have demonstrated an effect of visual similarity with Japanese stimuli when the stimuli were also acoustically dissimilar; Experiment 3 will assess whether the same effect obtains with English stimuli.

Subjects

Thirty different people from Prolific Academic (ProlificAC) volunteered. The mean age was 24.67 years ($SD = 3.397$, range 19–30), and 14 self-identified as female, 15 as male, and one did not answer the question.

Stimuli

A new set of stimuli were created using the visual similarity matrix for Latin-based alphabets of Simpson, Mousikou, Montoya, and Defior (2013). They had 332 subjects rate the visual similarity of uppercase letter pairs using a scale from 1 (*not at all similar*) to 7 (*very similar*). Using these data, a visual similarity score was computed for each word pair and this value represents the sum of the letter pairs score. For example, for the word CAT and the word KIT, the visual similarity score is based on the pairs "C K," "A I," and "T T." For identical letters pairs, a perfect similar score of 7 is given. The visually similar items were CORE, CUBE, CURB, DOPE, OGRE, and PURE. The visually dissimilar items were CROW, CURL, CZAR, SHIP, TALE, and UNDO. The visual similarity score was higher for the similar ($M = 20.34$, $SD = 1.96$) than the dissimilar ($M = 9.13$, $SD = 1.87$) words, $t(10) = 21.32$, $p < 0.001$.

The two sets of words did not differ in either semantic similarity or phonological similarity. WordNet is an on-line lexical database in which words are organized into synonym sets that represent the underlying lexical concept (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). Pedersen, Patwardhan, and Michelizzi (2004) calculated a number of measures of similarity between lexical concepts, and the one used here is the number of steps in the shortest path between two words. Thus, low values indicate a closer semantic relation than high values. The mean path length for the visually similar nouns was 7.4 ($SD = 2.71$) compared to 8.7 ($SD = 3.65$) for the visually dissimilar nouns; these means are not significantly different, $t(8) = 0.96$, $p = 0.35$. PSIMETRICA (Mueller, Seymour, Kieras, & Meyer, 2003) provides a measure of phonological dissimilarity. Both two-syllable words were omitted from this analysis. Using standard British pronunciation, the mean phonological dissimilarity was 0.371 ($SD = 0.048$) for the visually similar words compared to 0.375 ($SD = 0.065$) for the visually dissimilar words, $t(10) = 0.09$, $p = 0.93$. Using standard American pronunciation, the respective values were 0.342 (0.047) and 0.361 (0.056), $t(10) = 0.56$, $p = 0.58$. The two sets of words were equated on a number of other dimensions; details are in the Appendix.

Procedure

The procedure was the same as in Experiment 1.

Results and discussion

As is evident in the right panel of Fig. 1, there was a visual similarity effect, with better serial recall of visually dissimilar than visually similar words, replicating with English stimuli the findings of Logie et al. (2016). This observation was supported by the results of a two-levels of similarity (dissimilar vs. similar) × six serial-position repeated-measures ANOVA on the proportion of words correctly recalled in order. The main effect of similarity was significant, $F(1,29) = 36.45$, $MSE = 0.025$, $\eta_p^2 = 0.557$, $p < 0.001$, $BF_{10} = 521.80$, with a larger proportion of words correctly recalled in order in the dissimilar condition ($M = 0.618$, $SD = 0.156$) than in the similar condition ($M = 0.518$, $SD = 0.157$), $d = 1.102$. The main effect of position was also significant, $F(2.99, 86.95) = 86.94$, $MSE = 0.042$, $\eta_p^2 = 0.749$, $p < 0.001$, $BF_{10} = 7.735 \times 10^{63}$. The interaction was not significant, $F(4.16, 120.67) = 0.474$, $MSE = 0.010$, $\eta_p^2 = 0.016$, $p = 0.762$, $BF_{01} = 59.22$.

As predicted by the Feature Model, a visual similarity effect was observed when the words were acoustically dissimilar. As shown in Table 2, the difference between the visually dissimilar and similar items is larger in the model when the items sound dissimilar than when they sound similar, and the effect size was larger for this experiment than that observed in Experiment 1. However, this result should be treated with caution because the stimuli in Experiment 1 and those in the current experiment were not equated. Nonetheless, a prediction of the Feature Model is that if two sets of stimuli are appropriately equated, there should be a larger visual similarity effect for acoustically dissimilar items than for acoustically similar items.

Experiment 4

Experiment 3 found a visual similarity effect with a new set of stimuli that were not acoustically similar. The purpose of Experiment 4 was to test these stimuli in serial recognition. As shown in Table 2, the Feature Model predicts a larger difference in the proportion of mismatching features between visually similar and dissimilar items when the items are acoustically dissimilar, but it is not clear whether it will be sufficiently large to produce an effect in serial recognition. We will return to the issue of the diagnosticity of this prediction for a small effect in the *General discussion*.

Subjects

Sixty different volunteers from Prolific.AC participated. The mean age was 24.82 years ($SD = 3.69$, range 19–30); 31 self-identified as female, 28 subjects self-identified as male, and one did not answer the question.

Stimuli

The stimuli were the same as in Experiment 3.

Procedure

Experiment 4 was identical to Experiment 2, except for the stimuli.

Results and discussion

Hits and false alarms. There was no difference in the hit rate between the dissimilar ($M = 0.674$, $SD = 0.201$) and similar ($M = 0.660$, $SD = 0.188$) conditions, $t(59) = 0.508$, $p = 0.614$, $d = 0.066$, $BF_{01} = 6.258$. There was, however, a significant difference in the false-alarm rate, with more false alarms in the similar ($M = 0.301$, $SD = 0.208$) than in the dissimilar ($M = 0.215$, $SD = 0.177$) conditions, $t(59) = 3.002$, $p = 0.004$, $d = 0.388$, $BF_{10} = 7.921$.

d′ and C. There was a significant difference in d' between the dissimilar ($M = 1.637$, $SD = 1.109$) and similar ($M = 1.172$, $SD = 1.029$) conditions, $t(59) = 2.972$, $p = 0.004$, $d = 0.384$, $BF_{10} = 7.349$. There was no difference in C between the dissimilar ($M = 0.224$, $SD = 0.586$) and similar ($M = 0.089$, $SD = 0.479$) conditions, $t(59) = 1.637$, $p = 0.107$, $d = 0.211$, although the BF provides only slight evidence in favor of the null hypothesis and is not conclusive, $BF_{01} = 2.016$.

Pr and Br. There was a significant difference in Pr between the dissimilar ($M = 0.459$, $SD = 0.246$) and similar ($M = 0.359$, $SD = 0.298$) conditions, $t(59) = 2.708$, $p = 0.008$, $d = 0.349$, $BF_{10} = 3.901$. There was no difference in Br between the dissimilar ($M = 0.389$, $SD = 0.268$) and similar ($M = 0.448$, $SD = 0.223$) items, $t(59) = 1.531$, $p = 0.131$, $d = 0.198$, although as was the case with C, the BF provides only slight evidence in favor of the null hypothesis and is not conclusive, $BF_{01} = 2.352$.

The acoustically dissimilar stimuli used in this experiment produced a visual similarity effect in serial recognition, as predicted by the Feature Model. The basic idea is that serial recognition uses a global assessment of similarity and when items are acoustically similar, any increase in visual similarity is essentially swamped. In contrast, when the items are acoustically dissimilar, the increase in visual similarity is now detectable. Note that this arises in the Feature Model due to the way the modality effect was initially modelled (see Nairne, 1990).

General discussion

Experiment 1 replicated the results of Logie et al. (2000, 2016) by showing a visual similarity effect in immediate serial recall when the stimuli were acoustically similar. Using the same stimuli, Experiment 2 found no effect of visual similarity in immediate serial recognition. The Feature Model predicted the absence of a visual similarity effect in serial recognition when the items are acoustically similar because the number of features coding for

acoustic similarity is greater than the number coding for visual similarity. This has the net result of making the global measure of similarity between the noisy cues and the probe list virtually the same for both visually similar and dissimilar trials.

The stimuli in Experiments 3 and 4 were acoustically dissimilar, and a visual similarity effect was observed in immediate serial recall in Experiment 3, as predicted by the Feature Model, and replicating with English stimuli results previously shown with Japanese stimuli (Logie et al., 2016; Saito et al., 2008). More importantly, a visual similarity effect was also observed in immediate serial recognition in Experiment 4. According to the Feature Model, the visual similarity effect is observed in serial recognition when the stimuli are acoustically dissimilar because the features coding for visual similarity are no longer overwhelmed by a larger proportion of features coding for acoustic similarity. Tables 1 and 2 show values that index the size of the various similarity effects in serial recognition. For acoustic similarity, this index is approximately 0.05, and Chubala et al. (2019) reported an effect size of $d = 0.578$. For visual similarity, when the items were also acoustically similar, the index was 0.01 and the effect size in Experiment 2 was similarly close to 0, $d = 0.066$. When the items were acoustically dissimilar, the index doubled to 0.02, and now the observed effect size in Experiment 4 was $d = 0.384$. The effect is small, but this is consistent with the model's predictions.

The small predicted size of the visual similarity effect in serial recognition when the items are acoustically dissimilar raises the issue of the value of the prediction. It turns out that the effect was of a detectable size, but what if the experiment had not found a significant effect? In our view, the prediction would still have been diagnostic, but may have been more difficult to assess. One way, for example, is a meta-analysis of a number of studies, which could address the question of whether the effect size is non zero. Should such an analysis show the effect size is consistently around zero, it would have disconfirmed the prediction.

The prediction is diagnostic in a second way. The Feature Model made the correct qualitative prediction (i.e., when an effect is likely to be observed or not observed), but the visual similarity effect observed when the auditory items were dissimilar was larger than one might have thought based on the model. This suggests that although the basic implementation of serial recognition in the Feature Model is plausible, it may not be producing a sufficiently large effect. As noted earlier, there are too few studies to constrain modelling choices at this point, but further research will help in refining the model's account and address this potential deficiency.

The account of the Feature Model makes a number of other, testable predictions.[2] For example, it predicts that an acoustic similarity effect will be seen in serial recall when the items are visually similar. This can be seen in Table 2 by comparing the

---

[2] We thank an anonymous reviewer for drawing our attention to these.

"Ac Sim, Vis Sim," and "Ac Dis, Vis Sim" rows. It further predicts the same effect in serial recognition, as can be seen by computing the difference between the two rows, 0.107 – 0.067 = 0.04.

The idea that performance in serial recognition is supported by the proportion of feature mismatches allows the Feature Model to account for a number of other serial recognition results (see Chubala et al., 2019). First, we have already noted that the Feature Model accounts for the acoustic similarity effect seen in serial recognition. A second result is the presence of a concreteness effect in serial recognition. Although the Feature Model has not been applied to the concreteness effect previously, Chubala et al. suggested that Paivio's (1991) dual-coding account could be readily integrated. Concrete words are assumed to lead to a second, imagistic representation, not shared with abstract words, in addition to the verbal representation shared by both word types. A new type of feature could be added that supports this type of image. Because abstract words would lack discriminative information on these features, performance would be better for concrete than for abstract words. The effect would occur, within the Feature Model, in serial recognition because the locus of the effect is at the feature level, and therefore would affect the proportion of mismatching features.

We can also consider two instances in which an effect observed in serial recall is not observed in serial recognition. There is no word frequency effect in serial recognition, and according to the Feature Model this is because frequency is seen as having its effect during redintegration. Within the model, the redintegration stage occurs only for serial recall, and not for serial recognition. There is likewise no semantic relatedness effect in serial recognition. If this is viewed as something that occurs after the feature mismatch stage, then the model again accounts for the lack of an effect in serial recognition. This is the most tentative part of the account that we present here, as it is not clear that semantic relatedness is a redintegration effect. Nonetheless, it is plausible in that if one knows the first three items were all vegetables, then this affects searching for other exemplars of this category as potential items for recall.

According to the Feature Model, serial recall and serial recognition are far more different than is generally recognized in the literature, even if they do sometimes show similar effects (e.g., both show an acoustic similarity effect, both show a concreteness effect). Serial recognition is based more on an evaluation of global similarity between the noisy cues and the incoming test list. In contrast, serial recall is based more on individual comparisons between a specific noisy cue and possible candidates for recall. This suggests that any theory that assumes the tasks are very similar will likely fail to account for the pattern of results.

The only model of serial recognition of which we are aware is that of Farrell and McLaughlin (2007). Given that their basic idea, as implemented in the Feature Model, correctly predicted the results of the current experiments, it is likely that their model could also account for the results if appropriately modified for the stimuli used here. We therefore view the two models as complementary, in that the Feature Model already had a way of coding similarity (both acoustic and visual) whereas the Farrell and McLaughlin model already had a way of assessing global similarity and making a decision.

Of the many models of serial recall, only the Primacy Model (Page & Norris, 1998) has been applied to the visual similarity effect (Logie et al., 2016). As with the Feature Model, it accounts for the basic visual similarity effect observed with acoustically similar stimuli. It most likely also predicts a visual similarity effect with acoustically dissimilar items, although it is possible that the effect of visual similarity might be too small to have an effect within the model if the items differ acoustically. It is not clear whether the model could be extended to account for serial recognition.

The obvious limitation of the current work is that the implementation of serial recognition in the Feature Model is incomplete. We have incorporated the notion of Farrell and McLaughlin (2007) that subjects base their decision of same or different on a global measure of similarity, and we used the proportion of mismatching features as an index of performance. However, we do not implement anything further. For example, we do not address how a criterion is set or might be changed, and we do not say how the proportion of mismatching features is translated into a value used for the decision. Our rationale is that because there are too few constraints at the moment to develop the model further, it is better to base predictions on relative changes in the proportion of mismatching features. When more serial recognition data are available, then the options for the model will be more constrained.

Despite this limitation, the results do suggest that Farrell and McLaughlin's (2007) basic conception of serial recognition is viable, and the results also reinforce the idea that serial recall and serial recognition are more different than currently represented in the literature.

Appendix

Table 3 Parameter values used in the simulations

F	probability of overwriting	1.0
a	distance scaling	10.0
c	recovery constant	2.0
r	number of recovery attempts	2
n	number of items in list	6
	probability of a perturbation	0.05
MI	number of modality-independent features	20 or 40
MD	number of modality-independent features	2 or 10
Acoustic similarity (Table 1)		
sMI	number of guaranteed similar M1 features	0 or 8 (out of 20); 0 or 16 (out of 40)
sMD	number of guaranteed similar MD features	0 or 1 (out of 2); 0 or 3 (out of 10)
Visual similarity (acoustically similar items, top two rows of Table 1)		
sMI	number of guaranteed similar M1 features	16 (out of 40)
sMD	number of guaranteed similar MD features	3 or 8 (out of 10)
Visual similarity (acoustically dissimilar items, bottom two rows of Table 1)		
sMI	number of guaranteed similar M1 features	0 (out of 40)
sMD	number of guaranteed similar MD features	0 or 8 (out of 10)

Table 4 Visual similarity matrix for the visually similar and visually dissimilar words

Visually similar					
	CORE	CUBE	CURB	DOPE	OGRE	PURE
CORE	-	22.60	21.50	22.60	23.57	20.10
CUBE		-	22.10	19.00	19.27	20.70
CURB			-	16.10	18.17	19.60
DOPE				-	20.87	19.87
OGRE					-	19.03
PURE						-

Visually dissimilar					
	CROW	CURL	CZAR	SHIP	TALE	UNDO
CROW	-	11.63	11.3	7.03	7.00	9.97
CURL		-	12.43	8.37	7.63	8.43
CZAR			-	10.47	6.53	9.87
SHIP				-	9.87	9.47
TALE					-	7.00
UNDO						-

See text for details

Table 5 Descriptive statistics for the stimuli used in Experiments 3 and 4

Visually similar words

	CnC.M	LgWF	LgCD	LgHAL	OLD	OLDF	PLD	PLDF	Length	NPhon	NSyll	Celex	Orth	OrthZ	OrthF
CORE	3.96	2.70	2.46	9.89	1.00	8.29	1.00	10.20	4.00	3.00	1.00	4.26	24.00	2.77	154.76
CUBE	4.62	2.18	1.96	8.52	1.55	8.19	1.60	8.01	4.00	4.00	1.00	3.72	5.00	-0.68	12.02
CURB	4.68	2.32	2.24	7.62	1.70	7.98	1.50	8.10	4.00	3.00	1.00	3.64	5.00	-0.68	6.23
DOPE	3.00	2.91	2.71	7.55	1.15	8.17	1.00	7.70	4.00	3.00	1.00	3.70	16.00	1.32	42.39
OGRE	3.97	1.86	1.45	8.45	1.90	8.55	1.65	7.57	4.00	3.00	2.00	3.35	1.00	-1.41	0.06
PURE	1.87	3.10	2.98	9.96	1.40	7.93	1.75	8.60	4.00	4.00	1.00	4.66	11.00	0.41	29.23
Mean	3.68	2.51	2.30	8.67	1.45	8.19	1.42	8.36	4.00	3.33	1.17	3.89	10.33	0.28	40.78
SD	1.08	0.47	0.55	1.06	0.34	0.22	0.33	0.97	0.00	0.52	0.41	0.48	8.52	1.55	57.98
Min	1.87	1.86	1.45	7.55	1.00	7.93	1.00	7.57	4.00	3.00	1.00	3.35	1.00	-1.41	0.06
Max	4.68	3.10	2.98	9.96	1.90	8.55	1.75	10.20	4.00	4.00	2.00	4.66	24.00	2.77	154.76

Visually dissimilar words

	CnC.M	LgWF	LgCD	LgHAL	OLD	OLDF	PLD	PLDF	Length	NPhon	NSyll	Celex	Orth	OrthZ	OrthF
CROW	4.93	2.36	2.12	8.45	1.45	8.05	1.40	8.47	4.00	3.00	1.00	3.62	6.00	-0.50	26.22
CURL	4.38	2.08	2.05	7.12	1.50	8.00	1.25	7.82	4.00	3.00	1.00	3.72	10.00	0.22	4.01
CZAR	4.14	1.56	1.32	6.11	1.90	8.66	1.65	9.78	4.00	3.00	1.00	3.34	1.00	-1.41	0.83
SHIP	4.87	3.70	3.07	10.85	1.45	8.22	1.00	8.40	4.00	3.00	1.00	4.67	8.00	-0.14	19.80
TALE	3.23	2.79	2.64	8.73	1.00	8.54	1.00	9.07	4.00	3.00	1.00	4.25	19.00	1.86	68.03
UNDO	2.56	2.35	2.27	7.78	1.90	8.97	1.95	8.82	4.00	4.00	2.00	3.70	2.00	-1.23	4.05
Mean	4.02	2.47	2.25	8.17	1.53	8.41	1.38	8.73	4.00	3.17	1.17	3.88	7.67	-0.20	20.49
SD	0.94	0.73	0.59	1.62	0.34	0.38	0.38	0.67	0.00	0.41	0.41	0.48	6.53	1.19	25.38
Min	2.56	1.56	1.32	6.11	1.00	8.00	1.00	7.82	4.00	3.00	1.00	3.34	1.00	-1.41	0.83
Max	4.93	3.70	3.07	10.85	1.90	8.97	1.95	9.78	4.00	4.00	2.00	4.67	19.00	1.86	68.03

Note: From Brysbaert et al. (2014): CnC.M = concreteness

From Brysbaert & New (2009): LgWF = Log word frequency and LgCD = log contextual diversity from SUBTLEX$_{US}$

From Balota et al. (2007): LgHal = log HAL frequency; OLD = orthographic Levenshtein distance; OLDF = frequency of orthographic Levenshtein neighbors; PLD = phonological Levenshtein distance; OLDF = frequency of phonological Levenshtein neighbors; Length = number of letters; NPhon = number of phonemes; NSYll = number of syllables

From Medler & Binder (2005): Celex = zipf Celex frequency; Orth = number of orthographic neighbors; OrthZ = z-transformed ORTH; OrthF = frequency of orthographic neighbors

The similar and dissimilar words do not differ on any of these dimensions, all ts < 1 and all ps > 0.45 except for OLDF where $t = 1.228$, $p = 0.248$

References

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. doi:https://doi.org/10.1080/14640746608400055

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459. doi:https://doi.org/10.3758/BF03193014

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977-900. doi:https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904–911. doi:https://doi.org/10.3758/s13428-013-0403-5.

Castellà, J., & Campoy, G. (2018). The (lack of) effect of dynamic visual noise on the concreteness effect in short-term memory. *Memory*, 26, 1355-1363. doi:https://doi.org/10.1080/09658211.2018.1476550

Chubala, C. M., Neath, I., & Surprenant, A. M. (2019). A comparison of immediate serial recall and immediate serial recognition. *Canadian Journal of Experimental Psychology 73*, 5-27. doi:https://doi.org/10.1037/cep0000158

Chubala, C., Surprenant, A. M. Neath, I., & Quinlan, P. T. (2018). Does dynamic visual noise eliminate the concreteness effect in working memory? *Journal of Memory and Language, 102*, 97-114. doi:https://doi.org/10.1016/j.jml.2018.05.009

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology*, 55, 75–84. doi:https://doi.org/10.1111/j.2044-8295.1964.tb00899.x

Farrell, S., & McLaughlin, K. (2007). Short-term recognition memory for serial order and timing. *Memory & Cognition, 35*, 1724–1734. doi:https://doi.org/10.3758/BF03193505

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. doi:https://doi.org/10.3758/BRM.41.4.1149

Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *Quarterly Journal of Experimental Psychology, 54A*, 1-30. doi:https://doi.org/10.1080/02724980042000002

Gisselgård, J., Uddén, J., Ingvar, M., & Petersson, K. M. (2007). Disruption of order information by irrelevant items: a serial recognition paradigm. *Acta Psychologica, 124*, 356–369. doi:https://doi.org/10.1016/j.actpsy.2006.04.002

Jacobs, J. (1887). Experiments on "Prehension". *Mind*, 12, 75-79. doi:https://doi.org/10.1093/mind/os-12.45.75

JASP Team (2018). JASP (Version 0.9.2) [Computer software] https://jasp-stats.org/

Jefferies, E., Frankish, C. R., & Lambon Ralph, M. A. (2006). Lexical and semantic binding in verbal short-term memory. *Journal of Memory and Language, 54*, 81–98. doi:https://doi.org/10.1016/j.jml.2005.08.001

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795. doi:https://doi.org/10.1080/01621459.1995.10476572

Kinsbourne, M. (1972). Behavioral analysis of the repetition deficit in conduction aphasia. *Neurology, 22*, 1126-1132. doi:https://doi.org/10.1212/WNL.22.11.1126

Lian, A., Karlsen, P. J., & Winsvold, B. (2001). A re-evaluation of the phonological similarity effect in adults' short-term memory of words

and nonwords. *Memory*, 9, 281–299. doi: https://doi.org/10.1080/09658210143000074

Lin, Y.-C., Chen, H.-Y., Lai, Y. C., & Wu, D. H. (2015). Phonological similarity and orthographic similarity affect probed serial recall of Chinese characters. *Memory & Cognition, 43*, 538–554. doi.org:10.3758/s13421-014-0495-x

Logie, R. H., Della Sala, S., Wynn, V., & Baddeley, A. D. (2000). Visual similarity effects in immediate verbal serial recall. *Quarterly Journal of Experimental Psychology, 53A*, 626–646. doi:https://doi.org/10.1080/027249800410463

Logie, R. H., Saito, S., Morita, A., Varma, S., & Norris, D. (2016). Recalling visual serial order for verbal sequences. *Memory & Cognition, 44*, 590–607. doi:https://doi.org/10.3758/s13421-015-0580-9

Macmillan, N. A., & Creelman, C. D. (2005). Detection theory: A user's guide, second edition. Mahwah, NJ: Erlbaum.

Medler, D. A., & Binder, J. R. (2005). *MCWord: An on-line orthographic database of the English language*. Retrieved from http://www.neuro.mcw.edu/mcword/

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An online lexical database. *International Journal of Lexicography, 3*, 235–244. doi:https://doi.org/10.1093/ijl/3.4.235

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, and Cognition, 29*, 1353–1380. doi:https://doi.org/10.1037/0278-7393.29.6.1353

Murray, D. J. (1968). Articulation and acoustic confusability in short-term memory. *Journal of Experimental Psychology, 78*, 679–684. doi:https://doi.org/10.1037/h0026641

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition, 18*, 251–269. doi:https://doi.org/10.3758/BF03213879

Nairne, J. S., & Neath, I. (2013). Sensory and working memory. In A. F. Healy & R. W. Proctor (Eds.), Comprehensive handbook of psychology, second edition, Vol. 4: Experimental Psychology (pp. 419-445). New York: Wiley.

Neath, I. (2000). Modeling the effects of irrelevant speech on memory. *Psychonomic Bulletin & Review, 7*, 403-423. doi:https://doi.org/10.3758/BF03214356

Neath, I., & Nairne, J. S. (1995). Word-Length effects in immediate memory: Overwriting trace decay theory. *Psychonomic Bulletin & Review, 2*, 429-441. doi:https://doi.org/10.3758/BF03210981

Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review, 105*, 761–781. doi:https://doi.org/10.1037/0033-295X.105.4.761-781

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology, 45*, 255-287. doi:https://doi.org/10.1037/h0084295

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet:: Similarity - Measuring the relatedness of concepts. In S. Dumais, D. Marcu, & S. Roukos (Eds.), *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2004)*, p. 38-41, Boston. Retrieved from www.aclweb.org/anthology/N04-3012

Peterson, L. R., & Johnson, S. T. (1971). Some effects of minimizing articulation on short-term retention. *Journal of Verbal Learning & Verbal Behavior, 10*, 346–354. doi:https://doi.org/10.1016/S0022-5371(71)80033-6

Romani, C., McAlpine, S., & Martin, R. C. (2008). Concreteness effects in different tasks: Implications for models of short-term memory. *Quarterly Journal of Experimental Psychology, 61*, 292–323. doi:https://doi.org/10.1080/17470210601147747

Saito, S., Logie, R. H., Morita, A., & Law, A. (2008). Visual and phonological similarity effects in verbal immediate serial recall: A test with

kanji materials. *Journal of Memory and Language, 59*, 1–17. doi: https://doi.org/10.1016/j.jml.2008.01.004

Shallice, T., & Warrington, E. K. (1977). The possible role of selective attention in acquired dyslexia. *Neuropsychologia, 15*, 31–41. doi: https://doi.org/10.1016/0028-3932(77)90112-9

Simpson, I. C., Mousikou, P., Montoya, J. M., & Defior, S. (2013). A letter visual-similarity matrix for Latin-based alphabets. *Behavior Research Methods, 45*, 431–439. doi:https://doi.org/10.3758/s13428-012-0271-4

Thorn, A. S., & Gathercole, S. E. (1999). Language-specific knowledge and short-term memory in bilingual and non-bilingual children. *Quarterly Journal of Experimental Psychology, 52A*, 303–324. doi:https://doi.org/10.1080/713755823

Thorn, A. S. C., Gathercole, S. E., & Frankish, C. R. (2002). Language familiarity effects in short-term memory: The role of output delay and long-term knowledge. *Quarterly Journal of Experimental Psychology, 55A*, 1363–1383. doi:https://doi.org/10.1080/02724980244000198

Tse, C.-S., Li, Y., & Altarriba, J. (2011). The effect of semantic relatedness on immediate serial recall and serial recognition. *Quarterly Journal of Experimental Psychology, 64*, 2425– 2437. doi:https://doi.org/10.1080/17470218.2011.604787

Ueno, T., & Saito, S. (2013). The role of visual representations within working memory for paired-associate and serial order of spoken words. *Quarterly Journal of Experimental Psychology, 66*, 1858–1872. doi:https://doi.org/10.1080/17470218.2013.772645.

Wickelgren, W. A. (1965). Short-term memory for phonemically similar lists. *American Journal of Psychology, 78*, 567-74. doi:https://doi.org/10.2307/1420917