Original Article

QJEP

Quarterly Journal of Experimental Psychology
2024, Vol. 77(2) 298–307
© Experimental Psychology Society 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17470218231165863
qjep.sagepub.com
SAGE

The orthographic/phonological neighbourhood size effect and set size

Dominic Guitard[1], Leonie M Miller[2], Ian Neath[3] and Steven Roodenrys[2]

Abstract

A growing number of studies have shown that on serial recall tests, words with more orthographic/phonological neighbours are better recalled than otherwise comparable words with fewer neighbours, the so-called neighbourhood size effect. Greeno et al. replicated this result when using a large stimulus pool but found a reverse neighbourhood size effect—better recall of words with fewer rather than more neighbours—when using a small stimulus pool. We report three registered experiments that further examine the role of set size in the neighbourhood size effect. Experiment 1 used the large pool from Greeno et al. and replicated their finding of a large-neighbourhood advantage. Experiment 2 used the small pool from Greeno et al. but found no difference in recall between the large and small neighbourhood conditions. Experiment 3 also used a small pool but the small pool was randomly generated for each subject from the large pool used in Experiment 1. This resulted in a typical large neighbourhood advantage. We suggest that set size is not critical to the direction of the neighbourhood size effect, with a large neighbourhood advantage appearing with both small and large pools.

Keywords

Orthographic neighbourhood size effect; phonological neighbourhood size effect; set size; serial recall

Received: 18 November 2022; revised: 5 February 2023; accepted: 9 March 2023

Memory researchers frequently need to assemble specific stimulus sets when assessing predictions of various theories. For example, they might want a set of abstract words and a corresponding set of concrete words, or a set of high-frequency words and a corresponding set of low-frequency words. One well-known concern is that as the size of the stimulus set gets smaller, it becomes increasingly more likely that idiosyncratic properties of the particular words can influence performance and that the results may not generalise to other stimulus sets (e.g., Bireta et al., 2006; Caplan et al., 1992; Lovatt et al., 2000; Neath et al., 2003). In this article, we report a series of experiments that examine whether an unusual finding reported by Greeno et al. (2022) about the effect of set size on the neighbourhood size effect in serial recall generalises to other stimulus sets.

There are several different definitions of what constitutes an orthographic or phonological neighbour of a target word. One widely used definition of an orthographic neighbour is Coltheart's *N* (M. Coltheart et al., 1977), the number of words that differ from the target by a single letter (e.g., neighbours of *cat* include *bat, cot*, and *cap*). There is a corresponding definition for a phonological neighbour, a word that differs by a single phoneme. Other definitions permit the addition or deletion of letters and phonemes, such as that proposed by Yarkoni et al. (2008). Their measure, OLD (orthographic Levenshtein distance) is based on "the minimum number of substitution, insertion, or deletion operations required to turn one word into the other" (Yarkoni et al., 2008, p. 972). A similar measure, PLD (phonological Levenshtein distance), describes phonological neighbours. Words with more neighbours (i.e., a lower OLD or PLD value) are said to have a large orthographic/phonological neighbourhood whereas words with fewer

[1]School of Psychology, Cardiff University, Cardiff, UK
[2]School of Psychology, University of Wollongong, Wollongong, NSW, Australia
[3]Department of Psychology, Virginia Tech, Blacksburg, VA, USA

Corresponding author:
Steven Roodenrys, School of Psychology, University of Wollongong, Northfields Avenue, Wollongong, NSW 2522, Australia.
Email: steven@uow.edu.au

Table 1. The size of word pools used in neighbourhood size experiments, the mean number of orthographic neighbours in the large and small conditions, and performance measures.

Study	*N*	Neigh. size		Prop. correct		Notes
		Large	Small	Large	Small	
Allen and Hulme (2006)	16	11.31	4.31	0.59	0.53	High NF words from Roodenrys et al. (2002) Exp. 1
	16	10.81	4.06	0.52	0.46	Low NF words from Roodenrys et al. (2002) Exp. 1
Clarkson et al. (2017), Exp. 2	47	10.94	3.77	0.78	0.73	
Clarkson et al. (2017), Exp. 3	47	10.94	3.77	0.73	0.68	
Derraugh et al. (2017), Exp. 1	282	14.45	1.81	0.59	0.55	
Derraugh et al. (2017), Exp. 2	140	13.14	3.20	0.45	0.41	French stimuli
Greeno et al. (2022), Exp. 1[a]	48	10.94	3.92	0.61	0.52	From Clarkson et al. (2017) with 1 additional word
Greeno et al. (2022), Exp. 3[b]	12	12.25	3.58	0.54	0.56	Subset of Clarkson et al. (2017)
Guitard et al. (2018), Exp. 7	70	17.47	1.56	0.65	0.60	
Jalbert, Neath, Bireta and Surprenant (2011), Exp. 2	16	11.06	3.67	0.72	0.69	Low NF words from Roodenrys et al. (2002) Exp. 3
Jalbert, Neath and Surprenant (2011), Exp. 1	16	11.06	3.67	0.74	0.64	Low NF words from Roodenrys et al. (2002) Exp. 3
				Span score		
Goh and Pisoni (2003), Exp. 1[c]	66	14.32	9.89	2.83	3.30	
	8	2.76	2.80	3.51	3.55	
Goh and Pisoni (2003), Exp. 2[d]	66	14.32	9.89	3.26	3.72	
	8	2.76	2.80	4.03	3.94	
Roodenrys et al. (2002), Exp. 1	16	11.31	4.31	5.11	4.69	High NF
	16	10.81	4.06	4.53	4.43	Low NF
Roodenrys et al. (2002), Exp. 3	16	9.19	2.63	4.82	4.62	High NF
	16	11.06	3.67	4.72	4.27	Low NF

Note: Number of neighbours is Mean_Ortho_N from the ELexicon Project (Balota et al., 2007). NF: frequency of the neighbours.

[a] Estimated from Figure 3.

[b] Estimated from Figure 6.

[c] Estimated from Figure 1.

[d] Estimated from Figure 4.

neighbours are said to have a small orthographic/phonological neighbourhood.

The neighbourhood size effect refers to the finding that words which have a large number of orthographic or phonological neighbours are better recalled on immediate memory tests than otherwise comparable words which have fewer neighbours (e.g., Allen & Hulme, 2006; Clarkson et al., 2017; Derraugh et al., 2017; Guitard et al., 2018; Jalbert, Neath, Bireta, & Surprenant, 2011; Jalbert, Neath, & Surprenant, 2011; Roodenrys et al., 2002). As can be seen in Table 1, these studies have used stimulus sets that vary in size from small to large.

One commonly invoked account of the beneficial effect of neighbourhood size derives from Roodenrys (2009) and focuses on redintegration. The idea is that after presentation, the degraded cues representing the items can serve as input to an interactive network. Each cue partially activates or primes its neighbours. A word with more neighbours, such as *fir* which has 10 phonological neighbours, will partially activate more items than a word with fewer neighbours, such as *elm* which has only three neighbours.[1] The activation from the neighbours feeds back to the cue, with more feedback activation for large than for small neighbourhood words. The greater the amount of feedback activation, the greater the likelihood of successful redintegration. This type of explanation is consistent with work from the speech production literature, where a large-neighbourhood advantage is seen when producing words (e.g., Vitevitch, 2002; Vitevitch & Sommers, 2003) but not when perceiving words (e.g., Luce & Pisoni, 1998). In addition, this general redintegrative framework can be adapted to a variety of specific models and theories. For example, Derraugh et al. (2017) adapted it within the Feature Model (Nairne, 1990; Neath, 2000) whereas Clarkson et al. (2017) adapted it within the item/order hypothesis.

There are other theories of short-term memory that would suggest the effect of neighbourhood size does not just influence processes operating during recall. For many years, some have argued that verbal short-term memory makes use of structures and processes inherent in the comprehension and production of speech (e.g., Martin et al., 1999). These psycholinguistic accounts of short-term memory have varied from arguing that memory tasks rely entirely on co-opted language processes (e.g., Schwering & MacDonald, 2020) to arguing that they make use of linguistic representations of the words, along with additional attentional processes to maintain the novel order of the lists of words (e.g., Majerus, 2013). In either case, variation in activation between different linguistic representations might be expected to occur from the time of presentation onwards. Thus, the neighbourhood size effect may reflect differential activation across the entire task rather than just at retrieval.

There are some manipulations which reduce or abolish the neighbourhood size effect. For example, Jalbert, Neath, Bireta and Surprenant (2011) manipulated whether the lists were pure (i.e., contained only large or only small neighbourhood words) or mixed (i.e., contained both large and small neighbourhood words). A neighbourhood size effect was found for pure lists but not for mixed lists, a result replicated by Jalbert, Neath and Surprenant (2011) and Clarkson et al. (2017). Neighbourhood size is thus like a number of other memory results that differ as a function of pure or mixed lists (e.g., the generation effect; Serra & Nairne, 1993). As a second example, Jalbert, Neath and Surprenant (2011) showed that concurrent articulation eliminates the neighbourhood size effect, similar to the elimination of the acoustic similarity effect (e.g., Murray, 1967).

As can be seen in Table 1, there are two papers which reported *reverse* neighbourhood size effects, better recall of small than large neighbourhood words. Goh and Pisoni (2003) manipulated both neighbourhood size and set size.[2] With the large set size, they found better memory for words with small neighbourhoods than for words with large neighbourhoods, the reverse of the usual effect. With the small set size, they found no difference between the two neighbourhood size conditions. One reason for this divergent result may be their stimuli, which differ substantially from those used by other researchers. First, as can be seen in Table 1, the difference in neighbourhood size between the small and large neighbourhood words was much smaller than in other studies, 9.89 versus 14.32 (using Coltheart's *N*). Indeed, the majority of the small neighbourhood words in Goh and Pisoni's studies would be classified as having large neighbourhoods by the criteria used in the other studies. Second, there was considerable overlap in neighbourhood size in the small and large conditions. For example, 40 of the 66 small neighbourhood words had between 9 and 16 neighbours as did 44 of the 66 large neighbourhood words. In contrast, there was no overlap in neighbourhood size in the stimuli used by Derraugh et al. (2017) or Guitard et al. (2018). Third, the small and large neighbourhood words differed in imageability (M. Coltheart, 1981). In short, it is possible that the effects observed by Goh and Pisoni were driven by idiosyncratic characteristics of their stimulus set, including the small number of words in each pool that did not overlap with words in the other pool in terms of neighbourhood size.

The second experiment that found a *reverse* neighbourhood size effect was reported by Greeno et al. (2022). In their Experiment 1, a large pool of words was used and a typical neighbourhood size effect was observed with better recall of words with more neighbours compared to words with fewer neighbours. In their Experiment 3, they used 12 small and 12 large neighbourhood words hand-picked from the larger pool. This time, a reverse neighbourhood size effect was observed with better serial recall of lists of words with a small neighbourhood compared to words with a large neighbourhood. The redintegration account, described earlier, was formulated to account for results from experiments that used small set sizes (e.g., Roodenrys et al., 2002) and as a result predicts that a standard neighbourhood size effect should be observed with a small stimulus set. The results of Experiment 3 of Greeno et al. (2022) are therefore the opposite of what the redintegration account predicts.

Excluding the Goh and Pisoni (2003) stimuli for the reasons noted earlier, Table 1 indicates that four different large stimulus sets have been used, all of which resulted in a typical large neighbourhood advantage (Derraugh et al., 2017, Exps. 1 and 2; Clarkson et al., 2017; and Guitard et al., 2018). In contrast, only three small stimulus sets have been used. Roodenrys et al. (2002) had two small sets, and these were also used by Allen and Hulme (2006), Jalbert, Neath, Bireta and Surprenant (2011), and Jalbert, Neath, and Surprenant (2011). The third is the small set used by Greeno et al. (2022). The two sets from Roodenrys et al. both produced a large neighbourhood advantage, whereas the one set from Greeno et al. produced a small neighbourhood advantage. As Greeno et al. (2022, p. 13) noted, one issue with small stimulus pools is that the effects produced may be due to idiosyncratic properties of the small number of words tested. One example of an idiosyncratic small stimulus set concerns the time-based word length effect. Baddeley et al. (1975, Exp. 4) used a small, fixed set of items that were equated for number of syllables and number of phonemes but differed in pronunciation time. They found better recall of the words that could be said faster. There are a large number of studies that have used the same stimuli and have obtained the same results; however, there exists no other set of stimuli that show the same results (for a review, see Neath et al., 2003). A second example concerns the much more strongly supported syllable-based word length effect whereby lists composed entirely of words with fewer syllables are better recalled than lists of words with more syllables (but see

Guitard et al., 2018, for a discussion of whether this effect is actually driven by the number of syllables). Cowan et al. (2003) found that lists that contained 3-short and 3-long words were recalled worse than lists of 6-short words but better than lists of 6-long words. This result is replicable when using their small stimulus set, but a different result obtains with other stimulus sets (Bireta et al., 2006; Hulme et al., 2004).

Greeno et al. (2022) used one small set of words for all subjects. An alternative method is to randomly sample 12-large and 12-small neighbourhood words from a larger pool for each subject. The result is that each person receives a small set of words, but on average, each person will have a different subset. The advantage of using randomly generated pools is that it minimises the probability of unwanted systematic differences: although by chance there may be such an idiosyncratic set of items for one subject, it is highly unlikely any other subject will have the same variation. This method was used by Neath and Surprenant (2019) to demonstrate that other semantic effects are observed in serial recall even when the same six items are shown on every trial.

Below we report the results of three registered experiments. Experiment 1 used the same large pool of stimuli used in Experiment 1 of Greeno et al. (2022) and the prediction is that the standard neighbourhood size effect will be observed, with better recall of large than small neighbourhood words. Experiment 2 used the same small pool of stimuli used in Experiment 3 of Greeno et al. and a reverse neighbourhood size effect is expected. These two experiments, then, should replicate the results reported by Greeno et al. Experiment 3 was identical to Experiment 2 except that the small pool of stimuli was randomly generated for each subject by drawing from the larger pool in Experiment 1. This methodology provides a much stronger test of the neighbourhood effect with small stimuli pools. In this experiment, if a reverse neighbourhood size effect is observed, then set size is critical. If a standard neighbourhood size effect is observed, then Greeno et al.'s results may be attributed to idiosyncratic properties of the stimulus set tested.

Experiment 1

Experiment 1 was based on Experiment 1 of Greeno et al. (2022) but with the following differences: (a) Greeno et al. manipulated presentation modality (auditory and visual), but because they reported no interactions between modality and neighbourhood size, we used only visual presentation. (b) Greeno et al. used spoken recall whereas we used typed recall given the use of an online subject sample. (c) Greeno et al. showed the visual items for 350 ms followed by a 650 ms blank screen. The rate is thus 1 item per second but the item is not visible for the entire time. The reason was to match the presentation rate of the visual items to that of the auditory items, each of which lasted 350 ms. Because we were not using auditory presentation, we used a fixed rate of 1 item per second and the item remained visible for the full 1 second. (d) Greeno et al. had 24 lists per condition; we had 12 lists per condition to keep the experiment short to minimise fatigue and the amount of typing required. (e) Greeno et al. had 30 subjects whereas we had 50 subjects. The reason for the increase in the sample size is because of different power analyses. Greeno et al. performed their power analysis on detecting an interaction between modality and neighbourhood size based on previous studies. In contrast, we based our power analysis on detecting a main effect of neighbourhood size based on the visual condition in their Experiment 3.

Ethics

The experiments were approved by the Virginia Tech institutional Review Board (IRB).

Subjects

Fifty volunteers from Prolific participated and were paid £8.50 per hour (pro-rated). The inclusion criteria for all experiments were (a) native speaker of English; (b) age between 19 and 39; (c) an approval rating of 90 or higher on pervious Prolific experiments; and (d) normal or corrected to normal vision. The sample size was determined by a power analysis using Superpower (Version 0.2.0, Lakens & Caldwell, 2021) with estimates based on the first 12 trials of the visual condition of Experiment 3 of Greeno et al. (2022). The mean age was 32.32 years (*SD* = 5.91, range 19–39) and 32 self-identified as female and 18 as male.

Stimuli

The stimuli were the same as in Experiment 1 of Greeno et al. (2022).

Design

The experiment was a 2 neighbourhood size (large vs. small) × 6 serial position within-subjects design.

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." Six 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. After the final word had been shown, a message appeared prompting

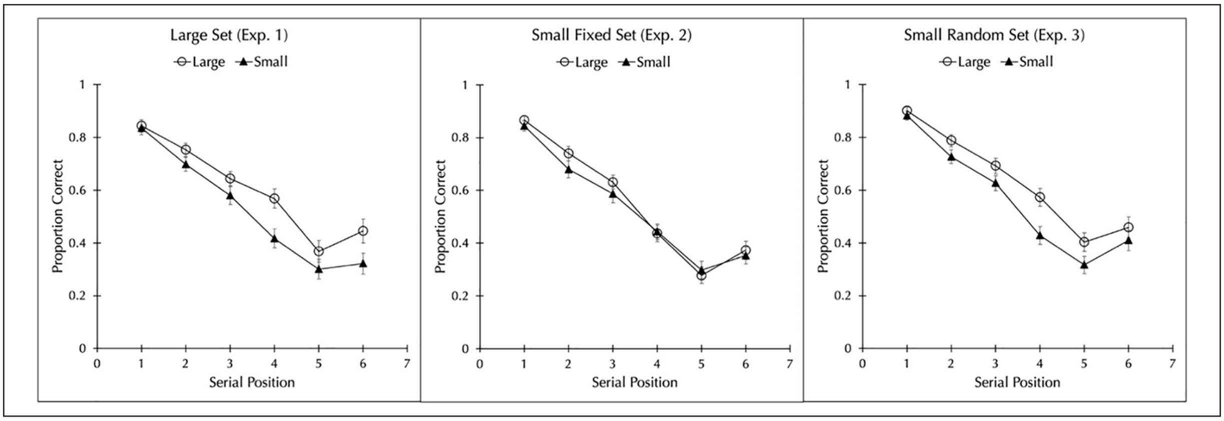

Figure 1. Proportion of large and small neighbourhood words correctly recalled in order as function of set size.
Error bars show standard error of the mean.

the subject to type in the first word. The instructions emphasised the importance of typing the first word first, the second word second, and so on. The instructions encouraged guessing but also indicated that the subject could click on a button labelled *skip*. There were 24 trials. Half the trials had large neighbourhood words and half had small neighbourhood words. The order of these trials was randomly determined for each subject. Subjects could take a break at any time by refraining from clicking on the "Start next trial" button. Note that because there were only 48 words, each pool was depleted after eight trials of that stimulus type. When this occurred, the pool was replenished. In effect, each word could appear a maximum of twice in the experiment whereas in the Greeno et al. study, each word could appear a maximum of three times.

Results and discussion

The data were analysed using both frequentist and Bayesian analysis of variance (ANOVA) using JASP (JASP Team, 2022). For the former, noninteger degrees of freedom indicate the Geenhouse-Geisser sphericity correction has been applied. For the latter, we report either a Bayes factor, BF_{10}, that indicates evidence for the alternative hypothesis or a Bayes factor, BF_{01}, that indicates evidence for the null 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).

The proportion of words correctly recalled was analysed by a 2 neighbourhood size (large vs. small) × 6 serial position repeated measures ANOVA.[3] There was a significant main effect neighbourhood size, $F(1,49)=44.648$, $MSE=0.021$, $\eta_p^2=0.477$, $p<.001$, $BF_{10}=86.92$, with better recall of large ($M=0.604$, $SD=0.174$) than small ($M=0.525$, $SD=0.177$) neighbourhood size words. These are similar to the values of 0.61 and 0.52 that were reported in the visual condition by Greeno et al. (2022).[4] There was also a significant effect of position, $F(2.40,117.41)=82.424$, $MSE=0.099$, $\eta_p^2=0.627$, $p<.001$, $BF_{10}=1.06\times10^{84}$, and a significant interaction, $F(3.98,195.05)=4.959$, $MSE=0.017$, $\eta_p^2=0.092$, $p<.001$, although the Bayes factor was only 2.65. As is apparent in the left panel of Figure 1, the effect of neighbourhood size is reduced at early positions, particularly at the first position, relative to later positions.

Error analysis. We analyse two types of item errors. An omission error is when a word from the list is not reported, whereas an intrusion error is when a word that was not in the list is reported. We also analyse order errors. An order error is when a word from the list is recalled in the wrong serial position; however, the raw number of order errors is misleading because it can vary with overall level of recall. For example, if three items from a six-item list are recalled, the maximum number of order errors is three, whereas if five of the six items are recalled, the maximum is five. To control for different numbers of opportunities for an order error, the number of order errors is divided by the number of items recalled in any position (Murdock, 1976; Saint-Aubin & Poirier, 1999).

There were more errors of each type in the small than in the large neighbourhood size condition. The proportion of intrusion errors was 0.171 ($SD=0.111$) for the large compared to 0.201 (0.128) for the small neighbourhood size conditions, $t(49)=3.493$, $p=.001$, $d=0.494$, $BF_{10}=27.94$. The proportion of omission errors was 0.128 ($SD=0.132$) for the large compared to 0.157 (0.153) for the small neighbourhood size conditions, $t(49)=3.583$, $p<.001$, $d=0.507$, $BF_{10}=35.59$. The proportion of conditionalized order errors was 0.114 ($SD=0.110$) for the large compared to 0.160 (0.140) for the small neighbourhood size conditions, $t(49)=3.746$, $p<.001$, $d=0.530$, $BF_{10}=55.73$. These order error rates are similar to those reported by

Greeno et al. (2022) for their visual condition, 0.12 and 0.16.[5]

Despite the many minor changes, Experiment 1 replicated the main results of Experiment 1 of Greeno et al. (2022). In particular, more large than small neighbourhood words were correctly recalled in order, and there were more conditionalized order errors for small than large neighbourhood words. In addition, Experiment 1 found more omission and intrusion errors for small than for large neighbourhood words.

Experiment 2

Experiment 2 was based on Experiment 3 of Greeno et al. (2022). The sole change from Experiment 1 is that instead of using a large pool of stimuli, Experiment 2 used a subset of 12 large and 12 small neighbourhood size words.

Subjects

Fifty different volunteers from Prolific participated. The mean age was 30.46 years ($SD=5.96$, range 20–39) and 39 self-identified as female and 11 as male.

Stimuli

The stimuli were the 12-large and 12-small neighbourhood size words used in Experiment 3 of Greeno et al. (2022).

Design

The design was the same as in Experiment 1.

Procedure

The procedure was identical to Experiment 1 except for the stimuli. As a small pool was used, each pool was depleted after two trials of that stimulus type. When this occurred, the pool was replenished. Each word appeared six times in the experiment.

Results and discussion

The proportion of words correctly recalled was analysed by a 2 neighbourhood size (large vs. small) × 6 serial position repeated measures ANOVA.[6] The main effect of neighbourhood size was not significant, $F(1, 49)=0.002$, $MSE=0.024$, $\eta_p^2=0.000$, $p=.965$, $\mathrm{BF}_{01}=11.11$, with equivalent recall of large ($M=0.554$, $SD=0.150$) and small ($M=0.555$, $SD=0.167$) neighbourhood size words. There was a significant effect of position, $F(3.05, 149.43)=165.367$, $MSE=0.049$, $\eta_p^2=0.771$, $p<.001$, $\mathrm{BF}_{10}=1.82\times10^{127}$, but no interaction, $F(4.28, 209.51)=1.123$, $MSE=0.016$, $\eta_p^2=0.022$, $p=.348$, $\mathrm{BF}_{01}=55.45$.

Error analysis. There were no differences in errors, although the Bayes factors did not offer much support for the null hypothesis. The proportion of intrusion errors was 0.163 ($SD=0.100$) for the large compared to 0.184 (0.130) for the small neighbourhood size conditions, $t(49)=1.842$, $p=.072$, $d=0.261$, $\mathrm{BF}_{01}=1.36$. The proportion of omission errors was 0.120 ($SD=0.134$) for the large compared to 0.129 (0.138) for the small neighbourhood size conditions, $t(49)=1.499$, $p=.140$, $d=0.212$, $\mathrm{BF}_{01}=2.29$. The proportion of conditionalized order errors was 0.189 ($SD=0.105$) for the large compared to 0.170 ($SD=0.126$) for the small neighbourhood size conditions, $t(49)=1.550$, $p=.128$, $d=0.219$, $\mathrm{BF}_{01}=2.12$.

Experiment 2 did not replicate Experiment 3 of Greeno et al. (2022). Greeno et al. reported better recall of small than large neighbourhood words, whereas there was no difference in this experiment; the Bayes factor indicated strong evidence in favour of the null hypothesis. However, the null results for order errors does replicate their finding. We discount differences in methodology as an explanation for the different outcome, given that our Experiment 1 replicated Experiment 1 of Greeno et al. We postpone further discussion until after presenting Experiment 3.

Experiment 3

Experiment 3 was identical to Experiment 2 except for the stimuli. Experiment 2 may be described as having a *fixed* small pool: All subjects saw the same small set of words, 12 of each type. In contrast, Experiment 3 may be described as having a *random* small pool: Once again, subjects saw only 12 words of each type, but the specific words were randomly drawn from the respective larger pools for each subject. In effect, each subject saw a different subset of the larger pools.

Subjects

Fifty different volunteers from Prolific participated. The mean age was 30.12 years ($SD=5.79$, range 19–39) and 39 self-identified as female and 11 as male.

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 Experiments 1 and 2.

Procedure

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

Results

A 2 neighbourhood size (large vs. small) × 6 serial position repeated measures ANOVA revealed a significant main effect of neighbourhood size, $F(1, 49)=18.132$, $MSE=0.042$, $\eta_p^2=0.270$, $p<.001$, $BF_{10}=36.37$, with better recall of large ($M=0.636$, $SD=0.170$) than small ($M=0.565$, $SD=0.159$) neighbourhood size words.[7] There was also a significant effect of position, $F(2.99, 146.35)=142.215$, $MSE=0.049$, $\eta_p^2=0.744$, $p<.001$, $BF_{10}=5.01 \times 10^{103}$, and a significant interaction, $F(4.35, 213.18)=3.245$, $MSE=0.016$, $\eta_p^2=0.062$, $p=.011$, $BF_{10}=4.88$. As is apparent in the right panel of Figure 1, the effect of neighbourhood size is reduced at early positions, particularly at the first position, relative to later positions.

Error analysis. There were numerically more errors of each type in the small than in the large neighbourhood size condition, but not all were statistically different. The proportion of intrusion errors was 0.138 ($SD=0.085$) for the large compared to 0.166 (0.114) for the small neighbourhood size conditions, $t(49)=2.185$, $p=.034$, $d=0.309$, but $BF_{10}=1.35$. The proportion of omission errors was 0.101 ($SD=0.122$) for the large compared to 0.130 (0.139) for the small neighbourhood size conditions, $t(49)=2.599$, $p=.012$, $d=0.367$, $BF_{10}=3.14$. The proportion of conditionalized order errors was 0.146 ($SD=0.119$) for the large compared to 0.179 (0.120) for the small neighbourhood size conditions, $t(49)=2.556$, $p=.014$, $d=0.361$, $BF_{10}=2.86$.

The results of Experiment 2 and 3 are very different: The small fixed pool in Experiment 2 resulted in no difference between small and large neighbourhood words whereas the small random pool in Experiment 3 resulted in the usual large neighbourhood advantage.

General discussion

The neighbourhood size effect has played an important role in the development of memory models and helps to further our understanding of basic human processing. For instance, the neighbourhood size effect is directly linked to Roodenrys et al. (2002) account of the complex interaction between the presented information and the redintegration of that information, which is widely incorporated in key hypotheses and simulation models of human memory. However, until the work of Greeno et al. (2022), the demonstrations of a neighbourhood size effect with small stimulus pool, on which the redintegration account of Roodenrys et al. (2002) was built, was restricted to the original stimuli. The results of Greeno et al., showing a detrimental effect of neighbourhood size, pose major challenges for memory models. The three registered experiments reported here demonstrate that the particularity of the Greeno et al. small pool was the key factor driving these results.

More exactly, the results of the experiments overall are clear and can be summarised as follows. Experiment 1, using a large pool, and Experiment 3, using a small pool, both found typical neighbourhood size effects with a better serial recall of words with large neighbourhoods than words with small neighbourhoods, using both frequentist and Bayesian statistics. In addition, Experiment 2, which used the same small pool as used by Greeno et al. (Experiment 3), found no effect of neighbourhood size. The difference between Experiments 2 and 3 is that in the former, all subjects received the same small pool whereas, in the latter, the small pool was determined randomly for each subject. Experiment 3 provides a further demonstration of a neighbourhood size effect with a small set size.

Additional evidence consistent with our results comes from neighbourhood size manipulations with nonwords. In this case, a nonword such as *rin* has an orthographic neighbourhood of words that includes *bin, ran*, and *rip*. Roodenrys and Hinton (2002) used a pool of 20 large and 20 small neighbourhood nonwords and found a typical beneficial effect of having a large neighbourhood. Similarly, Jalbert, Neath and Surprenant (2011) used a small pool of 12 in their Experiment 2 and a different small pool in their Experiment 3, observing typical neighbourhood size effects in each experiment.

There is one methodological difference between Experiment 2 here and Experiment 3 of Greeno et al. (2022), which is that Experiment 2 used typed recall instead of spoken. It seems unlikely that this is an important factor and should not challenge the interpretation of the results for a number of reasons. First, there was no difference between the results of Experiment 1 here, which used typed recall, and Experiment 1 of Greeno et al., which used spoken recall. Both of these experiments used a large pool, and replicate previous work using spoken (e.g., Roodenrys et al., 2002) and typed recall (e.g., Guitard et al., 2018), and there is no a priori reason why set size might interact with response modality. Second, Experiment 3 (here) used a small pool, but each set was randomly generated for each subject. This resulted in the same pattern of results observed in Experiment 1 (here), suggesting that the neighbourhood effect is present for small and large set sizes in this response mode. Third, there are many demonstrations in the literature that other effects in serial recall occur with either written or typed responses. For example, Saint-Aubin et al. (2020) found a large effect of word length in both spoken and typed recall, strong phonological similarity effects are found in written (e.g., V. Coltheart, 1993) and typed recall (e.g., Roodenrys et al., 2022), and Beaudry et al. (2018) found similar effects in spoken and written recall of imageability and word frequency. Ultimately, this is an empirical question, but based on the above, our prediction is that if Experiment 3 (here) were run with spoken recall, a neighbourhood size effect would still obtain.

Although it is apparent that the unusual set of stimuli is the small pool used by Greeno et al. (2022, Experiment 3), it is not entirely clear how they differ from other sets. One possibility is the distribution of phonemes within the words. Greeno et al. (2022) go to some lengths to justify the selection of the 12 items for each condition in their Experiment 3, including ensuring the two sets of stimuli have "the same dispersion of phonologically similar onset consonants" (p.9). However, an examination of their stimuli reveals they are equated on the initial letter of the word but not the initial phoneme. In the large neighbourhood set four words start with a hard "c," whereas in the small neighbourhood set only one has a hard "c" pronunciation. Such differences in phonological overlap within sets may be important. Similarly, analysis of phonological similarity with Psimetrica (Phonological Similarity Metric Analysis, Mueller et al., 2003) indicates the mean pairwise similarity of onset phonemes was greater in the large neighbourhood set.

One other point to note is that the current research focussed on serial recall and not serial recognition. Experiment 2 of Greeno et al. (2022) found that using the same large set of stimuli as in their Experiment 1, there was no effect of neighbourhood size in serial recognition. To our knowledge, there are no other published studies that examine whether neighbourhood size effects occur in serial recognition, but even if the null effect of Greeno et al. is replicated, the lack of effect is not particularly surprising or noteworthy. The reason is that a number of lexico-semantic effects that are robust with serial recall are known to be absent with serial recognition. Of particular interest, Chubala et al. (2019) used a procedure in which subjects did not know whether they would receive a serial recall or a serial recognition test until after the list had been presented. For the serial recall test, they found the usual effects of frequency and semantic relatedness effects, but no such effects were found for the serial recognition test. Thus, the absence of a particular manipulation in serial recognition is not necessarily informative of the effect of that manipulation in serial recall.

The results of Experiment 3 combined with those already extant in the literature reinforce the notion that one should be wary of any results from a small stimulus pool even when great care is taken to select stimuli. If the set of stimuli is small, the danger persists that some idiosyncratic factor differs enough to influence performance and lead to erroneous conclusions and overly complicated theoretical speculations. This is not to say that all research using small stimulus pools is suspect. Rather, the point is that small pools are inherently more likely to be influenced by idiosyncratic properties than a large pool, and therefore, one must exercise caution when relying on the results of only one set of items. Greater confidence in an outcome can be built by conducting many experiments, each with a different set. Alternatively, a more efficient method is to create a large pool and then randomly select a small subset for each subject. By chance, a particular subject might receive a set of items with unusual properties, but it is highly unlikely that any other subject will receive a similar set. Put another way, randomly selecting the items from a larger pool minimises the chances of some unwanted variation being confounded with the experimental factor.

Conclusion

The general pattern of results is consistent with the redintegration framework, which in turn is consistent with results from the speech production literature. The redintegration framework can be incorporated into a number of different theoretical accounts, including simulation models. Because the experiments reported here were designed to evaluate only set size, the data do not inform on which of those accounts provides a more complete explanation. In summary, this study strengthens the empirical support for a facilitative effect of neighbourhood size on serial recall, but it also highlights the danger associated with stimuli propriety. Methods such as those used here provide more reliable findings and a safeguard against unwarranted challenges to theories and the addition of more complicated mechanisms to models.

Authors' note

Order of authorship is alphabetical.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Dominic Guitard https://orcid.org/0000-0002-4658-3585
Steven Roodenrys https://orcid.org/0000-0002-3065-1766

Data accessibility statement

The data, stimuli, and registration are available from the Open Science Foundation, https://doi.org/10.17605/OSF.IO/PKJ58.

Notes

1. Number of neighbours is Phono_N from Balota et al. (2007).
2. They called words with a large neighbourhood "hard" and words with a small neighbourhood "easy" because they also differed on neighbourhood frequency and would thus be either harder or easier to perceive in an auditory identification task.

3. The typed responses were spell checked. Out of 7200 responses, 65 large and 95 small neighbourhood words were flagged. Of these, 40 large and 45 small neighbourhood words changed from incorrect to correct. Because correcting for spelling can be subjective, and because there were no substantive differences between the conclusions drawn from the raw and the spell-checked data, only analyses of the former are reported.
4. These values were estimated from Figure 2 of Greeno et al. (2022).
5. These values were estimated from Figure 3 of Greeno et al. (2022).
6. There were 18 large and 146 small neighbourhood words misspelled out of 7,200 responses. For the small neighbourhood words, THIEF was misspelled as THEIF 72 times and BEIGE was misspelled as BIEGE 21 times. Correcting just these two spelling errors changed 74 responses from incorrect to correct. Although correcting spelling can be subjective, we thought these two corrections were unambiguous and therefore the analyses reported include these two corrections. The supplementary analyses at the OSF reports analyses for the raw data and for when all misspellings are corrected. The only substantive change is whether the difference in intrusions is significant because a misspelled word counts as an intrusion.
7. The typed responses were spell checked. Out of 7,200 responses, 34 large and 68 small neighbourhood size words were flagged. Of these, 24 large and 42 small neighbourhood size words changed from incorrect to correct. The sole substantive difference is that the difference in the proportion of intrusion errors was no longer significant for the spell-checked data, 0.130 ($SD = 0.081$) for large and 0.150 ($SD = 0.112$) for small, $t(49) = 1.757$, $p = .085$, $d = 0.248$.

References

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

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. https://doi.org/10.1016/S0022-5371(75)80045-4

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*(3), 445–459. https://doi.org/10.3758/BF03193014

Beaudry, O., Saint-Aubin, J., Guérard, K., & Pâquet, M. (2018). Are lexical factors immune to response modality in backward recall? The effects of imageability and word frequency. *Canadian Journal of Experimental Psychology*, *72*(2), 105–116. https://doi.org/10.1037/cep0000126

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1037/cep0000158

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. https://doi.org/10.1080/09658211.2016.1179330

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33A*(4), 497–505. https://doi.org/10.1080/14640748108400805

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). Lawrence Erlbaum.

Coltheart, V. (1993). Effects of phonological similarity and concurrent irrelevant articulation on short-term-memory recall of repeated and novel word lists. *Memory & Cognition*, *21*(4), 539–545. https://doi.org/10.3758/bf03197185

Cowan, N., Baddeley, A. D., Elliott, E. M., & Norris, J. (2003). List composition and the word length effect in immediate recall: A comparison of localist and globalist assumptions. *Psychonomic Bulletin & Review*, *10*(1), 74–79. https://doi.org/10.3758/BF03196469

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. https://doi.org/10.1037/cep0000098

Goh, W. D., & Pisoni, D. B. (2003). Effects of lexical competition on immediate memory span for spoken words. *Quarterly Journal of Experimental Psychology*, *56A*(6), 929–954. https://doi.org/10.1080/02724980244000710

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. https://doi.org/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. https://doi.org/10.1037/xlm0000551

Hulme, C., Suprenant, A. M., Bireta, T. J., Stuart, G., & Neath, I. (2004). Abolishing the word-length effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 98–106. https://doi.org/10.1037/0278-7393.30.1.98

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

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

JASP Team. (2022). *JASP (Version 0.16.2)* [Computer software]. https://jasp-stats.org/

Lakens, D., & Caldwell, A. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science, 4*(1), 251524592095150. https://doi.org/10.1177/2515245920951503.

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

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Majerus, S. (2013). Language repetition and short-term memory: An integrative framework. *Frontiers in Human Neuroscience*, *7*, Article 357. https://doi.org/10.3389/fnhum.2013.00357

Martin, R. C., Lesch, M. F., & Bartha, M. C. (1999). Independence of input and output phonology in word processing and short-term memory. *Journal of Memory and Language*, *41*(1), 3–29. https://doi.org/10.1006/jmla.1999.2637

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*(6), 1353–1380. https://doi.org/10.1037/0278-7393.29.6.1353

Murdock, B. B. (1976). Item and order information in short- term serial memory. *Journal of Experimental Psychology: General*, *105*(2), 191–216. https://doi.org/10.1037/0096-3445.105.2.191

Murray, D. J. (1967). The role of speech responses in short-term memory. *Canadian Journal of Psychology*, *21*(3), 263–276. https://doi.org/10.1037/h0082978

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

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

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. https://doi.org/10.3758/BF03196502

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. https://doi.org/10.3758/s13421-018-0883-8

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). Psychology Press.

Roodenrys, S., Guitard, D., Miller, L. M., Saint-Aubin, J., & Barron, J. M. (2022). Phonological similarity in the serial recall task hinders item recall, not just order. *British Journal of Psychology*, *113*(4), 1100–1120. https://doi.org/10.1111/bjop.12575

Roodenrys, S., & Hinton, M. (2002). Sublexical or lexical effects on serial recall of nonwords? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 29–33. https://doi.org/10.1037/0278-7393.28.1.29

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. https://doi.org/10.1037/0278-7393.28.6.1019

Saint-Aubin, J., Beaudry, O., Guitard, D., Pâquet, M., & Guérard, K. (2020). The word length effect in backward recall: The role of response modality. *Memory*, *28*(5), 692–700. https://doi.org/10.1080/09658211.2020.1762896

Saint-Aubin, J., & Poirier, M. (1999). The influence of long-term memory factors on immediate serial recall: An item and order analysis. *International Journal of Psychology*, *34*, 347–352. https://doi.org/10.1080/002075999399675

Schwering, S. C., & MacDonald, M. C. (2020). Verbal working memory as emergent from language comprehension and production. *Frontiers in Human Neuroscience*, 14, Article 68. https://doi.org/10.3389/fnhum.2020.00068

Serra, M., & Nairne, J. S. (1993). Design controversies and the generation effect: Support for an item-order hypothesis. *Memory & Cognition*, *21*(1), 34–40. https://doi.org/10.3758/BF03211162

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 735–747. https://doi.org/10.1037/0278-7393.28.4.735

Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition*, *31*(4), 491–504. https://doi.org/10.3758/BF03196091

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. https://doi.org/10.1177/1745691611406923

Yarkoni, T., Balota, D. A., & Yap, M. J. (2008). Beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971