Memory & Cognition (2021) 49:939–954
https://doi.org/10.3758/s13421-021-01137-6

Re-assessing age of acquisition effects in recognition, free recall, and serial recall

Molly B. Macmillan[1] · Ian Neath[1] · Aimeé M. Surprenant[1]

Accepted: 31 December 2020 / Published online: 8 February 2021

Abstract

Age of acquisition (AoA) refers to the age at which a person learns a word. Research has converged on the conclusion that early AoA words are processed more efficiently than late AoA words on a number of perceptual and reading tasks. However, only a few studies have investigated whether AoA affects memory on recognition, serial recall, and free recall tests, and the results are equivocal. We took advantage of the recent increase in the number of high-quality norms and databases to construct a pool of early and late AoA words that were equated on numerous other dimensions. There was a late AoA advantage in recognition using both pure (Experiment 1) and mixed (Experiment 2) lists, no effect of AoA on serial recall of either pure (Experiment 3) or mixed (Experiment 4) lists, and no effect of AoA on free recall of either pure (Experiment 5) or mixed lists (Experiment 6). We conclude that AoA does reliably affect memory on some memory tasks (recognition), but not others (serial recall, free recall), and that no current account of AoA can explain the findings.

Keywords Age of acquisition · Memory · Recognition · Serial recall · Free recall

Age of acquisition (AoA) refers to the age at which a person learns a word and results from a number of studies have converged on the conclusion that early acquired words tend to be processed more efficiently than late acquired words on a number of psycholinguistic tasks such as lexical decision and word naming (for reviews, see Gilhooly & Watson, 1981; Johnston & Barry, 2006; Juhasz, Yap, Raoul, & Kaye, 2019). However, it is not clear whether AoA affects memory on standard tests such as recognition, serial recall, or free recall, because the extant findings are contradictory. The purpose of the current set of experiments is to re-assess whether AoA affects these memory tasks.

One reason that AoA came to prominence in a number of research areas is that the results seemed to challenge current accounts and suggested new explanations. For example, Carroll and White (1973) demonstrated that AoA affected object naming latencies above and beyond word frequency. One implication of these findings is the suggestion that items may be stored chronologically in long-term memory rather than in terms of frequency. As a second example, Brown and Watson (1987) suggested that the phonological forms of early acquired words were represented as a single unit, while later acquired words were represented across multiple units. Under this model, an additional assembly step is required to produce the phonological representation of late-acquired but not early-acquired words, resulting in the observed retrieval discrepancies between the two word types.

Two later publications argued that the effects commonly ascribed to word frequency in picture naming (Morrison, Ellis, & Quinlan, 1992) and word reading (Morrison & Ellis, 1995) tasks should be attributed to confounded AoA effects (Johnston & Barry, 2006). This premise challenged lexical models that incorporated word frequency as a central explanatory tenet. Although existing connectionist models could easily accommodate frequency effects, AoA effects posed a theoretical problem (Ellis & Lambdon Ralph, 2000). As a result, newer models evolved, incorporating order-of-learning and network plasticity into the existing frameworks. The introduction of these newer models became a major impetus for the study of AoA effects across a range of processing tasks. AoA publications have provided insight into the relationship between orthographic, phonologic, and semantic representations and suggested a role for age of acquisition in the organization of semantic networks (Juhasz, 2005).

One consequence of the theoretical debate was a growing literature examining the effects of AoA on a number of

✉ Ian Neath
ineath@mun.ca

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

cognitive tasks. In both picture-naming (e.g., Meschyan & Hernandez, 2002; Morrison et al., 1992; Pérez, 2007) and word-naming paradigms (e.g., Brysbaert & Cortese, 2011; Cortese & Khanna, 2007), response latencies are faster and accuracy is higher for early-acquired than for late-acquired words. These effects remain even after accounting for word frequency and other possible confounding variables such as word length and imageability. In word-pronunciation tasks, people repeated earlier acquired words more rapidly than later acquired words (Roodenrys, Hulme, Alban, Ellis, & Brown, 1994). Although this result might suggest an articulation rather than a processing advantage in word-naming tasks, the AoA effect disappears when a delay is introduced between the presentation of the word and when participants are asked to report it (Gerhand & Barry, 1998). The lack of a significant AoA effect with a delay suggests that earlier acquired words are not easier to articulate, but are in fact processed more rapidly.

Additionally, performance on lexical decision tasks suggests that early-acquired words are processed more rapidly than words acquired later in life (e.g., Brysbaert & Cortese, 2011; Cortese & Khanna, 2007; Juhasz et al., 2019). This effect has been found in studies using rated estimates, objective measures, and frequency trajectories as proxies for AoA. The effect has also been demonstrated in multiple languages and remains significant even after controlling for objective and subjective frequency, word length, neighbourhood size, and other psycholinguistic variables (Johnston & Barry, 2006). Lastly, evidence from eye-fixation studies converges to support an AoA effect in lexical processing (Juhasz & Rayner, 2003, 2006). When participants were asked to read complete sentences, the single-fixation duration and total gaze duration on earlier acquired words were significantly shorter than for late-acquired words.

Whereas the results from lexical processing studies are quite clear, those from memory studies are less so. For example, Gilhooly and Gilhooly (1979) used regression analyses and found no effect of AoA on either recognition (Experiment 4) or free recall (Experiment 3) using mixed lists (lists that contain both early and late AoA words). Similarly, Rubin (1980) used correlational analyses and found no effect of AoA on free recall, again using mixed lists. Coltheart and Winograd (1986) created word pools that differed in AoA, but were equated for frequency, imagery, and length. They found no effect of AoA on either free recall with pure lists (lists that contain only early or only late AoA words) or on recognition using mixed lists. Dewhurst, Hitch, and Barry (1998) also found no effect of AoA on free recall when pure lists were used. Roodenrys et al. (1994) used a factorial manipulation of frequency and AoA; with pure lists, they found that frequency affected memory span but AoA did not.

In contrast, a number of studies have concluded that AoA does affect memory. Morris (1981) used a regression analysis and observed an effect of AoA on free recall with mixed lists, with late-acquired words being recalled better than early-acquired words. He attributed the difference in results compared with those of Gilhooly and Gilhooly (1979) as being due to when frequency was entered into the regression. Dewhurst et al. (1998) also found a late-word advantage in free recall when mixed lists were used, but Almond and Morrison (2014) found an early-word advantage on free recall when pure lists were used. For recognition using mixed lists, Dewhurst et al. found that performance was better for late-acquired than early-acquired words, but only on remember judgements and not on know judgments. Cortese, Khanna, and Hacker (2010) and Cortese, McCarty, and Schock (2015) used regression analyses on recognition of approximately 2,500 one syllable and two syllable words, respectively. In both studies, AoA was positively correlated with recognition performance, reflecting a late AoA advantage.

One long-standing problem, which may be contributing to the contradictory results summarized above, is that AoA is correlated with many other variables. For example, of the approximately 11,600 words that occur in the test-based AoA norms of Brysbaert and Biemiller (2017), the concreteness norms of Brysbaert, Warriner, and Kuperman (2014), the frequency and contextual diversity norms of Brysbaert and New (2009), and the various measures available in the E-Lexicon project (Balota et al., 2007), AoA correlates 0.30 with number of letters, 0.34 with number of phonemes, −0.34 with concreteness, −0.54 with frequency, −0.33 with contextual diversity, and 0.30 with both orthographic and phonological Levenshtein distance. It is no wonder that Gilhooly and Gilhooly (1979) concluded that such correlations make it "impossible to carry out factorial experiments in which confounded variables are balanced out or experimentally manipulated, while still retaining a reasonable number of words per condition" (p. 215). Thirty years later, Cortese et al. (2010) noted the difficulty in selecting "items that vary only by one dimension (e.g., AoA, but not length, imageability, frequency, etc.)" (p. 598). However, the recently developed databases now make it possible to create a set of stimuli that differ in AoA but that are equated on multiple other dimensions known to affect memory, including length, concreteness, frequency, contextual diversity, and orthographic and phonological Levenshtein distance. In addition, for the few studies that provide their stimuli in the report, these same databases can be used to reevaluate whether those studies to determine if confounds could be affecting the results.

Table 1 summarizes studies that have examine the effect of AoA on recognition, serial recall, or free recall and which also reported the stimuli. In this table, the AoA values come from test-based norms (Brysbaert & Biemiller, 2017), and the value indicates the grade in school when the word is typically learned. For all but one study, early AoA words were learned around Grades 2–3 and late AoA words were learned around

Table 1 Mean age of acquisition (AoA) values, according to the Brysbaert and Biemiller (2017) test-based norms, for published memory studies that provided the stimuli and for the stimuli in Experiments 1–6 of this paper

	Early AoA			Late AoA			AoA effect on memory test		
	M	*SD*	Range	*M*	*SD*	Range	Recognition	Serial recall	Free recall
Coltheart and Winograd (1986) Experiment 2	2.53	0.90	2–4	5.59	1.69	2–8	None (mixed list)		
Dewhurst et al. (1998) Experiment 1	2.88	1.83	2–8	6.00	3.09	2–12	Late advantage (mixed list)		
Dewhurst et al. (1998) Experiment 2	2.53	0.90	2–4	5.59	1.69	2–8	Late advantage (mixed list)		
Roodenrys et al. (1994) Experiment 1	2.75	1.04	2–4	5.75	1.67	4–8		None (pure list)	
Roodenrys et al. (1994) Experiment 3	2.29	0.73	2–4	5.29	1.86	4–8		None (pure list)	
Almond and Morrison (2014)	2.27	1.55	0–10	3.33	1.77	2–8			Early advantage (pure list)
Coltheart and Winograd (1986) Experiment 1	2.53	0.90	2–4	5.59	1.69	2–8			None (pure list)
Dewhurst et al. (1998) Experiment 3	2.80	1.74	2–8	6.15	3.04	2–12			None (pure list)
Dewhurst et al. (1998) Experiment 3									Late advantage (mixed list)
Experiment 1	3.20	0.98	2–4	10.77	2.02	8–14	Late advantage (pure list)		
Experiment 2	3.15	1.00	2–4	10.03	1.78	8–14	Late advantage (mixed list)		
Experiment 3								None (pure list)	
Experiment 4								None (mixed list)	
Experiment 5									None (pure list)
Experiment 6									None (mixed list)

Note. The AoA value indicates the grade in school when the word is typically learned

Grades 5–6. For the other study, that of Almond and Morrison (2014), the early words were learned in Grade 2 and the late words were learned in Grade 3. For all studies, there is some overlap between the early and late AoA words in terms of AoA.

Recognition Table 1 includes three studies of recognition, two of which found a late-word advantage (Dewhurst et al., 1998, Experiments 1 and 2) and one which found no effect of AoA (Coltheart & Winograd, 1986, Experiment 2). All three studies used mixed lists, in which both early and late AoA words appeared. Oddly, Experiment 2 of Dewhurst et al. (1998) used the same stimuli as Experiment 2 of Coltheart and Winograd (1986), but the results are different. One possible reason is that Dewhurst et al. analyzed their recognition data in terms of *d′*, whereas Coltheart and Winograd reported only proportion correct. For both sets of stimuli, however, the ranges of the early and late AoA words overlap (Grades 2–4 vs. Grades 2–8). Moreover, the early and late AoA words also differ in frequency, as measured by $SUBTLEX_{US}$ (Brysbaert & New, 2009) and $SUBTLEX_{UK}$ (van Heuven, Mandera, Keuleers, & Brysbaert, 2014), as well as a number of other dimensions. It is therefore possible that the effects ascribed to AoA are due to word frequency or to a combination of factors.

Serial recall Table 1 includes two studies that used immediate serial recall, neither of which found an effect of AoA (Roodenrys et al., 1994, Experiments 1 and 3). They used a memory-span task in which the first four lists had three items. All subjects recalled all three words in order on each of these lists. Then, four more lists were presented that were longer by one word. This continued until the subject made errors on at least three of the lists at a given length. The measure they analyzed is the longest list length with no errors on any of the four lists plus 0.25 for each longer list recalled correctly. The two experiments used different stimuli, but the stimuli used in Experiment 3 did not differ on any dimension we assessed that is likely to affect serial recall other than AoA.[1]

[1] The early and late AoA words did differ in valence, $t(26) = 2.45$, $p = .02$, with the early words being more positive ($M = 6.04$, $SD = 1.24$) than the later words ($M = 4.94$, $SD = 1.14$), using the Warriner, Kuperman, and Brysbaert (2013) norms. However, Bireta, Guitard, Neath, and Surprenant (2021) have argued that valence does not affect immediate serial recall.

Free recall Table 1 includes four studies that used free recall. Two studies using pure lists found no effect of AoA, Coltheart and Winograd (1986, Experiment 1) and Dewhurst et al. (1998, Experiment 3). Dewhurst et al. manipulated both frequency and AoA, but this resulted in a number of differences between the early and late AoA words. For example, in the high-frequency group, the late AoA words had higher frequency than the early AoA words. A third study that also used pure lists, Almond and Morrison (2014), found an advantage for early AoA words. As noted above, the Almond and Morrison stimuli differ substantially from the other studies in the range of AoA assessed; for example, many of their late AoA words would fall into the early AoA category of other researchers. Moreover, the early AoA words differ from the late AoA words in frequency (both $SUBTLEX_{US}$ and $SUBTLEX_{UK}$) and come close to being significantly shorter as measured by the number of syllables and phonemes ($p = .08$ and .09, respectively, according to the E-Lexicon database; Balota et al., 2007). The only study that used a mixed list in free recall, Dewhurst et al. (Experiment 3), found a late-word advantage.

Given these conflicting findings, we postpone discussion of theoretical considerations of whether AoA should be expected to affect recognition, serial recall, or free recall until after we report the results of our experiments

.

Overview of experiments

The purpose of the following experiments was to take advantage of databases not available to previous researchers and construct a set of early and late AoA words, as defined by a test-based measure, that (1) had no overlap in AoA between the early and late words and (2) had a larger difference in mean AoA than most previous studies. In addition, the early and late AoA pools were equated on numerous other dimensions known to affect memory performance. Two such stimulus sets were created. The first, larger, pool was used in Experiment 1 for testing pure lists in recognition and the second, smaller, pool was used in all the other experiments. The reason for using two pools was that the serial and free recall tests require typed responses, and therefore the length of the words was kept short. This pool yielded too few words for a pure list recognition experiment, however; to create a larger pool, longer words were permitted because typing is not required. Both pools were created the same way, the only difference being the smaller pool was restricted to words of one or two syllables. The initial pool consisted of all words with an AoA of 4 or less (early AoA) or 8 or more (late AoA) using the Brysbaert and Biemiller (2017) test-based norms. These pools were then reduced in size until the words were equated on the dimensions shown in the Appendix. Where possible, multiple measures of a dimension (e.g., frequency) were used to provide converging evidence that the early and late AoA words did not differ. For semantic relatedness, we used WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990), an online lexical database in which words are organized into synonym sets that represent the underlying lexical concept. Different senses of a word (e.g., *racket* as in tennis and *racket* as in an unpleasant noise) are represented in different synonym sets (known as synsets). Pedersen, Patwardhan, and Michelizzi (2004) calculated a number of measures of similarity between synsets, and the one used here is the number of steps in the shortest path between two words. For words with more than one sense, the value used was the lowest from examining all senses. Low values indicate a closer relation than did high values. For each set, a path length was obtained for all possible pairs, and then mean path length was computed.

For each test—recognition, serial recall, and free recall—there is one experiment with pure lists and one with mixed lists. The reason is that some variables that correlate with AoA, such as frequency, interact with list type. For example, in recognition, low-frequency words are recognized more accurately than high-frequency words in both pure (Gorman, 1961) and mixed (Schulman, 1967) lists. In serial recall, high-frequency words are better recalled than low-frequency words in pure lists (Roodenrys et al., 1994), but in mixed lists there is no effect of frequency (Hulme, Stuart, Brown, & Morin, 2003). In free recall, high-frequency words are better recalled than low-frequency words in pure lists (Deese, 1960; Peters, 1936), but in mixed lists, all three possible patterns have been observed, but the most common is low frequency being better recalled than high frequency (DeLosh & McDaniel, 1996; May & Tryk, 1970).

Experiment 1

The purpose of Experiment 1 was to assess whether AoA affects recognition performance when pure lists are used. We could find no published studies that examined this. We therefore took the design of Neath, Hockley, and Ensor (2021), who found effects of contextual diversity, frequency, and concreteness in recognition of mixed lists, but changed the design such that subjects completed two study–test cycles. For half the subjects, the first study–test cycle used early AoA words and the second used late AoA words, and for the other half of the subjects, the order was reversed.

Method

Subjects Forty-four volunteers from ProlificAC were paid £8.00 per hour (prorated) for their participation. The inclusion criteria for all studies were (1) native speaker of English; (2) age between 19 and 39 years; and (3) at least a 90% approval

rating on prior participation. The mean age was 28.00 years (*SD* = 5.38, range: 20–39 years); 29 subjects self-identified as female and 15 self-identified as male. The sample size was determined by a power analysis. A sample of 44 has power of 0.90 to detect an effect size of d = 0.5 (Faul, Erdfelder, Buchner, & Lang, 2009).

Stimuli The stimuli were 139 early and 139 late AoA words that were equated on a number of other dimensions (see Table 4 in the Appendix for details).

Procedure After indicating consent, the subjects were reminded of the instructions. They saw a list of 64 words, either all early or all late AoA. For each subject, the words were selected randomly from the appropriate pool. Each word appeared for 1 s in the middle of the screen in 28-point Helvetica font. Subjects were asked to read each word silently for an upcoming recognition test. After all 64 words were shown, there was a short distractor task. An uppercase letter (either *B, F, G, J,* or *R*) was shown rotated either 90°, 180°, or 270° and as either a normal or a mirror image. The task was to indicate if the letter was normal or mirror reversed. There was 24 of these trials. Following this, they saw a list of 128 words, half of which were seen in the study phase and half of which were new. For each word, subjects were asked to click on a button from 1 to 6 to indicate their confidence in their response. The display informed the subject that responses 1–3 indicated the word had been shown in Part 1 (an *old* response), whereas the responses 4–6 indicated the word had not been shown (a *new* response). Within these ranges, 1 and 6 meant "very confident"; 2 and 5 meant "confident"; and 3 and 4 meant "not very confident." Following this, subjects were encouraged to take a short break. They then repeated the study–test sequence (study list of 64 words, 24 letter task trials, 128-word recognition test) using the other type of words.

Results and discussion

Both frequentist and Bayesian analyses were conducted using JASP (JASP Team, 2019). For the latter, a Bayes factor (BF_{10}) between 3 and 20 indicates positive evidence for the alternate hypothesis (and therefore evidence against the null hypothesis); BF_{10} between 20 and 150 indicates strong evidence, and BF_{10} greater than 150 indicates very strong evidence (Kass & Raftery, 1995). BF_{01} indicates evidence for the null hypothesis and is interpreted on the same scale. Default priors were used.

The confidence ratings were used to construct hit and false-alarm rates and also to construct *z*-ROC curves for each subject for each condition, from which d_a was computed (Macmillan & Creelman, 2005). Table 2 shows the means, standard deviations, effect sizes, and Bayes factors for various performance measures.

Table 2 Descriptive statistics and performance measures for Experiment 1

Measure	Early AoA		Late AoA		Cohen's *d*	BF_{10}
	M	*SD*	*M*	*SD*		
Hit	0.673	0.135	0.688	0.127	0.161	0.278
FA	0.313	0.150	0.269	0.139	0.404	3.808
d'	1.031	0.577	1.214	0.561	0.479	11.993
C	0.029	0.348	0.078	0.340	0.203	0.379
d_a	1.095	0.540	1.241	0.560	0.435	6.069
Slope	0.749	0.183	0.837	0.266	0.313	1.147

Note. Slope indicates the slope of the *z*-ROC function

There was a significant effect of AoA: Mean d_a was higher for late AoA words than for early AoA words, $t(43) = 2.887$, $p = .006$, $BF_{10} = 6.069$. Thirty subjects had higher d_a for late words compared with 14 who had higher d_a for early words, which is significant by a two-tailed sign test, $p = .023$. There was no evidence of a mirror effect: Although the false-alarm rate was higher for early than for late AoA words, $t(43) = 2.680$, $p = .010$, $BF_{10} = 3.808$, there was no difference in the hit rate, $t(43) = 1.068$, $p = .292$, $BF_{01} = 3.597$. We postpone further discussion of these results until after Experiment 2.

Experiment 2

Experiment 1 found a late AoA advantage in recognition when pure lists were used. The purpose of Experiment 2 was to assess whether AoA affects recognition performance in the same way when mixed lists were used.

Method

Subjects Forty-four different volunteers from ProlificAC were paid £8.00 per hour (prorated) for their participation. The mean age was 29.32 years (*SD* = 5.99, range: 19–39 years); 28 subjects self-identified as female, and 16 self-identified as male.

Stimuli The stimuli were 68 early and 68 late AoA words that were equated on a number of other dimensions (see Table 5 in the Appendix for details).

Procedure The procedure was similar to that of Experiment 1, except that there was only one study–test cycle and the study list contained 32 early and 32 late AoA words. For each subject, the words were selected randomly from the main pool and were shown in random order. At test, there were 128

trials—64 old trials using the words from the list and 64 trials using 32 early and 32 late AoA words, which had not been shown.

Results and discussion

Table 3 shows the means, standard deviations, effect sizes, and Bayes factors for various performance measures. As in Experiment 1, there was an effect of AoA on recognition: Mean d_a was higher for late AoA words than for early AoA words, $t(43) = 4.326$, $p < .001$, $BF_{10} = 264.339$. Thirty-three subjects had higher d_a for late words compared with 11 who had higher d_a for early words, which is significant by a two-tailed sign test, $p = .001$.

This advantage for late AoA words in mixed lists replicates the findings from the experimental studies of Dewhurst et al. (1998, Experiments 1 and 2) and is in contrast to the null results from the experimental study of Coltheart and Winograd (1986, Experiment 2). As noted earlier, Experiment 2 of Dewhurst et al. used the same stimuli as Experiment 2 of Coltheart and Winograd, so one possibility for the differing results is the use of signal-detection measures in the former study versus proportion correct in the latter.

As in Experiment 1, there was no evidence of a mirror effect: The false-alarm rate was higher for early than for late AoA words, $t(43) = 3.718$, $p < .001$, $BF_{10} = 48.573$, but there was no difference in the hit rate, $t(43) = 0.151$, $p = .881$, $BF_{01} = 6.061$. Neath et al. (2021) found mirror effects obtained for contextual diversity, frequency, and concreteness only when the stimuli were confounded; when confounds were removed, the mirror effect was absent. AoA affected only false alarms, the same result that Neath et al. (2021) found for manipulations of contextual diversity and frequency; both of these dimensions correlate with AoA. In contrast, concreteness affected only hits; false alarms were unaffected.

Experiment 2 replicated the finding of Experiments 1 and 2 of Dewhurst et al. (1998) of a late AoA advantage in recognition when mixed lists are used, and Experiment 1 found the same result for pure lists. This pattern is also consistent with the regression analyses of Cortese et al. (2010) and Cortese et al. (2015). Of the two studies noted above that did not find an effect of AoA on recognition, one may be explained by not using a signal-detection analysis, and the second may be explained by how the different factors were entered into the regression equation. Based on this, we conclude that AoA affects recognition and the advantage accrues to late AoA words.

Table 3 Descriptive statistics and performance measures for Experiment 2

Measure	Early AoA		Late AoA		Cohen's *d*	BF_{10}
	M	*SD*	*M*	*SD*		
Hit	0.686	0.130	0.689	0.140	0.023	0.165
FA	0.340	0.137	0.280	0.147	0.561	48.573
d'	0.979	0.465	1.197	0.560	0.441	6.619
C	−0.038	0.351	0.055	0.381	0.351	1.834
d_a	1.041	0.505	1.281	0.571	0.652	264.339
Slope	0.865	0.271	0.819	0.252	0.161	0.278

Note. Slope indicates the slope of the *z*-ROC function

Experiment 3

Only one paper has examined the effect of AoA on serial recall. Roodenrys et al. (1994, Experiments 1 and 3) found no effect of AoA using pure lists, but they used a memory-span task in which the list lengths varied. Experiment 3 was designed to assess whether AoA affects serial recall in pure lists, but used fixed-length lists rather than varying the list length because performance can differ between fixed-length and varying-length lists (e.g., Crowder, 1969; Pollack, Johnson, & Knaff, 1959).

Method

Subjects Forty-four different volunteers from ProlificAC were paid £8.00 per hour (prorated) for their participation. The mean age was 28.00 years (*SD* = 5.53, range: 19–39 years); 22 subjects self-identified as female, and 22 self-identified as male.

Stimuli The stimuli were the same as in Experiment 2.

Procedure After indicating consent, the subjects were reminded of the instructions. They saw a list of six words presented one at a time for 1 s in the middle of the screen in 28-point Helvetica font. Immediately after the last item disappeared, the subjects were prompted to type in the first word, then the second word, and so on. Subjects were encouraged to guess or they could click on a button labelled "skip." There was no time limit on the recall period. After all six responses had been made, the subject could click on a "Start Next Trial" button when ready.

There were 24 trials, half with early and half with late AoA words. For each subject, the words for the upcoming trial were randomly selected without replacement from the appropriate pool, and then randomly ordered. The order of the trials—early versus late AoA—was randomly determined for each subject.

Results and discussion

The proportion of words correctly recalled in order was analyzed by a 2 AoA × 6 serial position analysis of variance

(ANOVA).[2] In this and subsequent experiments, noninteger degrees of freedom for the frequentist ANOVA indicate the Greenhouse–Geisser sphericity correction was applied. For the Bayesian ANOVA, main-effect models were evaluated with respect to a random-effects error model, and interaction models were evaluated with respect to a main-effects model.

The main effect of AoA was not significant: The proportion of words correctly recalled in order was the same for early (M = 0.534, SD = 0.196) and late (M = 0.525, SD = 0.181) AoA words, $F(1, 43) = 0.630$, $MSE = 0.015$, $\eta_p^2 = 0.014$, $p = .432$, $BF_{01} = 9.90$. There was the usual significant effect of serial position, $F(2.91, 125.24) = 140.632$, $MSE = 0.053$, $\eta_p^2 = 0.766$, $p < .001$, $BF_{10} = 6.67 \times 10^{116}$. The upper left panel of Fig. 1 shows serial position functions, which are typical of immediate serial recall. There was no interaction, $F(4.28, 184.03) = 1.556$, $MSE = 0.012$, $\eta_p^2 = 0.035$, $p = .184$, $BF_{01} = 39.29$. Twenty-two subjects recalled more early words, 20 recalled more late words, and two were tied; this difference is not significant by a two-tailed sign test, $p = .878$.

The data were also scored using free-recall criteria; that is, a word was counted as correctly recalled regardless of whether it was recalled in the correct position. The main effect of AoA was again not significant: The proportion of words correctly recalled regardless of position was the same for early AoA (M = 0.620, SD = 0.148) and late AoA words (M = 0.632, SD = 0.155), $F(1, 43) = 1.566$, $MSE = 0.012$, $\eta_p^2 = 0.035$, $p = .218$, $BF_{01} = 8.55$. There was the usual significant effect of serial position, $F(2.70, 116.23) = 81.129$, $MSE = 0.061$, $\eta_p^2 = 0.654$, $p < .001$, $BF_{10} = 9.32 \times 10^{78}$. The upper-right panel of Fig. 1 shows the serial position functions. There was no interaction, $F(4.11, 176.71) = 0.879$, $MSE = 0.016$, $\eta_p^2 = 0.020$, $p = .480$, $BF_{01} = 58.97$. Twenty-three subjects recalled more early words, 20 recalled more late words, and one was tied; this difference is not significant by a two-tailed sign test, $p = .761$.

These results replicate those of Roodenrys et al. (1994), suggesting that for this manipulation, varying versus fixed list length is not a factor. Moreover, scoring without regard to position led to the same conclusion: AoA does not affect immediate serial recall of pure lists.

Experiment 4

Experiment 4 was identical to Experiment 3, except that it used mixed lists instead of pure lists. Half of the lists had early AoA words at odd positions and late AoA words at even positions, and the remaining lists had the reverse.

Method

Subjects Forty-four different volunteers from ProlificAC were paid £8.00 per hour (prorated) for their participation. The mean age was 26.91 years (SD = 4.93, range: 19–39); 33 subjects self-identified as female, and 11 self-identified as male.

Stimuli The stimuli were the same as in Experiments 2 and 3.

Procedure The procedure was identical to that of Experiment 3, except that each list contained three early and three late AoA words, which alternated. Half the lists began with an early AoA word, and the remaining half began with a late AoA word.

Results and discussion

Composite lists were created for analysis (see Hulme et al., 2003). The early AoA words from the odd positions were combined with the early AoA words from the even positions to form composite early lists. The same was done with late AoA words to form composite late lists.

The proportion of words correctly recalled in order was analyzed by a 2 AoA × 6 serial position ANOVA.[3] The main effect of AoA was not significant: The proportion words correctly recalled in order was the same for early (M = 0.538, SD = 0.138) and late (M = 0.542, SD = 0.143) AoA words, $F(1, 43) = 0.176$, $MSE = 0.009$, $\eta_p^2 = 0.004$, $p = .677$, $BF_{01} = 10.53$. There was the usual significant effect of serial position, $F(2.90, 124.59) = 118.019$, $MSE = 0.052$, $\eta_p^2 = 0.733$, $p < .001$, $BF_{10} = 1.379 \times 10^{95}$. The lower left panel of Fig. 1 shows the serial position functions. There was no interaction, $F(3.50, 150.37) = 0.629$, $MSE = 0.026$, $\eta_p^2 = 0.014$, $p = .621$, $BF_{01} = 61.99$. Twenty-one subjects recalled more late AoA words, 18 recalled more early words, and five were tied; this difference is not significant by a two-tailed sign test, $p = .749$.

The data were also scored using free-recall criteria. The main effect of AoA was again not significant: The proportion words correctly recalled, ignoring order, was the same for early (M = 0.617, SD = 0.123) and late (M = 0.627, SD = 0.119) AoA words, $F(1, 43) = 1.101$, $MSE = 0.015$, $\eta_p^2 = 0.025$, $p = .300$, $BF_{01} = 8.93$. There was the usual significant effect of serial position, $F(3.37, 144.86) = 74.977$, $MSE = 0.044$, $\eta_p^2 = 0.636$, $p < .001$, $BF_{10} = 3.92 \times 10^{69}$. The lower

[2] The responses were checked for spelling and typing errors. Of the 6,336 responses, 154 (2.43%) were flagged by the spellchecker, 60 early and 92 late AoA words. Correcting the spelling resulted in 30 early words becoming correct compared with 46 late words becoming correct. Because this did not change the results of the analyses, and because correcting spelling is not entirely objective, only analyses from the uncorrected responses are presented.

[3] The responses were checked for spelling and typing errors. Of the 6,336 responses, 102 (1.61%) were flagged by the spellchecker, 50 early and 52 late AoA words. Correcting the spelling resulted in 20 early words becoming correct compared with 30 late words becoming correct. Because this did not change the results of the analyses, and because correcting spelling is not entirely objective, only analyses from the uncorrected responses are presented.

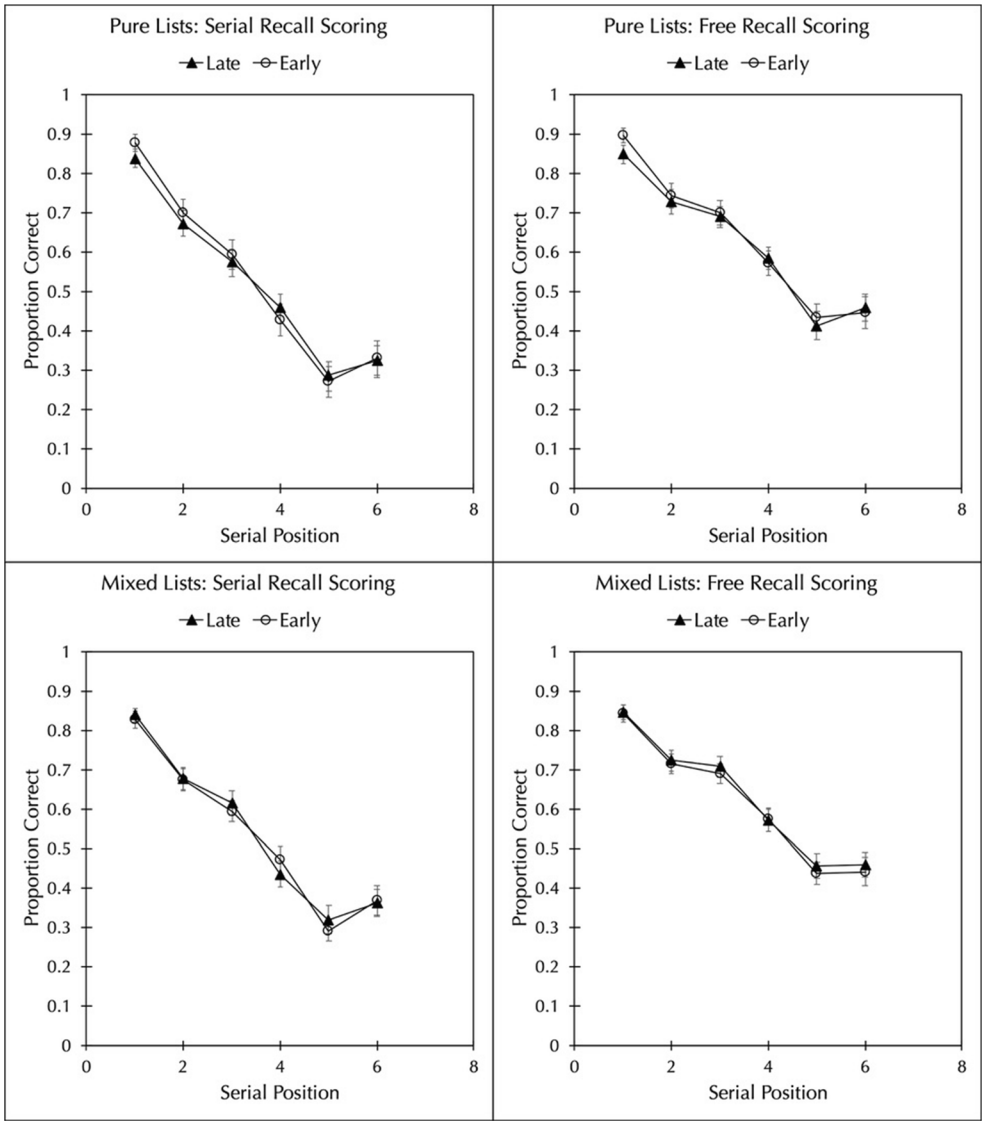

Fig. 1 The proportion of late and early AoA words recalled in Experiment 3 (top row) and Experiment 4 (bottom) row when scored using strict serial recall criteria (left panels) or when scored ignoring position (right panels). Error bars show the standard error of the mean

right panel of Fig. 2 shows the serial position functions. There was no interaction, $F(4.26, 183.14) = 0.114$, $MSE = 0.019$, $\eta_p^2 = 0.003$, $p = .982$, $BF_{01} = 141.71$. Twenty-three subjects recalled more early words, 20 recalled more late words, and one was tied; this difference is not significant by a two-tailed sign test, $p = .761$.

Given the results of the two experiments reported by Roodenrys et al. (1994) and those of Experiments 3 and 4, the conclusion is that AoA has no effect on serial recall regardless of whether the lists are pure or mixed.

Experiment 5

Experiment 5 examined free recall of pure lists. Two studies, Coltheart and Winograd (1986, Experiment 1) and Dewhurst et al. (1998, Experiment 3), reported no effect of AoA on free recall of pure lists whereas one study, Almond and Morrison (2014), reported an early-word advantage.

Method

Subjects Forty-four different volunteers from ProlificAC were paid £8.00 per hour (prorated) for their participation. The mean age was 28.02 years ($SD = 6.53$, range: 19–39 years) and 32 subjects self-identified as female and 12 self-identified as male.

Stimuli The stimuli were the same as in Experiments 2–4.

Procedure The procedure was similar to that of Experiment 3, except for the following. Each list contained 12 words, either

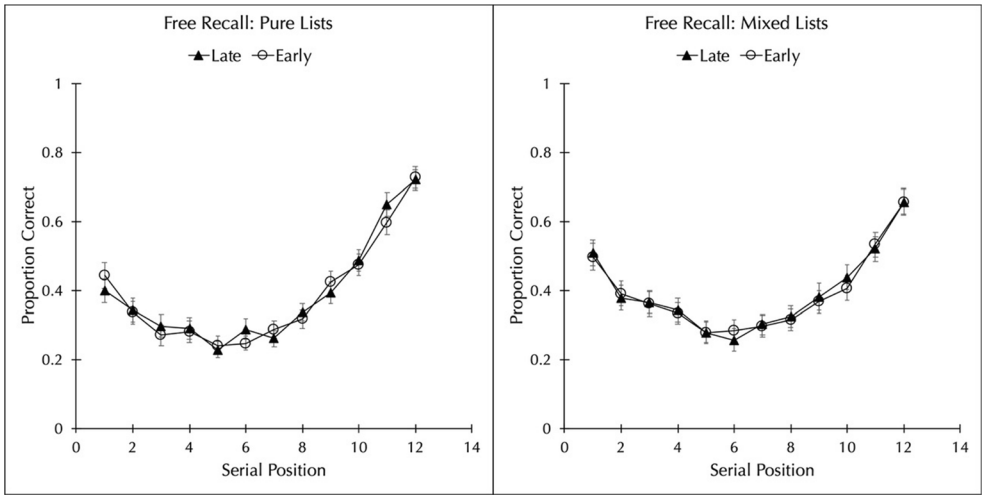

Fig. 2 The proportion of late and early AoA words recalled in Experiment 5 (left panel) and Experiment 6 (right panel). Error bars show the standard error of the mean

all early or all late AoA words, and the instructions asked the subjects to type in as many of the words as they could remember in any order. There were 10 lists of each type, with the order randomly determined for each subject.

Results

The proportion of words correctly recalled was analyzed by a 2 AoA × 12 serial position ANOVA.[4] There was no effect of AoA: The proportion of early words recalled (M = 0.387, SD = 0.094) did not differ from the proportion of late words recalled (M = 0.391, SD = 0.093), $F(1, 43) = 0.327$, MSE = 0.013, $\eta_p^2 = 0.008$, $p = .357$, $BF_{01} = 14.08$. There was the usual effect of position, $F(3.32, 142.53) = 43.049$, $MSE = 0.158$, $\eta_p^2 = 0.500$, $p < .001$, $BF_{10} = 3.75 \times 10^{102}$. As can be seen in the left panel of Fig. 2, there are primacy and recency effects typical of free recall. The interaction was not significant, $F(8.30, 356.72) = 0.914$, $MSE = 0.027$, $\eta_p^2 = 0.021$, $p = .507$, $BF_{01} = 713.25$. Nineteen subjects recalled more early words, 21 recalled more late words, and four were tied; this difference is not significant by a two-tailed sign test, $p = .874$.

The results replicate both Coltheart and Winograd (1986, Experiment 1) and Dewhurst et al. (1998, Experiment 3) in finding no effect of AoA on free recall of pure lists. They contrast with the results of Almond and Morrison (2014) who reported an early-word advantage. As noted earlier, the early and late Almond and Morrison stimuli also differed in frequency, with the advantage going to the early words. As shown in Table 1, the difference in AoA was also very small, 2.27 versus 3.33. Given that recall was likely influenced by frequency in the Almond and Morrison study, our conclusion is that there is no evidence that AoA affects free recall of pure lists.

Experiment 6

Experiment 6 examined free recall of mixed lists. One study, Dewhurst et al. (1998, Experiment 3), reported a late-word advantage for free recall of mixed lists.

Method

Subjects Forty-four different volunteers from ProlificAC were paid £8.00 per hour (prorated) for their participation. The mean age was 30.07 years (SD = 5.26, range: 20–39 years); 29 subjects self-identified as female, and 15 self-identified as male.

Stimuli The stimuli were the same as in Experiments 2–5.

Procedure The procedure was similar to that of Experiment 5, except that each list contained six early and six late AoA words. Half the lists alternated early and late AoA words beginning with an early word and the remaining half began with a late AoA word.

Results

As in Experiment 4, composite lists were created for analysis. The proportion of words correctly recalled was analyzed by a

[4] The responses were checked for spelling and typing errors. Of the 4,949 responses, 52 (1.05%) were flagged by the spellchecker, 22 early and 30 late AoA words. Correcting the spelling resulted in 16 early words becoming correct compared with 19 late words becoming correct. Because this did not change the results of the analyses, and because correcting spelling is not entirely objective, only analyses from the uncorrected responses are presented.

2 AoA × 12 serial position ANOVA.[5] There was no effect of AoA: The proportion of early words recalled (M = 0.394, SD = 0.128) did not differ from the proportion of late words recalled (M = 0.397, SD = 0.126), $F(1, 43) = 0.137$, MSE = 0.016, $\eta_p^2 = 0.003$, $p = .713$, $BF_{01} = 14.70$. There was the usual effect of position, $F(3.11, 133.52) = 21.830$, $MSE = 0.191$, $\eta_p^2 = 0.337$, $p < .001$, $BF_{10} = 2.09 \times 10^{55}$. The serial position functions are shown in the right panel of Fig. 2. The interaction was not significant, $F(7.21, 310.18) = 0.229$, MSE = 0.034, $\eta_p^2 = 0.005$, $p = .980$, $BF_{01} = 3036.06$. Twenty-three subjects recalled more early words, 21 recalled more late words, and there were no ties; this difference is not significant by a two-tailed sign test, $p = .880$.

The result differs from that reported by Dewhurst et al. (1998, Experiment 3), who found a late-word advantage. However, their result may be influenced by word frequency rather than AoA. Their stimuli are better described as forming a 2 × 2 design, with both frequency and AoA as factors. In the mixed lists, the low-frequency words were better recalled than the high-frequency words, a pattern that is very common (e.g., DeLosh & McDaniel, 1996; Duncan, 1974; May & Tryk, 1970). If one examines just the low-frequency words, it turns out that the early AoA words and the late AoA words differ in frequency, with the early words being the more frequent. This could lead to a recall advantage for the late AoA words. The same is true for the high-frequency words. It is possible, then, that their results are due to the influence of frequency. Given that the stimuli in Experiment 6 did not differ in frequency, our conclusion is that AoA does not affect free recall of mixed lists.

General discussion

The existing literature on whether AoA affects common memory tasks such as recognition, serial recall, and free recall is not clear. One reason may be that older studies could not take advantage of the recent norms and databases that allow the researcher to better control stimuli. Using these norms, we created two sets of stimuli where the early and late AoA words did not overlap in terms of AoA. In addition, the early and late words had a larger mean difference in AoA than in previous studies, and the early and late AoA words were equated on a number of other dimensions known to affect memory performance. Using these stimuli, we re-assessed whether AoA affects each of these tests.

The first conclusion is that AoA affects recognition, resulting in a late-word advantage. Experiment 1 found a late-word advantage on an old/new recognition test using pure lists, and Experiment 2 found the same result with mixed lists. This was true for both d_a and d'. Dewhurst et al. (1998, Experiments 1 and 2) also found a late-word advantage in d'. Coltheart and Winograd (1986, Experiment 2) reported no effect of AoA on recognition, but their conclusion is based on proportion correct, which does not distinguish between sensitivity and bias. In Experiments 1 and 2, the effect of AoA was entirely on the false-alarm rate; there was no difference in the hit rate.

The second conclusion is that there is no effect of AoA on serial recall, regardless of whether pure (Experiment 3) or mixed (Experiment 4) lists are used. Roodenrys et al. (1994, Experiments 1 and 3) also found no effect of AoA on serial recall with pure lists, but they used a span task, whereas we used fixed-length lists.

The third conclusion is that there is no effect of AoA on free recall, regardless of whether pure (Experiment 5) or mixed (Experiment 6) lists are used. Both Coltheart and Winograd (1986, Experiment 1) and Dewhurst et al. (1998, Experiment 3) also found no effect of AoA on free recall of pure lists, but Almond and Morrison (2014) reported an early-word advantage. As noted earlier, it is possible that the Almond and Morrison result is due to the combination of a number of differences between the early and late AoA words, including differences in frequency, accompanied by a very small difference in AoA (2.27 vs. 3.33). Dewhurst et al. (1998, Experiment 3) found a late-word advantage in free recall of mixed lists. As noted earlier, their result may also be due to a difference in frequency between the early and late AoA words. A low-frequency advantage is commonly found when high-frequency and low-frequency words are mixed in the same list, and the late AoA words were of lower frequency than the early AoA words, according to both the Brysbaert and New (2009) and van Heuven et al. (2014) norms.

One advantage of using the same stimulus set for Experiments 2–6 is that it suggests that the null results observed in Experiments 3–6 are most likely not due to an insufficient manipulation of AoA: The same stimuli produced a late AoA advantage in Experiment 2. We note, however, that although we controlled for a large number of dimensions, it is always possible that we overlooked one or more dimensions that may have affected performance. For this reason, we encourage other researchers to create their own stimulus sets rather than using the ones we created, and we also encourage them to publish the stimulus sets in their reports.

The results of Experiments 1–6 provide more evidence that AoA differs from related variables such as frequency despite the fact that the two variables correlate because the pattern of effects differs. Both variables affect recognition in the same way: Late AoA words (Experiments 1 and 2; Dewhurst et al., 1998, Experiments 1–2) are recognized better than early AoA words in pure and mixed lists, just as low-frequency words are recognized better than high-frequency words (Gorman, 1961;

[5] The responses were checked for spelling and typing errors. Of the 4,845 responses, 56 (1.16%) were flagged by the spellchecker, 25 early and 31 late AoA words. Correcting the spelling resulted in 19 early words becoming correct compared with 18 late words becoming correct. Because this did not change the results of the analyses, and because correcting spelling is not entirely objective, only analyses from the uncorrected responses are presented.

Schulman, 1967) in pure and mixed lists. Whereas AoA has no effect on serial recall of pure lists (Experiment 3; Roodenrys et al., 1994), frequency has a robust effect showing a high-frequency advantage (Neath & Surprenant, 2019; Roodenrys et al., 1994). For mixed lists, however, neither AoA (Experiment 4) nor word frequency (Hulme et al., 2003; Morin, Poirier, Fortin, & Hulme, 2006) affect serial recall. In free recall, AoA has no effect on pure lists (Experiment 5), whereas there is a robust high-frequency advantage (Deese, 1960; Peters, 1936). Finally, AoA has no effect on mixed lists (Experiment 6), whereas the most common pattern with frequency is a low-frequency advantage (DeLosh & McDaniel, 1996; Duncan, 1974; May & Tryk, 1970). These differences suggest that explanations based on word frequency may not fare well in explaining AoA effects.

How, then, can the results be explained? Almond and Morrison (2014) explained the recall advantage they observed for early-acquired over late-acquired words by suggesting that early-acquired words are stored in more interconnected cognitive and neuronal networks relative to words acquired later in life. As a result, people are able to form more interitem associations between early-acquired than late-acquired words. As such, activating the cognitive representation of one early-acquired word primes access to the other such words to a greater extent than occurs among late-acquired words. Almond and Morrison further posited that, because early-acquired words are more rapidly and efficiently processed than late-acquired words, if study time is not equated between the two words types, participants will devote more attention to late-acquired words at encoding. This attentional imbalance may result in the appearance of a spurious recall advantage for late-acquired words.

The Almond and Morrison (2014) account could explain the late AoA advantage in recognition seen in Experiments 1 and 2. The idea is that the early words appear more familiar, and therefore lead to a higher false-alarm rate. However, this does not explain why there was no effect of AoA in serial or free recall. Having more interitem associations should have led to an early AoA advantage in both serial and free recall.

Dewhurst et al. (1998) offered a different explanation. They interpreted the discrepancy in processing fluency between the two word types as the primary source of AoA effects in both recall and recognition memory. According to the item-order hypothesis of free recall (DeLosh & McDaniel, 1996), list items that require more attentional resources to process interfere with the encoding of order information. As order information is used to guide retrieval, items that are processed more fluently are better recalled in pure lists. However, in mixed lists, items that require more elaborate processing have an advantage at retrieval because of the distinctiveness of their features. Dewhurst et al. (1998) suggested that their results may be partially explained by applying the item-order hypothesis to AoA, which predicts a recall advantage for late-acquired words in mixed lists and for early-acquired words in pure lists. Whereas Dewhurst et al. (1998) found an advantage for late-acquired words when recall was tested for mixed lists, there was no effect of AoA on recall of pure lists. In recognition memory, Dewhurst et al. (1998) suggested that the disparity in processing fluency between early-acquired and late-acquired words may have resulted in more distinctive episodic traces associated with the late-acquired words. According to the distinctiveness-fluency framework (Rajaram, 1996), the greater distinctiveness of late-acquired words would enhance the amount of conscious recollection associated with these words. This explanation corresponds well with Dewhurst et al.'s findings, as the recognition advantage for late-acquired words was located specifically in the recollection component of recognition memory.

The results of Experiments 1 and 2 are consistent with the account of Dewhurst et al. (1998). There was a late AoA advantage in both pure and mixed lists. In recognition, the same process that is posited to help late AoA words should apply regardless of whether the lists are mixed or pure. The reason is that with such long lists, there is little if any role for order information. Their account does not address why there was no effect of AoA on serial recall. If anything, an item-order account would predict that the focus on order information in serial recall would lead to enhanced recall of early AoA items to the detriment of late AoA items. Finally, their account predicts a late AoA advantage on free recall of mixed lists, for the same reasons it predicts a low frequency advantage on mixed lists (DeLosh & McDaniel, 1996), but Experiment 6 found no effect.

Cortese et al. (2010) hypothesized that late AoA words are more semantically distinct than early AoA words (see also Gullick & Juhasz, 2008). The reason is that during vocabulary acquisition, early AoA words serve as the reference point to which later words are compared. Such an account would predict a late AoA advantage in recognition, which is what was observed in Experiments 1 and 2. Although Cortese et al. do not address free or serial recall, a straightforward prediction is possible at least for tests involving pure lists. In serial recall, lists of related words are better recalled than lists of unrelated words, the so-called semantic relatedness effect (Tehan, 2010; Tse, 2009). Therefore, the semantic distinctiveness account predicts an early AoA advantage in serial recall, but Experiment 3 found no such effect. The semantic relatedness effect also occurs in free recall (Crowder, 1979; Glanzer & Schwartz, 1971). Similarly, the semantic distinctiveness account predicts an early AoA advantage in free recall, but Experiment 5 found no effect of AoA on free recall.

The six experiments reported here help clarify when AoA will affect memory by using stimulus sets in which the early and late AoA words were equated on more dimensions than previously possible, in which there was a large difference in AoA between the early and late items, and in which there was no overlap in AoA. The results indicate that AoA does affect recognition, but does not affect serial or free recall; this pattern of results poses a problem for current explanations of the locus

of the AoA effect on memory. The extant accounts of AoA all predict the late advantage for recognition, but none offer an explanation of why AoA does not affect either serial or free recall.

Author note This research was supported, in part, by grants from the Natural Sciences and Engineering Research Council of Canada to I.N. and A.M.S. Authors are listed alphabetically.

Open practices statement The stimuli are provided in the manuscript. Additional experiments are described in a supplemental report which, along with all the raw data, is available at the Open Science Foundation (https://doi.org/10.17605/OSF.IO/2CAGB).

Appendix

Note. CELEX: Log base 10 CELEX frequency; Orth: number of orthographic neighbours (Coltheart's *N*); OrthZ: a *z* score based on ORTH (see Storkel, 2004); OrthF: frequency of orthographic neighbours; C2: constrained bigram frequency; C3: constrained trigram frequency; C2Z: a *z* score based on C2; C3Z: a z score based on C3; U2: constrained bigram frequency; U3: constrained trigram frequency; U2Z: a *z* score based on U2; U3Z: a *z* score based on U3 (from Medler & Binder, 2005); LgWF: log base 10 $SUBTLEX_{US}$ frequency; LgCD: log base 10 $SUBTLEX_{US}$ contextual diversity (from Brysbaert & New, 2009); zipf UK: zipf frequency SUBLEXUK; zipf BNC: zipf British National Corpus frequency (from van Heuven et al., 2014); LgHAL: log base 10 HAL frequency; OLD: orthographic Levenshtein distance; OLDF: frequency of the orthographic Levenshtein neighbours; PLD: phonological Levenshtein distance; PLDF: frequency of the phonological Levenshtein neighbours; NPhon: number of phonemes; NSyll: number of syllables; NLet: number of letters (from Balota et al., 2007); AoA: tested age of acquisition (from Brysbaert & Biemiller, 2017); Cnc.M: mean concreteness; Cnc.SD: standard deviation of the concreteness rating; Known: proportion of respondents indicating they knew the word (from Brysbaert et al., 2014); V.M: mean valence rating; A.M: mean arousal rating; D.M: mean dominance rating (from Warriner et al., 2013); and WordNET: mean path length (from Pedersen et al., 2004). ***Bold italic*** indicates a significant difference.

Table 4 Descriptive properties of the stimuli used in Experiment 1

	Early AoA		Late AoA		*t*	*p*
	M	*SD*	*M*	*SD*		
CELEX	3.506	0.709	3.447	0.684	0.502	.616
Orth	0.669	1.452	0.669	1.452	0.000	1.000
OrthZ	−0.490	0.452	−0.491	0.453	0.012	.990
OrthF	0.698	2.170	0.697	2.176	0.082	.935
C2	594.942	431.160	600.504	431.176	0.241	.810
C3	107.101	174.409	86.820	86.382	1.137	.257
C2Z	−0.535	0.635	−0.528	0.648	0.376	.707
C3Z	−0.380	0.649	−0.414	0.512	0.654	.514
U2	19,472.719	9,824.884	2,0038.086	9,190.743	0.493	.622
U3	1,979.223	2,084.938	2,127.885	1,704.342	0.091	.928
U2Z	−0.356	0.956	−0.262	0.949	0.702	.483
U3Z	−0.311	0.754	−0.206	0.771	0.332	.740
LgWF	3.406	0.520	3.328	0.481	0.726	.468
LgCD	1.965	0.511	1.872	0.460	0.913	.362
zipf UK	3.410	0.679	3.377	0.643	0.269	.788
zipf BNC	3.418	0.818	3.483	0.787	0.644	.520
LgHAL	7.135	1.790	7.209	1.866	0.282	.778
OLD	2.801	0.760	2.777	0.723	0.094	.925
OLDF	6.714	0.698	6.737	0.735	0.281	.779
PLD	2.689	0.872	2.625	0.810	0.204	.839
PLDF	6.635	0.940	6.762	0.934	0.990	.323
NPhon	5.871	1.372	5.799	1.395	0.000	1.000
NSyll	2.029	0.496	2.014	0.466	0.428	.669

Table 4 (continued)

	Early AoA		Late AoA		*t*	*p*
	M	*SD*	*M*	*SD*		
NLet	7.338	1.497	7.317	1.470	0.063	.949
AoA	***3.209***	***0.982***	***10.770***	***2.019***	***21.153***	***.000***
Cnc.M	3.952	0.914	3.815	0.944	0.532	.595
Cnc.SD	0.983	0.436	1.057	0.364	0.948	.344
Known	0.991	0.020	0.989	0.022	0.698	.486
V.M	5.445	1.073	5.346	1.036	0.047	.963
A.M	4.039	0.789	3.987	0.738	0.349	.727
D.M	5.405	0.748	5.336	0.783	0.255	.799
WordNET	11.833	1.937	12.015	2.057	0.763	.446

Early AoA words: accordion, agreement, anchor, arctic, baboon, backfire, ballroom, barber, bargain, barley, bathrobe, beefsteak, bitterness, blessing, blowout, boatman, bodyguard, bouquet, broomstick, bumper, bundle, buzzer, catholic, champion, chilly, collie, compound, corkscrew, curb, daybreak, divorce, draft, drape, druggist, expression, falcon, faucet, firewood, focus, footwear, forecast, fortune, foxhole, frontier, giraffe, glare, greyhound, grouch, hairdo, hardware, hardwood, hatchet, hillside, improvement, instant, insult, ketchup, keyhole, kickoff, kidney, kimono, lesson, lifeguard, luggage, lumberjack, marshmallow, mattress, member, membership, merchant, mining, mitten, modern, moonbeam, muscle, nervousness, noodle, northwest, ordinary, oyster, peephole, percent, platter, playmate, playpen, porthole, puppet, redwood, refund, reindeer, relative, roomful, sawdust, scold, scrub, seesaw, shaker, shield, sideshow, silence, slang, sleigh, southeast, spaceship, sparkle, speech, spinach, spook, spoonful, spree, standstill, storage, storeroom, suggestion, sunflower, surfboard, tadpole, teamwork, teaspoon, termite, thinker, tinfoil, tracer, transfer, translation, trapeze, trombone, tulip, unemployed, violinist, vision, warrior, wartime, watchtower, whitewash, width, woodwork, worship, wreckage

Late AoA words: absolute, access, airship, alliance, anguish, artichoke, attire, awning, backdrop, badger, bereavement, bistro, boardwalk, boutique, bowler, brisket, broadside, brochure, buffet, buffoon, bunker, buttermilk, cardigan, cashmere, chalice, chateau, circuit, clambake, classical, cleaver, coaster, commerce, commitment, component, corruption, counsel, coverage, cuisine, debris, detail, dialogue, dinghy, dolphin, doodle, dreamer, embrace, enterprise, facade, feline, ferret, flagship, flair, flamboyant, floss, fluke, footwork, fortress, franchise, gauntlet, gazelle, geisha, gender, gesture, ginseng, grid, grotesque, guinea, hospice, hostile, hydrate, instinct, jasmine, kingpin, landfall, ledger, lifeline, limestone, lithium, loophole, mainland, mantle, meatball, mentor, microwave, midwife, migraine, mocha, mousse, movement, mulch, niche, nursemaid, objective, option, parchment, parish, passion, payload, peanuts, penthouse, perspective, physics, pillbox, pipeline, printout, province, radius, rhinestone, saffron, saline, scaffold, scope, scowl, seaboard, searchlight, sequence, shank, shrapnel, skylight, soccer, software, source, species, sphere, spittoon, squid, stairwell, stethoscope, stooge, super, tabloid, tambourine, theory, threshold, trauma, treadmill, vermin, vermouth, waiver

Table 5 Descriptive properties of the stimuli used in Experiments 2–6

	Early AoA		Late AoA		*t*	*p*
	M	*SD*	*M*	*SD*		
CELEX	0.720	0.428	0.763	0.465	0.569	.570
Orth	4.691	4.086	4.926	4.240	0.330	.742
OrthZ	−0.239	0.712	−0.209	0.759	0.245	.807
OrthF	14.533	22.427	18.081	28.223	0.812	.418
C2	129.397	62.125	139.015	71.156	0.840	.403
C3	16.382	10.009	15.868	9.248	0.311	.756
C2Z	−0.271	0.739	−0.193	0.671	0.642	.522
C3Z	−0.107	0.830	−0.157	0.704	0.382	.703
U2	9,505.015	5,198.689	10,264.809	5,426.010	0.834	.406
U3	725.074	1,094.561	759.441	970.924	0.194	.847
U2Z	−0.121	0.926	0.014	0.884	0.874	.384
U3Z	−0.007	1.154	0.039	0.996	0.246	.806
LgWF	2.291	0.493	2.355	0.545	0.721	.472
LgCD	2.119	0.465	2.145	0.482	0.326	.745
zipf UK	3.603	0.511	3.606	0.659	0.030	.976
zipf BNC	3.536	0.575	3.654	0.611	1.161	.248

Table 5 (continued)

	Early AoA		Late AoA		*t*	*p*
	M	*SD*	*M*	*SD*		
LgHAL	7.629	1.295	7.931	1.568	1.225	.223
OLD	1.722	0.213	1.713	0.274	0.210	.834
OLDF	7.787	0.448	7.855	0.396	0.945	.346
PLD	1.565	0.316	1.547	0.350	0.321	.749
PLDF	7.800	0.737	7.947	0.690	1.199	.233
NPhon	3.926	0.719	4.000	0.712	0.599	.550
NSyll	1.279	0.452	1.338	0.477	0.738	.462
NLet	4.706	0.459	4.676	0.471	0.369	.713
AoA	***3.147***	***0.996***	***10.029***	***1.770***	***27.939***	***.000***
Cnc.M	3.966	0.785	3.832	0.822	0.973	.333
Cnc.SD	1.044	0.374	1.130	0.326	1.428	.156
Known	1	0	1	0	—	—
V.M	5.095	1.204	4.953	1.030	0.698	.486
A.M	4.017	0.914	4.169	0.904	0.925	.357
D.M	5.390	0.923	5.150	0.769	1.557	.122
WordNET	11.161	1.647	11.408	1.554	0.902	.369

Early AoA words: acre, beam, boost, brook, buggy, butch, cargo, chow, cigar, comic, crank, curb, curl, curse, darn, delay, drift, drip, elbow, faker, fatty, flake, flask, flick, flop, focus, glee, glide, honk, lava, major, moan, munch, nerve, noon, panda, pine, plum, proof, prune, quake, quart, quiz, racer, rake, razor, reset, riot, roomy, scoop, scout, seam, sleet, sling, snarl, snuff, stomp, stool, stoop, swirl, tough, tulip, twirl, twist, usher, verse, vine, vowel

Late AoA words: aide, alias, alien, bats, beret, binge, booty, bowls, canon, chaos, chic, chuck, cleft, cove, craze, cult, dean, dorm, dowry, drone, duct, dude, dune, felon, fifth, finch, floss, given, gram, grid, gross, guru, hutch, japan, khaki, lance, lied, loner, lotus, lure, madam, manor, maze, means, oval, pasta, plaza, plumb, prank, prime, prior, probe, quest, ramp, realm, sage, salon, scope, scowl, sinus, snare, spur, stud, super, torso, traps, weeds, womb

References

Almond, N. M., & Morrison, C. M. (2014). Episodic intertrial learning of younger and older participants: Effects of age of acquisition. *Aging, Neuropsychology, and Cognition, 21*, 606–632. https://doi.org/10.1080/13825585.2013.849653

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

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

Brown, G. D. A., & Watson, F. L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition, 15*, 208–216. https://doi.org/10.3758/BF03197718

Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods, 49*, 1520–1523. https://doi.org/10.3758/s13428-016-0811-4

Brysbaert, M., & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *Quarterly Journal of Experimental Psychology, 64*, 545–559. https://doi.org/10.1080/17470218.2010.503374

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

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

Carroll, J. B., & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology, 25*, 85–95. https://doi.org/10.1080/14640747308400325

Coltheart, V., & Winograd, E. (1986). Word imagery but not age of acquisition affects episodic memory. *Memory & Cognition, 14*, 174–179. https://doi.org/10.3758/BF03198377

Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *Quarterly Journal of Experimental Psychology, 60*, 1072–1082. https://doi.org/10.1080/17470210701315467

Cortese, M. J., Khanna, M. M., & Hacker, S. D. (2010). Recognition memory for 2,578 monosyllabic words. *Memory, 18*, 595–609. https://doi.org/10.1080/09658211.2010.493892

Cortese, M. J., McCarty, D. P., & Schock, J. (2015). A mega recognition memory study of 2897 disyllabic words. *Quarterly Journal of Experimental Psychology, 68*, 1489–1501. https://doi.org/10.1080/17470218.2014.945096

Crowder, R. G. (1969). Behavioral strategies in immediate memory. *Journal of Verbal Learning and Verbal Behavior, 8,* 524–528. https://doi.org/10.1016/S0022-5371(69)80098-8

Crowder, R. G. (1979). Similarity and order in memory. In G. Bower (Ed.), *Psychology of learning and motivation* (Vol. 13, pp. 319–353). New York: Academic.

Deese, J. (1960). Frequency of usage and number of words in free recall: The role of association. *Psychological Reports, 7*, 337–344. https://doi.org/10.2466/PR0.7.6.337-344

DeLosh, E. L., & McDaniel, M. A. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1136–1146. https://doi.org/10.1037/0278-7393.22.5.1136

Dewhurst, S. A., Hitch, G. J., & Barry, C. (1998). Separate effects of word frequency and age of acquisition in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 284–298. https://doi.org/10.1037/0278-7393.24.2.284

Duncan, C. P. (1974). Retrieval of low-frequency words from fixed lists. *Bulletin of the Psychonomic Society, 4*, 137–138. https://doi.org/10.3758/BF03334222

Ellis, A. W., & Lambdon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1103–1123. https://doi.org/10.1037/0278-7393.26.5.1103

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

Gerhand, S., & Barry, C. (1998). Word frequency effects in oral reading are not merely age-of-acquisition effects in disguise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 267–283. https://doi.org/10.1037/0278-7393.24.2.267

Gilhooly, K. J., & Gilhooly, M. L. (1979). Age-of-acquisition effects in lexical and episodic memory tasks. *Memory & Cognition, 7*, 214–223. https://doi.org/10.3758/BF03197541

Gilhooly, K. J., & Watson, F. L. (1981). Word age-of-acquisition effects: A review. *Current Psychological Reviews, 1*, 269–286. https://doi.org/10.1007/BF02684489

Glanzer, M., & Schwartz, A. (1971). Mnemonic structure in free recall: Differential effects on STS and LTS. *Journal of Verbal Learning and Verbal Behavior, 10*, 194–198. https://doi.org/10.1016/S0022-5371(71)80013-0

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology, 61*, 23–29. https://doi.org/10.1037/h0040561

Gullick, M. M., & Juhasz, B. J. (2008). Age of acquisition's effect on memory for semantically associated word pairs. *Quarterly Journal of Experimental Psychology, 61*, 1177–1185. https://doi.org/10.1080/17470210802013391

Hulme, C., Stuart, G., Brown, G. D. A., & Morin, C. (2003). High- and low-frequency words are recalled equally well in alternating lists: Evidence for associative effects in serial recall. *Journal of Memory and Language, 49*, 500–518. https://doi.org/10.1016/S0749-596X(03)00096-2

JASP Team. (2019). JASP (Version 0.11.10) [Computer software]. https://jasp-stats.org

Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition, 13*, 789–845. https://doi.org/10.1080/13506280544000066

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin, 131*, 684–712. https://doi.org/10.1037/0033-2909.131.5.684

Juhasz, B. J., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1312–1318. https://doi.org/10.1037/0278-7393.29.6.1312

Juhasz, B. J., & Rayner, K. (2006). The role of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition, 13*, 846–863. https://doi.org/10.1080/13506280544000075

Juhasz, B. J., Yap, M. J., Raoul, A., & Kaye, M. (2019). A further examination of word frequency and age-of-acquisition effects in English lexical decision task performance: The role of frequency trajectory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*, 82–96. https://doi.org/10.1037/xlm0000564

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

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

May, R. B., & Tryk, H. E. (1970). Word sequence, word frequency, and free recall. *Canadian Journal of Psychology, 24*, 299–304. https://doi.org/10.1037/h0082866

Medler, D. A., & Binder, J. R. (2005). *MCWord: An on-line orthographic database of the English language.* Medical College of Wisconsin, Language Imaging Laboratory. www.neuro.mcw.edu/mcword/

Meschyan, G., & Hernandez, A. (2002). Age of acquisition and word frequency. *Memory & Cognition, 30*, 262–269. https://doi.org/10.3758/BF03195287

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

Morin, C., Poirier, M., Fortin, C., & Hulme, C. (2006). Word frequency and the mixed-list paradox in immediate and delayed serial recall. *Psychonomic Bulletin & Review, 13*, 724–729. https://doi.org/10.3758/BF03193987

Morris, P. E. (1981). Age of acquisition, imagery, recall, and the limitations of multiple-regression analysis. *Memory & Cognition, 9*, 277–282. https://doi.org/10.3758/BF03196961

Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology, 21*, 116–133. https://doi.org/10.1037/0278-7393.21.1.116

Morrison, C. M., Ellis, A. W., & Quinlan, P. T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory & Cognition, 20*, 705–714. https://doi.org/10.3758/BF03202720

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*, 455–472. https://doi.org/10.3758/s13421-018-0883-8

Neath, I., Hockley, W. E., & Ensor, T. M. (2021). *Contextual diversity, word frequency, and concreteness mirror effects revisited.* Manuscript submitted for publication.

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

Pérez, M. A. (2007). Age of acquisition persists as the main factor in picture naming when cumulative word frequency and frequency trajectory are controlled. *Quarterly Journal of Experimental Psychology, 60*, 32–42. https://doi.org/10.1080/17470210600577423

Peters, H. N. (1936). The relationship between familiarity of words and their memory value. *American Journal of Psychology, 48*, 572–584. https://doi.org/10.2307/1416508

Pollack, I., Johnson, L. B., & Knaff, P. R. (1959). Running memory span. *Journal of Experimental Psychology, 57*, 137–146. https://doi.org/10.1037/h0046137

Rajaram, S. (1996). Perceptual effects on remembering: Recollective processes in picture recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 365–377. https://doi.org/10.1037/0278-7393.22.2.365

Roodenrys, S., Hulme, C., Alban, J., Ellis, A. W., & Brown, G. D. A. (1994). Effects of word frequency and age of acquisition on short-term memory span. *Memory & Cognition, 22*, 695–701. https://doi.org/10.3758/BF03209254

Rubin, D. C. (1980). 51 properties of 125 words: A unit analysis of verbal behavior. *Journal of Verbal Learning and Verbal Behavior, 19*, 736–755. https://doi.org/10.1016/S0022-5371(80)90415-6

Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychonomic Science, 9*, 211–212. https://doi.org/10.3758/BF03330834

Storkel, H. L. (2004). Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech, Language, and Hearing Research, 47*, 1454–1468. https://doi.org/10.1044/1092-4388(2004/108)

Tehan, G. (2010). Associative relatedness enhances recall and produces false memories in immediate serial recall. *Canadian Journal of Experimental Psychology, 64,* 266–272. https://doi.org/10.1037/a0021375

Tse, C.-S. (2009). The role of associative strength in the semantic relatedness effect on immediate serial recall. *Memory, 17,* 874–891. https://doi.org/10.1080/09658210903376250

van Heuven, W. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67*, 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*, 1191–1207. https://doi.org/10.3758/s13428-012-0314-x