Journal of Memory and Language 63 (2010) 46–63

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

The influence of item properties on association-memory

Christopher R. Madan [a,*], Mackenzie G. Glaholt [b], Jeremy B. Caplan [a,c]

[a] Department of Psychology, University of Alberta, Canada
[b] Department of Psychology, University of Toronto at Mississauga, Canada
[c] Centre for Neuroscience, University of Alberta, Canada

ARTICLE INFO

Article history:
Received 29 May 2009
revision received 16 February 2010
Available online 28 March 2010

Keywords:
Association-memory
Associative symmetry
Imageability
Word frequency
Episodic memory

ABSTRACT

Word properties like imageability and word frequency improve cued recall of verbal paired-associates. We asked whether these enhancements follow simply from prior effects on item-memory, or also strengthen associations between items. Participants studied word pairs varying in imageability or frequency: pairs were "pure" (high–high, low–low) or "mixed" (high–low, low–high) where "high" and "low" refer to imageability or frequency values and are probed with forward (A–?) and backward (?–B) cues. Probabilistic model fits to the data suggested that imageability primarily improved retrieval of associations, but frequency primarily improved recall of target items. All pair types exhibited a high correlation between forward and backward probe accuracy, a measure of holistic learning (Kahana, 2002), which extends the boundary conditions of holistic association-memory and challenges Paivio's (1971) suggestion that holistic learning depends critically on imagery. In sum, item properties can boost association-memory beyond simply boosting target retrievability.

Introduction

Much of everyday memory function involves deliberate retrieval of information associated with cues in the environment. For example, meeting a friend cues your memory (cued recall probe) causing you to recall her husband's name (cued recall target). This kind of memory requires both retrieval of an association (relationship) and production of a target item (item recall). This kind of memory function is typically studied with cued recall of paired-associates (Calkins, 1896; Underwood, 1966). The participant is required to learn and retrieve the association. However, cued recall performance could be influenced by memory for the association between the items in the pair, as well as by memory for the individual items themselves (e.g., Hockley & Cristi, 1996). Some material types are remembered better than others in cued recall, and although one would be tempted to attribute such differences to effects of the underlying encoding and retrieval of associations, it is also possible that such effects can be entirely attributed to differences in retrievability of the target items. Consider two properties of words that are known to improve cued recall accuracy: frequency of usage, or *word frequency* (Clark, 1992; Clark & Burchett, 1994) and subjective ratings of conduciveness to mental imagery production, or *imageability* (Lockhart, 1969; Paivio, 1965, 1968; Paivio, Smythe, & Yuille, 1968; Wood, 1967). Both of these item properties have well established effects on memory for items: high-frequency words are better recalled but worse recognized (the "word frequency paradox"; e.g., Gorman, 1961; Gregg, 1976; Hall, 1954, 1979; Shepard, 1967) and high-imageable words are better recalled and better recognized (e.g., Gorman, 1961). It is possible that the cued-recall advantage for high-frequency words and high-imageable words follows directly from the greater retrievability of target items. Alternatively,

* Corresponding author. Psychology Department, Biological Sciences Building, University of Alberta, Edmonton, Alberta, Canada T6G 2E9. Fax: +1 780 492 1768.
E-mail addresses: cmadan@ualberta.ca (C.R. Madan), mackenzie@psych.utoronto.ca (M.G. Glaholt), jcaplan@ualberta.ca (J.B. Caplan).

0749-596X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2010.03.001

high-frequency words and high-imageable words may in fact lead to better encoding and retrieval of the associations between paired items.

Our first objective was to determine the locus of item-property effects in cued recall performance. Specifically, item-property effects could be related to memory for associations or memory for the items themselves. Our second objective was to ask whether these item properties influence the relationship between forward and backward associations, a measure that is diagnostic of holistic association-memory (Kahana, 2002). Our third objective was to directly test Paivio's (1971) conceptual-peg hypothesis, which suggests that pairs consisting of two low-imageability words cannot be learned holistically as the association cannot be 'pegged' to an image suggested by either word.

Disentangling item- versus association-memory effects in cued recall

A related question has been asked about the effect of word frequency on serial recall and free recall with respect to whether word frequency improves (a) memory for serial order or (b) the simple retrievability of the list items (Hulme, Stuart, Brown, & Morin, 2003; Ward, Woodward, Stevens, & Stinson, 2003). These studies compared memory for pure lists with memory for lists that alternated between items of high and low-frequency. The obvious sign of enhancement of item recall alone would be a zig-zag pattern wherein alternating lists literally alternate between the pure-high and pure-low accuracy levels. However, in serial list learning, manipulations of word frequency have suggested that the enhanced memory for high-frequency words results from improved serial-order memory (e.g., item–item association strengths) rather than from item recall.

We follow a similar logic to disentangle item- versus association-memory effects in cued recall of pairs. In Experiment 1a, participants studied pairs composed of words that had high versus low-imageability values or high versus low-frequency values. For each manipulation (word frequency or imageability), pairs were either pure (composed of two items of the same class, i.e., high–high or low–low), or mixed (composed of items differing in class, i.e., high–low or low–high). Given a studied pair (*A*–*B*) cued recall probes can be in the forward direction (A–?) or in the backward direction (?–*B*). Half the cued recall probes for each pair type were forward and half were backward. Subsets of our design have been carried out, but as some conditions were missing, these results are open to interpretation. Briefly, studies of imageability have found symmetric mean performance in both pure pairs and mixed pairs (Bower, 1972; Crowder, 1976; Paivio, 1971; Wollen & Lowry, 1971). Word frequency studies have produced symmetric mean performance in pure pairs, but asymmetric mean performance in mixed pairs (Crowder, 1976; Nelson & McEvoy, 2000; Paivio, 1971). The results of these studies suggest that imageability primarily acted through modulating the strength of the association, or through item retrievability. Word frequency was found to likely act through item retrievability. However, there is still some interpretational ambiguity as these studies not have sufficient conditions to make definitive conclusions or contrast all possible mechanisms. Our complete design allowed us to fit the data using a simple model to obtain separate estimates of the effects of each item property on association-memory and target-item recall probabilities.

The model was based on the assumption that accurate cued recall relies on two kinds of mechanisms: (a) successful recall is dependent on individual properties of the cued recall probe and target items, individually, and (b) successful recall is also dependent on retrieval of the relationship between paired items. Here we make the simplifying assumption that retrieval operations are independent; thus retrieval probabilities combine multiplicatively. The model estimates the degree to which properties of the probe, the target, and the relationship between items influence probability of recall. To understand how the model works, consider two extreme cases, depicted in Fig. 1. Fig. 1a presents simulated data generated by a model that treated high- and low-valued items and pairs identically except that it assumed that high-valued target words would be better recalled. In this example, the difference between pure high (HH) and pure low (LL) pairs is clarified by the asymmetries observed in the mixed pairs (HL and LH). Accuracy is high whenever the target item is high-valued, regardless of the identity of the paired (probe) item. In contrast, Fig. 1b was generated by a model that treated high- and low-valued items identically except that it assumed that as the number of high-valued words in a pair increased, probability of retrieving the association would increase, regardless of which item was the probe or target. In this example, the same level of cued-recall advantage for pure high versus pure low pairs is clarified by the mixed pairs, which show no asymmetries.

As an additional measure of item retrieval effects, we also decided to look at intrusion rates for the high and low items in cued recall. We hypothesized that if the high items are more retrievable than low items, participants will recall them more often as cued recall errors. However, if high and low items are equally retrievable, they should intrude in cued recall in equal proportions.

Holistic associations

Separate from the question of whether associations are learned better or worse, one can ask whether material type can influence the holistic nature of the learned association. Consider the association between a pair of items, *A*–*B*. One can decompose the association into two directional associations, a forward association, $A \rightarrow B$, and a backward association, $A \leftarrow B$. The Independent Associations Hypothesis is that the association between paired items, *A*–*B*, is composed of two separate, unidirectional associations (Wolford, 1971). Accordingly, $A \rightarrow B$ is learned in a statistically independent step from $A \leftarrow B$. The consequence is that performance for forward probes of a pair, *A*-? is expected to be independent of performance for backward probes of the same pair, ?–*B*. In contrast, the Associative Symmetry Hypothesis assumes that pairs are learned as a compound, holistic unit (Asch & Ebenholtz, 1962; Köhler, 1947). Thus, forward and backward probes should be sensitive to the same variability in learning.

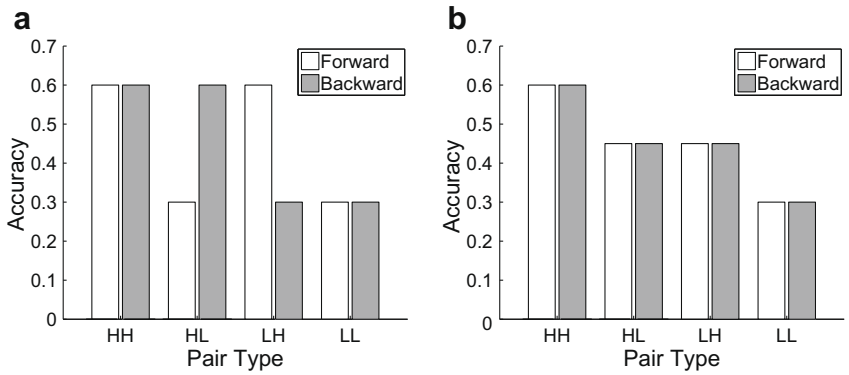

Fig. 1. Simulations of single-effect memory effects on mean accuracy. Pair types: high–high (HH), high–low (HL), low–high (LH), and low–low (LL) for two models: (a) Simulated effect on target-item recall ($t = 2.00$). (b) Simulated effect on association-memory ($r_1 = 1.33, r_2 = 1.50$). See main text, modeling section, for an explanation of the model parameters.

Gestalt psychologists initially claimed that associative symmetry implies equivalent performance on forward and backward cued recall tests (e.g., Asch & Ebenholtz, 1962), a finding that has been replicated numerous times (see Kahana, 2002, for a review). However, Kahana (2002) pointed out that symmetry of mean performance is orthogonal to whether a pair is learned holistically. He argued that the direct support for holistic learning would be an observation of a high, near-unity correlation between forward and backward tests at the level of individual pairs, over successive tests. Holistic learning would be observed as a high correlation if forward and backward probes measure the same underlying associative strength. Indeed, such a high forward–backward correlation has been observed in several studies of verbal paired-associates learning (e.g., Caplan, Glaholt, & McIntosh, 2006; Kahana, 2002; Rehani & Caplan, in preparation; Rizzuto & Kahana, 2000; Rizzuto & Kahana, 2001) and object–location learning (Sommer, Rose, & Büchel, 2007).

To understand why the mean performance and correlation measures reflect different memory phenomena, consider a hypothetical cued recall experiment using a brief study list consisting of two pairs, SHROUD–RUMOUR and HELMET–MALICE. A participant might recall SHROUD–? and ?–MALICE correctly, but ?–RUMOUR and HELMET–? incorrectly. Mean performance for forward and backward probes is equal (both 50%) but the tests are not positively correlated, which is suggestive of non-holistic learning (Independent Associations Hypothesis). However, another participant recalls SHROUD–? and ?–RUMOUR correctly but HELMET–? and ?–MALICE incorrectly. Here, mean performance is still symmetric (50% for both forward and backward probes); however forward and backward probe performance is correlated at the level of pairs, suggesting holistic learning (Associative Symmetry Hypothesis). In other words, forward and backward probes test the same learned information.

What is interesting is that while this illustrates that the two measures (correlation versus mean performance symmetry) are mathematically independent, it is not known whether they are separable in human behaviour in tasks of associative learning (e.g., asymmetric mean performance with no reduction in forward–backward correlation). In previous research, Caplan (2005) and Caplan et al. (2006) partly dissociated these two measures in serial lists. Rehani and Caplan (in preparation) found symmetry in mean performance of pairs with reduced correlations. In the present study, we included successive testing (each pair tested twice to examine the relationship between forward and backward cued-recall performance) and asked whether the two measures are coupled empirically. We aimed to induce asymmetries in mean performance via mixed pairs, as has been achieved by prior studies (e.g., Ebbinghaus, 1885/1913; Horowitz, Norman, & Day, 1966; Lockhart, 1969; Paivio, 1965; Paivio, 1968; Paivio, 1971; Wollen & Lowry, 1971) and asked whether the forward–backward correlation is disrupted when mean performance is highly asymmetric.

The conceptual-peg hypothesis and the Gestalt

One well developed theory of the influence of imageability on memory for verbal pairs is Paivio's (1971) conceptual-peg hypothesis. A core assumption of this hypothesis was that imageability of words facilitates the participant's ability to form an image combining the paired items, and that this image functions as a Gestalt. He further argued that if even one of the paired words is imageable (concrete), it can act as a 'peg' to which the low-imageable (abstract) item can be attached. However, if both words are abstract, no image can be formed. Because he assumed that the holistic representation was an image, the conceptual-peg hypothesis implies that pairs comprised of two low-imageable items, when learned, will not be learned holistically. Paivio and colleagues provided evidence based on measures of mean performance, which, as we elaborated in the previous section, cannot directly test whether or not associations are learned holistically (Kahana, 2002). We tested the conceptual-peg hypothesis by asking whether the correlation between forward and backward probes of pure abstract pairs was reduced compared to pairs containing high-imageable words (Experiments 1a and 1b).

To summarize the goals of the present study, we asked several questions regarding the effects of item properties

on memory for associations. First, by presenting pure and mixed pairs of words simultaneously we asked whether single-item properties influence item-learning, association-learning, or both. We tested the hypothesis that these single-item properties influence not only recall of target items but also the recall of the associations between items. Second, when asymmetric mean performance is found, as expected in cued recall performance of the mixed pairs (Experiment 1a), this could signal a disruption of the holistic association. We tested whether greater asymmetry in cued recall implies a greater reduction in the correlation between forward and backward cued recall. Third, we tested the conceptual-peg assumption that the holistic representation is an image by asking whether pure low-imageable pairs exhibit reduced forward–backward correlation compared to high–high-imageable pairs (Experiments 1a, as well as a follow-up experiment, 1b).

Experiment 1a

Methods

Participants

Fifty-nine undergraduate students from the University of Alberta participated in the two-session study for partial fulfillment of an introductory psychology course requirement. Data from three participants were not included in our analyses because these participants failed to appear for the second session. Sixteen male and 40 female participants (mean age ± *sd* = 21.4 ± 4.7) were included in the analyses. Participants were required to have English as their first language, to have normal or corrected-to-normal vision, and to provide written informed consent.

Materials

Study sets were constructed from four pools of nouns: high-frequency, low-frequency, high-imageability, and low-imageability (see Table 1 for item properties and Appendices A and B for the words themselves). Each pool contained 110 English words, ranging between four and six letters in length (inclusive). Between each pair of pools of a given type (i.e., high-frequency and low-frequency), the words were matched on letter length, mean positional bigram frequency, and orthographic neighbourhood size using phonological data and frequency counts from the CELEX Lexical Database (Baayen, Piepenbrock, & Gulikers, 1995). Imageability ratings were also matched (imageability values used were the average of four sources: Bird, Franklin, & Howard, 2001; Cortese & Fugett, 2004; Stadthagen-Gonzalez & Davis, 2006; Wilson, 1988). Note that for the frequency session, imageability was constrained to an intermediate range while frequency was manipulated, and vice-versa for the imageability session. For each participant words were drawn at random with an equal numbers of pairs in each of the following pair types: high–high (HH), high–low (HL), low–high (LH), and low–low (LL).

LSA properties

All possible word pairs from both sessions were assessed pair-wise for pre-existing semantic similarities using the latent semantic analysis (LSA) method (Landauer & Dumais, 1997). LSA $cos(\theta)$ is an index of similarity and was low overall. In the word frequency session however, high–high pairs were slightly more semantically related than the mixed and low–low word pairs ($d' = 0.86$).

LSA $cos(\theta)$ for all possible pairs of high–high, low–low and high–low items was calculated[1] for the imageability session (mean ± *sd* = high–high: 0.10 ± 0.14; low–low: 0.12 ± 0.15; mixed: 0.078 ± 0.080) and for the word frequency session (high–high: 0.17 ± 0.16; low–low: 0.05 ± 0.11, mixed: 0.077 ± 0.080). Note that two words from the low-frequency word pool were not found in the LSA database and were excluded from these LSA calculations.

Procedure

The task was comprised of two sessions on different days: a word frequency session and an imageability session in a within-subject design, with session order counterbalanced across participants.

In each session, each participant participated in one practice set (excluded from analyses) followed by 10 experimental sets involving 8 pairs each. Each set in the task consisted of the five phases (Fig. 2): the study phase, a distractor task, a cued recall (Test 1), another distractor task, and finally another cued recall (Test 2). Pairs were presented in random order but were subject to the constraint that every two consecutive study sets included four pairs of each pair type (HH/HL/LH/LL). In the distractor and cued-recall phases, participants typed their responses on the keyboard. Responses were recorded on the computer and later scored for accuracy. All stimuli were presented in a white "Courier New" font, which ensured fixed letter width, on a black background. Paired nouns were presented simultaneously in the centre of the computer screen for 3200 ms, followed by a 150 ms blank inter-stimulus interval.

The distractor task consisted of five equations in the form of A + B + C = ___, where A, B, and C were randomly selected digits between two and eight. Each equation remained in the centre of the screen for 5000 ms. The participant was asked to type the correct answer during this fixed interval, after which the screen was cleared for 200 ms.

Cued recall consisted of a probe word and a blank line, either to the right or left of the word (forward and backward testing directions, respectively). The participant was instructed to recall the word that was paired with the probe during the study phase, type it on the computer keyboard, and then press the "ENTER" key. The probe remained on the screen and the participant was given up to 15,000 ms to respond. If the participant pressed "ENTER," the experiment would proceed to the next probe. A 400 Hz beep was presented for 500 ms to signal that the response was submitted, after which the screen was cleared for 250 ms. If participants could not recall a target item they were instructed to type "PASS". Misspellings or variants of the correct word were scored as incorrect re-

[1] Pairwise LSA comparisons were calculated using http://lsa.colorado.edu/. The "General Reading up to 1st year college" semantic space was used with all 300 factors.

Table 1
Word pool statistics. IMAG = imageability rating; FREQ = word frequency (per million); ON = orthographic neighbourhood size; ONFREQ = average orthographic frequency (per million) of orthographic neighbours; PN = phonological neighbourhood size; PNFREQ = phonological frequency (per million) of phonological neighbours; CONBG = summed frequency that any two letter-pairs in the word occur together in the place that they are in the current word. See Westbury and Hollis (2007) for more information on these measures.

		IMAG	FREQ	*Letters*	*ON*	*ONFREQ*	*PN*	*PNFREQ*	*CONBG*
Imageability (IMAG) manipulated									
High	Mean	5.74	21	4.98	4.8	21	10.6	31	3422
	St.Dev.	0.39	14	0.79	5.1	37	8.6	85	2255
	Min	5.00	6	4.00	0.0	0	0.0	0	419
	Max	6.80	56	6.00	19.0	200	31.0	481	16072
Low	Mean	2.96	20.2	5.00	4.9	26	10.8	38	3621
	St.Dev.	0.44	15.0	0.79	5.4	44	9.1	106	2428
	Min	1.50	6.0	4.00	0.0	0	0.0	0	345
	Max	3.49	60.0	6.00	19.0	240	34.0	755	18257
	t	−49.8☆☆☆	−0.2	0.2	0.2	1.0	0.2	0.6	0.6
Word frequency (FREQ) manipulated									
High	Mean	4.26	224	4.58	7.5	36	14	52	5268
	St.Dev.	0.45	278	0.73	5.7	46	10	119	3147
	Min	3.50	61	4.00	0.0	0	0	0	676
	Max	4.99	1787	6.00	24.0	247	65	755	21244
Low	Mean	4.26	1.9	4.57	7.4	29	14.5	42	4590
	St.Dev.	0.44	1.5	0.72	5.0	38	8.6	104	2770
	Min	3.50	0.0	4.00	0.0	0	0.0	0	252
	Max	4.97	5.0	6.00	23.0	180	35.0	663	18281
	t	0.0	8.4☆☆☆	0.1	0.1	1.3	−0.0	0.7	1.7†

† p<.10.
☆ p<.05.
☆☆ p<.01.
☆☆☆ p<.001.

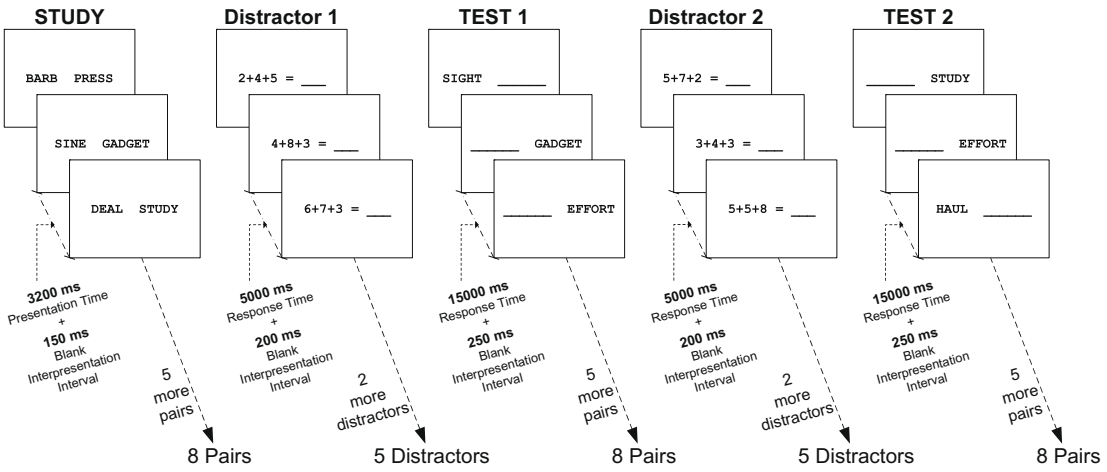

Fig. 2. A single set in the task. Each box illustrates the computer screen at a particular stage in the task (text has been enlarged relative to the screen size to improve clarity of the figure). Each phase was directly followed by the next of the five phases, without pause. In both of the test phases, each pair presented during the study phase was tested only once, half of which were in the forward direction.

sponses. Both response initiation and termination (“ENTER” key press) were logged. Response time measures yielded no additional information (e.g., no speed-accuracy trade-off) and as such will not be discussed further.

All pairs were tested twice in cued recall (see Fig. 2), following from the successive testing method suggested in Kahana (2002). In each test, half of the pairs were tested in the forward direction, while half of the pairs were tested in the backward direction. Testing direction was counterbalanced over the two tests such that half of the pairs were tested in the same direction over both tests and half of the pairs were tested in different directions each time.

At the end of each set, the task paused briefly and the participant was instructed to press “ENTER” to begin the next set. The task was designed using the Python programming language and the pyEPL experimental library (Geller, Schleifer, Sederberg, Jacobs, & Kahana, 2007).

Responses were also analyzed using a common spell-checking search algorithm used by the UNIX program aspell (Philips, 1990; Philips, 2000). All incorrect responses were

processed by the algorithm and were marked as corrected if the correct response was found in the list of possible corrections. Since analyses with responses both before and after spell-checking were not substantially different, we report only analyses using the strict spelling criterion for accuracy.

All analyses are reported with Greenhouse–Geisser correction for non-sphericity where appropriate. Effects were considered significant based on an alpha level of 0.05 and post-hoc pair-wise comparisons are always Bonferroni-corrected. Non-significant 'trend' effects ($p < .1$) are also reported.

Results

Cued recall accuracy

For both the word frequency and imageability sessions, repeated-measures ANOVAs were conducted on mean accuracy. Table 2 gives an overview of the pair types and their respective probe and target items.

Imageability session

Mean accuracy as a function of pair type and test direction is presented in Fig. 3a. Accuracy in the various conditions is as follows: (a) In HH pairs, for both forward and backward test directions, the probe and target item are both high-imageability items. In the forward direction, accuracy was .56, while in the backward direction accuracy was .55. (b) In HL pairs in the forward direction the probe item is high-imageability while the target item is low-imageability, accuracy was found to be .44. In the backward direction, when the probe is low-imageability and the target is high-imageability, accuracy was .37. (b) In LH pairs, the forward direction provides low-imageability probe and requires a high-imageability target, while in the backward direction the opposite is true. Accuracy in these instances was found to be .41 and .42, respectively. (d) In LL pairs, testing in both the forward and backward directions relies on a low-imageabilty probe and a low-imageability target. In both testing directions, accuracy was .33.

Table 2
Factorial design of the pair types and test directions used in our study. Types of probe, relationship, and target are listed for all possible pair type × testing direction combinations.

Pair type	Testing direction	Probe	Relationship	Target
HH	Forward	H	HH (Pure)	H
	Backward	H	HH (Pure)	H
HL	Forward	H	HL (Mixed)	L
	Backward	L	HL (Mixed)	H
LH	Forward	L	LH (Mixed)	H
	Backward	H	LH (Mixed)	L
LL	Forward	L	LL (Pure)	L
	Backward	L	LL (Pure)	L

We performed a mixed $4 \times 2 \times 2 \times 2$ repeated-measures ANOVA using the within-subjects factors PAIR TYPE (HH, HL, LH, LL), TEST DIRECTION (Forward, Backward), TEST NUMBER (Test 1, Test 2), and the between-subjects factor SESSION ORDER (Imageability first, Word Frequency first). Our main interest was whether mixed pairs were asymmetric and whether pair types differed in accuracy.

Participants were more accurate on Test 2 than Test 1 [$F(1, 54) = 34.2, MSe = 0.005, p < .001$], but none of the interactions with TEST NUMBER were significant [all F's < .5]. The main effect of PAIR TYPE was significant [$F(3, 153) = 49.9, MSe = 0.021, p < .001$]. Post-hoc pair-wise comparisons revealed that HH pairs were significantly more accurate than LL pairs [$p < .001$]. HH pairs were also significantly more accurate than both mixed pair types [both p's < .001]. Mixed pair types were not different [$p > .5$]. HL and LH pairs were both significantly more accurate than LL pairs [both p's < .001]. The PAIR TYPE × TEST DIRECTION interaction was significant [$F(3, 145) = 4.3, MSe = 0.022, p < .001$]. Post-hoc analyses reveal that the interaction was driven by HL pairs, where accuracy is higher for the forward probes than the backward probes of the pair [$p < .05$]. Symmetric mean performance in the pure pairs replicates numerous findings (Bower, 1972; Crowder, 1976; Paivio, 1971; Wollen & Lowry, 1971).

To more directly test the pure item- and association-effect models, we conducted a second repeated-measures

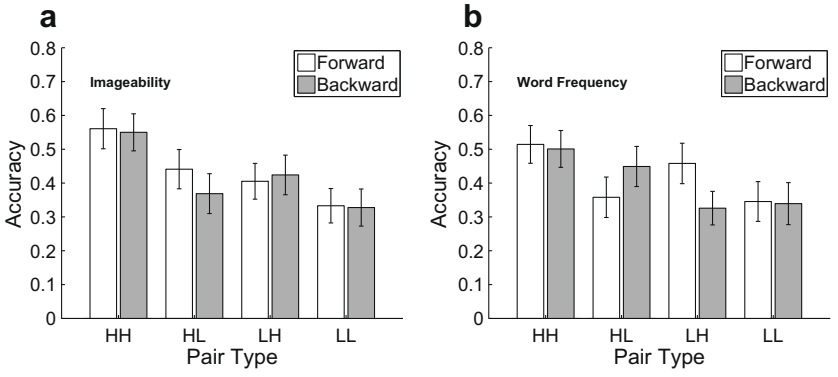

Fig. 3. Cued recall accuracy for both sessions of Experiment 1a. Pair types: high–high (HH), high–low (HL), low–high (LH), and low–low (LL). Error bars are 95% confidence intervals of participant mean performance. Since there were no significant interactions with TEST NUMBER we collapsed across Test 1 and Test 2. Accuracy as a function of condition for: (a) the imageability manipulation; and (b) the word frequency manipulation. Please see Table 2 for an overview of the pair types and test directions used in the study.

ANOVA with the design TARGET[2] × ASSOCIATION[2] and accuracy as the measure. TARGET was either 'high' (e.g., HL in the backward direction) or 'low' (i.e., HL in the forward direction), ASSOCIATION was either 'pure' (HH and LL) or 'mixed' (HL and LH). If imageability influenced cued recall solely by enhancing single-item memory, there would be only a main effect of TARGET, with no other significant effects (as in Fig. 1a). Conversely, the pure association-effect model (Fig. 1b) would produce an interaction of TARGET × ASSOCIATION only. However TARGET $[F(1,54) = 5.0, MSe = 0.012, p < .05]$ (H > L) and ASSOCIATION $[F(1,54) = 131.5, MSe = 0.008, p < .001]$ (pure > mixed) were both significant main effects, and the interaction was also significant $[F(1,54) = 57.3, MSe = 0.008, p < .001]$ (H–Pure > H–Mixed = L–Mixed > L–Pure). This suggests that a dual-effect model (imageability influencing both memory for items and memory for associations) is necessary to explain the full accuracy pattern.

Word frequency session

Mean accuracy for each pair type in each testing direction is presented in Fig. 3b. We report the mean accuracy to be as follows: (a) In HH pairs, recall accuracy in the forward and backward directions was .51 and .50, respectively.[2] (b) In HL pairs, when given a H item and asked to recall a L item (forward testing), accuracy was .36. In backward testing, accuracy was .45. (c) In LH pairs, forward testing accuracy was .46, while backward testing accuracy was .33. (d) In LL pairs, both the probe and target items are low-frequency. In the forward testing direction, accuracy was .35. In the backward testing direction, accuracy was .34.

We again performed a mixed $4 \times 2 \times 2 \times 2$ repeated-measures ANOVA, analogous to that done for the imageability session. Again, Test 2 outperformed Test 1 $[F(1,54) = 15.1, MSe = 0.008, p < .001]$, but no significant interactions involving TEST NUMBER were found. There was a main effect of PAIR TYPE $[F(2,130) = 23.8, MSe = 0.028, p < .001]$, as well as an interaction of PAIR TYPE × TEST DIRECTION $[F(3,149) = 23.7, MSe = 0.011, p < .001]$. Follow-up pairwise comparisons found that HH and LL were significantly different from the mixed pairs. Pure pairs (HH and LL) produced symmetric accuracy (equivalent forward and backward performance), replicating prior research (Crowder, 1976; Nelson & McEvoy, 2000; Paivio, 1971). HH pairs were more accurately recalled than LL pairs. In contrast, mixed pairs were recalled asymmetrically: there was a backward-probe advantage for HL pairs and a forward-probe advantage for LH pairs (Fig. 3b), similar to prior findings of a high-frequency recall advantage in item-memory tests (e.g., Gregg, 1976). Thus, when the target is a high-frequency word (see HL–Backward and LH–Forward in Fig. 3b) the accuracy is higher than that for a low-frequency word (see HL–Forward and LH–Backward in Fig. 3b).

For the TARGET [2] × ASSOCIATION [2] ANOVA, the main effect of TARGET was significant $[F(1,54) = 5.3, MSe = 0.008, p < .05]$ with high-frequency targets recalled more than low-frequency targets H > L. The main effect of ASSOCIATION was not significant. However, the interaction between the two factors was significant $[F(1,54) = 129.8, MSe = 0.008, p < .001]$ (H–Pure = L–Mixed > H–Mixed = L–Pure), suggesting that a dual-effect (hybrid) model would be required to account for the full pattern of results.

Intrusions

As suggested in the Introduction, intrusion rates can be used as a measure of item retrievability (e.g., sampling probability – how likely an item is to be sampled from one's lexicon). Intrusion rates are reported as means across all participants.

Imageability session

When participants responded incorrectly during cued recall in the imageability session, they had an equal proportion of high imageability and low-imageability word intrusions $[t(55) = 0.34, p > .5$; High: $M = .067$, Low: $M = .064]$. These results show that the manipulation of imageability had no significant effect on intrusion rates. Hence, based on the intrusion rates, there was no evidence that imageability has an effect on item retrievability.

Word frequency session

In contrast, when participants responded incorrectly during cued recall in the word frequency session, they were more likely to respond with a high-frequency word than a low-frequency word $[t(55) = 3.12, p < .01$; High: $M = .084$, Low: $M = .059]$. These findings show that when incorrect, participants are more likely to recall a high-frequency word than a low-frequency word. Additionally, this suggests that high-frequency words are retrieved more easily than low-frequency words.

Modeling cued recall accuracy

To quantify the relative effects of item properties on item recall probability versus association-memory, we fit the mean accuracy data with a simple model. In this "item-relationship" model, we assume that successful cued recall requires successful access of three separate mechanisms in order to recall the correct item, similar to previous multi-step models of association-memory (Kelley & Wixted, 2001; McGuire, 1961), as follows. In the model, accuracy as a function of pair type and test direction, *Acc*(Pair Type, Test Direction), is the product of the probability of effectively accessing the probe item, the probability of effectively retrieving the association (including having encoded the association well), and the probability of effectively producing the target item:

$$Acc(\text{PairType}, \text{TestDirection}) = P(Probe_i) \times P(Relat_j) \times P(Target_k) \quad (1)$$

[2] While we noted that HH pairs were more semantically similar than mixed and LL pairs in the word frequency session, it does not appear that this substantially affected mean performance in our study. In HH pairs, collapsing across test directions, mean accuracy was .51. Consider the following: The highest $cos(\theta)$ for LL pairs that was tested behaviourally was .36. If we restrict our analyses to HH pairs with equal or less intra-pair similarity than this maximum (e.g., only HH pairs with $cos(\theta) < .36$), the mean performance was only reduced to .50. This suggests that the difference in $cos(\theta)$ does not affect the qualitative nature of the results we report.

where $P(Probe_i)$ and $P(Target_k)$ denote the probabilities of effectively handling the probe item and effectively retrieving the target item, respectively, where $i = \mathrm{H}, \mathrm{L}$ and $k = \mathrm{H}, \mathrm{L}$. $P(Relat_j)$ denotes the probability of retrieving the pair depending on the relationship between the two items, where $j = \mathrm{HH}, \mathrm{HL}, \mathrm{LH}, \mathrm{LL}$. This results in the following system of equations:

$$
\begin{aligned}
Acc(\mathrm{HH}, \mathrm{Forward}) &= P(Probe_{\mathrm{H}}) \times P(Relat_{\mathrm{HH}}) \times P(Target_{\mathrm{H}}) \\
Acc(\mathrm{HH}, \mathrm{Backward}) &= P(Probe_{\mathrm{H}}) \times P(Relat_{\mathrm{HH}}) \times P(Target_{\mathrm{H}}) \\
Acc(\mathrm{HL}, \mathrm{Forward}) &= P(Probe_{\mathrm{H}}) \times P(Relat_{\mathrm{HL}}) \times P(Target_{\mathrm{L}}) \\
Acc(\mathrm{HL}, \mathrm{Backward}) &= P(Probe_{\mathrm{L}}) \times P(Relat_{\mathrm{HL}}) \times P(Target_{\mathrm{H}}) \\
Acc(\mathrm{LH}, \mathrm{Forward}) &= P(Probe_{\mathrm{L}}) \times P(Relat_{\mathrm{LH}}) \times P(Target_{\mathrm{H}}) \\
Acc(\mathrm{LH}, \mathrm{Backward}) &= P(Probe_{\mathrm{H}}) \times P(Relat_{\mathrm{LH}}) \times P(Target_{\mathrm{L}}) \\
Acc(\mathrm{LL}, \mathrm{Forward}) &= P(Probe_{\mathrm{L}}) \times P(Relat_{\mathrm{LL}}) \times P(Target_{\mathrm{L}}) \\
Acc(\mathrm{LL}, \mathrm{Backward}) &= P(Probe_{\mathrm{L}}) \times P(Relat_{\mathrm{LL}}) \times P(Target_{\mathrm{L}})
\end{aligned}
$$

(Note that these equations parallel Table 2.)

Because our focus is on the relative effects of stimulus properties on each of these three stages, we define the following parameters which will be used as free parameters in the model fits:

$$p = \frac{P(Probe_{\mathrm{H}})}{P(Probe_{\mathrm{L}})} \tag{2}$$

$$r_1 = \frac{P(Relat_{\mathrm{HH}})}{P(Relat_{\mathrm{HL,LH}})} \tag{3}$$

$$r_2 = \frac{P(Relat_{\mathrm{HL,LH}})}{P(Relat_{\mathrm{LL}})} \tag{4}$$

$$t = \frac{P(Target_{\mathrm{H}})}{P(Target_{\mathrm{L}})} \tag{5}$$

Note that if any of these parameters has the value 1, this would represent a null effect of stimulus property on the respective stage of the model. Thus far, our item-relationship model is underdetermined (there are multiple ways to explain the data using various combinations of parameters). For this reason, we only worked with further-constrained model variants wherein some subset of the parameters p, r_1, r_2, and t were fixed to 1 and from one to three parameters were free at a time.

After constraining the model, the system of equations could be solved algebraically for each participant and then parameter values and model fits were summarized across participants. An additional tuning parameter was derived algebraically in order to properly scale the model fits to the behavioural data. We calculated both *AIC* and *BIC* as measures of model fitness.[3]

[3] *AIC* and *BIC* refer to the Akaike Information Criterion and Bayesian Information Criterion, respectively. Both *AIC* and *BIC* are measures of model fitness, but also take into account the number of free parameters (degrees of freedom). In both cases, lower is better and the absolute measures are meaningless, thus we report all scores as ΔAIC and ΔBIC relative to the best-fitting model considered. As a rule of thumb, if the difference between two model fits is less than two, neither of the models' fit to the data is significantly better. Here we used the "special case of least-squares estimation with normally distributed errors" variant of the *AIC*/*BIC* formulas using all participant × condition combinations (Burnham & Anderson, 2002; Burnham & Anderson, 2004).

Approach to model selection

We started by considering three highly constrained models, assuming that the respective item property affected either only item-memory (e.g., item retrieval; Fig. 1a) or only association-memory (Fig. 1b).

In the first highly constrained model, we test how much of the cued recall performance is accounted for by only item retrieval effects in our target-only model (only the t parameter). In the second model we further test models of pure item-memory in a probe-only model (p parameter). In the third highly constrained model, we test the relationship-only model, $p \equiv t \equiv 1$, leaving two free parameters, r_1 and r_2. The pure item-effect models had fewer degrees of freedom, but we focus on whether each model captures the qualitative features of the data and account for this difference in our model selection process through the use of ΔAIC and ΔBIC when fitting the model to the previously presented behavioural findings.

We then tested three hybrid models. One hybrid model contained the two item retrieval parameters of p and t (two free parameters). The final two hybrid variants involved the relationship parameters and one of the item parameters, which had as free parameters r_1 and r_2 in addition to either p or t (three free parameters in total).

Each model variant was fit to each participant individually. We report the 95% confidence intervals across participants for best-fitting parameter values.[4]

Imageability session

Refer to Fig. 4 and Table 3 for the complete set of model fits and their respective parameters, with the exception of the full model as it is underdetermined by the data.

We first tested our initial hypothesis, asking whether target-retrievability could account for the bulk of the behavioural pattern (target-only model; Fig. 4a). However, compared to all other model variants considered, this model fit by far the worst. Confirming the qualitative similarity of the data to the model examples shown previously, the best-fitting pure probe-effect and pure target-effect models fit worse than the pure relationship-effect model.

This suggests that the associations themselves were better learned and retrieved for pure high–high imageability pairs than mixed pairs (r_1) which, in turn, were retrieved better than low–low-imageable pairs (r_2). Note that due to model mimicry, both relationship + probe and relationship + target effect variants converge upon identical fits to the empirical data, though they employ differing mechanisms and thus different parameter fits (Fig. 4e). The probe + target model produced a worse fit than the hybrid models that also included the relationship.

In sum, the imageability accuracy pattern can be explained mainly by imageability enhancing memory for the association. This finding is also consistent with the results of the intrusion analysis.

[4] As all parameters are ratios, they center around a value of 1. To accurately compute symmetric 95% confidence intervals for the model parameters, the parameters for each participant were first log-transformed. We then calculated the 95% confidence intervals across participants before exponentially-transformed the fits back for reporting.

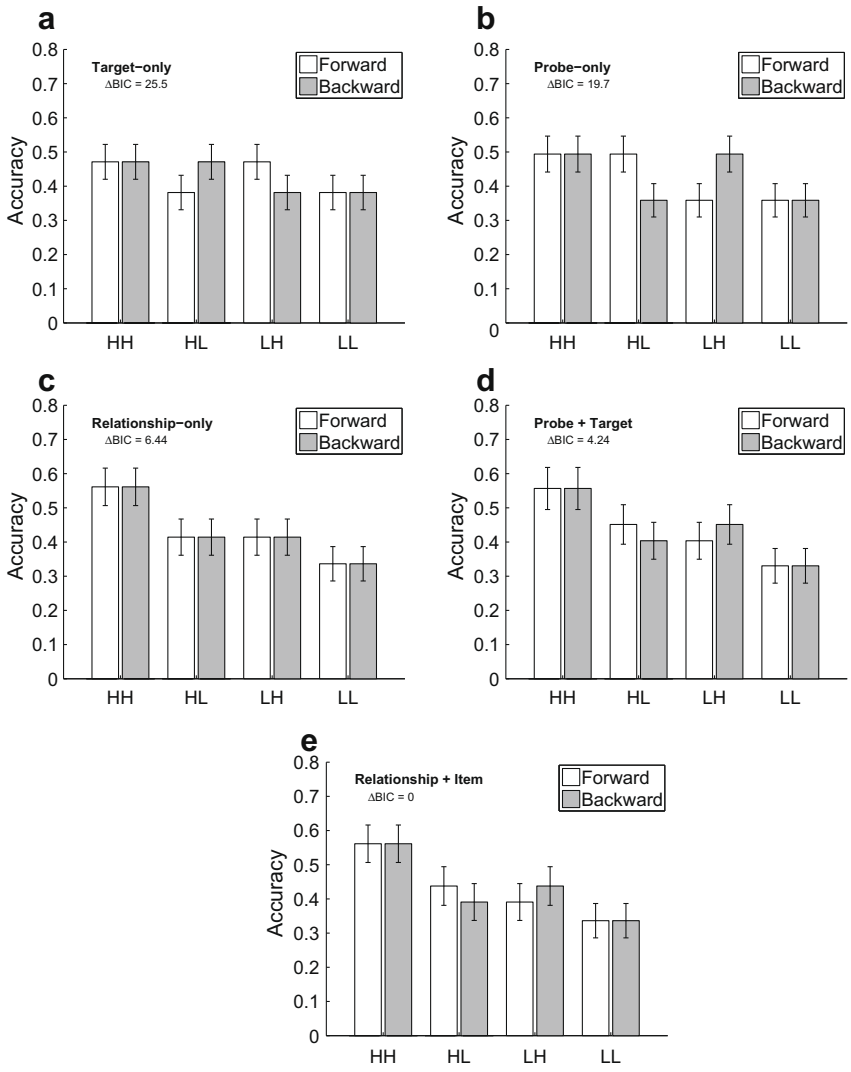

Fig. 4. Modeling of the mean accuracy in the imageability session. Pair types: high–high (HH), high–low (HL), low–high (LH), and low–low (LL). Error bars are 95% confidence intervals for model fits of participant mean performance (a) Target-only model ($t = [1.21, 1.42]$). (b) Probe-only model ($p = [1.35, 1.55]$). (c) Relationship-only model ($r_1 = [1.28, 1.58], r_2 = [1.14, 1.51]$). (d) Item-effect hybrid model (probe+target hybrid model: $p = [1.35, 1.55], t = [1.21, 1.42]$). (e) Relationship + item hybrid models (relationship + target hybrid: $r_1 = [1.36, 1.72], r_2 = [1.16, 1.57], t = [0.81, 1.01]$; relationship + probe hybrid: $p = [0.99, 1.23], r_1 = [1.22, 1.57], r_2 = [1.05, 1.42]$).

Word frequency session

Refer to Fig. 5 and Table 3 for the complete set of model fits and their respective parameters, with the exception of the full model as it is underdetermined by the data.

Here we again directly test our *a priori* hypothesis to see if cued recall performance directly follows from the known item-memory enhancement of word frequency. However, here the target-only model is by far the best-fitting single-effect model (Fig. 5a), in contrast to the imageability session.

An examination of the hybrid models shows that all models that include the t parameter fit the data quite well. However, none of the hybrid models are a substantially better fit than the others. The caveat to this approach is that the relationship + probe model does not allow the t parameter to vary from 1, but is quantitatively equivalent to the relationship + target model. Considering the fit of the target-only model, a strong argument could be made that the relationship + probe model is fairly implausible. For the relationship + target hybrid model, the relationship parameters, r_1 was found to be significantly different from 1, suggesting that the additional parameter does explain the behavioural performance better than the target-only model. (This is further addressed in the Discussion.) While the probe + target model was found to be the best-fitting model, we believe this is chiefly because it (a) contained the t parameter, and (b) included an additional parameter to help explain the mean performance, but only added one additional parameter (in contrast the relationship + target hybrid model).

Table 3
Model fits for both imageability and word frequency. All model variants are shown, with the exception of the full model (as it is underdetermined by the data). All free parameter fits are presented as 95% confidence intervals. Note that the relationship + target and relationship + probe algebraically produce identical fits due to model mimicry.

	ΔAIC	ΔBIC	p	r_1	r_2	t
Imageability manipulated						
Target-only	25.7	25.5	1	1	1	[1.21, 1.42]
Probe-only	19.8	19.7	[1.35, 1.55]	1	1	1
Relationship-only	6.52	6.44	1	[1.28, 1.58]	[1.14, 1.51]	1
Probe + Target	4.31	4.24	[1.35, 1.55]	1	1	[1.21, 1.42]
Relationship + Target	0	0	1	[1.36, 1.72]	[1.16, 1.57]	[0.81, 1.01]
Relationship + Probe	0	0	[0.99, 1.23]	[1.22, 1.57]	[1.05, 1.42]	1
Word frequency manipulated						
Target-only	7.08	7.00	1	1	1	[1.43, 1.80]
Probe-only	26.5	26.4	[1.03, 1.30]	1	1	1
Relationship-only	19.6	19.5	1	[1.18, 1.47]	[1.17, 1.47]	1
Probe + Target	0	0	[1.03, 1.30]	1	1	[1.43, 1.80]
Relationship + Target	0.56	0.64	1	[1.01, 1.29]	[0.95, 1.23]	[1.28, 1.52]
Relationship + Probe	0.56	0.64	[0.66, 0.78]	[1.41, 1.81]	[1.34, 1.69]	1

Several interpretations of the data are possible with the present modeling framework. All but one of these models includes a t parameter that is allowed to freely vary from 1. The model that does not encompass this we suggest is less plausible because it conflicts the most with the cued-recall pattern, as well as the previously established high-frequency advantage for item recall. Finally, the model fits suggest that even if item-retrievability (t parameter) explains part of the high-frequency advantage in cued recall, an additional process is needed to account for the full pattern of behaviour.

Correlation of accuracy on successive tests

To test for the holistic property of memory for pairs (high correlation between forward and backward probe accuracy), we calculated Yule's $\mathcal{Q}$ as our measure of correlation between Test 1 and Test 2 probes. Yule's $\mathcal{Q}$ is a correlation measure appropriate for dichotomous data (for a review, see Kahana, 2002). When calculating $\mathcal{Q}$, we collapsed across participants and collected all the raw data into a single contingency table.[5] $\mathcal{Q}$ was then log-odds ratio transformed for statistical tests (Bishop, Fienberg, & Holland, 1975; Hayman & Tulving, 1989).

We calculated three kinds of correlation between Test 1 and Test 2 performance. The "Same" correlation estimates the highest correlation (i.e., due to test-retest reliability) by calculating the correlation between Test 1 and Test 2 when both probes were in the same direction (Forward–Forward and Backward–Backward). The "Different" correlation represents the correlation between forward and backward tests. This correlation is our measure of interest as it compares cued recall of the pair in both directions; thus it is our test of associative symmetry. The "Control" correlation, introduced by Caplan (2005), is a bootstrap which estimates the lowest possible expected correlation by measuring the correlation between unrelated pairs within the same set, one pair from Test 1 and a different pair from Test 2. This controls for subject and study-set variability (Simpson's Paradox; cf. Hintzman, 1980). Thus, "Same" and "Control" set the effective range of the "Different" correlation.

Participants did perform better on Test 2 than on Test 1, likely due to output-encoding. Similar findings have also been reported by Rizzuto and Kahana (2000), Rizzuto and Kahana (2001) and Sommer et al. (2007) using related successive testing methods. However, both of these prior groups found that although the testing effect increased the "Different" correlation slightly, the high correlation is not largely due to the testing effect.

As illustrated in Fig. 6a (imageability session) and Fig. 6b (frequency session), pair-wise comparisons revealed no significant differences between the "Same" correlation across pair types ($p > .1$). This is also true between all of the "Different" correlations. However, within each pair type "Same" correlations were higher than "Different" correlations (all p's < .01). All "Same" and "Different" correlations were also significantly higher than the "Control" correlation ($p < .01$).

Importantly, we did not find that the successive testing correlations were reduced in any of the pair types. The extremely high "Same" correlation (testing test-retest reliability) is typical of paired associate experiments (Caplan, 2005; Caplan et al., 2006; Rizzuto & Kahana, 2000; Rizzuto & Kahana, 2001). Following Caplan (2005), we median-split the participants based on their average "Same" correlation and for the low-same-correlation group we still fail to find a significant interaction [$p > .1$ for both sessions]. This suggests that the holistic nature of the associations was not significantly disrupted in any of the conditions.

Relationship between mean accuracy asymmetry and forward–backward correlation

To ask whether the mean accuracy was associated with a disruption of the holistic property (high $\mathcal{Q}_{\text{DIFFERENT}}$) we exploited individual variability. We calculated Spearman's

[5] We initially also computed the log-odds transformed Yule's $\mathcal{Q}$ value for each participant individually. However, here the "Different" correlation was much more variable due to too few data points for each participant. Nonetheless, we still found no significant differences in correlations across pair types.

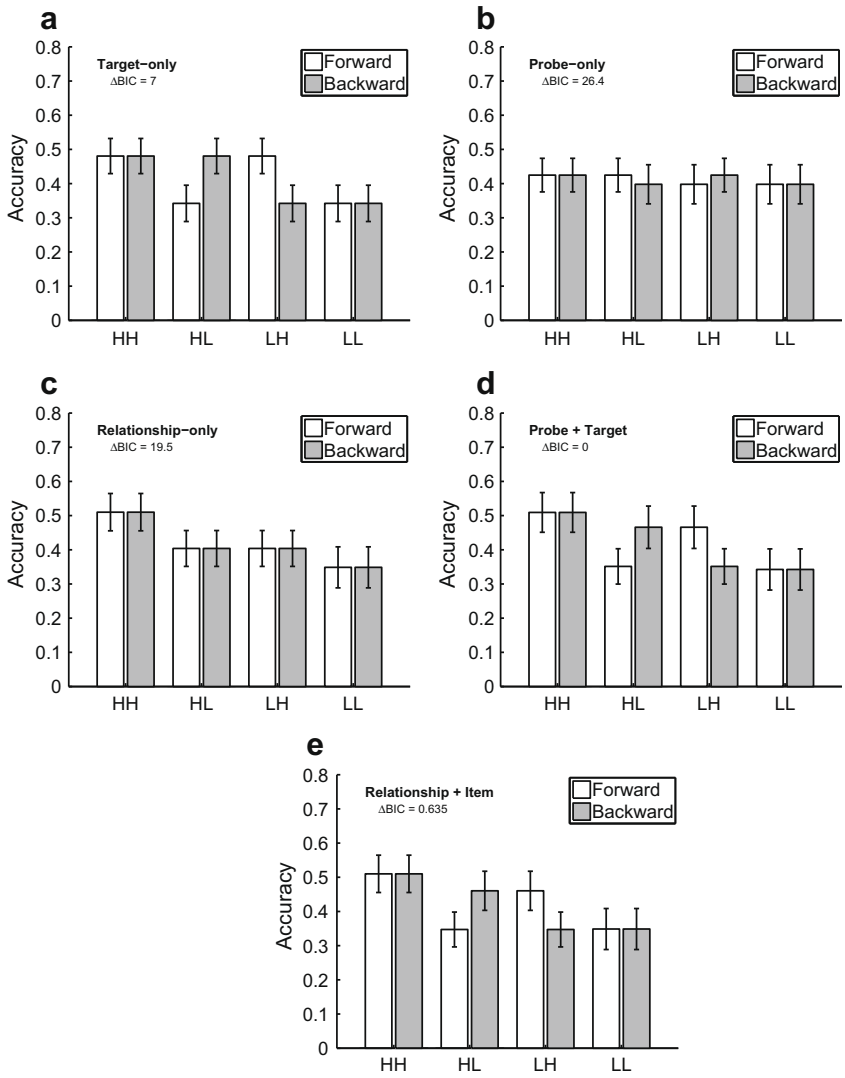

Fig. 5. Modeling of the mean accuracy in the word frequency session. Pair types: high–high (HH), high–low (HL), low–high (LH), and low–low (LL). Error bars are 95% confidence intervals for model fits of participant mean performance. (a) Target-only model ($t = [1.43, 1.80]$). (b) Probe-only model ($p = [1.03, 1.30]$). (c) Relationship-only model ($r_1 = [1.18, 1.47], r_2 = [1.17, 1.47]$). (d) Item-effect hybrid model (probe + target hybrid model: $p = [1.03, 1.30], t = [1.43, 1.80]$). (e) Relationship + item hybrid models (relationship + target hybrid: $r_1 = [1.01, 1.29], r_2 = [0.95, 1.23], t = [1.28, 1.52]$; relationship + probe hybrid: $p = [0.66, 0.78], r_1 = [1.41, 1.81], r_2 = [1.34, 1.69]$).

rank correlation (ρ) between a measure of mean-asymmetry ($|(\text{Forward}-\text{Backward})/(\text{Forward}+\text{Backward})|$) and a measure of holistic-disruption ($\mathcal{Q}_{\text{DIFFERENT}}/\mathcal{Q}_{\text{SAME}}$), calculated for each participant.[6] When calculating the Q values for individual participants, there were frequently cells with missing data, which required a correction for these missing values. To avoid division by zero errors when calculating Q we add half an observation to each cell.

The correlation between mean-asymmetry and holistic-disruption was not significant for any of the pair types in either the imageability [HH: $\rho(55) = -.22$; HL: $\rho(55) = .13$; LH: $\rho(55) = -.06$; LL: $\rho(55) = -.14$; all p's > .1] or word frequency sessions [HH: $\rho(55) = .13$; HL: $\rho(55) = -.31$; LH: $\rho(55) = -.04$; LL: $\rho(55) = -.01$; all p's >.1]. Thus, asymmetric mean accuracy did not reliably indicate a disruption of holistic learning.

Experiment 1b

Paivio (1971) held that low–low-imageability pairs would not be learned holistically. In Experiment 1a, we were unable to detect a disruption of holistic pair learning (using the successive testing measure), for the low–low-imageable

[6] Given that it would not make sense to average $\mathcal{Q}$ values and the "same" correlation is close to unity, we felt that the ratio of the "Different" $\mathcal{Q}$ value and "Same" $\mathcal{Q}$ value was the most sensible measure. Here we are using the "Same" $\mathcal{Q}$ value as a reference point to control for test-retest reliability and thus measure "different" $\mathcal{Q}$ value proportionally to the "Same" $\mathcal{Q}$ value. However, when comparing performance in the forward and backward test directions, neither is being used as a reference point and instead we need to control for overall accuracy effects.

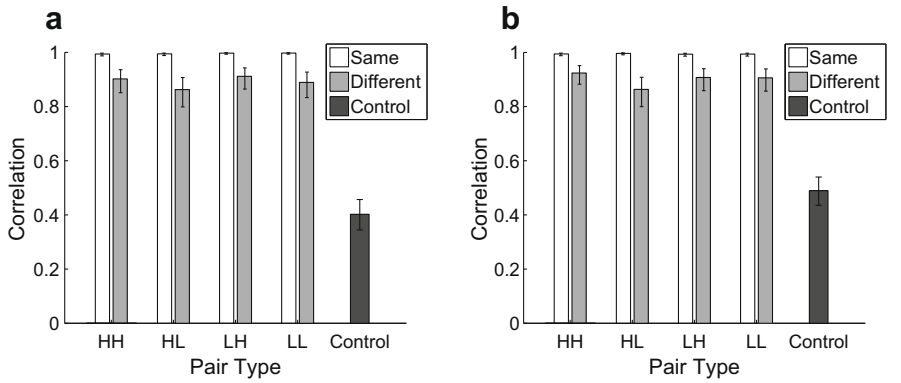

Fig. 6. Correlation (Yule's $\mathcal{Q}$) of Test 1 and Test 2 performance. "Same" denotes correlations between successive testing when both tests were in the same direction (Forward–Forward and Backward–Backward). "Different" denotes correlations when the two tests were in different directions (Forward–Backward and Backward–Forward). "Control" denotes the correlation between Test 1 and Test 2 for unrelated pairs presented in the same study set. Error bars are 95% confidence intervals. (a) Test 1–Test 2 correlation for the imageability manipulation. (b) Test 1–Test 2 correlation for the word frequency manipulation.

condition. However, Sommer et al. (2007) found that for object–location pairs, if the presentation of the two items was sequential rather than simultaneous, which increasingly challenges the participant to link the paired items, it is possible to induce a disruption in the forward–backward correlation. By adapting Experiment 1a to use a sequential presentation and only pure high- and low-imageability pairs, we further tested the conceptual-peg hypothesis. Using only pure high- and low-imageability pairs increased sensitivity to a possible disruption of holistic learning.

Methods

Participants

Forty-one undergraduate students participated in the one-session follow-up experiment for partial fulfillment of an introductory psychology course requirement (mean age ± *sd* = 20.9 ± 4.9; 11 male and 30 female). None of the participants from Experiment 1a participated in Experiment 1b.

Materials

The materials were identical to those in Experiment 1a; however, only the high- and low-imageability pools were used.

Procedure

The procedure was identical to the imageability session of Experiment 1a except the presentation of the paired items was sequential as opposed to simultaneous. Also there was no word frequency session in this experiment. Only pure high- and low-imageability pairs were presented. Each noun was presented in the center of the screen for 2000 ms, with a blank intra-pair interval of 50 ms and an inter-pair interval of 1000 ms.

Results

The analysis methods for the second experiment were nearly identical to those used in Experiment 1a. However, because only pure pairs and one experimental session were used, the ANOVA had the design PAIR TYPE[2] × TEST DIRECTION[2] × TEST NUMBER[2]. Model fits could not be carried out as there were no mixed pairs.

Cued recall accuracy

Mean accuracy as a function of pair type and test direction is plotted in Fig. 7. The main effect of TEST NUMBER was significant $[F(1,39) = 15.8, MSe = 0.064, p < .001]$. The main effect of PAIR TYPE was significant $[F(1,40) = 110.9, MSe = 0.018, p < .001]$. The main effect of PAIR TYPE was due to HH pairs begin more accurately recalled than LL pairs, replicating the findings of Experiment 1a.

Correlation of accuracy on successive tests

As evident in Fig. 8, pair-wise comparisons found that for both HH and LL pair types, "Same", "Different", and "Control" correlations were all significantly different [all *p*'s < .001]. The "Same" correlation for the HH pairs was no different than that for LL pairs. The "Different" correlation for HH pairs was significantly higher for HH pairs than for LL pairs $[t(40) = 2.01, p < .05]$. Note, however, that this

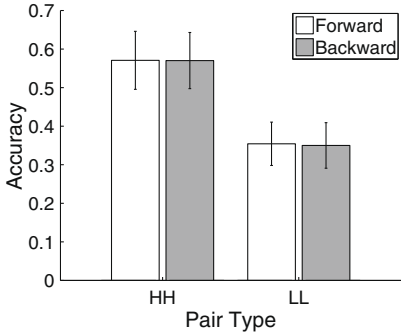

Fig. 7. Cued recall accuracy in the imageability session of Experiment 1b. Pair types: high–high (HH) and low–low (LL). Error bars are 95% confidence intervals of participant mean performance.

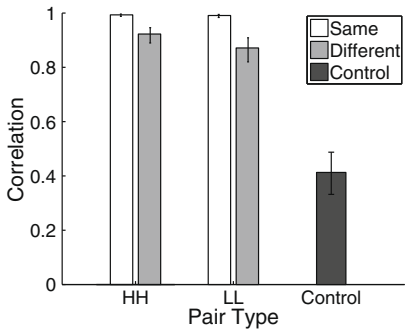

Fig. 8. Correlation (Yule's $\mathcal{Q}$) of Test 1 and Test 2 performance in the imageability session of Experiment 1b. "Same" denotes correlations between successive testing when done in the same direction (Forward–Forward and Backward–Backward). "Different" denotes correlations when Test 1 and Test 2 were in different directions (Forward–Backward and Backward–Forward). "Control" denotes the correlation between unrelated pairs found in the same set. Error bars are 95% confidence intervals of participant mean performance.

difference, while statistically significant, was small in magnitude. (95% confidence intervals for "Different" correlations: HH: $\mathcal{Q}(40) = [.89, .96]$, LL: $\mathcal{Q}(40) = [.82, .92]$.)

Discussion

In two related experiments, we investigated the effects of two item properties, imageability and word frequency, on cued recall performance. Imageability primarily influenced memory for the associations, whereas word frequency primarily influenced target-item recall. Additionally, we replicated the finding of a high correlation between forward and backward cued recall accuracy, which follows from the notion that pairs are learned holistically, as opposed to forward and backward associations being learned in separate steps. This correlation remained high in the face of asymmetric mean performance (Experiment 1a, mixed pairs, word frequency manipulation) and low-imageability pairs presented simultaneously (Experiment 1a), as well as sequentially (Experiment 1b). This suggests that holistic learning of pairs does not only apply to symmetric pairs or pairs that are readily conducive to imagery. We now discuss the implications of each of these findings.

Effects of item properties on item- versus association-memory

Evident in the ANOVAs and model fits, no single-effect model provided a satisfying account of the observed mean accuracy patterns. For both imageability and word frequency manipulations, a hybrid model fit the data better, implicating both item- and association-memory effects, quantitatively and qualitatively. Nonetheless, the primary effect of word frequency and imageability manipulations differed. For the word frequency manipulation, the bulk of the variability across pair type and test direction was best described by the target-only model. For the imageability manipulation, the bulk of the variability across pair type and test direction was better described by the association-only model. Beyond specifying the effects of imageability and word frequency *per se*, these distinct empirical patterns demonstrate a range of possible effects of item properties on cued recall performance and suggest that our combination of experimental design and model-based data analysis could be applied to elucidate the effects of a range of additional stimulus properties on memory for associations in future studies.

To appreciate the possible multifaceted effects of word properties, we fit all possible model variants except the full model as it is underdetermined by the data. Furthermore, we implemented *AIC* and *BIC* as blind model selection criteria, prior to drawing upon additional sources of insight into model plausibility, such as cued recall error measures, converging evidence from previous studies, or even additional memory tests (e.g., associative recognition). The present modeling approach was informative in the following specific ways.

Mechanism of imageability enhancement of association-memory

In the present study, manipulations of imageability presented as enhancements of associative memory (r_1 and r_2 parameters in our model), rather than the enhancements of item-memory. Analyses of cued recall errors also suggest that imageability does not affect item retrieval in the paired-associate paradigm used in this study; but naturally, if the association is more retrievable, the target item will be retrieved more often. Additionally, our finding that imageability is primarily reliant on association-memory converges with previous research that has used associative recognition as a test of memory rather than cued recall (Hockley, 1994; Hockley & Cristi, 1996).

Furthermore, research manipulating imagery instructions suggests that imagery depends on association-memory in paradigms similar to our own. Imageabilty of items has been previously shown to enhance cued recall performance (Lockhart, 1969; Paivio, 1965; Paivio, 1968; Paivio et al., 1968; Wood, 1967) and instructions to use imagery also enhance cued recall performance (Bower, 1970; Bower, 1972; Foth, 1973). However, this enhancement seems to rely on items being combined within an image rather than simply forming a separate image to each constituent item ("separation imagery", Hockley & Cristi, 1996). Thus, it may be the combination of items within an image that produces the relational memory enhancement though this does not separately enhance the retrievability of the high-imageable constituent items themselves (our model's t parameter).

Mechanism of word frequency enhancement of item-memory

Research using manipulations of word frequency in associative recognition, a more direct test of association-memory, have found no significant differences due to word frequency manipulations (Clark, 1992; Hockley, 1994; Hockley & Cristi, 1996).

Our modeling results suggest that cued recall performance is almost exclusively driven by the word frequency of the target word ($t > 1$). Our hybrid models all fit the data equally well, causing us to instead look to convergent evidence in order to further evaluate the plausibility of

each model. As discussed earlier, the relationship + probe model is the least plausible account as it does not implicate better target retrievability for high-frequency words. One argument in favor of the relationship + target model is that we do find a boost of associative memory ($r_1 > 1$ in the best-fitting model). Previous research has also suggested that it may be easier to learn word pairs where both words are high-frequency (e.g., Clark & Shiffrin, 1992; Guttentag & Carroll, 1997), which would explain the enhancement of the r_1 parameter in the best-fitting relationship + target hybrid model. This also converges with associative recognition research using high-frequency and very low-frequency words, which suggests that high-frequency words have better associability than very low-frequency words (Chalmers & Humphreys, 2003).

Another possible argument is that word frequency might act as a proxy for contextual distinctiveness (e.g., high-frequency words are more likely to appear in a variety of contexts than low-frequency words; see Nelson & McEvoy, 2000, for an in-depth discussion). This argument directly follows from previous studies finding that high-frequency words are better recalled but worse recognized than low-frequency words (c.f. Gregg, 1976). Within our modeling approach, this argument would suggest a disadvantage for high-frequency probes ($p < 1$) or more competition between associations learned during the experiment and pre-existing associations with previously learned contexts for high-frequency words (r_1 and $r_2 < 1$). However, our findings are inconsistent with this feature driving our measure of memory. Nonetheless, our present findings suggest that manipulations of word frequency mainly affect item retrievability (t in our model) and minimal differences were found between the various hybrid models.

In converging research, Morton's (1969) logogen model and the 'Source of Activation Confusion' theory (SAC; Reder et al., 2000) also suggest advantages for high-frequency words. Logogens are recognition units or detectors responsible for identifying individual words (Morton, 1969). In the logogen model, less sensory evidence is required to identify a high-frequency word than a low-frequency word (Morrison & Ellis, 1995). In other words, low-frequency words have to reach a higher threshold than high-frequency words in order to be identified. This finding could also correspond to an enhanced ability of recalling high-frequency words. SAC theory is based upon our inability to determine the source of the activation when producing words; source must be inferred (Reder et al., 2000). By definition, high-frequency words have more pre-experimental presentations than low-frequency words; this could cause high-frequency words to have a higher baseline activation. This could further suggest that high-frequency words are easier and more likely recalled due to their increased availability (Diana & Reder, 2006; Reder et al., 2000). Our finding of high-frequency words intruding more than low-frequency words is consistent with both of these accounts.

One finding that is inconsistent with our word-frequency effect can be found in tests of serial recall when manipulating word frequency. Hulme et al. (2003) found that in serial recall, lists of alternating high and low-frequency words do not exhibit a zig-zag accuracy pattern, but instead present as a smooth curve. As mentioned in the Introduction, this result suggests word frequency is enhancing serial-order memory by strengthening inter-item associations, rather than through an enhancement of item-memory for the high-frequency words alone. However, in this study, the researchers also pre-familiarized participants with the words being used in the experiment, possibly diminishing the effect of word frequency on differential item accessibility. Furthermore, a study by Poirier and Saint-Aubin (1996) found that in immediate serial recall, word frequency enhanced item-memory (specifically item retrieval), but had no effect on memory for order. Kahana and Caplan (2002) suggested that asymmetric performance in cued recall of serial lists distinguished serial list-learning from paired-associate learning, challenging attempts to model both association- and serial-list-learning using the same model mechanisms (Caplan, 2004; Caplan, 2005; Caplan et al., 2006; Lewandowsky & Murdock, 1989).

Mean performance and forward–backward correlation are independent measures of association learning

We found symmetric mean performance overall (lack of main effects of TEST DIRECTION) and symmetry in pairs containing items drawn from the same pool and even in pairs combining high- and low-imageable words, replicating a large body of evidence that suggests that in general, memory for pairs is symmetric (Bower, 1972; Caplan et al., 2006; Crowder, 1976; Horowitz, Brown, & Weissbluth, 1964; Horowitz et al., 1966; Kahana, 2002; Paivio, 1971; Rehani & Caplan, in preparation; Rizzuto & Kahana, 2000; Rizzuto & Kahana, 2001; Sommer et al., 2007; Wollen & Lowry, 1971). Kahana (2002) argued that symmetry in mean performance does not directly support Associative Symmetry (Asch & Ebenholtz, 1962) and likewise, asymmetry in mean performance does not directly challenge Associative Symmetry. He introduced the correlation between forward and backward cued-recall performance, at the level of individual pairs, as a direct test of associative symmetry. Importantly, disruptions of holistic learning have previously been shown in sequentially presented object–location pairs (Sommer et al., 2007).

In the present findings, although some mixed pairs were recalled with asymmetric mean performance (Fig. 3b), such pairs were learned no less holistically (Fig. 6b). A caveat is necessary. The "Different" correlations were significantly lower than what might be expected from a perfectly holistic association (namely, one would expect it to be equal to the "Same" correlation). This is consistent with prior values of the forward–backward correlation (Caplan, 2005; Caplan et al., 2006; Kahana, 2002; Rehani & Caplan, in preparation; Rizzuto & Kahana, 2000; Rizzuto & Kahana, 2001). Nonetheless, the correlation is still quite high, nearly as high as the high end of the range set by the two boundary correlations (the "Same" and "Control" correlations). Thus, it is more accurate to conclude that, as in prior measures of the forward–backward correlation, the Associative Symmetry Hypothesis is not perfectly supported, but a close approximation of this hypothesis is supported. The key finding we report is that this high correlation is not disrupted by our manipulations.

Beyond the non-difference finding, it should be observed that if a difference between mixed and pure pairs exists (i.e., the present power is insufficient to detect differences), then according to the confidence intervals, the difference would have to be quite small. Thus, this difference would be smaller in magnitude than previous studies of associative symmetry have observed in cued recall portions of serial lists, though larger than those found in studies of pair learning (Caplan, 2005; Caplan et al., 2006).

The measures of mean symmetry and forward–backward correlation are distinct in principle, but our findings suggests they are also distinct in empirically observable human behaviour. We found that it is possible to manipulate (a) the strength of an association between paired items, (b) the strength of the paired items themselves, as well as (c) the symmetrically of mean accuracy when testing the pair itself – all without disrupting the correlation-measure of the holistic nature of the association. In addition, in individual-difference analyses, the mean performance measure and the correlation measure were not strongly correlated. These findings extend the boundary conditions of associative symmetry to asymmetric pairs and high- as well as low-imageability and high- and low-frequency word pairs. Holistic learning of pairs may be a general phenomenon in human paired-associate learning. Thus, constituent item-properties can affect overall item- and association-memory levels (mean performance modeling) but leave the nature of the association unaffected (the relationship between forward and backward associations).

The conceptual-peg hypothesis

Paivio's conceptual-peg hypothesis (Paivio, 1965; Paivio, 1971; Paivio et al., 1968) suggested that as long as one of the paired words is high-imageability, the pair can be learned holistically, through a single, holistic, interactive image. Thus, if a pair consists of two low-imageability words, the pair cannot be learned as a Gestalt. Contradicting this notion, we found that low–low-imageability pairs remained nearly as holistic as high–high-imageability and mixed pairs, despite not containing a highly imageable item to use as a 'peg.' While we did observe a slight reduction in the correlation when presentation was sequential (Experiment 1b), this was not as drastic as one would expect if imagery, and thus, the holistic code, were disrupted.

The p parameter in the model could be equivalently interpreted as the probability of the probe item accessing the learned association, along the lines of Paivio's suggestion that the higher imageable item ('peg') has preferred access to the Gestalt – the image that forms the basis of the learned association. In our model framework, the peg hypothesis would materialize as a pure-probe or relationship + probe hybrid model fitting the data optimally (and the p parameter fitting to values significantly different than 1). Our modeling results are inconsistent with this prediction. Although the relationship + probe hybrid model yields identical simulated data as the relationship + target hybrid model, the former yielded a p parameter value that was not significantly different than 1, failing to support the prediction of a reliable difference in association-access by high- versus low-imageable items.

Analogues of the ratio parameters in distributed memory models

Findings like those presented here will have important implications for possible loci of stimulus properties in memory models. The specific consequences will depend on the particular model. Most models formally de-couple item- and association-learning processes, including TODAM (Murdock, 1982; Murdock, 1983; Murdock, 1999), CHARM (Metcalfe Eich, 1982), and the Matrix Model (Humphreys, Bain, & Pike, 1989); see Clark and Gronlund (1996) for a review of the aforementioned memory models. While there are many possible ways to modulate item- and association-memory during both study and test in memory models, there are some general statements one can make. For instance, our p and t parameters depend on the properties of the probe and target items, respectively. In our procedure, because we probe unpredictably in both the forward and backward directions, the effects of stimulus properties on the p and t parameters would have to reflect phenomena that influence behaviour at test. In contrast, stimulus properties affecting the r_1 and r_2 parameters could exert their influence at study or test since the composition of the pair is known at study. Furthermore, for models that rely on associative mechanisms like the vector outer-product or convolution, a parameter that affects item representations might explain behavioural data that implicate modulations of the p or t as well as r parameters, but would be hard to reconcile with a behavioural pattern that included an effect on p or t but not on the r parameters. Other possible mechanisms include the orthogonality or similarity of the representations of words of one class or the other (potentially influencing r, p, and t parameters), differential encoding strength (r parameters), or different retrieval-strength thresholds for different stimulus types as well as influences on redintegration processes (t parameter).

Conclusion

Our findings elucidate the effects of item properties on cued recall in several ways: (a) Properties of single-items can either affect primarily item-memory (i.e., word frequency) or primarily memory for associations (i.e., imageability). (b) Single-item properties cannot by themselves disrupt the holistic-like association learning. (c) Models of association-memory must be able to accommodate overall highly correlated forward and backward associations and particular models may have multiple ways of accomodating the differential effects on item- versus association-memory effects.

Acknowledgments

We thank Chris Westbury for assistance in creating the word pools as well as Anthony R. McIntosh and Zainab Fatima for help with early planning of the experiments. Partly supported by Natural Sciences and Engineering Research Council (NSERC) of Canada.

Appendix A. Imageability manipulation word pools

High					Low				
AISLE	CABIN	FEAST	MOTOR	SQUASH	ACCORD	EVENT	JOIN	QUEUE	STRIFE
ANGLE	CANAL	FLAME	MUSEUM	STABLE	AFFAIR	EXCUSE	LUCK	QUOTE	SURGE
ANKLE	CARD	FLESH	ONION	STAIN	AIDE	FAKE	MALICE	RANK	TENURE
ARMOUR	CART	FLOCK	PARCEL	STAKE	AMITY	FARE	MIDST	RATIO	THIRST
AUTUMN	CAVE	FUEL	PLATE	STEEL	ASPECT	FATE	MINOR	REALM	TONE
BADGE	CHAP	GATE	POLE	STOOL	BLAME	FAULT	MODE	REFORM	TREATY
BARREL	CHAPEL	GIFT	PRINCE	STOVE	BLISS	FEAT	MOTIVE	REGARD	TRUST
BASKET	CHERRY	GUINEA	QUEEN	TAIL	BOON	FLANK	MUCK	REIGN	TURNER
BEAM	CIGAR	HAMMER	RIFLE	TANK	BOUND	FRIGHT	MUSLIM	REMARK	ULSTER
BEARD	CREST	HELMET	RIOT	TOILET	BOUT	FUND	ORIGIN	RUMOUR	UNEASE
BIBLE	CROWD	HUNTER	ROAST	TOMB	BREACH	FUSS	OUTPUT	SAKE	URGE
BIKE	CRUST	INFANT	ROCKET	TONGUE	BRIEF	GAIN	OUTSET	SAVE	VERGE
BILL	DECK	ISLE	ROPE	TROOPS	BRINK	GALE	PACE	SCORE	VIRTUE
BISHOP	DEER	ITALY	ROSE	TWIN	CHOSE	GRANT	PAUSE	SCORN	VOID
BLOUSE	DEVIL	LACE	RUBBER	TWIST	CLAIM	GREEK	PENCE	SERVE	WAIT
BOLT	DISC	LADIES	SALT	VEIL	CLAUSE	GUESS	PHASE	SHROUD	WAKE
BONE	DISH	LENS	SCARF	WOUND	COPE	HEATH	PHRASE	SLACK	WARD
BRANDY	DRIVE	LIMB	SKETCH	WRECK	COUP	HINDU	PLEA	SLIGHT	WISHES
BREAST	DRUM	LIMP	SKULL		DOSE	HIRE	PLIGHT	SOUL	
BUCKET	DUMMY	LINEN	SLOPE		DOUBLE	IDEAL	PLOY	SPAN	
BULLET	DWARF	LOCK	SMOKE		DREAD	INTENT	PRIOR	STANCE	
BURIAL	ELEVEN	LUNG	SPAIN		ENTRY	IRONY	PROOF	STAY	
BUTTON	ESSAY	MEAL	SPONGE		EQUITY	ISLAM	PROSE	STEIN	

Appendix B. Word frequency manipulation word pools

High					Low				
AREA	FELT	LOSS	ROAD	THIRD	ANNEX	COOP	HAUL	PAVE	TART
BACK	FLAT	MANY	ROLE	TIME	BALE	CRANK	HISS	PAWN	TEASE
BASE	FORCE	MASS	SAFE	TODAY	BARB	CZECH	HITCH	PEEP	THROB
BOTTOM	FORM	MIND	SECOND	TOUCH	BARD	DEED	HOES	PELT	TINT
CALL	GOLD	MINE	SHAPE	TRUTH	BILE	DIKE	HOOT	PEST	TONG
CARE	GROUP	MISS	SHARP	TURN	BLAZE	DILL	HULL	PIKE	TOTE
CASE	GROWTH	MODERN	SIGHT	UNDER	BLINK	DOZE	LASH	PLAZA	TROUPE
CLASS	HALF	MONTH	SMALL	VIEW	BOTANY	DUCT	LATCH	POMP	TUNES
CLOSE	HELL	MORAL	SOFT	VISIT	BRACE	FRILL	LEEK	POUT	UNREST
COST	HELP	NAME	SORT	VOICE	BUFF	FUSE	LOBE	QUACK	VALE
COURSE	HOLD	NEWS	SPACE	WARM	BUMP	GADGET	LURE	ROOK	VEAL
DEAL	HOPE	NONE	SPIRIT	WEALTH	BURROW	GAUGE	MALT	ROUGE	WADE
DEAR	HOUR	NUMBER	STAFF	WEEK	BUST	GERM	MASH	SASH	WAIL
DEEP	INCOME	PALE	STATE	WELL	CERISE	GONG	MICA	SCARE	WAMPUM
DESIRE	KIND	PART	STEP	WIDE	CHEAT	GRATE	MITT	SEAM	WARDER
DOWN	LEAD	PIECE	STOCK	WORK	CHEER	GRIP	MOAN	SHAG	WELD
EAST	LEADER	POWER	STUDY	YEAR	CHIRP	GROOM	MUFF	SINE	WELDER
EFFORT	LENGTH	PRESS	STYLE	YOUTH	CHIVE	GROUCH	MUSH	SLOUCH	YELL
ENERGY	LEVEL	PRICE	TALK		CHORE	GROWL	NECTAR	SPREE	
FAIR	LIFE	PRIME	TASK		CLANG	GULLET	NOOK	STARCH	
FAST	LIST	QUIET	TERM		CLASP	GUST	OPAL	STOAT	
FEEL	LONG	REPORT	TEST		CLONE	HAIL	PANT	STOLE	
FELLOW	LORD	REST	THICK		CLUCK	HASH	PARDON	SWORE	

References

Asch, S. E., & Ebenholtz, S. M. (1962). The principle of associative symmetry. *Proceedings of the American Philosophical Society, 106*(2), 135–163.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania (Distributor).

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers, 33*(1), 73–79.

Bishop, Y., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Bower, G. H. (1970). Imagery as a relational organizer in associative learning. *Journal of Verbal Learning and Verbal Behavior, 9*(5), 29–533.

Bower, G. H. (1972). Mental imagery and associative learning. In L. Gregg (Ed.), *Cognition in learning and memory* (pp. 51–88). Pittsburgh: Wiley.

Burnham, K. E., & Anderson, D. R. (2002). *Model selection and multimodel interference* (2nd ed.). New York: Springer-Verlag.

Burnham, K. E., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*, 261–304.

Calkins, M. W. (1896). Association: An essay analytic and experimental. *The Psychological Review, 2*(Monograph Supplement 1), 35–56.
Caplan, J. B. (2004). Unifying models of paired associates and serial learning: Insights from simulating a chaining model. *NeuroComputing, 58-60*, 739–743.
Caplan, J. B. (2005). Associative isolation: Unifying associative and order paradigms. *Journal of Mathematical Psychology, 49*(5), 383–402.
Caplan, J. B., Glaholt, M., & McIntosh, A. R. (2006). Linking associative and list memory: Pairs versus triples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(6), 1244–1265.
Chalmers, K. A., & Humphreys, M. S. (2003). Experimental manipulation of prior experience: Effects on item and associative recognition. *Memory & Cognition, 11*, 233–246.
Clark, S. E. (1992). Word frequency effects in associative and item recognition. *Memory & Cognition, 20*, 231–243.
Clark, S. E., & Burchett, R. E. R. (1994). Word frequency and list composition effects in associative recognition and recall. *Memory & Cognition, 22*, 55–62.
Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3*, 37–60.
Clark, S. E., & Shiffrin, R. M. (1992). Cuing effects and associative information in recognition memory. *Memory & Cognition, 20*, 580–598.
Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments, & Computers, 36*(3), 384–387.
Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.
Diana, R. A., & Reder, L. M. (2006). The low-frequency encoding disadvantage: Word frequency affects processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(4), 805–815.
Ebbinghaus, H. (1885/1913). Memory: A contribution to experimental psychology. New York: Teachers College, Columbia University.
Foth, D. L. (1973). Mnemonic technique effectiveness as a function of word abstractness and mediation instructions. *Journal of Verbal Learning and Verbal Behavior, 12*(3), 239–245.
Geller, A. S., Schleifer, I. K., Sederberg, P. B., Jacobs, J., & Kahana, M. J. (2007). PyEPL: A cross-platform experiment-programming library. *Behavior Research Methods, Instruments, & Computers, 65*(1), 50–64.
Gorman, A. N. (1961). Recognition memory for names as a function of abstractness and frequency. *Journal of Experimental Psychology, 39*(4), 950–958.
Gregg, V. H. (1976). Word frequency, recognition and recall. In J. Brown (Ed.), *Recall and recognition*. London: Wiley.
Guttentag, R. E., & Carroll, D. (1997). Recollection-based recognition: Word frequency effects. *Journal of Memory and Language, 37*, 502–516.
Hall, J. F. (1954). Availability and associative symmetry. *The American Journal of Psychology, 67*(1), 138–140.
Hall, J. F. (1979). Recognition as a function of word frequency. *The American Journal of Psychology, 92*(3), 497–505.
Hayman, C. A. G., & Tulving, E. (1989). Contingent dissociation between recognition and fragment completion: The method of triangulation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(2), 228–240.
Hintzman, D. L. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review, 87*, 398–410.
Hockley, W. E. (1994). Reflections of the mirror effect for item and associative recognition. *Memory & Cognition, 22*, 713–722.
Hockley, W. E., & Cristi, C. (1996). Tests of encoding tradeoffs between item and associative information. *Memory & Cognition, 24*, 202–216.
Horowitz, L. M., Brown, Z. M., & Weissbluth, S. (1964). Availability and the direction of associations. *Journal of Experimental Psychology, 68*(6), 541–549.
Horowitz, L. M., Norman, S. A., & Day, R. S. (1966). Availability and associative symmetry. *Psychological Review, 73*(1), 1–15.
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.
Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review, 96*(2), 208–233.
Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition, 30*(6), 823–840.
Kahana, M. J., & Caplan, J. B. (2002). Associative asymmetry in probed recall of serial lists. *Memory & Cognition, 30*(6), 841–849.
Kelley, R., & Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 701–722.
Köhler, W. (1947). Gestalt psychology. New York: The New American Library.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240.
Lewandowsky, S., & Murdock, B. B. (1989). Memory for serial order. *Psychological Review, 96*, 25–57.
Lockhart, R. S. (1969). Retrieval asymmetry in the recall of adjectives and nouns. *Journal of Experimental Psychology, 79*(1), 12–17.
McGuire, W. J. (1961). A multiprocess model for paired-associate learning. *Journal of Experimental Psychology, 62*, 335–347.
Metcalfe Eich, J. (1982). A composite holographic associative recall model. *Psychological Review, 89*, 627–661.
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: Learning, Memory, and Cognition, 21*(1), 116–133.
Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*(2), 165–178.
Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*, 609–626.
Murdock, B. B. (1983). A distributed memory model for serial-order information. *Psychological Review, 90*, 316–338.
Murdock, B. B. (1999). Item and associative interactions in short-term memory: Multiple memory systems? *International Journal of Psychology, 34*(516), 427–433.
Nelson, D. L., & McEvoy, C. L. (2000). What is this thing called frequency? *Memory & Cognition, 28*(4), 509–522.
Paivio, A. (1965). Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 4*, 32–38.
Paivio, A. (1968). A factor-analytic study of word attributes and verbal learning. *Journal of Verbal Learning and Verbal Behavior, 7*(1), 41–49.
Paivio, A. (1971). Imagery and verbal processes. New York: Holt, Rinehart and Winston.
Paivio, A., Smythe, P. E., & Yuille, J. C. (1968). Imagery versus meaningfulness of nouns in paired-associate learning. *Canadian Journal of Psychology, 22*, 427–441.
Philips, L. (1990). Hanging on the metaphone. *Computer Language Magazine, 7*(12), 38–44.
Philips, L. (2000). The double metaphone search algorithm. *C/C++ Users Journal, 18*(6), 38–43.
Poirier, M., & Saint-Aubin, J. (1996). Immediate serial recall, word frequency, item identity, and item position. *Canadian Journal of Experimental Psychology, 50*, 408–412.
Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(2), 294–320.
Rehani, M., & Caplan, J. B. (in preparation). Interference explains dissociations between memory for serial lists versus associations.
Rizzuto, D. S., & Kahana, M. J. (2000). Associative symmetry vs. independent associations. *NeuroComputing, 32–33*, 973–978.
Rizzuto, D. S., & Kahana, M. J. (2001). An autoassociative neural network model of paired-associate learning. *Neural Computation, 13*, 2075–2092.
Shepard, R. N. (1967). Recognition memory for words, sentences and pictures. *Journal of Verbal Learning and Verbal Behaviour, 6*, 156–163.
Sommer, T., Rose, M., & Büchel, C. (2007). Associative symmetry versus independent associations in the memory for object–location associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(1), 90–106.
Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavioral Research Methods, Instruments & Computers, 38*(4), 598–605.
Underwood, B. J. (1966). Experimental psychology. New York: Appleton-Century-Crofts.
Ward, G., Woodward, G., Stevens, A., & Stinson, C. (2003). Using overt rehearsals to explain word frequency effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(2), 186–210.

Westbury, C. F., & Hollis, G. (2007). Putting humpty together again: Synthetic approaches to nonlinear variable effects underlying lexical access. In G. Libben & G. Jarema (Eds.), *The mental lexicon: Core perspectives*. Amsterdam: Elsevier Science.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, version 2. *Behavioral Research Methods, Instruments & Computers, 20*, 6–11.

Wolford, G. (1971). Function of distinct associations for paired-associate performance. *Psychological Review, 78*(4), 303–313.

Wollen, K. A., & Lowry, D. H. (1971). Effects of imagery on paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 10*, 276–284.

Wood, G. (1967). Mnemonic systems in recall. *Journal of Educational Psychology, 58*(6), 1–27.