Set Size and the Orthographic/Phonological Neighbourhood Size Effect in Serial Recognition: The Importance of Randomization

Dominic Guitard[1]

Leonie M. Miller[2]

Ian Neath[3]

Steven Roodenrys[2]

[1] School of Psychology, Cardiff University, Cardiff, United Kingdom.

[2] School of Psychology, University of Wollongong, NSW, Australia

[3] Department of Psychology, Virginia Tech, Blacksburg, VA, USA

Author Note

The data and stimuli are available from the Open Science Foundation, https://doi.org/10.17605/OSF.IO/E3HWJ. Order of authorship is alphabetical. Correspondence can be sent to any author at GuitardD@cardiff.ac.uk, leoniem@uow.edu.au, ineath@vt.edu, or steven@uow.edu.au.

Dominic Guitard, School of Psychology, Cardiff University, 70 Park Place, Cardiff, CF10 3AT, United Kingdom, GuitardD@cardiff.ac.uk, https://orcid.org/0000-0002-4658-3585

Leonie M. Miller, School of Psychology, University of Wollongong, Wollongong, NSW 2522 Australia, leoniem@uow.edu.au, +61 (0)242 214 454, https://orcid.org/0000-0002-1492-0954

Ian Neath, Department of Psychology, Virginia Tech, 890 Drillfield Road, Blacksburg, VA 24061, ineath@vt.edu, +1 (540) 231-1811

Steven Roodenrys, School of Psychology, University of Wollongong, Wollongong, NSW 2522 Australia, steven@uow.edu.au, +61 (0)242 214 072, https://orcid.org/0000-0002-3065-1766

Abstract

The neighbourhood size effect refers to the finding of better memory for words with more orthographic/phonological neighbours than otherwise comparable words with fewer neighbours. Although many studies have replicated this result with serial recall, only one has used serial recognition. Greeno et al. (2022) found no neighbourhood size effect when a large stimulus pool was used and a reverse effect -- better performance for small neighbourhood words -- when a small stimulus pool was used. We re-examined these results but made two methodological changes. First, for the large pool, we randomly generated lists for each subject rather than creating one set of lists that all subjects experienced. Second, for the small pool, we randomly generated a small pool for each subject rather than using one small pool for all subjects. In both cases, we observed a neighbourhood size effect consistent with results from the serial recall literature. Implications for methodology and for theoretical accounts of both the neighbourhood size effect and serial recognition are discussed.

Keywords: Orthographic neighbourhood size effect; phonological neighbourhood size effect; set size; serial recognition.

Public Significance Statement: A challenge in experimental psychology is to design experiments that will produce the most robust findings possible. Our study examined the impact of stimulus randomization on a serial recognition task, in which people indicate whether the words in two lists are in the same order or a different order. Full randomization of stimuli yielded results consistent with other related tasks, but different to findings using more limited controls of

the stimuli, thus underscoring the importance of methods that minimize the likelihood of the *un*generalizable effect.

Set Size and the Orthographic/Phonological Neighbourhood Size in Serial Recognition: The Importance of Randomization

A long-standing issue in psychological research is the extent to which results observed with one set of stimuli will generalize to other sets of similarly constructed stimuli. In the memory literature, one well-known example of a failure to generalize to different stimulus sets concerns the time-based word length effect. Baddeley et al. (1975, Exp. 3) created two sets of 10 words that were equated for number of syllables and number of phonemes but differed in pronunciation time. They found better recall of words that could be said faster than words that took longer to say, and this result has been replicated many times with these stimuli (e.g., Cowan et al., 1992; Longoni et al., 1993; Lovatt et al., 2000; Nairne et al., 1997). However, no other stimulus sets produce the same result. For example, Neath et al. (2003) reported four experiments that were identical except for the stimuli. Experiment 1 used the Baddeley et al. stimuli and replicated the original result. However, Experiment 2 used stimuli created by Caplan et al. (1992), Experiment 3 used stimuli created by Lovatt et al., and Experiment 4 used a newly created set of stimuli: None of these stimulus sets produced a short word advantage. The implication is that there is some idiosyncratic property of the Baddeley et al. stimulus set that caused their results.

One approach to ensuring that results generalize to different stimulus sets is to run lots of experiments each with a different set of stimuli. For many manipulations, however, there is an alternative (Keppel, 1976). Neath and Surprenant (2019) examined whether concreteness effects obtain when a closed set is used. They began with a large pool of concrete and abstract words that were equated on multiple other dimensions. For each subject in the closed set condition, they

randomly sampled 6 abstract and 6 concrete words from the larger pool, and these words were used throughout the experiment. In effect, they tested 30 different closed sets thereby avoiding the problem of relying on just one set of stimuli. If by chance one of the randomly generated sets did differ in ways other than concreteness, no other subject would have received those particular stimuli. Such randomization then, minimizes the likelihood of unwanted systematic variation between conditions.

A related problem can occur with large pools. For example, Tse and Altarriba (2022) created a single large pool of words when manipulating both valence (positive vs. negative) and concreteness (concrete vs. abstract). They found better serial recall of concrete positive than concrete negative words. In contrast, Bireta et al. (2021) created three different large pools of words manipulating only valence. In all three experiments, there was no effect of valence and the Bayes factors (BF) all indicated evidence supporting the null hypothesis. Guitard et al. (in press b) tested whether the difference in results between the two studies was due to randomization. Whereas Bireta et al. randomly generated each list for each subject, Tse and Altarriba created 5 concrete positive and 5 concrete negative lists and all subjects experienced the same lists. Guitard et al. noted that when the concrete positive and concrete negative lists were ranked by mean frequency, four of the five highest frequency lists were positive, and four of the five lowest frequency lists were negative. Although the larger pools were equated for frequency, the individual lists turned out to differ. Guitard et al. then re-ran Tse and Altarriba's experiment using their stimuli but randomly constructed each list for each subject. The effect of valence disappeared replicating the null effect of valence reported by Bireta et al. Randomly generating each list for each subject minimizes the chance of some unwanted systematic variation between conditions. One purpose of the current work is to assess whether similar methodological factors

involving randomization affected the results reported by Greeno et al. (2022) when examining the neighbourhood size effect on serial recognition.

There are a number of different definitions of what constitutes an orthographic or phonological neighbour of a target word. One commonly used definition of a neighbour is a word that differs by the substitution of a single letter (orthographic) or phoneme (phonological) in the target word (Coltheart et al., 1977). For example, orthographic neighbours of *cat* include *bat*, *cot*, and *cap*. Other definitions allow for adding or subtracting letters/phonemes, but in general, the same results obtain regardless of the specific definition: In serial recall, words which have a large number of neighbours are better recalled than words which have fewer neighbours for both large pools (e.g., Clarkson et al. 2017; Derraugh et al., 2017; Guitard et al., 2018) and small pools (e.g., Allen & Hulme, 2006; Jalbert et al., 2011a,b; Roodenrys et al., 2002).

Roodenrys (2009) proposed one explanation for this effect. As in many other accounts it is assumed that after list presentation, what is stored in memory are degraded representations of the list items. In the short-term/working memory literature, redintegration is the process whereby degraded representations are made interpretable or reconstructed using other information (for specific examples, see chapters 3 and 6 of Surprenant & Neath, 2009). Within Roodenrys' proposed framework, these degraded representations are used as input to an interactive activation network such that each representation partially activates its neighbours and the partial activation from the neighbours feeds back to the representation thereby accomplishing redintegration.

The activation feedback hypothesis posits that words with more neighbours will receive more feedback activation than words with fewer neighbours. Not only does this boost recall, but it is consistent with results from the speech production literature, in which words from larger neighbourhoods are produced more quickly than words from smaller neighbourhoods (e.g.,

Vitevitch, 2002; Vitevitch & Sommers, 2003). This general idea can be adapted to a number of different models and theories. For example, Derreaugh et al. (2017) incorporated it within the Feature Model (Nairne, 1990; Neath, 2000) whereas Clarkson et al. (2017) incorporated it within the item/order hypothesis. Regardless of the specific implementation, however, it is an unresolved question whether the advantage accruing to large neighbourhood words from more activation feedback occurs only at recall, when it facilitates redintegration, or whether it could also occur during presentation of the items.

This general account predicts a large neighbourhood advantage for serial recall regardless of the set size. Greeno et al. (2022) reported two studies using serial recall, one using a large pool and one using a small pool. They found a large neighbourhood advantage for the large pool but a reverse effect -- a small neighbourhood advantage -- for the small pool. Guitard et al. (in press a) noted that as with the Baddeley et al. (1975, Exp. 3) study, only one small pool was tested: Every subject received the same small set of items. Therefore, Guitard et al. re-ran the experiment but randomly generated a small pool for each subject, in essence testing 50 different small sets. They found a large neighbourhood advantage, the same result seen with a large pool. They ascribed Greeno et al.'s unusual results to some idiosyncratic property of the one small pool that was used. One speculation was that while both conditions had four words starting with the letter “c”, they differed in the actual phoneme. In the large neighbourhood condition, all four began with a hard 'C' sound whereas in the small neighbourhood condition, only one word began with that phoneme. This may have rendered the large neighbourhood words more confusable than the small.

Greeno et al. (2022) also reported two experiments that used serial recognition rather than serial recall. In serial recognition, a short list of items is presented one at a time followed by a

second list. The second list contains the same words as the first list and on half of the trials, the order of the words is also identical. On the other half of the trials, two adjacent items are transposed and the subject's task is to indicate whether the words in the two lists are in the same order. Greeno et al. found that with a large pool, there was no neighbourhood size effect in contrast to the robust effect seen in serial recall, and with a small pool, there was a reverse effect.

One possible explanation for the finding that neighbourhood size has different effects on serial recall and serial recognition is that the two tests differ considerably despite their surface similarity. For example, Chubala et al. (2018) found that dynamic visual noise -- a matrix of squares each of which randomly changes between black and white several times a second -- reduced performance in a serial recognition test but had no effect on a serial recall test. Importantly, the subjects did not know whether the test would be serial recognition or serial recall until after presentation. As a second example, Chubala et al. (2019) found that lists of semantically related words were better remembered in serial recall than lists of unrelated words, but semantic relatedness had no effect on serial recognition. They also found that lists of high frequency words were better remembered in serial recall than lists of low frequency words but frequency had no effect on serial recognition. Once again, subjects did not know the type of test until after presentation. It could be the case that the absence of a neighbourhood size effect, at least with a large pool, is just another instance of this type of difference. A second possible explanation, however, involves choices of randomization. When all subjects receive the same lists, or when only a single small pool is tested, it is possible that the results are due to unwanted systematic differences between the two conditions. Given that no studies other than those of Greeno et al. (2022) have examined the effect of neighbourhood size in serial recognition, the present studies re-examined the effect of neighbourhood size on this task.

To our knowledge, only one model of serial recognition has been proposed. Farrell and McLaughlin (2007) assumed that items are represented along a temporal dimension, but the exact location of each item on that dimension becomes less certain over time. At test, the first list is a noisy representation whereas the second list is noise free. A difference score is calculated between the order of items in the two sequences. If the score is sufficiently small, a response of "same" is given whereas if the score is sufficiently large, a response of "different" is given. One implication of this account is that redintegration is not needed in serial recognition: The two list representations are directly compared without needing to first identify or redintegrate each individual item. On this view, effects that require redintegration will not be observed in serial recognition whereas those that do not require redintegration should be observed. This accords well with the finding of Chubala et al. (2019) of no frequency effect in serial recognition because explanations of the frequency effect in serial recall typically invoke redintegration (e.g., Hulme et al., 1997; Lewandowsky & Farrell, 2000; Roodenrys et al., 2002; Saint-Aubin & Poirier, 2005; see Chubala et al., 2019, 2020, for more discussion about this implication). In terms of the neighbourhood size effect, the prediction of the Farrell and McLaughlin account depends on whether the feedback activation occurs only at test or whether it also occurs during list presentation. If it occurs only at retrieval, there should be no neighbourhood size effect because the two list representations are compared prior to redintegrating the individual items and therefore the choice about whether the representations are similar or different occurs before the activation feedback can boost the large neighbourhood words. On the other hand, if feedback activation can occur during list presentation, then a neighbourhood size effect should be observed.

The purpose of the current experiments, then, is twofold. One purpose is to determine whether the results reported by Greeno et al. (2022) replicate when the stimuli are randomized rather than fixed for each subject. A second purpose is to further refine the theoretical account of neighbourhood size effects and in particular whether there is evidence for feedback activation during presentation or whether such feedback occurs only during redintegration.

Experiment 1

Experiment 1 re-assessed whether a large neighbourhood advantage would be seen on a serial recognition task when a large pool was used. The experiment was based on Experiment 2 of Greeno et al. (2022) but differed in a number of ways. First, Greeno et al. manipulated modality, whether the lists were seen or heard. Because they reported no main effect of modality nor any interactions involving modality, we used only visual presentation. Second, Greeno et al. used 6-item lists whereas we used 5-item lists. The reason is that in unpublished pilot work in one of our labs, we found serial recognition performance with 6-item lists to be quite low, raising the possibility of floor effects. We therefore followed Chubala et al. (2019) in using 5-item lists. Third, Greeno et al. presented the visual words at a rate of one item every second but the words were displayed for only 350 ms; the reason was to match the duration of their auditory items. We again followed Chubala et al. and had the words visible for the full second. Fourth, Greeno et al. had a sample size of 30, whereas we had a sample size of 60. The reason for the difference is Greeno et al. estimated the sample size based on a frequentist analysis of variance whereas we used simulations to estimate the likelihood of getting informative Bayes factors. The final difference is that Greeno et al. created fixed lists of items and all subjects received the same lists. In contrast, we randomly constructed each list for each subject. Using randomized rather than

fixed lists minimizes the likelihood of any unwanted systematic variation between the two conditions.

Method

Ethics. The research was approved by Cardiff University's School of Psychology Ethics Committee.

Sample Size. The critical statistical test is a within-subjects Bayesian *t* test comparing *d'* for large and small neighbourhood words. The Bayes factor design analysis package (Schönbrodt & Stefan, 2018) was used to estimate an appropriate sample size. An effect size of $d = 0.5$ was used as the effect size for the alternative hypothesis and $d = 0.0$ was used for the null hypothesis. For each hypothesis, 10,000 simulations were conducted using a non-directional Bayesian within-subjects *t* test with default priors. For the alternative hypothesis, the simulations indicated that with 60 subjects 88.9% of the samples indicated evidence for the alternative hypothesis (BF > 3), 10.5% were inconclusive (0.333 < BF < 3), and 0.7% indicated evidence for the null hypothesis (BF < 0.333). For the null hypothesis, simulations indicated that 82.1% of the samples showed evidence for the null hypothesis (BF < 0.333); 16.8% were inconclusive (0.333 < BF < 3), and 1.0% showed evidence for the alternative hypothesis (BF > 3). A sample size of 60, then, should result in an informative Bayes factor whether there is a neighbourhood size effect or a null result.

Subjects. Sixty volunteers from Cardiff University's psychology subject pool volunteered to participate in exchange for course credit. The mean age was 19.57 years ($SD = 1.91$, range 18-30), 49 self-identified as female and 11 as male, and all were native speakers of English.

Stimuli. The stimuli were the same as in Experiment 2 of Greeno et al. (2022) and consisted of the 47 large and 47 small neighborhood size words from Clarkson et al. (2017), to

which Greeno et al. added the word NUT to the large neighborhood pool and the word RIB to the small neighborhood pool. Each word was a single syllable CVC (see the Appendix).

Procedure. After reading an informed consent form and agreeing to participate, the subjects were reminded of the instructions. A trial began when the subject clicked on a button labelled “Start next trial”. Five words were randomly drawn without replacement from the appropriate pool (i.e., large or small neighbourhood size) and were shown one at a time for 1 s in the centre of the screen in 28 point Helvetica. Two seconds after the final word had been shown, a second list was shown. Half the time this second list was identical to the first and half the time two adjacent items were transposed. A message then appeared prompting the subject to indicate whether the order of the words was the same or different and they responded by clicking on an appropriately labelled button.

There were 60 trials. Half the trials had large neighbourhood words and half had small neighbourhood words. For each type of list there were 15 same and 15 different lists. For different trials, words 2 and 3, 3 and 4, and 4 and 5 were transposed equally often; the first word was never transposed. The order of these trials was randomly determined for each subject. On each trial, words were randomly sampled without replacement from the appropriate larger pool. After 9 trials of a given condition, 45 words would have been used from that pool and only 3 words remained. At this point, the pool was restored to 48 words and sampling without replacement began again for another 9 trials of that condition. Each word, therefore, typically appeared either 3 or 4 times during the experiment.

Data Analysis

The data were analyzed using JASP (JASP Team, 2023) and we report a Bayes factor, BF_{10}, that indicates evidence for the alternative hypothesis. We interpret a value between 3 and

10 as indicating substantial evidence; a value between 10 and 30 indicating strong evidence; values between 30 and 100 indicating very strong evidence; and values greater than 100 indicating decisive evidence (Wetzels et al., 2011). BF_{01} indicates evidence for the null hypothesis using the same scale.

A hit was defined as correctly responding "different" to a different list and a false alarm was defined as incorrectly responding "different" to a same list. From these we calculated *d′*, the ability to discriminate between same and different trials, and *C*, a measure of response bias (see Macmillan & Creelman, 2004). Because *d′* cannot be calculated if there is a hit or false alarm rate of 1 or 0, we transformed the data to eliminate values of 1 and 0 as recommended by Snodgrass and Corwin (1988), as did Greeno et al. (2022). The transformed hit rate is calculated according to the formula $\frac{\#H + 0.5}{\#Diff + 1}$, where #H is the number of hits and #Diff is the number of different trials. The false alarm rate was calculated similarly.

Results and Discussion

Table 1 shows the hit and false alarm rates, *d′*, and *C* for the large and small neighbourhood size conditions, as well as the Bayes factors comparing the two conditions on each measure. The left panel of Figure 1 also shows the *d′* values. There was a neighbourhood size effect with *d′* higher for large neighbourhood lists than for small neighbourhood lists; the effect size was $d = 0.477$. This does not replicate the null result in serial recognition reported by Greeno et al. (2022, Exp. 2) when using a large pool. The two studies differ most notably in whether a fixed set of lists was used or whether the lists were randomly generated. When fixed lists are used, any inadvertent systematic difference will affect every subject because all subjects receive the same lists. When randomly generated lists are used, any inadvertent systematic difference will affect only one subject.

Table 1.

Mean hit rate, false alarm rate, *d'*, and *C* in Experiment 1 as a function of whether the lists had large neighbourhood words or small neighbourhood words.

	Large Neighbourhood		Small Neighbourhood		
Measure	*M*	*SD*	*M*	*SD*	BF_{10}
Hit	0.729	0.195	0.696	0.212	0.629
FA	0.210	0.143	0.261	0.165	9.338
d'	1.676	1.089	1.375	1.057	51.090
C	0.094	0.318	0.042	0.390	0.253

The presence of a neighborhood size effect in serial recognition suggests that within the context of Roodenrys' (2009) activation feedback account and Farrell and McLaughlin's (2007) model, feedback activation can occur during list presentation. Within the model, the decision about same versus different occurs without the need for redintegration; if activation feedback occurred only during redintegration, there would have been no effect.

Experiment 2

The sole change from Experiment 1 was that instead of using a large pool of stimuli, Experiment 2 used the small pool used by Greeno et al. (2022) in their Experiment 4. Experiment 1 found a large neighbourhood advantage when a large pool was used, and the interactive activation feedback account predicts the same result for a small pool.

Method

Subjects. Sixty different volunteers from Cardiff University's psychology subject participated. The mean age was 19.13 years (*SD* = 1.07, range 18-23) and 54 self-identified as female, 5 as male, and 1 as other. All were native speakers of English.

Stimuli. The stimuli were the 12 large and 12 small neighbourhood size words used in Experiment 4 of Greeno et al. (2022) (see the Appendix).

Design. The design was the same as in Experiment 1.

Procedure. The procedure was identical to Experiment 1 except for the stimuli. Because a small pool was used, each pool needed to be replenished after 2 trials of that stimulus type.

Results and Discussion

Table 2.

Mean hit rate, false alarm rate, *d'*, and *C* in Experiment 2 as a function of whether the lists had large neighbourhood words or small neighbourhood words.

	Large Neighbourhood		Small Neighbourhood		
Measure	*M*	*SD*	*M*	*SD*	BF_{10}
Hit	0.666	0.168	0.695	0.138	0.417
FA	0.297	0.179	0.303	0.179	0.149
d'	1.114	0.921	1.150	0.779	0.156
C	0.064	0.337	0.014	0.333	0.241

Table 2 shows the hit and false alarm rates, *d'*, and *C* for the large and small neighbourhood size conditions, as well as the Bayes factors comparing the two conditions on each measure. The middle panel of Figure 1 also shows the *d'* values. For *d'*, there is no effect of neighbourhood size: $BF_{01} = 6.410$, indicating substantial evidence for the null hypothesis, and the effect size was $d = 0.092$. This does not replicate the reverse neighbourhood result reported

by Greeno et al. (2022) in their Experiment 4. Moreover, this result does not replicate the robust neighbourhood size effect observed in Experiment 1, which used a large set size. We postpone discussion of the results of this experiment until after presenting the results of Experiment 3.

Experiment 3

Experiment 4 of Greeno et al. (2022) and Experiment 2 here may both be described as having a *fixed* small pool: All subjects saw the same small set of words, 12 of each type. As noted in the introduction, a different way to assess small sets of stimuli that minimizes the chance of an unwanted systematic variation is to randomly generate a small pool for each subject (Neath & Surprenant, 2019). This can be described as a *random* small pool. Guitard et al. (in press a) showed that whereas the fixed small pool of stimuli resulted in no difference between small and large neighbourhood words in serial recall, a large neighbourhood advantage was observed when random small pools were used. The purpose of Experiment 3, then, was to assess whether this is the case for serial recognition. Experiment 3 was identical to Experiment 2 except that instead of using a fixed small pool, the small pool of words was randomly determined for each subject; in effect, each subject saw a different small pool of items.

Method

Subjects. Sixty different volunteers from Cardiff University's psychology subject participated. The mean age was 19.28 years (*SD* = 0.88, range 18-22) and 54 self-identified as female, 5 as male, and 1 as other. All were native speakers of English.

Stimuli. The stimuli were 12 large neighbourhood size words and 12 small neighbourhood size words randomly drawn, for each subject, from the larger pool used in Experiment 1.

Design. The design is the same as in Experiment 2.

Procedure. The procedure is identical to Experiment 2 except that the 12 large and 12 small neighbourhood size words that comprised the small pools were randomly determined for each subject.

Results and Discussion

Table 3.

Mean hit rate, false alarm rate, *d'*, and *C* in Experiment 3 as a function of whether the lists had large neighbourhood words or small neighbourhood words.

	Large Neighbourhood		Small Neighbourhood		
Measure	*M*	*SD*	*M*	*SD*	BF_{10}
Hit	0.678	0.200	0.655	0.177	0.321
FA	0.226	0.143	0.285	0.170	14.128
d'	1.438	1.118	1.128	0.980	36.920
C	0.153	0.277	0.101	0.304	0.305

Table 3 shows the hit and false alarm rates, *d'*, and *C* for the large and small neighbourhood size conditions, as well as the Bayes factors comparing the two conditions on each measure. The right panel of Figure 1 also shows the *d'* values. There is a neighbourhood size effect with *d'* higher for large neighbourhood lists than for small neighbourhood lists. The Bayes factor indicated very strong evidence in favour of the alternative hypothesis and the effect size was $d = 0.462$. This is in contrast to the null result observed in Experiment 2. The sole difference between the two experiments is that all subjects in Experiment 2 saw the same fixed small pool of items whereas each subject in Experiment 3 saw a small pool that was randomly

generated specifically for them. In effect, each subject had a different small pool. This result parallels that observed with serial recall by Guitard et al. (in press a): No effect of neighbourhood size with the fixed small pool but a robust large neighbourhood advantage with a random small pool.

General Discussion

The three experiments reported here make three primary contributions, one methodological, one empirical, and one theoretical. For the first, the results emphasize yet again the importance of randomizing stimuli to prevent unwanted systematic variations. Greeno et al. (2022) reported the only experiments that have looked at neighbourhood size effects in serial recognition to date, one with a large pool and one with a small pool. With the large pool, they found no neighbourhood size effect, but all subjects saw the same words in the same order. Experiment 1 used the same large pool but generated each list randomly for each subject and a large neighbourhood advantage was observed, just as in serial recall. We attribute the different results to randomization, which minimizes the chance of unwanted systematic differences influencing the results. With the small pool, they found a reverse neighborhood size effect, with better performance for small rather than large neighborhood size words. Experiment 2 failed to replicate this result; rather, there was no difference, with the Bayes factor supporting the null hypothesis. Again, the selection of words to each list was randomized for each subject in our Experiment 2. Experiment 3 also used a small pool, but the small pool was randomly generated for each subject and again a large neighbourhood advantage was found. In effect, Experiment 3 tested 60 different small pools as compared to the single small pool used in Experiment 2. We again attribute the different results to randomization. If by chance the randomly selected small and large neighborhood words differed systematically along some other dimension, that would

affect only one subject. If all subjects had received the same pool, all subjects would have been affected by the unwanted systematic variation.

For the empirical contribution, the results suggest that orthographic/neighborhood size affects serial recognition in a similar way as serial recall. It was a possibility that the two tests would yield different results, given that a number of manipulations have been identified that affect only one but not the other test. The results make it clear that neighborhood size differs from other manipulations that affect only serial recall (such as semantic relatedness and frequency) or only serial recognition (such as dynamic visual noise). Moreover, neighborhood size effects obtain in both tasks regardless of whether the stimulus pool is large or small as long as appropriate randomization occurs.

For the theoretical contribution, the results are consistent with Roodenrys' (2009) activation feedback account, but also inform on where the locus of the effect may reside. The sole model of serial recognition of which we aware (Farrell & McLaughlin, 2007) posits that people compare a noisy representation of List 1 to a noise free representation of List 2. If the difference is sufficiently large, a "different" judgment is made, otherwise a "same" judgment is made. The two list representations are compared without needing to first identify or redintegrate each individual item. By analogy, a person who does not know Chinese characters can assess whether two characters are the same or different without knowing the meaning of the individual items.

According to the activation feedback account, whether a neighbourhood size effect obtains depends on when the activation feedback occurs. If activation feedback occurs only during redintegration, neighbourhood size should not affect serial recognition because the same/different decision is made without the need for redintegration. On the other hand, if

activation feedback can also occur during presentation, then a neighbourhood size effect should occur because the List 1 representation of large neighbourhood words should be less noisy than that of small neighbourhood words. The results are consistent with the idea that activation feedback can occur during presentation. As List 1 is presented, words with more neighbours receive more activation feedback than words with fewer neighbours, resulting in a less noisy list representation. This idea, that activation feedback can occur during presentation, is not without precedent, albeit from a different literature. In the reading literature, for example, models of visual word recognition involve some form of competition between visually similar words, and words with more neighbours receive more activation (Norris, 2013). The task of a reading a word is akin to seeing a word during presentation in that both are the initial processing of the stimulus. As a second example, in a lexical decision task, words with more orthographic and/or phonological neighbours are judged to be words more quickly than words with fewer neighbors (Andrews, 1997; Yates et al., 2004).

The results are also consistent with the idea that one important difference between serial recall and serial recognition is the extent to which redintegration is involved. While redintegration is a key feature of many models of serial recall (e.g., Lewandowsky, 1999; Nairne, 1990; Neath, 2000; Roodenrys & Miller, 2008), the Farrell and McLaughlin (2007) model suggests redintegration is not necessary in serial recognition because a same/different decision can be made without the need for this stage. Although this view -- that serial recall and serial recognition differ in whether redintegration occurs -- has not received many empirical tests, it does account for the findings that concreteness, acoustic similarity, and neighbourhood size affect serial recall and serial recognition in the same way, whereas frequency and semantic

relatedness affect only serial recall (Chubala et al., 2019). The results reported here suggest that this redintegration explanation should be further explored.

Summary

A large neighbourhood advantage was seen in serial recognition when a large pool was used when the lists were randomly determined. Similarly, a large neighbourhood advantage was also seen when a small pool was used when the small pools were generated by randomly sampling a subset of stimuli from the larger pool for each subject. This is the same pattern of results reported in serial recall when similar randomization occurs (Guitard et al., in press a). This randomization of stimuli is necessary to minimize the risk of idiosyncratic results due to unwanted systematic differences between the conditions. These findings have important theoretical implications for the activation feedback account of Roodenrys (2009), suggesting that activation feedback can occur during encoding and that redintegration is absent in serial recognition.

References

Allen, R., & Hulme, C. (2006). Speech and language processing mechanisms in verbal serial recall. *Journal of Memory and Language*, *55*(1), 64-88. Doi: 10.1016/j.jml.2006.02.002

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review, 4*(4), 439-461. doi:10.3758/BF03214334

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning & Verbal Behavior, 14*(6)*,* 575–589. Doi:10.1016/S0022- 5371(75)80045-4

Bireta, T. J., Guitard, D., Neath, I., & Surprenant, A. M. (2021). Valence does not affect serial recall. *Canadian Journal of Experimental Psychology, 75*(1), 35-47. doi:10.1037/cep0000239

Bireta, T. J., Neath, I., & Surprenant, A. M. (2006). The syllable-based word length effect and stimulus set specificity. *Psychonomic Bulletin & Review*, *13*(3), 434-438. doi:10.3758/BF03193866

Caplan, D., Rochon, E., & Waters, G. S. (1992). Articulatory and phonological determinants of word length effects in span tasks. *Quarterly Journal of Experimental Psychology, 45A*(2), 177-192. doi:10.1080/14640749208401323

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*(1), 5-27. doi:10.1037/cep0000158

Chubala, C. M., 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:10.1016/j.jml.2018.05.009

Clarkson, L., Roodenrys, S., Miller, L. M., & Hulme, C. (2017). The phonological neighbourhood effect on short-term memory for order. *Memory, 25*(3), 391-402. doi:10.108 0/09658211.2016.1179330

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535-555). Hillsdale, NJ: Erlbaum.

Cowan, N., Day, L., Saults, J. S., Keller, T. A., Johnson, T., & Flores, L. (1992). The role of verbal output time in the effects of word length on immediate memory. *Journal of Memory and Language, 31*(1), 1-17. doi:10.1016/0749-596X(92)90002-F

Derraugh, L. S., Neath, I., Surprenant, A. M., Beaudry, O., & Saint-Aubin, J. (2017). The effect of lexical factors on recall from working memory: Generalizing the neighbourhood size effect. *Canadian Journal of Experimental Psychology, 71,* 23-31. doi:10.1037/cep0000098

Farrell, S., & McLaughlin, K. (2007). Short-term recognition memory for serial order and timing. *Memory & Cognition, 35*(7), 1724-1734. doi:10.3758/ BF03193505.

Greeno, D. J., Macken, B., & Jones, D. M. (2022). The company a word keeps: The role of neighbourhood density in verbal short-term memory. *Quarterly Journal of Experimental Psychology 5*(11), 2159-2176. doi:10.1177/17470218221080398

Guitard, D., Gabel, A. J., Saint-Aubin, J., Surprenant, A. M., & Neath, I. (2018). Word length, set size, and lexical factors: Re-examining what causes the word length effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 1824-1844. doi:10.1037/ xlm0000551

Guitard, D., Miller, L. M., Neath, I., & Roodenrys, S. (in press a). The

orthographic/phonological neighbourhood size effect and set size. *Quarterly Journal of Experimental Psychology*.

Guitard, D., Miller, L. M, Neath, I., & Roodenrys, S. (2023). Set size and the orthographic/phonological neighbourhood size effect in serial recognition: The importance of randomization. doi:10.17605/OSF.IO/E3HWJ

Guitard, D., Neath, I., & Saint-Aubin, J. (in press b). Additional evidence that valence does not affect serial recall. *Quarterly Journal of Experimental Psychology*. doi:10.1177/17470218221126635

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, M., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(5), 1217-1232. doi:10.1037/0278-7393.23.5.1217

JASP Team (2023). JASP (Version 0.17.1) [Computer software] https://jasp-stats.org/

Jalbert, A., Neath, I., Bireta, T. J., & Surprenant, A. M. (2011a). When does length cause the word length effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 338-353. doi:/10.1037/a0021804

Jalbert, A., Neath, I., & Surprenant, A. M. (2011b). Does length or neighborhood size cause the word length effect? *Memory & Cognition, 39*(7), 1198-1210. doi:10.3758/s13421-011-0094-z

Keppel, G. (1976). Words as random variables. *Journal of Verbal Learning and Verbal Behavior*, *15*(3), 263–265. doi:10.1016/0022-5371(76)90024-4

Lewandowsky, S. (1999). Redintegration and response suppression in serial recall: A dynamic network model. *International Journal of Psychology*, *34*(5-6), 434-446.

doi:10.1080/002075999399792

Lewandowsky, S., & Farrell, S. (2000). A redintegration account of the effects of speech rate, lexicality, and word frequency in immediate serial recall. *Psychological Research, 63*(2), 163-173. doi:10.1007/ PL00008175

Longoni, A. M., Richardson, J. T., & Aiello, A. (1993). Articulatory rehearsal and phonological storage in working memory. *Memory & Cognition, 21*(1), 11-22. doi:10.3758/BF03211160

Lovatt, P., Avons, S. E., & Masterson, J. (2000). The word-length effect and disyllabic words. *Quarterly Journal of Experimental Psychology, 53A*(1), 1-22. doi:10.1080/027249800390646

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. New York, NY: Psychology Press. doi:10.4324/9781410611147

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition, 18*(3), 251-269. doi:10.3758/BF03213879

Nairne, J. S., Neath, I., & Serra, M. (1997). Proactive interference plays a role in the word-length effect. *Psychonomic Bulletin & Review, 4*(4), 541-545. doi:10.3758/BF03214346

Neath, I. (2000). Modeling the effects of irrelevant speech on memory. *Psychonomic Bulletin & Review, 7*(3), 403-423. doi:10.3758/BF03214356

Neath, I., & Surprenant, A. M. (2019). Set size and long-term memory/lexical effects in immediate serial recall: Testing the impurity principle. *Memory & Cognition, 47*(3), 455-472. doi:10.3758/s13421-018-0883-8

Neath, I., Bireta, T. J., & Surprenant, A. M. (2003). The time-based word length effect and stimulus set specificity. *Psychonomic Bulletin & Review, 10*(2), 430-434.

doi:10.3758/BF03196502

Norris D. (2013). Models of visual word recognition. *Trends in Cognitive Sciences*, *17*(10), 517-524. doi:10.1016/j.tics.2013.08.003

Roodenrys, S. (2009). Explaining phonological neighbourhood effects in short-term memory. In A. S. C. Thorn & M. P. A. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 177-197). Hove, UK: Psychology Press.

Roodenrys, S., & Miller, L. M. (2008). A constrained Rasch model of trace redintegration in serial recall. *Memory & Cognition*, *36*(3), 578-587. doi:10.3758/MC.36.3.578

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. M. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(6)*, 1019-1034. doi:10.1037/0278-7393.28.6.1019

Saint-Aubin, J., & Poirier, M. (2005). Word frequency effects in immediate serial recall: Item familiarity and item co-occurrence have the same effect. *Memory, 13*(3-4), 325-332. doi:10.1080/09658210344000369

Schönbrodt, F. D., & Stefan, A. M. (2019). BFDA: An R package for Bayes factor design analysis (version 0.5). https://github.com/nicebread/BFDA

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*(1), 34-50. doi:10.1037/0096-3445.117.1.34

Surprenant, A. M., & Neath, I. (2009). *Principles of memory.* Psychology Press.

Tse, C., & Altarriba, J. (2022). Independent effects of word concreteness and word valence on immediate serial recall. *British Journal of Psychology, 113*(3), 820-834.

doi:10.1111/bjop.12566

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6*(3), 291-298. doi: 10.1177/1745691611406923

Yates, M., Locker, L., Jr., & Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review, 11*(3), 452-457. doi:10.3758/BF03196594

Appendix

In Experiment 1, the stimuli were the same as in Experiment 2 of Greeno et al. (2022). These were the 47 large and 47 small neighborhood size words from Clarkson et al. (2017) and two words added by Greeno et al., NUT in the large neighborhood pool and RIB in the small neighborhood pool.

Large neighborhood: bark, bead, bin, boot, buzz, cone, cop, cork, cull, duck, fade, fizz, ham, haze, hike, hood, kite, lag, lard, lice, lick, maim, meek, mole, node, nut, peach, pearl, pod, poise, poke, pun, rake, rim, ripe, robe, sane, sap, shack, siege, sock, tan, tart, thorn, tile, vine, weep, weird

Small neighborhood: badge, beige, carve, chase, chef, chime, churn, couch, dab, dodge, fang, fog, forge, geese, germ, gig, gown, gush, herb, ledge, loaf, lurch, merge, mesh, morgue, moth, noose, notch, nudge, peg, pierce, rib, shove, sieve, soothe, thatch, thief, thud, torch, turf, verse, vogue, void, web, wharf, yarn, zip, zoom

In Experiment 2, the stimuli were the same stimuli as in Experiment 4 of Greeno et al. (2022).

Large neighborhood: bark, boot, cone, cop, cork, cull, pearl, pod, tan, tart, thorn, tile

Small neighborhood: badge, beige, chase, chef, chime, couch, peg, pierce, thatch, thief, torch, turf

In Experiment 3, for each subject 12 words were randomly selected from the large neighborhood pool and 12 from the small neighborhood pool used in Experiment 1. The pools formed can be determined by examining the raw data at the Open Science Foundation (Guitard et al., 2023).

Figure 1: Mean *d′* values as a function of neighborhood size for Experiment 1 (left panel), which used a large set size; Experiment 2 (middle panel), which used a small fixed set size; and Experiment 3 (right panel), which used a small random set size. Error bars show the standard error of the mean.