Testing the myth of the encoding–retrieval match

Winston D. Goh · Sharon H. X. Lu

Abstract The view that successful memory performance depends importantly on the extent to which there is a match between the encoding and retrieval conditions is common-place in memory research. However, Nairne (*Memory*, *10*, 389–395, 2002) proposed that this idea about trace–cue compatibility being the driving force behind memory retention is a myth, because one cannot make unequivocal predictions about performance by appealing to the encoding–retrieval match. What matters instead is the relative diagnostic value of the match, and not the absolute match. Three experiments were carried out in which participants memorised word pairs and tried to recall target words when given retrieval cues. The diagnostic value of the cue was varied by manipulating the extent to which the cues subsumed other memorised words and the level of the encoding–retrieval match. The results supported Nairne's (*Memory*, *10*, 389–395, 2002) assertion that the diagnostic value of retrieval cues is a better predictor of memory performance than the absolute encoding–retrieval match.

Keywords Encoding–retrieval match · Cue overload · Cued recall · Encoding specificity · Memory

One of the fundamental ideas in memory research is that the match between the encoding and retrieval conditions affects memory performance. This idea is central to the encoding specificity principle (Thomson & Tulving, 1970; Tulving, 1983; Tulving & Osler, 1968; Tulving & Thomson,

1973), the proceduralist approach according to which remembering involves re-enacting encoding operations (Kolers, 1973), and the transfer-appropriate processing framework (Morris, Bransford, & Franks, 1977). In each case, it is assumed that successful remembering is a joint function of encoding and retrieval processes; more recent research appeals to this general principle when referring to the extent of the "overlap" between encoding and retrieval processes (e.g., Kent & Lamberts, 2008), reinstatement of encoding states during retrieval (e.g., Danker & Anderson, 2010), and the encoding–retrieval match (e.g., Dewhurst & Knott, 2010; Hannon & Daneman, 2007).

The idea that the "match" is all-important for performance can be found in many cognitive psychology textbooks. For example, "The phenomenon of transfer-appropriate processing shows that memory performance is enhanced if the type of task at encoding matches the type of task at retrieval" (Goldstein, 2011, p. 185), and "Both constructs [encoding specificity and transfer-appropriate processing] emphasise that good performance depends on maximising the similarity between the encoding and the retrieval of the material" (Reed, 2010, p. 151). Findings from studies investigating the effects of matching physical (e.g., Godden & Baddeley, 1975) as well as emotional (e.g., Eich, Macaulay, & Ryan, 1994) conditions on memory retention are often cited as support.

However, Nairne (2002) argued that similarity between the encoding and retrieval conditions (the encoding–retrieval match) is not all that matters in determining the accuracy of memory performance. In fact, he argued that knowing the status of the encoding–retrieval match by itself predicts next to nothing about subsequent retention. It is not matching features from the encoding and retrieval conditions per se that are needed, but the presence of features that help discriminate the correct target from incorrect competitors—that is, the *diagnostic value* of the retrieval

W. D. Goh (✉) · S. H. X. Lu
Department of Psychology, Faculty of Arts & Social Sciences,
National University of Singapore,
Block AS4, 9 Arts Link,
Singapore 117570, Singapore
e-mail: psygohw@nus.edu.sg

cue. These ideas are similar to theoretical accounts of memory processes that posit a critical role for the discriminability of a memory record relative to all possible traces at the retrieval stage (e.g., Norman & Bobrow, 1979). Thus, if a target word *NAIL* is encoded and stored in the context of the word *FINGER*, a subsequent retrieval cue such as *tool* will be ineffective, but a cue such as *part of the human body* will probably be quite effective. This illustrates the "match," in the sense that the *tool* meaning of the word *NAIL* was probably not encoded in the context of *FINGER*. However, if competitors for the target word *NAIL* included *TOE* and *HAND*, the effectiveness of the cue *part of the human body* would probably be diminished. Adopting Watkins and Watkins's (1975) terminology, the cue can be said to be "overloaded," as it does not provide any diagnostic information about the target occurrence, because the competitors are also subsumed by the cue.

The importance of retrieval cue diagnosticity had been demonstrated in some previous studies. For example, using a release-from-proactive-interference paradigm, Gardiner, Craik, and Birtwistle (1972) had participants study lists of words drawn from the same general category (*flowers*), with the items in the first three lists belonging to a subcategory (*wild flowers*) and the items in the fourth list belonging to a related but different subcategory (*garden flowers*). In the control condition, participants were not informed of the subcategory shift during the fourth list, and recall across all four lists showed the standard build-up of proactive interference in which performance continued to decline across lists. In two experimental conditions, participants were given a cue indicating the subcategory shift, either prior to encoding or prior to retrieval during the fourth list. Both conditions showed a release effect, in that performance on the fourth list improved. The concept of *cue overload* can be used to explain the findings. Across successive lists, the number of words subsumed under the category name of *flowers* increased, and hence reduced the efficacy of the category name as a retrieval cue in recalling to-be-remembered words within a particular list. However, once the subcategory name was presented, the diagnostic power of the cue increased, as the new cue subsumed only the words in the current list.

Similarly, findings from the short-term cued recall paradigm (e.g., Goh & Tan, 2006; Tehan & Humphreys, 1996) showed that accurate recall depended on the extent to which retrieval cues uniquely specified the target or subsumed other competitors. Participants were presented with two blocks of words in which they were to forget the words in the first block and recall a word from the second block, as specified by a subsequent retrieval cue. On some trials, the cue (e.g., *type of juice*) subsumed both a target word (*CARROT*) in the second block and a foil word (*ORANGE*) in the first block, increasing the likelihood of proactive interference from the foil word. On other trials, the cue *type of vegetable* uniquely specified the target word, increasing accurate recall.

From the preceding examples, it is clear that two factors influence memory performance—cue overload and the encoding–retrieval match. Nairne's (2002) contention is that memory researchers tend to treat the two factors as though they were independent, with increasing match always helping performance and increasing overload always impairing performance. In Nairne's (2002) view, however, the two factors are actually intertwined, as illustrated by his basic sampling model shown in the following equation:

$$P_r(E_1|X_1) = \frac{s(X_1, E_1)}{\sum s(X_1, E_1)}$$

The likelihood of recalling event E_1 depends on how well a retrieval cue X_1 matches E_1 to the exclusion of all other competitors E_i, where $s(X_1, E_1)$ represents the similarity of X_1 to E_1 in terms of the degree of matching attributes. The argument is that memory performance can be conceptualised as a joint function of the degree of the encoding–retrieval match (as represented by the numerator of the equation) and the degree of cue overload (as represented by the denominator of the equation), where performance is proportional to the former and inversely related to the latter. Hence, increasing the numerator (or the match) per se does not guarantee an increase in the likelihood of recall—which also depends on the extent to which the match is accompanied by an increase in the denominator (or the overload). If the increase in match is greater than the increase in overload, performance should be enhanced. Depending on the circumstances, there can also be no effect if the net increases in the match and the overload are about the same, and performance may even decline if the increase in cue overload outweighs that in the encoding–retrieval match.

As an analogy, the relationship between the encoding–retrieval match and remembering can be likened to intensity and brightness: What determines the perception of the brightness of a light is the amount of light falling in the centre relative to the surroundings, not the absolute amount of light. In a similar vein, Nairne (2002) argued that the critical factor is not the absolute encoding–retrieval match, but rather the relative diagnostic value of the match, which is the extent to which the retrieval cue uniquely specifies the target. In this sense, he argued that the encoding–retrieval match is a "myth," because although successful retrieval of a target depends importantly on the extent that to which there is a match between the encoding and retrieval conditions, the *absolute* match per se does not predict memory perfor-

mance, but rather the *relative diagnostic value* of the retrieval cue.

The aims of the present study were to examine these arguments in more detail and to empirically test the predictions articulated in Nairne's (2002) thought experiments. Most previous studies have examined either the effects of match or the effects of overload alone. From the earlier equation, it is clear that testing the predictions will require a combined manipulation of both match and overload variables in the same design.

Experiment 1

One thought experiment can be summarised as follows. Participants memorise a series of events E_1, E_2, E_3, … E_n and are asked to recall E_1. E_1 has features X_1 and X_2 that could be provided as retrieval cues. Providing both X_1 and X_2 as cues should logically increase the degree of the encoding–retrieval match as compared to presenting a single retrieval cue. Hence, one would expect better recall for the condition with two retrieval cues than the condition with just one. However, suppose that X_2 is also found in E_2, E_3, … E_n. X_2 now provides no diagnostic value for differentiating E_1 from its competitors. In this case, the two-cue condition would not have any advantage over presention of a single cue. Performance might even decline, since X_2 subsumes all of the memorised events.

We tested these predictions using a cued recall task. The logic of the design is summarised in Table 1. In each trial, participants studied 10 semantically unrelated cue–target pairs, such as *abort–DONKEY*. The degree of the encoding–retrieval match and the degree of cue overload were varied. In the high-match conditions, two retrieval cues were provided at test, the original studied cue *abort* and a second cue, which was the name of the semantic category that *DONKEY* belonged to, *a four-footed animal*. In the low-match conditions, only the original studied cue was provided. In the high-overload conditions, the second retrieval cue provided at test subsumed all of the targets in the study list; that is, all 10 cue–target pairs had targets that were from the same semantic category as *DONKEY*—for instance, *TIGER*, *ELEPHANT*. In the low-overload conditions, the second retrieval cue did not subsume any of the competitor targets; that is, they were all unrelated to *DONKEY*—for instance, *FINGER*, *PAPAYA*.

If the predictions of the thought experiment are correct, only the low-overload condition would benefit from an increase in the encoding–retrieval match by the provision of a second retrieval cue, since that cue uniquely specifies the target *DONKEY* and maximises the diagnostic value of the retrieval environment.

Method

Participants A group of 40 introductory psychology students participated for course credit.

Design and materials A 2 (overload: high, low) x 2 (match: high, low) within-subjects design was employed.

A total of 40 semantic categories were selected from the Van Overschelde, Rawson, and Dunlosky (2004) and Yoon, Feinberg, Hu, Gutchess, Hedden, Chen et al., (2004) norms. The category name was used as the second retrieval cue in the high-match conditions. From each category (e.g., *a four-footed animal*), an exemplar with a low response frequency (e.g., *DONKEY*) was selected to be a critical target—that is, a target that would be tested in the recall phase. *Response frequency* refers to the proportion of responses for that exemplar out of the total responses for the category. High-response exemplars were not selected as critical targets, in order to minimise guessing when given the category name as a retrieval cue. From each category, nine high-response exemplars (e.g., *TIGER*, *ELEPHANT*) were selected to be foil targets—that is, targets that would be studied but not tested during recall. When studied within the same list, these foil targets would generate high cue overload, as they are subsumed by the category retrieval cue. Each target was then paired with a semantically unrelated cue word. Thus, there were altogether 40 category lists of 10 cue–target pairs (1 critical pair and 9 foil pairs) each.

All words were rated for familiarity by 20 participants who did not take part in the study but were from the same population as the experimental sample. A 7-point scale was used, where 1 = *do not know the meaning of the word at all*, 7 = *know the meaning of the word*, and 2–6 represent intermediate levels of familiarity. Words that did not achieve an average familiarity rating of at least 6 were replaced and checked again. The mean and standard deviation of the familiarity ratings of all selected words were 6.82 and 0.52, respectively. The lists were then divided into four sets of 10 lists each, which were equated for average response frequency, word frequency, and number of syllables, as summarised in Table 2.

Procedure A balanced Latin square procedure was used to rotate the sets across the four conditions in the study. For any participant, a set was assigned to a single condition and was never repeated across conditions. In the high-overload conditions, all cue–target pairs were sampled from within the same category list. In the low-overload conditions, each cue–target pair was sampled from a different category list within the set.

A single trial comprised a study phase followed by a recall phase. In the study phase, 10 cue–target pairs were

Table 1 Experiment 1 design schematic

	Overload Condition	
	High (targets are same-category exemplars)	Low (targets are different-category exemplars)
Study Phase		
	cable–dog	lane–finger
	coral–cat	killer–papaya
	secret–horse	traffic–snow
	conference–lion	knob–pamphlet
10 Cue–Target Pairs	**abort–donkey**	glass–sock
(The critical pair to be tested at recall is in boldface.)	biscuit–bear	**abort–donkey**
	receipt–tiger	bond–denim
	hexagon–cow	address–rifle
	spectacle–elephant	situation–dentist
	handphone–deer	aroma–tennis
Recall Phase		
Match Condition		
High: Original Studied Cue and Category Name Cue	abort; a four footed animal	abort; a four footed animal
Low: Original Studied Cue	abort	abort

displayed centred on the computer monitor, one at a time, with the cue appearing to the left of the target, at a rate of 2 s per pair. To avoid primacy and recency effects, the critical cue–target pair always appeared randomly in either the fourth, fifth, sixth, or seventh serial position within the sequence. In the recall phase, the test cue(s) appeared (one cue if the trial was from the low-match condition, and two cues if the trial was from the high-match condition), and participants typed in their response before moving to the next trial. Participants were told that in the recall phase on

some trials, an additional cue would be provided to assist recall.

The 40 trials, 10 from each condition, were randomly presented throughout the experiment. A further 18 filler trials that were similar to the experimental trials were also randomly interspersed. These fillers were created such that their critical pairs were presented in either the first three or last three positions of the sequence, so as to prevent participants from noticing that the position of the critical pair to be tested always occurred in the middle of the

Table 2 Mean stimulus properties across category list sets in Experiment 1

Stimulus	Set				F ratio
	1	2	3	4	
Log Category Response Frequency					
Foil targets	0.38 (0.13)	0.33 (0.15)	0.36 (0.06)	0.32 (0.14)	<1
Critical targets	0.09 (0.15)	0.04 (0.03)	0.06 (0.02)	0.05 (0.01)	<1
Log Word Frequency					
Foil targets	1.17 (0.47)	1.62 (0.48)	1.45 (0.30)	1.33 (0.52)	1.74
Foil cues	1.41 (0.16)	1.50 (0.47)	1.60 (0.15)	1.52 (0.22)	<1
Critical targets	0.55 (0.65)	0.61 (0.72)	0.93 (0.71)	0.71 (0.68)	<1
Retrieval cue 1	0.48 (0.54)	1.12 (0.82)	1.14 (0.91)	0.91 (0.63)	1.75
Number of Syllables					
Foil targets	2.02 (0.68)	1.70 (0.34)	1.81 (0.38)	1.84 (0.52)	<1
Foil cues	2.10 (0.34)	2.12 (0.36)	1.92 (0.40)	2.03 (0.26)	<1
Critical targets	2.00 (0.67)	2.10 (0.74)	2.30 (0.82)	1.80 (0.79)	<1
Retrieval cue 1	2.30 (1.25)	2.00 (0.94)	2.10 (0.74)	2.00 (0.47)	<1

*SD*s appear in parentheses. A one-way between-subjects ANOVA was run across sets to check for equivalence, all *p*s > .05

sequence. No subsequent analyses were performed on these fillers.

Results and discussion

The correct recall proportions are summarised in Fig. 1. A two-way within-subjects ANOVA revealed a significant interaction, $F(1, 39) = 13.89$, $MSE = 0.01$, $p < .01$. Simple-effects analyses showed that when cue overload was low, recall was better in the high-match ($M = .37$, $SD = .19$) than in the low-match ($M = .24$, $SD = .20$) condition, $F(1, 39) = 17.14$, $MSE = 0.02$, $p < .001$; recall did not differ between the high-match ($M = .28$, $SD = .20$) and low-match ($M = .28$, $SD = .20$) conditions when cue overload was high, $F < 1$. In the high-match conditions, recall was better with low overload than with high overload, $F(1, 39) = 8.92$, $MSE = 0.02$, $p < .01$; recall did not differ between the low-overload and high-overload conditions when match was low, $F(1, 39) = 2.94$, $MSE = .02$, n.s.

It should be noted that the increase in cue overload could have theoretically decreased performance in the high-match, high-overload condition, since the increase in match provided by the second retrieval cue was accompanied by an increase in overload, as the cue subsumed the competitor targets. The results suggest that increasing the encoding–retrieval match had a facilitative effect on recall when overload was low, but no effect when overload was high.

Nevertheless, this pattern of results is consistent with the predictions of Nairne's (2002) thought experiment: Providing a second retrieval cue in the form of the category name of the target should increase the encoding–retrieval match relative to a single cue. However, this advantage only occurs if the second cue confers additional diagnostic information. If the second cue subsumes competing targets, as was the case in the high-overload conditions, then
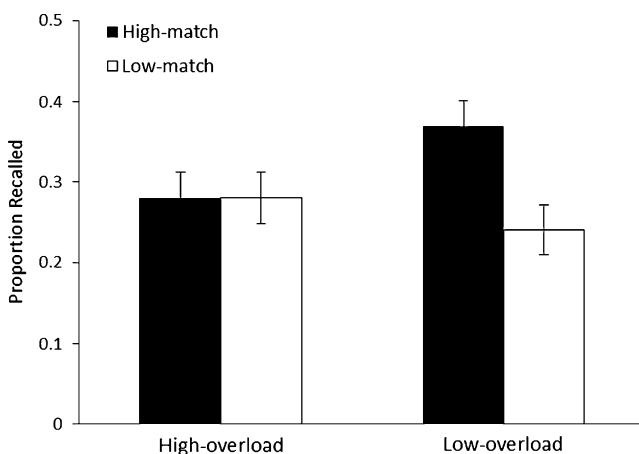
performance is no better than with just a single cue. This supports the contention that the relative diagnostic value of the cue is the important factor in predicting cue efficacy, rather than the absolute value of the encoding–retrieval match.

Experiment 2

One potential problem with Experiment 1 was that when all of the targets came from a single semantic category in the high-overload conditions, participants would have quickly realised this and utilised that information, whether or not that information was subsequently provided at test, in the form of the second retrieval cue. Hence, the lack of a difference between the high- and low-match conditions when cue overload was high might simply have been an artefact of this procedure, rather than evidence of the efficacy (or lack thereof) of the information provided by the second retrieval cue. Experiment 2 addressed this problem by making the overload manipulation less obvious.

Method

Participants A group of 43 introductory psychology students participated for course credit.

Design, materials, and procedure The basic design and procedure of Experiment 1 was used, except that only three foil pairs, instead of nine, had targets that belonged to the same category list as the critical target for each trial in the high-overload conditions. To avoid clustering of these related targets within a trial, which might have otherwise made the semantic category more salient, one of these foil pairs was randomly assigned to the primacy region (first to third serial positions), another to the middle region (fourth to seventh positions), and the last to the recency region (eighth to tenth positions). Each of the other six foil pairs was sampled from a different category list. Participants were also asked at the end of the experiment about what they thought the study was about and how the words were manipulated.

Results and discussion

One participant was dropped from the analyses due to computer failure during the study. Only 2 participants mentioned noticing that some of the words within a trial were from the same semantic category. Analyses with and without these 2 participants revealed identical findings, so the participants were kept in the database.

Fig. 1 Proportions recalled (± *SE*s) across overload and match conditions in Experiment 1

The correct recall proportions are summarised in Fig. 2. A two-way within-subjects ANOVA revealed a significant interaction, $F(1, 41) = 5.23$, $MSE = 0.02$, $p < .05$. Simple-effects analyses showed that when cue overload was low, recall was better in the high-match ($M = .31$, $SD = .19$) than in the low-match ($M = .20$, $SD = .22$) condition, $F(1, 41) = 12.15$, $MSE = 0.02$, $p < .01$; recall did not differ between the high-match ($M = .24$, $SD = .19$) and low-match ($M = .23$, $SD = .15$) conditions when cue overload was high, $F < 1$. In the high-match conditions, recall was also better with low than with high overload, $F(1, 41) = 5.76$, $MSE = 0.02$, $p < .05$; recall did not differ between the low-overload and high-overload conditions when match was low, $F < 1$.

The pattern of results replicated that found in Experiment 1—increasing the encoding–retrieval match had a facilitative effect on recall when overload was low, but no effect when overload was high. Importantly, the manipulation check revealed that the majority of participants did not notice the use of targets from the same semantic category in the high-overload conditions, and thus the artefact that might have been operating in Experiment 1 could be ruled out. As before, the pattern of results supported the view that the relative diagnostic value of the cue is the driving force behind recall, rather than the absolute value of the encoding–retrieval match.

Experiment 3

Experiments 1 and 2 demonstrated that high match can be negated by high overload, suggesting that the efficacy of retrieval cues must be considered in relation to the specific retrieval environment, rather than just in relation to the absolute level of the encoding–retrieval match, or even the absolute level of cue overload. Theoretically, it should also be possible to demonstrate that under certain retrieval

Fig. 2 Proportions recalled (± *SE*s) across overload and match conditions in Experiment 2

contexts, the encoding–retrieval match can facilitate recall despite a large amount of overload. Any increase in the match should increase the numerator of the equation presented in the introduction, regardless of the amount of overload represented in the denominator. Therefore, one hypothesis is that increasing the encoding–retrieval match should cause performance to increase if the degree of cue overload, whether high or low, does not change across experimental conditions.

One way to test this hypothesis is to use the experimental paradigms from past studies that have demonstrated robust effects of the encoding–retrieval match. As the encoding specificity principle has often been used as the basis for maximising memory performance, the Thomson and Tulving (1970) paradigm is a good candidate. Their general procedure required participants to memorise a set of cue–target pairs that were preexperimentally weakly associated, such as *train–BLACK*. Weakly associated pairs were studied so that recall by guessing would be minimised. At test, some participants were re-presented with the originally studied weak cue *train*; others were given an extralist cue that was preexperimentally strongly associated with the target *BLACK*, such as *white*; and a final group was given no cue at all, or a weakly associated extralist cue in some studies (e.g., Newman, Cooper, Parker, Sidden, Gonder-Frederick, Moorefield and Nelson 1982). Results showed that the original studied cue *train*, though weakly associated with the target, was much better in eliciting recall of *BLACK* than the other conditions, including strongly associated cues such as *white*. This was the primary basis for the encoding specificity argument—when the target *BLACK* was studied in the context of *train*, the target word was encoded in a specific manner that was distinct from the preexperimental encoding of *BLACK* in the context of *white*.

It is important to note that most experimental work on encoding specificity has not controlled explicitly for the degree of cue overload. Although each cue–target pair was associated, there were presumably no associations across pairs similar to the manipulations in Experiments 1 and 2. It is then entirely possible that the degree of cue overload was essentially held constant across the experimental conditions. Hence, re-presenting the original studied cue would maximally increase the encoding–retrieval match and result in the best performance.

Experiment 3 introduced a novel manipulation of cue overload to this basic design in order to test its joint effects with the encoding–retrieval match. The logic of the design is summarised in Table 3. Participants studied weakly associated cue–target pairs such as *park–GROVE*, *airplane–BIRD*, and *roof–TIN*. At test, participants were re-presented with the originally studied weak intralist cue—for instance, *park* for recalling *GROVE*; a strong extralist cue—for instance, *feather* for recalling *BIRD*; or a weak extralist cue—for
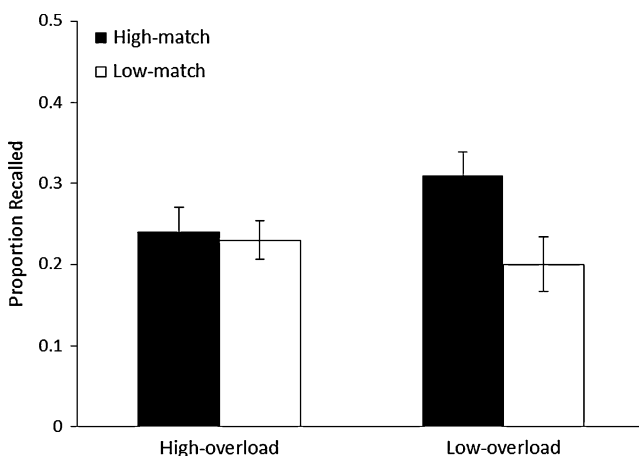
Table 3 Experiment 3 design schematic

	Cue Type Condition		
	Intralist	Strong Extralist	Weak Extralist
High-Overload Condition			
	park–grove	**airplane–bird**	**roof–tin**
	beaver–tree	morning–light	castle–knight
	pedestrian–car	quilt–soft	ozone–protection
Study Phase	chair–bench	gloves–hat	ban–shield
Cue–Target Pairs	recreation–playground	eraser–pen	foil–metal
	gymnastics–swing	nap–pillow	server–plate
	cooler–picnic	pressure–weight	jacket–suit
	alligator–lake	pinch–tickle	legion–war
Recall Phase			
Retrieval Cue	park	feather	armour
Low-Overload Condition			
	marker–buoy	**blackboard–class**	**beard–santa**
	meadow–sheep	loop–ring	chess–pawn
	tide–shore	hobby–swim	oil–vinegar
Study Phase	scroll–script	mood–temper	captain–boss
Cue–Target Pairs	guardian–adult	permit–ticket	million–lottery
	sales–profit	verse–chorus	moustache–razor
	rainbow–spectrum	market–stock	ghost–fear
	scout–uniform	salad–potato	analysis–data
Recall Phase			
Retrieval Cue	marker	course	stocking

Critical pairs in these examples are in boldface

instance, *armour* for recalling *TIN*. In the high-overload conditions, the foil cue–target pairs that were studied with the critical pair had targets that were strongly associated with the recall cue in the respective test conditions—for example, the targets *TREE* and *BENCH* for the intralist cue *park*; *LIGHT* and *SOFT* for the strong extralist cue *feather*; and *KNIGHT* and *PROTECTION* for the weak extralist cue *armour*. In the low-overload conditions, foil targets (e.g., *SHEEP*, *SHORE* or *RING*, *SWIM*) were unrelated to the recall cues (e.g., *marker*, *course*).

One critical difference between the present design and those of the first two experiments should be noted. In Experiments 1 and 2, the first retrieval cue was always an intralist cue, and hence the match provided by this first cue was effectively identical across all experimental conditions. The introduction of the second extralist cue theoretically increased the match but was simultaneously countered by an increase in cue overload in some conditions.

In Experiment 3, when the encoding–retrieval match was increased (i.e., from no match in the two extralist conditions [*feather–BIRD*, *armour–TIN*] to a match in the intralist condition [*park–GROVE*]), the degree of overload would not change across the three cue-type conditions. Overload remained either high or low because all retrieval cues (e.g., *park*, *feather*, *armour*) were manipulated so as to either subsume all of the targets of the foil pairs (e.g.,

TREE, *CAR*, *BENCH*; *LIGHT*, *SOFT*, *HAT*; and *KNIGHT*, *PROTECTION*, *SHIELD*, respectively) or were unrelated (e.g., *marker*, *course*, *stocking*) to the targets of the foil pairs (e.g., *SHEEP*, *SHORE*, *SCRIPT*; *RING*, *SWIM*, *TEMPER*; and *PAWN*, *VINEGAR*, *BOSS*, respectively) across the relevant comparison conditions. Within each of the two overload conditions, using the equation provided in the introduction, the denominator was kept constant, while the numerator changed across the intralist, strong-extralist, and weak-extralist cue conditions. Hence, overall recall rates should drop in the high-overload conditions due to the subsuming of foil targets under the recall cues (resulting in a larger denominator as compared to the low-overload conditions), but there should still be some evidence of facilitation, due to the use of an originally studied cue, relative to the extralist cues—that is, the encoding specificity effect, even in the high-overload conditions. In this situation, the use of an intralist cue should provide enough diagnostic information relative to extralist cues to facilitate recall.

Method

Participants A group of 40 introductory psychology students who did not take part in the previous experiments participated for course credit.

Design and materials A 2 (overload: high, low) x 3 (cue type: intralist, strong extralist, weak extralist) within-subjects design was employed.

A total of 54 critical cue–target pairs, 9 in each of the six experimental conditions, were created based on the Nelson, McEvoy, and Schreiber (1998) norms. For the intralist and weak extralist conditions, a target word (e.g., *GROVE*) that was weakly associated with each recall cue (e.g., *park*) was first selected. For the strong extralist conditions, a strongly associated target was selected. In the three high-overload conditions, seven words (e.g., *TREE*, *BENCH*) strongly associated with each recall cue were then selected to be the subsumed targets of the foil pairs. In the three low-overload conditions, the seven foil targets were not associated with the recall cues. All targets were then paired with a weakly associated word to form the cue–target pairs (e.g., *park–GROVE*, *beaver–TREE*, *chair–BENCH*). A further constraint was that for both extralist conditions, each studied cue (e.g., *roof*) of a critical cue–target pair was semantically unrelated to the corresponding extralist recall cue (e.g., *armour*). This was done to minimise indirect retrieval of the target via associations between the studied and recall cues. Each of the six experimental conditions thus had nine lists of cue–target pairs, with each list comprising 1 critical pair and 7 foil pairs.

The words were then checked for familiarity using the same method as in Experiment 1. The selected words had an average familiarity rating and standard deviation of 6.93 and 0.23, respectively. Due to various constraints faced in selecting the words, it was not possible to divide them into lists and rotate them through the conditions using Latin square procedures. Lists were thus fixed across conditions and were equated on average forward associative strength between the recall cue and the subsumed targets for the three high-overload conditions. The average strength values ranged from 0.059 to 0.062 (overall $M = 0.06$, $SD = 0.02$), and a one-way between-subjects ANOVA run across the three conditions was not significant, $F < 1$. The average strength values were equated to ensure that recall differences could not be attributed to differences in the amount of overload generated by the subsumed targets across the three cue-type conditions. The low-overload conditions were already equated, as it was ensured that the recall cues and targets of the foil pairs were not associated.

Procedure Care was taken to ensure that the procedures in the original studies on encoding specificity were followed as closely as possible, in particular the instructions and the number of word pairs to be studied before cued recall was tested. Participants were asked to memorise all target words and to pay attention to the cue word that accompanied each target, as it might help them to remember the target. They were also informed that recall cues at test were related to the targets but might or might not be from the study lists. The participants were to recall the target that they thought was related to the cue.

The 54 experimental lists were randomly presented without replacement in 18 blocks of 3 lists each.[1] Within each list of one critical pair and seven foil pairs, the presentation order was fixed such that the critical pair is presented first followed by the foil pairs in descending order of their targets' associative strength to the corresponding condition's eventual recall cue. This was done to maximise the degree of activation of the recall cue, and thus maximise cue overload at test.

Figure 3 shows a schematic of a block structure. A single block comprised a study phase followed by a recall phase. In the study phase, six filler cue–target pairs were first presented, followed by a random selection of three lists from three of the six experimental conditions. The purpose of the fillers was to minimise primacy effects, since the critical pair of each list was always the first pair in the within-list sequence. This resulted in 30 pairs to be studied on each block, which closely approximated the original procedure of Thomson and Tulving (1970), which had 24 pairs for each study block. Each pair was presented one at a time on the monitor at a rate of 2 s per pair.

In the recall phase, the three recall cues for the presented conditions were shown one at a time in a random order. There was also a recall cue from one of the filler trials, since otherwise participants might have noticed that the first few pairs were never tested. Participants typed their response to each cue before the next one was shown. Responses to the filler recall cues were not analysed. After all responses to the four cues were made, the study phase of the next block of 30 pairs was initiated.

---

[1] The original Thomson and Tulving (1970) paradigm had participants recall as many words as they could from 24-word lists, using all intralist and extralist cues from the respective conditions. It was not possible for us to construct 24-word lists with the constraint of the retrieval cue subsuming all of the competitor targets in the high-overload conditions. Hence, to approximate the memory load of 24 word pairs, we had to place three of our 8-word lists in a single study block. Another difference between our procedure and Thomson and Tulving's was that the latter used a free recall condition as a control condition rather than a weak extralist cue condition. Most previous studies on encoding specificity have employed the free recall control, although Newman et al. (1982) included a weak extralist control condition in one of their experiments. Since our manipulations on cue overload involved overloading the retrieval cue presented in the recall phase by subsuming the competitor targets, it was more appropriate to use a cuing condition for comparison rather than a free recall situation, where we would not be able to direct participants to recall a specific critical target.

Fig. 3 Example of a block
structure in Experiment 3. The
left panel shows the pool of
word pairs, with each condition
having nine lists with eight word
pairs apiece. The right panel
shows what happens in a single
block. After the filler pairs are
presented, for this example we
select one list each from the
high-overload strong extralist
cue, low-overload strong extra-
list cue, and low-overload intra-
list cue conditions to be
presented in the study phase. In
the recall phase, we present the
corresponding recall cues from
the selected conditions
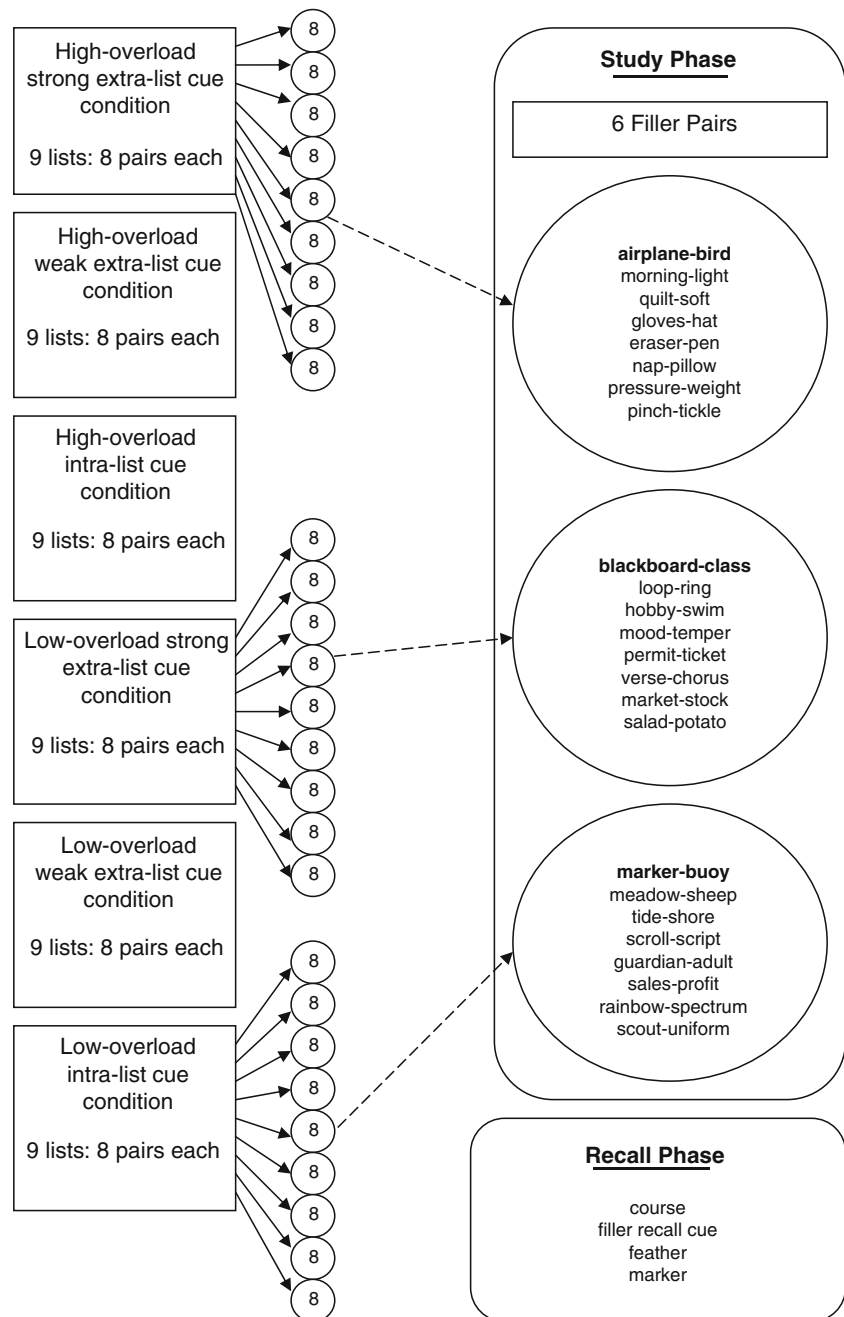
Results and discussion

The correct recall proportions are shown in Fig. 4. A two-
way within-subjects ANOVA revealed a significant main
effect of overload, $F(1, 39) = 90.24$, $MSE = 0.02$, $p < .001$.
Recall was better with low overload ($M = .31$, $SD = .11$)
than with high overload ($M = .14$, $SD = .07$). The main
effect of cue type was also significant, $F(2, 78) = 38.33$,
$MSE = 0.03$, $p < .001$. Planned contrasts showed that the
intralist cues ($M = .32$, $SD = .16$) elicited better recall than
did either the strong extralist cues ($M = .25$, $SD = .11$), $F(1,$
$39) = 4.83$, $MSE = 0.02$, $p < .05$, or the weak extralist cues
($M = .09$, $SD = .08$), $F(1, 39) = 62.40$, $MSE = 0.02$, $p <$
$.001$. Strong extralist cues were better than weak extralist
cues, $F(1, 39) = 66.58$, $MSE = 0.01$, $p < .001$. The
interaction was not significant, $F < 1$.

The pattern of results replicated the basic findings of
Thomson and Tulving (1970) and Newman et al. (1982). The
low-overload conditions were essentially similar to the ones
in these previous studies, and the results showed that re-
presenting the original studied cue, even though it was
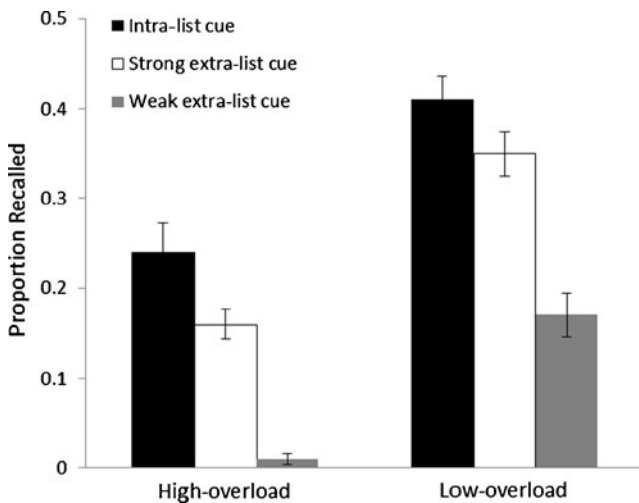weakly associated with the target, elicited better recall than

Fig. 4 Proportions recalled (± *SE*s) across overload and cue-type conditions in Experiment 3

extralist cues, even those strongly associated with the target. In the novel high-overload conditions, the same pattern emerged, even though overall recall dropped. This demonstrates that if cue overload can be experimentally held constant across the relevant comparison conditions, then increasing the encoding–retrieval match (e.g., by presenting intralist cues) will increase the diagnostic value of the cue, which in turn will facilitate recall. This pattern holds even for the high-cue-overload condition.

General discussion

The present experiments were designed to evaluate Nairne's (2002) claims regarding the myth that the encoding–retrieval match primarily controls memory performance. He instead argued that memory performance is directly proportional to the encoding–retrieval match and inversely related to the amount of cue overload, as suggested by his simple ratio model represented by the equation in the introduction.

As the efficacy of recall depends on the joint effects of match and overload, increasing the encoding–retrieval match did not necessarily result in a recall advantage, as shown in Experiments 1 and 2. Providing a second retrieval cue that subsumed the target should theoretically have increased the match because there were now two features, the original studied cue and an additional feature that was present in the target. However, if this second cue also subsumed the target's competitors, the increased match was now countered by an increase in cue overload. This essentially reduced the diagnostic value of the two-cue retrieval condition, leading to recall levels that were no better than those in the one-cue conditions. Only when

the second cue uniquely specified the target was there an increase in the relative diagnostic value for the two-cue retrieval condition over the single-cue condition, leading to better recall.

Experiment 3 demonstrated that the encoding and retrieval conditions could be manipulated such that increases in the encoding–retrieval match would result in improved recall in spite of increases in cue overload. This occurred only because the degree of overload, regardless of whether it was high or low, was held constant across the relevant comparison conditions. Using intralist cues provided information that overlapped with the original encoding, conferring a recall advantage over the extralist cues, and this advantage remained in spite of the general reduction of recall performance due to high cue overload.

One point that should be noted is that the present experiments have demonstrated only the first two of the three outcomes that Nairne argued could occur when the encoding–retrieval match is increased: It could improve performance, produce no effect, or even lower performance. Theoretically, retention could be impaired if the net increase in cue overload was larger than the increase in the encoding–retrieval match. Perhaps a stronger cue-overload manipulation using three or more cues sharing features by competitor targets in a future study might be able to confirm this prediction.

It should also be noted that the present results do not invalidate the encoding specificity principle or the generalisation of this principle in the transfer-appropriate processing framework. Surprenant and Neath (2009, p. 45) pointed out that the latter has often been taken to be transfer "similar" processing rather than "appropriate" processing. Processing that is different between the encoding and retrieval stages can still enhance memory performance, as long as it is appropriate for discriminating correct from incorrect targets. For example, Surprenant and Neath argued that Gardiner et al's (1972) findings can be considered a case in which encoding and retrieval processes do not match, but the retrieval process is more appropriate for retrieving the correct items. In particular, when participants were given a subcategory name prior to retrieval, the cue was appropriate, as it discriminated the items in the *garden flowers* subcategory from those in the preceding *wild flowers* subcategory. This was a case in which the use of the subcategory cue resulted in matching of the appropriate features belonging to the subcategory and, importantly, a simultaneous reduction of cue overload, since the cue now did not subsume the flowers from the previous lists. This was essentially the opposite effect of some of our manipulations in Experiments 1 and 2, where the category cue increased the match but simultaneously

increased the overload, and therefore led to no improvement in performance. Both examples underscore the importance of considering the joint effects of the encoding–retrieval match and cue overload—in one case, the cue was appropriate, as it increased the diagnostic value of the retrieval environment, and in the other, it was not appropriate. This is consistent with the idea of transfer-appropriate processing.

The key point here is that memory retrieval is essentially a discrimination problem (see Hunt, 2003, 2006; Nairne, 2005, for reviews): To what extent can a particular memory trace be distinguished from other traces? This question cannot be answered by considering only the match between encoding and retrieval conditions. The efficacy of a retrieval cue is only relative to a particular retrieval context; the same cue (e.g., the name of a semantic category) might be effective in one condition (e.g., the high-match low-overload condition) but might be relatively impotent in another (e.g., the high-overload conditions). Hence, the diagnostic value of a retrieval cue is a relative concept and is a property of a cue *in context* (Nairne, 2006); it depends critically on what the competing memory traces are. It can be argued, therefore, that an encoding–retrieval match is not intrinsically or universally beneficial, but should be effective in so far as the matching features do not overlap with the encoded features of other possible retrieval candidates. The critical factor for successful retrieval thus appears to depend on the extent to which a retrieval cue can provide diagnostic information about the target. For example, if one were asked to search for a particular individual and were told that this person was a boy in school uniform, this information would most likely be very useful for identifying this person in a room full of adults. However, the same cue would provide practically no discriminative information about that particular person in the setting of a school cafeteria full of schoolboys (Goh & Tan, 2006).

In conclusion, it is apparent from the present findings that increasing the encoding–retrieval match need not necessarily facilitate recall. It could facilitate recall, when cue overload was kept constant (Exp. 3), but it could also have no effect on recall, when the match increase was countered by a cue overload increase (Exps. 1 and 2). Is the encoding–retrieval match a myth? The present findings suggest a "yes" answer, because in line with Nairne's (2002) proposal, an encoding–retrieval match by itself cannot be used to make unequivocal predictions about memory performance. As we have argued throughout, the relative extent to which a retrieval cue uniquely specifies a given target is what determines successful memory performance.

References

Danker, J. F., & Anderson, J. R. (2010). The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding. *Psychological Bulletin, 136,* 87–102. doi:10.1037/a0017937

Dewhurst, S. A., & Knott, L. M. (2010). Investigating the encoding–retrieval match in recognition memory: Effects of experimental design, specificity, and retention interval. *Memory & Cognition, 38,* 1101–1109. doi:10.3758/MC.38.8.1101

Eich, E., Macaulay, D., & Ryan, L. (1994). Mood dependent memory for events of the personal past. *Journal of Experimental Psychology. General, 123,* 201–215. doi:10.1037/0096-3445.123.2.201

Gardiner, J. M., Craik, F. I. M., & Birtwistle, J. (1972). Retrieval cues and release from proactive inhibition. *Journal of Verbal Learning and Verbal Behavior, 11,* 778–783. doi:10.1016/S0022-5371(72)80012-4

Godden, D. R., & Baddeley, A. D. (1975). Context - dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66,* 325–331. doi:10.1111/j.2044-8295.1975.tb01468.x

Goh, W. D., & Tan, H. (2006). Proactive interference and cuing effects in short-term cued recall: Does foil context matter? *Memory & Cognition, 34,* 1063–1079. doi:10.3758/BF03193253

Goldstein, E. B. (2011). *Cognitive psychology* (3rd ed.). Belmont: Wadsworth.

Hannon, B., & Daneman, M. (2007). Prospective memory: The relative effects of encoding, retrieval, and the match between encoding and retrieval. *Memory, 15,* 572–604. doi:10.1080/09658210701407281

Hunt, R. R. (2003). Two contributions of distinctive processing to accurate memory. *Journal of Memory and Language, 48,* 811–825. doi:10.1016/S0749-596X(03)00018-4

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. Worthen (Eds.), *Distinctiveness and memory* (pp. 3–25). New York: Oxford University Press.

Kent, C., & Lamberts, K. (2008). The encoding–retrieval relationship: Retrieval as mental simulation. *Trends in Cognitive Sciences, 12,* 92–98. doi:10.1016/j.tics.2007.12.004

Kolers, P. A. (1973). Remembering operations. *Memory & Cognition, 1,* 347–355. doi:10.3758/BF03198119

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16,* 519–533. doi:10.1016/S0022-5371(77)80016-9

Nairne, J. S. (2002). The myth of the encoding–retrieval match. *Memory, 10,* 389–395. doi:10.1080/09658210244000216

Nairne, J. S. (2005). The functionalist agenda in memory research. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer* (pp. 115–126). Washington, DC: American Psychological Association.

Nairne, J. S. (2006). Modeling distinctiveness: Implications for general memory theory. In R. R. Hunt & J. Worthen (Eds.), *Distinctiveness and memory* (pp. 27–46). New York: Oxford University Press.

Nelson, D.L., McEvoy, C.L., & Schreiber, T.A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Available at http://w3.usf.edu/FreeAssociation/

Newman, S. E., Cooper, M. H., Parker, K. O., Sidden, J. A., Gonder-Frederick, L. A., Moorefield, K. M., et al. (1982). Some tests of the encoding specificity and semantic integration hypotheses. *The American Journal of Psychology, 95,* 103–123.

Norman, D. A., & Bobrow, D. G. (1979). Descriptions: An intermediate stage in memory retrieval. *Cognitive Psychology, 11,* 107–123. doi:10.1016/0010-0285(79)90006-9

Reed, S. K. (2010). *Cognition: Theories and applications* (8th ed.). Belmont: Wadsworth.

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

Tehan, G., & Humphreys, M. S. (1996). Cuing effects in short-term recall. *Memory & Cognition, 24,* 719–732. doi:10.3758/BF03201097

Thomson, D. M., & Tulving, E. (1970). Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology, 86,* 255–262. doi:10.1037/h0029997

Tulving, E. (1983). *Elements of episodic memory.* New York: Oxford University Press.

Tulving, E., & Osler, S. (1968). Effectiveness of retrieval cues in memory for words. *Journal of Experimental Psychology, 77,* 593–601. doi:10.1037/h0026069

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352–373. doi:10.1037/h0020071

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50,* 289–335. doi:10.1016/j.jml.2003.10.003

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory, 1,* 442–452. doi:10.1037/0278-7393.1.4.442

Yoon, C., Feinberg, F., Hu, P., Gutchess, A. H., Hedden, T., Chen, H.-Y. M., et al. (2004). Category norms as a function of culture and age: Comparisons of item responses to 105 categories by American and Chinese adults. *Psychology and Aging, 19,* 379–393. doi:10.1037/0882-7974.19.3.379