UNIVERSITY *of York*

This is a repository copy of *Schematic information influences memory and generalisation behaviour for schema-relevant and -irrelevant information*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/187767/

Version: Published Version

Article:

White Rose
university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Original articles

Schematic information influences memory and generalisation behaviour for schema-relevant and -irrelevant information
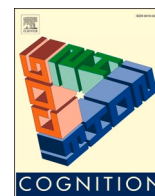
Jamie P. Cockcroft [a,b], Sam C. Berens [c], M. Gareth Gaskell [a], Aidan J. Horner [a,b,*]

[a] *Department of Psychology, University of York, UK*
[b] *York Biomedical Research Institute, University of York, UK*
[c] *School of Psychology, University of Sussex, UK*

ARTICLE INFO

ABSTRACT

Schemas modulate memory performance for schema-congruent and -incongruent information. However, it is assumed they do not influence behaviour for information irrelevant to themselves. We assessed memory and generalisation behaviour for information related to an underlying pattern, where a schema could be extracted (schema-relevant), and information that was unrelated and therefore irrelevant to the extracted schema (schema-irrelevant). Using precision measures of long-term memory, where participants learnt associations between words and locations around a circle, we assessed memory and generalisation for schema-relevant and -irrelevant information. Words belonged to two semantic categories: human-made and natural. For one category, word-locations were clustered around one point on the circle (clustered condition), while the other category had word-locations randomly distributed (non-clustered condition). The presence of an underlying pattern in the clustered condition allows for the extraction of a schema that can support both memory and generalisation. At test, participants were presented with old (memory) and new (generalisation) words, requiring them to identify a remembered location or make a best guess. The presence of the clustered pattern modulated memory and generalisation. In the clustered condition, participants placed old and new words in locations consistent with the underlying pattern. In contrast, for the non-clustered condition, participants were less likely to place old and new non-clustered words in locations consistent with the clustered condition. Therefore, we provide evidence that the presence of schematic information modulates memory and generalisation for schema-relevant and -irrelevant information. Our results highlight the need to carefully construct appropriate schema-irrelevant control conditions such that behaviour in these conditions is not modulated by the presence of a schema. Theoretically, models of schema processing need to account for how the presence of schematic information can have consequences for information that is irrelevant to itself.

1. Introduction

Schemas are mental representations that alter our memory of the past, perception of the present, and predictions of the future. Schemas are thought to be formed when we experience multiple related events that have a common structure (Anderson, 1984; Bartlett, 1932; Head & Holmes, 1911; Piaget, 1926; Posner & Keele, 1968). In this way, schemas may capture the general structure of events that have occurred, abstracting away from the specific content of individual events.

Schemas can relate to the locations of items in the real world. When entering someone's home, semantically related items are typically grouped spatially. If you know where the soap and toilet paper are

located, you can use this information to predict where a towel will be located. However, this schematic information will be of little use when predicting the location of items from an unrelated category. For example, house plants can be placed anywhere in a home, and as such, the presence of a "bathroom" schema should be irrelevant to where a house plant is located, or where you predict one might be. In this example, the towel could be located in the bathroom with other bathroom-related items (schema-congruent) or in the living room (schema-incongruent). Whereas the location of the towel could either be schema-congruent or -incongruent, the location of a specific house plant is schema-irrelevant, as it should not be included as part of the "bathroom" schema.

The presence of a schema affects the encoding and retention of related events. Typically, schemas improve memory performance for congruent and, in some situations, incongruent information, relative to schema-irrelevant information (Frank, Montaldi, Wittmann, & Talmi, 2018; Greve, Cooper, Tibon, & Henson, 2019). However, less is known about the effects of a schema on irrelevant information. If we experience multiple related events that are intermixed with unrelated events, does the extraction of a schema for the related events also affect performance for the unrelated, irrelevant events? For example, does the presence of a "bathroom" schema affect our memory of where the house plant was located? Current theories do not make clear predictions about schema-irrelevant information (Henson & Gagnepain, 2010; McClelland, McNaughton, & O'Reilly, 1995; van Kesteren, Ruiter, Fernández, & Henson, 2012), though most would assume that such events should be unaffected by the presence of a schema. If information is unrelated to a schema, then the schema should not modulate its encoding or retention. Here we asked whether memory and generalisation behaviour for both schema-relevant and -irrelevant information is modulated by the presence of schematic information.

1.1. Schemas and memory

Information that is either congruent (Atienza, Crespo-Garcia, & Cantero, 2011; Brewer & Treyens, 1981; Mandler & Johnson, 1977; van Kesteren, Fernández, Norris, & Hermans, 2010) or incongruent (Frank et al., 2018; Hunt & Worthen, 2006; Tulving & Kroll, 1995) with schematic information is often better remembered than information unrelated to a schema. Brewer and Treyens (1981) had participants recall items present in an office they were asked to wait in for 35 s. Items in the office were either congruent (i.e., items expected given the context), such as a desk, or incongruent (i.e., items that would be unusual given the context), such as a picnic basket. In our bathroom example, this would equate to finding a towel in the bathroom (schema-congruent) versus finding a microwave in the bathroom (schema-incongruent). It was found that schema-congruent items were better recalled than incongruent items.

Although memory performance for schema-congruent information is typically greater than for schema-incongruent information, research also suggests that schemas can boost memory performance for schema-incongruent information relative to unrelated information. Frank et al. (2018) had participants learn A-B, B-C, C-D, D-A pairings, with the first A-B pairing providing schematic context and the second B-C pairing including schema-congruent or -incongruent information (the final D element was also schema-congruent). For example, the A-B pairing could be Farm-Tractor, with C then being congruent (e.g., Farmer) or incongruent (e.g., Lawyer). These two conditions were compared to a control condition where all item pairings were unrelated (e.g., Torch-Professor, Professor-Lego). Consistent with Brewer and Treyens (1981), memory performance was greater in the schema-congruent relative to -incongruent condition. However, they also saw (in some circumstances) greater memory performance in the schema-incongruent relative to the unrelated control condition. Consequently, schemas may benefit the encoding and retention of congruent and incongruent information under specific conditions.

Though a facilitation effect for both schema-congruent and -incongruent information is relatively well established (see Rojahn & Petti-grew, 1992, for a meta-analysis), what drives this facilitation is still a matter of debate (Quent, Henson, & Greve, 2021; Sakamoto & Love, 2004; van Kesteren et al., 2012). Recent theoretical and experimental work has attempted to explain why schemas benefit both congruent and incongruent information. The schema-linked interactions between medial prefrontal and medial temporal regions (SLIMMs) model (van Kesteren et al., 2012), an extension of the PIMMS model (Henson & Gagnepain, 2010), predicts that schema-incongruent information results in high prediction error and that this potentiates encoding of the event in medial temporal regions. Conversely, schema-congruent information is

detected by the medial prefrontal cortex (mPFC) and this region strengthens pre-existing neocortical associations between the relevant elements. As a result, memory can be positively affected when information is congruent or incongruent with a relevant schema, relative to when information is schema-irrelevant or of "neutral" congruence.

Greve et al. (2019) tested this prediction by having participants learn the value of a set of objects through trial-and-error learning where certain objects were of higher value (e.g., umbrella compared to shoes). During test, participants were presented with both old and new displays that contained new combinations of the same objects. Participants had to decide if it was an old or new display and which set of objects was more valuable based on previous learning. Critically, the value of the items either remained the same across learning (congruent), changed on the final trial (incongruent) or changed on every trial (unrelated). Across four experiments, they found that recognition was better for congruent and incongruent information compared to unrelated information. More specifically, there was an advantage for first-encountered episodes in the congruent trials despite there being no distinguishing characteristics from the other trials at this point. This suggests schema congruency benefited these trials through post-encoding processes by potentially prioritising these memories for consolidation. In contrast, memory for the final trial, where the trial switched values in the incongruent condition, was greater in the incongruent than the unrelated condition suggesting the memory benefit for these trials was driven by prediction error. The findings demonstrate how the congruency and incongruency benefit may result from dissociable processes allowing for memory enhancement in both circumstances.

1.2. Schemas and generalisation

Schemas are also thought to be critical to our ability to generalise to novel, but related, events. Sweegers and Talamini (2014) examined how the presence of an association between certain facial characteristics (e. g., wearing a hat, face shape) and a location in hexagonal space could be learned and used to make inferences for novel faces. Along with benefiting later recall of old items, the presence of face-location associations could also be used to make novel inferences about the location for unseen faces. This was observed shortly after studying the material. In another domain, Mirković and Gaskell (2016) had participants learn new vocabulary using a word-picture matching task. When tested on their ability to generalise suffixes, participants showed they had extracted the suffix rules and were able to use these rules to generalise to novel word-picture pairs. Studies such as these are plentiful, using tasks such as inferential reasoning (e.g., Zeithamova, Dominick, & Preston, 2012), associative weather prediction (e.g., Kumaran, Summerfield, Hassabis, & Maguire, 2009) and novel affix completion (e.g., Tamminen, Davis, Merkx, & Rastle, 2012). Across these studies, it has been shown that schematic representations based on relational information can be used to make generalisations about novel stimuli. In this way, schemas do not simply function to benefit memory encoding and recall, but also help guide our behaviour for future instances.

1.3. Models of schema processing

The extraction of schematic representations is often linked to systems consolidation (Dudai, 2012). Perhaps the most influential model of systems consolidation is the Complementary Learning Systems account (CLS; McClelland et al., 1995). CLS proposes two distinct learning systems – a fast encoding system in the hippocampus that stores pattern separated representations of events, and a slower learning system in the neocortex that extracts higher-order meaning. The more abstract representations in the neocortex are thought to be schematic and support generalisation. Critically, CLS predicts that schema formation should be relatively slow (potentially requiring periods of offline consolidation during sleep; Born & Wilhelm, 2012; Davis & Gaskell, 2009; McClelland et al., 1995; Zola-Morgan & Squire, 1990). Once formed, schemas are

believed to be independent of the individual event representations that supported their formation. Given that this class of model requires learning a latent schematic representation prior to retrieval, we refer to them as "encoding-based" mechanisms. These mechanisms are conceptually related to prototype models of categorisation that propose separate 'prototypical' or 'average' representations that are used to assess category membership of novel exemplars (Rosch, 1973; Smith & Minda, 2000).

Alternatively, generalisation behaviour consistent with the presence of a schema may be supported by the retrieval of individual event representations (Kumaran & McClelland, 2012; Schapiro, Turk-Browne, Botvinick, & Norman, 2017), conceptually similar to exemplar models of categorisation (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986). These models allow for generalisation "on the fly" at the point of retrieval (and we therefore refer to them as "retrieval-based" models). As retrieval-based models do not rely on the process of systems consolidation, the ability to generalise can occur more rapidly (relative to encoding-based models). However, generalisation is contingent on the retrieval of individual events and, as such, forgetting of these events will result in decreased generalisation performance. The key difference between encoding-based and retrieval-based models is how memories are used. For encoding-based models, schematic representations are relied upon when generalising. In contrast, retrieval-based models argue that there is no need to form an independent schematic representation; instead, we generalise based on sampling individual events.

More recent neurocognitive models propose both rapid retrieval-based and slower encoding-based mechanisms. For example, Schapiro et al. (2017) have shown that different pathways in the hippocampus could theoretically support both encoding-based and retrieval-based mechanisms simultaneously. Similarly, the REMERGE (Kumaran & McClelland, 2012) model builds on the CLS model, proposing that the hippocampus can initially support generalisation before systems consolidation has occurred via retrieval-based mechanisms. If this retrieval-based mechanism was combined with the original encoding-based mechanism of CLS, such a hybrid model could accommodate findings that neocortical reorganisation can take days, weeks, months or years (Born & Wilhelm, 2012; Davis & Gaskell, 2009; McClelland et al., 1995; Zola-Morgan & Squire, 1990), with recent research suggesting that schema effects can be observed immediately following learning (e. g., Sweegers & Talamini, 2014; Tompary & Thompson-Schill, 2021). Therefore, both encoding- and retrieval-based mechanisms could operate at different time scales and under different conditions. Though we do not directly test the assumptions of these models in the present study, our findings relate to this broader theoretical debate (see Section 7).

1.4. Precision measures of schema processing

Most studies related to schema-processing rely on binary decisions related to both memory and generalisation behaviour. Though this provides insight into whether retrieval or generalisation can occur, it is less informative about patterns of responses across trials. For example, if an underlying pattern (or schema) across trials is present, do participants recreate that pattern when generalising to novel events? Using an approach that assesses the pattern of responses across trials would allow us to more clearly identify biases in behaviour for both schema-relevant and -irrelevant information.

Precision (or continuous) memory measures provide a non-binary output that allows us to look at patterns of responses across trials. Used extensively to study working memory (Bays, Catalao, & Husain, 2009; Luck & Vogel, 2013; Peich, Husain, & Bays, 2013; Sun et al., 2017; Zhang & Luck, 2008) and long-term memory (Berens, Richards, & Horner, 2020; Harlow & Donaldson, 2013; Korkki, Richter, Jeyar-athnarajah, & Simons, 2020; Nilakantan, Bridge, VanHaerents, & Voss, 2018; Richter, Bays, Jeyarathnarajah, & Simons, 2019; Tompary, Zhou, & Davachi, 2020), precision memory experiments associate a stimulus (e.g., a word or object) with a continuous property (e.g., colour or

location around a circle). At test, participants are required to retrieve the associated property of the stimulus. In the case of a location around a circle, performance is measured as the degree of error (real location vs. retrieved location). With a continuous measure like this, we can assess the distribution of retrieved locations across trials. For example, we can compare the distribution of memory trials for a set of stimuli whose locations were dictated by an underlying pattern to see if the retrieved distribution matches the encoded distribution.

More recently, precision measures have been used to assess schema processing. The idea here is that an underlying pattern can dictate the associated properties of a set of stimuli. For example, when learning word-location associations, the locations can conform to a von Mises (circular Gaussian) distribution, such that they are clustered in a specific location around the circle (see Richards et al., 2014 for a conceptually similar approach in rodents). Brady, Schacter, and Alvarez (2018) associated objects with colours (on a colour wheel), where the colours of exemplars of a category (e.g., lamps) were clustered. They showed that participants were systematically biased towards the average of the colour category when retrieving the colour of previously presented objects. Further, using object-location associations, Richter et al. (2019) showed that learning of new associations was modulated by the congruence of the new object-locations with the underlying pattern, but only following a period of sleep, consistent with the slow extraction of schematic representations predicted by CLS.

The effect of an underlying pattern on forgetting has also been assessed. Berens et al. (2020) required participants to learn word-location associations around a circle. Words came from one of two semantic categories (i.e., human-made and natural), with one category having locations clustered (i.e., locations were more likely to appear in one area of the circle) while the other was non-clustered (i.e., no relationship between word meanings and locations). Using this paradigm, measures of memory accessibility (i.e., proportion of word-locations retrieved) and precision (i.e., degree of location accuracy given successful word-location retrieval) were assessed. They found that the presence of a pattern differentially influenced memory accessibility and precision. Specifically, accessibility was higher, but precision was lower, in the clustered relative to the non-clustered condition. Consequently, schematic information affects distinct memory components differently – benefiting overall accessibility at the cost of precision.

The above precision memory studies focused on how an underlying pattern modulates retrieval performance and new learning. Using both previously presented and novel (semantically related) stimuli, Tompary et al. (2020) investigated how these underlying patterns modulate memory (old stimuli) and generalisation (novel stimuli) behaviour. Participants learned to associate objects with locations around a circle. The locations of images were drawn from two cosine distributions, with the means of these distributions on opposite sides of the circle (separated by 180°). They found that schema use, relative to the use of episodic memory, increased with time, but interestingly, schema memory also showed evidence of decay. This is in line with evidence elsewhere showing that schema benefits on memory performance can decrease with time (Antony et al., 2021; Berens et al., 2020). Therefore, precision measures have been used to assess memory and generalisation behaviour in the presence of a schema. However, they have not been used to assess behaviour for schema-irrelevant information.

1.5. Schemas and irrelevant information

Across the studies discussed above, many consider information that is either congruent or incongruent with pre-existing knowledge. Most do not study how schematic information influences memory and generalisation behaviour for schema-irrelevant information. Irrelevant here relates to information not in the same semantic category as the schematic items. In our earlier example, the presence of a "bathroom" schema should have little impact on where you predict the house plant will be located. Whereas a schema-incongruent item (e.g., a towel in the

living room) conflicts with an existing schema and therefore can change or update the schema, a schema-irrelevant item (e.g., a house plant) is neither congruent nor incongruent with the schema. Our ability to remember where a house plant is located, or predict where a house plant would be located, should therefore be unaffected by the presence or absence of schematic information related to bathroom items. In the case of the Berens et al. (2020) study, one semantic category (e.g., human-made – the experimental equivalent to bathroom items in our example) was associated with an underlying pattern (the clustered condition), whereas the other semantic category (i.e., natural – the experimental equivalent to house plants in our example) was not (the non-clustered condition). The location of words in the non-clustered condition are not relevant to the "human-made" pattern in this case, so we define these items as "schema-irrelevant".

Though evidence suggests that schemas can bias memory by increasing false alarms (Neuschatz, Lampinen, Preston, Hawkins, & Toglia, 2002) and increasing the number of false memories (Kleider, Pezdek, Goldinger, & Kirk, 2008), it is not often considered how schemas influence information that is not relevant to themselves. Though some studies have included irrelevant information in their paradigm (e.g., Frank et al., 2018; Greve et al., 2019), this was used as a control condition to compare performance relative to congruent and incongruent information, as opposed to examining how the presence of schematic information could bias behaviour for this irrelevant information. Indeed, our schema-irrelevant (non-clustered) condition was first created as a control condition before we focussed our attention on behavioural biases specifically in this condition.

Returning to precision measures of memory and generalisation, Tompary et al. (2020) did not include a control condition where locations for one semantic group were randomly distributed. Instead, they used two clustered distributions that were separated by 180°, and as such, it is difficult to disentangle the effects of one cluster against another. In the present experiments, we used the clustered and non-clustered conditions introduced in Berens et al. (2020) and introduced novel semantically related items (as in Tompary et al., 2020). This allowed us to focus on behaviour in the non-clustered condition, where the words are from a separate semantic category to the clustered condition, and the locations of these words are randomly distributed. As such, word-locations in the non-clustered condition are technically irrelevant to extracting the underlying pattern (or schema) in the clustered condition.

Critical to the present studies is the use of an underlying pattern across a set of word-location associations to provide insight into schema processing. Ghosh and Gilboa (2014) recently proposed specific features that define a schema: (1) an associative network structure that represents units of information and their interrelations, (2) are based on multiple episodic events, (3) lack specificity in unit details, and (4) have a degree of adaptability. Concerning the present experiment, a participant may rely on a schema that maps the associations between words and locations. For instance, a schema that captures how certain semantically related words are clustered to a particular area of the circle. Here we have a set of related word-location associations (related semantically and by location), conforming to the first criteria. Participants are encoding multiple related word-location associations, conforming to the second criteria. If a pattern is extracted (e.g., the average word-location association for a given semantic category), this conforms to the third criteria. Finally, although we do not assess "adaptability" per se (i.e., the extent to which an existing schema can be flexibly updated), we assess behaviour shortly after encoding. Therefore, if behaviour is consistent with schema processing, schematic representations must have developed rapidly (i.e., during encoding). Consequently, our paradigm conforms to the stringent criteria outlined by Ghosh and Gilboa (2014) and readily fits with less stringent definitions of schemas (see Preston & Eichenbaum, 2013; van Kesteren et al., 2012). We return to whether schemas are being extracted and used in our paradigm in the General Discussion.

1.6. Overview of experiments

We explored how the presence of a pattern influences memory and generalisation when one condition possesses a pattern and the other does not. We used an experimental design similar to Berens et al. (2020), but with novel items at test. Participants learned word-location associations around a circle. Word stimuli came from two semantic categories: human-made (e.g., chair, computer) and natural (e.g., leaf, giraffe). The locations associated with these words were either clustered or non-clustered. By including the non-clustered condition, we explored how memory and generalisation behaviour for semantically related information (i.e., words belonging to the clustered category) and semantically unrelated (i.e., words belonging to the non-clustered category) was modulated by the presence of an underlying pattern. Specifically, participants may form a 'schematic' representation for the semantic category associated with the clustered condition, allowing them to make predictions about the possible locations of novel words belonging to the same category. In contrast, for the semantic category associated with the non-clustered condition, there was no underlying pattern; this allowed us to observe how the presence of a pattern in the clustered condition influences memory and generalisation behaviour for 'schema-irrelevant' words. Across four experiments, we manipulated delay between Study and Test, and whether we collected data in person (in the lab) or online, providing evidence that schematic information can bias memory and generalisation behaviour in the schema-irrelevant (non-clustered) condition.

2. Experiment 1

In Experiment 1, we asked two questions: (1) does the presence of a pattern increase memory performance in the clustered relative to non-clustered condition, and (2) can participants generalise, such that they place novel words in locations similar to the pattern in the clustered relative to the non-clustered condition? To answer these questions, we had two pre-registered hypotheses: (1) participants' overall memory performance will be greater in the clustered relative to non-clustered condition (as measured by 'Total Information', see Methods), and (2) the distribution of locations for novel words will be more similar to the underlying pattern (von Mises distribution) in the clustered relative to non-clustered condition (whereas the distribution in non-clustered condition will be more uniform; as measured by Kullback-Leibler divergence). The preregistration for Experiment 1 is available at: https://osf.io/h6wba/.

2.1. Methods

2.1.1. Participants

2.1.1.1. Power analysis. Two power calculations were conducted to estimate the required sample size to explore the pre-registered hypotheses. First, to estimate the required sample size for the effect of clustering on total information, G*Power (3.1.9.2; Faul, Erdfelder, Lang, & Buchner, 2007) was used to perform an *a priori* power analysis. A power analysis was computed for a paired samples *t*-test comparing total information in the clustered and non-clustered conditions. The effect size for the analysis was estimated from a pilot investigation reported in Berens et al. (2020); this pilot study estimated an effect of $d = 0.33$, with the clustered condition showing significantly greater total information than the non-clustered. This effect size estimate, along with an α (one-tailed) $= 0.05$ and power $= 0.80$ were used. A suggested sample size of $N = 59$ usable datasets was estimated.

Second, data simulations were conducted to estimate the required sample size to compare the distribution of locations for clustered and non-clustered novel words to the experimental distributions. Data simulations were run to identify: (1) the minimum number of responses

required to get a reliable estimate of Kullback-Leibler divergence (D_{KL}), and (2) to determine the required sample size to gain 80% power. The number of participants and words varied on each iteration of 100 simulations. The simulation assumed that each participant reproduced the spatial distribution of clustered and non-clustered locations with varying accuracy. Specifically, the reproduced distributions took the form of a von Mises probability density with a mean parameter drawn from a von Mises ($\mu = 0$, $\kappa = 5.5$). The concentration parameter for each distribution was then sampled independently from a gamma distribution with a mean of 2 (i.e., the true concentration) and a standard deviation of 2. These hyperparameters were estimated from a previous pilot study by Berens et al. (2020). Non-parametric density functions were then estimated from the simulated responses in both conditions separately. Using D_{KL}, the probability density across circular locations was then compared to the experimentally imposed von Mises distribution ($\mu = 0$ and $\kappa = 2$). A generalised linear mixed-effects (GLME) model, using the same parameters as described below (see Section 2.1.4), was fit to the D_{KL} measures of both clustered and non-clustered responses with varying intercepts based on each 'participant'. No random slopes were computed for these simulations. It was found that a minimum of 11 words and 9 participants were required. Given the above, a final sample of 60 usable datasets was pre-registered.

2.1.1.2. Final sample. Sixty-nine participants (63 female) were recruited for the study. The mean age was 19.59 years ($SD = 2.22$ years). The mixture model failed to converge for 6 participants, meaning the final sample consisted of 63 participants (58 female) with a mean age of 19.63 years (SD = 2.30 years). The over-recruitment resulted from a minor coding error leading to incorrect rejection of valid model fits for three participants. Participants were fluent English-speakers with normal or corrected-to-normal vision and were recruited from the University of York student population and took part in exchange for course credit. Notably, for Experiments 1 and 2, there are few participants identifying as male that make up the sample. However, this is not the case for Experiments 3 and 4 where there is a relatively equal number of males and females. Given the results in Experiments 3 and 4 replicate those in Experiments 1 and 2, we do not believe the gender imbalance affects our conclusions. Ethical approval for all experiments was granted by the Department of Psychology Ethics Committee at the University of York. Exclusion criteria for the data are detailed below (see Section 2.1.4.4).

2.1.2. Materials

2.1.2.1. Word lists. Eight stimulus sets were generated consisting of a list of 240 English nouns (https://osf.io/a8536/). These words belonged to one of two semantic categories: human-made object nouns (120 words) and natural object nouns (120 words). Each group of 120 words was split into four sets of 30, ensuring the semantic properties across sets were reasonably equated. For each participant, one of these sets was assigned to the "new word" (generalisation) condition, and the remaining three sets (90 words) were placed into the "old word" (memory) condition. The assignment of the 4 sets to the novel word condition was counterbalanced across participants.

To develop the word lists and ensure sufficient semantic distance between categories, semantic representations of 324 words were extracted from a pre-trained word2vec model (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). This pre-trained model of numerical word representations contained over 3 million English words based on the Google News dataset. The semantic similarity between word representations was then estimated via Euclidean distance. Simulations were run to ensure a small semantic distance between words of the same category (e.g., natural), whilst ensuring a large semantic distance between cross-category pairs. To meet these criteria, simulations were run using 10,000 iterations to identify a word list containing a total of 240

(120 human-made and 120 natural) words. The final list had a mean semantic distance of 4.24 ($SD = 0.47$) within and 4.44 ($SD = 0.42$) between categories, suggesting the two lists were sufficiently distinct in terms of semantic grouping. The distributions of semantic distances within each group were comparable as compared using the Kolmogorov-Smirnov test ($D = 0.01$). After generating the lists, we ensured word length and frequency of use in natural language, as quantified using the Zipf-scale of the SUBTLEX-UK database (van Heuven, Mandera, Keuleers, & Brysbaert, 2014), were comparable across lists. Finally, to split the two lists into eight sets of 30, a further 10,000 iterations were run. Sub-lists were generated by controlling for the mean and variance in Euclidean distance, the difference in distributions using the Kolmogorov–Smirnov test, word frequency and word length. The relevant scripts can be found here: https://osf.io/bxru4/.

Though there is some degree of semantic overlap between the two categories, it is important to note evidence for distinct superordinate categories from neuropsychological studies (Damasio, Grabowski, Tranel, Hichwa, & Damasio, 1996; Warrington & Shallice, 1984). In this study, word2vec was used as a tool to ensure selection of words within vs. across lists minimised semantic distance within-category and maximised semantic distance across-categories. Notably, a linear support vector machine was able classify items as either human-made or natural with a high degree of accuracy (98%), showing that the categories were highly separable. The code for this can be found here: https://osf.io/y7jum/.

2.1.3. Procedure

2.1.3.1. Study phase. Participants learned associations between different locations around a circle and a specific word displayed on each trial. During the study phase, 180 words were presented. One of the semantic categories was assigned to the clustered condition (counterbalanced across participants). Word-locations in this condition were clustered by sampling from a von Mises distribution with a fixed width ($\kappa = 2.0$) and a fixed mean (randomly selected for each participant). The other semantic category was assigned to the non-clustered condition. Word-locations in this condition were randomly distributed around the circle by sampling from a uniform distribution. Participants were not informed about the presence of the semantic categories or the clustering manipulation. They were only told that they would need to remember each individual word-location association.

All stimuli were presented using MATLAB (2019) and the COGENT 2000 toolbox (www.vislab.ucl.ac.uk/cogent/index.html) on a desktop PC. Participants sat approximately 50 cm away from the screen so that the circle subtended ~16 visual degrees. Each study trial (shown in Fig. 1) started with a fixation cross (1 s), followed by a location marker (2 s). The location marker and circle were then removed and the study word displayed (4 s). Subsequently, with the word still present, the circle and marker, the latter of which was redrawn at a random location around the circle, were presented. Participants were asked to reposition the cursor at the cued location using a mouse; this response window lasted 6 s. Repositioning the marker during study ensured participants deliberately attended to the word-location association as opposed to passively viewing. If participants did not respond within the 6 s time window, or selected an area >10° from the presented location, the trial was repeated, with a red fixation cross at the beginning of the trial to alert them to this repetition. The average number of repetitions across all experiments was 0.17 trials ($SD = 0.45$, Proportion = 0.002) and 0.15 trials ($SD = 0.43$, Proportion = 0.002) for the non-clustered and clustered conditions, respectively.

Before starting the study phase, participants were given practise trials to ensure they understood the task and knew how to make responses. The practise trials used similar parameters as described above, but with abstract nouns (e.g., beauty, jealousy, integrity) that held no semantic clustering and no relation to words within the study lists. There
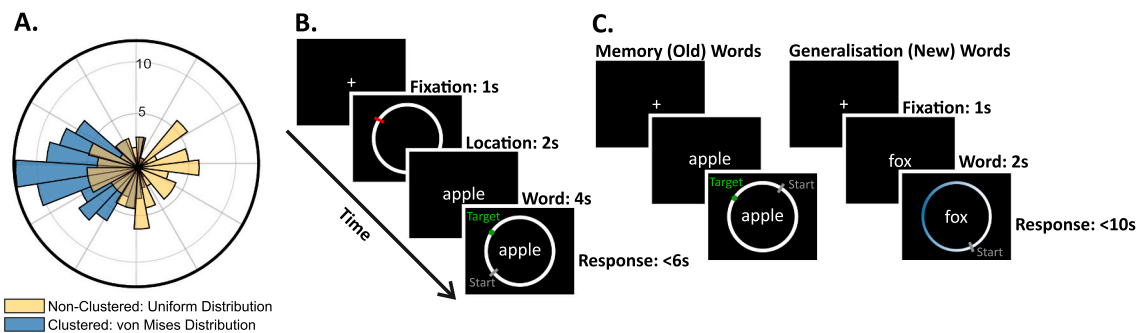
Fig. 1. Experimental Design. (A) **Clustering**: Locations were selected to follow either a von Mises distribution (clustered condition) or a uniform distribution (non-clustered condition) around the circle. The polar plot shows an example of distributed locations for one participant. The clustered and non-clustered conditions were associated with either the human-made or natural word category (counterbalanced across participants) and the centre of the clustered distribution was randomised for each participant. Numbers on the polar plot show the number of words located in that area of the circle. (B) **Study Phase**: Participants were presented with a fixation cross (1 s), followed by the location (2 s), then the word alone (4 s), and then presented with the word, the circle, and a randomly placed marker to make a response (6 s). Participants moved the marker from the start location back to the location just presented. (C) **Test Phase**: Participants were first presented with a fixation cross (1 s), the word alone (2 s) and then asked to replace the marker from the randomly generated start position back to the remembered location (memory trials) or make an inference based on experience (generalisation trials, 10s). In the example above, natural words were assigned to the clustered condition, and the blue shading in the generalisation trial shows the area of the circle they are likely to generalise to in the clustered condition. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

was a total of 10 practise trials. Following Study, participants took part in an immediate Test phase.

2.1.3.2. Test phase. At test, participants were required to recall the 180 previously presented word-location associations and to select locations for novel words (60 words). These novel words came from the same semantic groupings as above. The old and new words were intermixed, and presentation order was randomised.

On each test trial (Fig. 1), a fixation cross appeared (1 s), followed by the word (2 s) and then the circle and marker appeared, with the marker being presented at a random location around the circle. Participants had a 10s response window to move the marker (via the mouse) back to the remembered location, or to make a best guess if they had forgotten. Participants were not told about the presence of novel words at test, with the trial structure being identical. Participants were told to make a best guess for any words they had forgotten the location for.

2.1.3.3. Introspection questionnaire. Following test, participants completed an Introspection Questionnaire. The questionnaire addressed their perceptions on task difficulty, asked them to report their strategies for words they had forgotten, whether they noticed any words presented at Test that were not presented at Study, their strategies for these words and whether they felt a pattern was present in the presentation of word-location associations. The questionnaire is located here: https://osf.io/7fgzm/.

2.1.4. Data handling

2.1.4.1. Mixture model estimation. Using mixture modelling, we estimated accessibility (i.e., word-location retrieval probability) and precision (i.e., how precisely are locations remembered given they are accessible) for individual participants. We calculated the replacement error for each response (i.e., the angular difference between the correct location and remembered location). These angular errors are assumed to come from one of two distributions: (1) a circular uniform distribution representing guesses, and (2) a von Mises distribution representing accessible word-location associations, whose variance represents the 'precision' that locations are remembered. These two distributions have associated prior probabilities, which are statistics reflecting the overall proportion of responses belonging to either distribution. For the von Mises distribution, the prior (p) represents retrieval probability (i.e., accessibility). This distribution also has two other parameters: mean (μ)

and concentration (κ). The value of μ was fixed at zero, assuming the average error of responses was zero. The value of κ represents the variance, or precision, in responses. Higher κ values indicate a narrower distribution (higher precision), lower κ values indicate a wider distribution (lower precision).

Mixture modelling was conducted using the HoopStats toolbox developed in Berens et al. (2020), which can be found here: https://osf.io/8mzyc/. First, an Expectation Maximisation (EM) algorithm was used to estimate p and κ, for each participant and the clustered and non-clustered items, separately. The overall fit of this model was compared to a reduced model where all angular errors are assumed to be from a uniform distribution (i.e., no mnemonic information is present). This comparison was conducted using the Bayesian Information Criterion (BIC). If the BIC was less than -10 (i.e., evidence in favour of the two-distribution model), the parameters returned from the EM were accepted. If, however, the BIC was greater than -10, representing a poorly fit model, an alternative fitting procedure was implemented. This failure to meet criterion often occurs when low accessibility is present in the data ($p \lesssim 0.2$). For this alternative model, the p value was systematically varied over several steps, with κ being estimated from the corresponding responses with the smallest angular error. Using this method, valid model fits could be found that were otherwise missed by the EM algorithm. If this alternate model produced a better fit than the single uniform distribution, again using the BIC < -10 criterion, these parameters were accepted. If BIC > -10, or the estimates of κ were modelled on fewer than 8 trials, the participant's entire dataset was excluded. Note that this alternative fitting procedure was not used in all cases, as it only involves searching regions of the parameter space that correspond to low levels of memory accessibility (where the original mixture model was likely to fail).

2.1.4.1.1. Conversion to entropy measures. Once both the p and κ parameters were estimated for clustered and non-clustered trials, they were converted into information entropy measures I_p and I_κ. Small values of I_p indicate lower levels of accessibility. Similarly, small values of I_κ indicate poor precision. Conversion of p and κ to I_p and I_κ allows for a more direct comparison between the two, as they describe performance using the same metric: information gain (in nats) relative to random responses. Additionally, we computed a combined measure of memory performance, "Total Information" (I_t), which is directly proportional to both I_p and I_κ ($I_t = \frac{I_p * I_k}{\log(2\pi)}$). I_t reflects the total amount of mnemonic information present at the point of retrieval, which is a function of both the proportion of word-location pairs that were

accessible and the precision of these accessible word locations. Hypothesis 1 uses this measure of Total Information to assess overall memory performance between the clustered and non-clustered conditions.

2.1.4.2. Kernel density estimation. Kernel density estimates were computed to characterise the distribution of location responses; this was identical to Berens et al. (2020). The primary purpose of the kernel density estimates was to compute the D_{KL} between participant's responses and the pattern of studied locations. They were also used to: (1) plot the distribution of angular errors for memory trials and (2) plot the distribution of responses relative to the experimentally imposed von Mises distribution for memory and generalisation trials. To do this, a von Mises probability density function, with a concentration of $\kappa = 2$, was centred around each response. This distribution acted as a smoothing kernel that spread a small portion of the overall density around the local area. As such, the density estimates at a given angle were taken as the mean probability density value across all these distributions. The responses were either angular errors for each condition (for memory trials) or angular differences between the responses and the centre of the experimentally imposed cluster (for memory and generalisation trials).

2.1.4.3. Kullback-Leibler divergence. Once the spatial distribution of responses was estimated through the kernel density function, D_{KL} was computed to assess the similarity between specific distributions. D_{KL} measures divergence between two distributions, with higher values representing greater divergence (i.e., less similarity) between the two; this was computed via numerical integration, as in Berens et al. (2020), rather than by using a discrete approximation. First, we assessed how divergent the distributions for clustered and non-clustered novel words (i.e., generalisation trials) were to the reference distribution (i.e., the experimentally imposed von Mises distribution associated with the clustered condition). The distribution of clustered novel words was predicted to be less divergent to the underlying von Mises pattern relative to the non-clustered condition (Hypothesis 2). Second, we assessed how divergent the distributions for clustered and non-clustered novel words were to a uniform distribution (i.e., no pattern). The distribution of non-clustered novel words was predicted to be less divergent from a uniform distribution relative to the clustered condition (Hypothesis 2).

2.1.4.4. Exclusion criteria. All exclusion criteria were pre-registered. If an additional exclusion was included that was not pre-registered, this is explicitly identified throughout the report.

At Study, participants repeated trials where they were $> 10°$ away from the location presented. If a single trial was repeated 5 times or more it was removed from later statistical analyses to ensure that the extra encoding of these word-location pairs did not impact retrieval. This cut-off was selected based on an observation made during the in-lab piloting for Berens et al. (2020), where very few participants needed to repeat a trial on more than five occasions, with only ~10% of participants requiring >5 repetitions on a given trial before they would place the marker within 5° of the presented location. For all four experiments in this manuscript, only 13 trials across participants and experiments were repeated >5 times showing few trials were removed for this reason. Along with this, if no response was given at Test, the trial was excluded from later analyses to ensure only trials where participants gave an explicit response were included. Across all experiments reported, on average, participants did not respond to 1.75 trials (*SD* = 4.23).

Datasets would only be included for analysis when the following criteria were met: (1) both the study and test trials were complete, (2) the number of old words with no response did not exceed 20 trials for the clustered and non-clustered separately, (3) the number of novel words not responded to did not exceed 15 trials for the clustered and non-

clustered separately, (4) the dataset was not corrupted, and (5) the mixture model could be fit adequately to the data (see: Section 2.1.4.1, above).

2.1.5. Statistical analysis

All statistical analyses reported in the main results sections of all experiments were preregistered. Where exploratory analyses were run, these are clearly labelled as such. We computed three separate GLME models. The models were used to predict (1) Total Information (I_t), (2) D_{KL} in comparison to the experimentally imposed von Mises distribution, and (3) D_{KL} in comparison to a uniform distribution. The first model relates to Hypothesis 1, assessing whether overall memory performance differs between the clustered and non-clustered conditions. The second and third models relate to Hypothesis 2, testing whether participants can position novel words from the same semantic category according to the underlying pattern.

For each model, we compared the clustered and non-clustered conditions for each measure of interest. All models were fit to the data using a log link function, a gamma distribution to model the spread of the data, and were estimated using the maximum likelihood fitting procedure in the MATLAB (2019) Statistics and Machine Learning Toolbox. The models included the independent variable of clustering (0 = Non-Clustered; 1 = Clustered). In addition to this fixed effect, a set of random effect parameters (2 per participant) were included. One random effect allowed the intercepts to vary based on participant, the other allowed the effect of clustering to vary by participant. All elements of the associated random effects covariance matrix were estimated from the data. For the D_{KL} model that assessed the uniformity of clustered and non-clustered responses, the model did not converge. As a result, the random slopes for clustering were removed for this comparison across experiments using this analysis.

All mean values represent the mean estimate of the population derived from the GLME. Further, the Cohen's *d* values reported were calculated as reported in Berens et al. (2020) and estimated only on the fixed factors. All analyses use two-tailed tests unless otherwise specified. For all results, Bayes Factors were computed. We pre-registered that Bayes Factors would only be reported for non-significant results to aid in interpreting the outcome of these tests by assessing whether there was greater support for the null relative to the alternative hypothesis. However, we feel these can be informative for both significant and non-significant results. A further deviation from the pre-registration was how Bayes Factors were computed. Previously we specified Bayes Factors would be computed in JASP. However, to increase reproducibility, the computation used in Berens et al. (2020) was implemented for all Bayes Factors reported. For these analyses, a prior Cauchy distribution of $r = 0.707$ centred at 0 was used; these parameters were identical to the pre-registration.

2.2. Results

2.2.1. Memory

Fig. 2 shows the Total Information metric for Experiments 1–4, as well as the probability density estimates for angular error. The angular error plots demonstrate differences, across the delay periods, in the degrees of error around the circle. These plots demonstrate possible differences between conditions based on accessibility and precision. Specifically, the higher peaks in the clustered condition suggest greater accessibility, whilst the narrower distributions for the non-clustered condition suggest greater memory precision. An analysis of these metrics is presented in the Exploratory Analysis: Across-Experiment Comparisons section 6.

Hypothesis 1 related to whether clustering benefits overall memory performance. Consistent with this, in Experiment 1, total information was significantly greater in the clustered relative to the non-clustered condition, $t(124) = 1.99$, $p = .049$, $d = 0.35$, $BF_{01} = 0.87$. Though significant, the Bayes Factor was inconclusive.
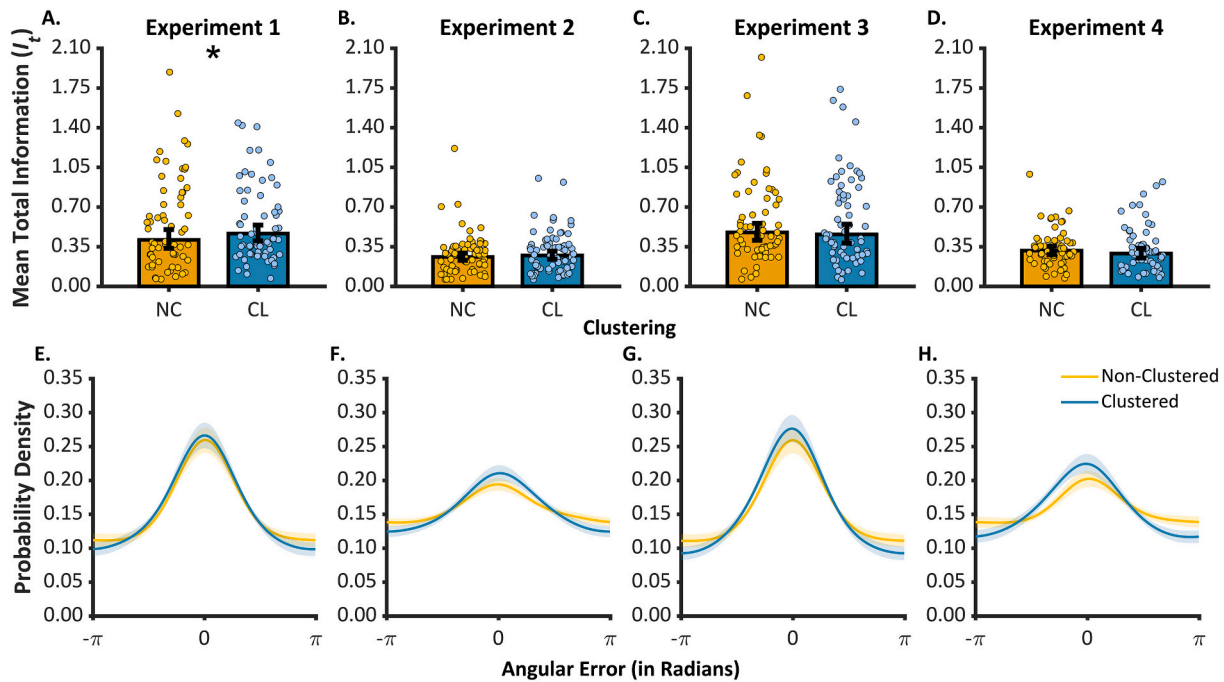
Fig. 2. Memory performance across Experiments 1–4. *A-D*: Mean Total Information (I_t) across Experiments 1–4 as a function of clustering (clustered and non-clustered). Individual data points represent participant scores. *E-G*: Spatial distribution of angular errors across Experiments 1–4, 0 here represents 0 degrees of error. Error bars represent 95% confidence intervals around the mean for all plots. *$p < .05$. CL = Clustered. NC = Non-Clustered.

2.2.2. Generalisation

Fig. 3 shows generalisation behaviour for the novel words for Experiments 1–4. Hypothesis 2 was that the distribution of selected locations for clustered generalisation trials would be more similar to (less divergent from) the experimentally imposed von Mises distribution than the distribution of locations for non-clustered generalisation trials. Consistent with this, the clustered responses were significantly less divergent from the von Mises distribution than the non-clustered

responses, $t(124) = 4.26, p < .001, d = 0.76, BF_{01} = 0.001$. This suggests participants were able to make reasonable guesses or predictions about where novel words would be located based on the learnt locations from the same semantic category. Specifically, participants placed novel words in the clustered category in similar locations to the old clustered items relative to novel words in the non-clustered category.

We also predicted that the distribution of locations for non-clustered novel words would be more similar to (less divergent from) a uniform
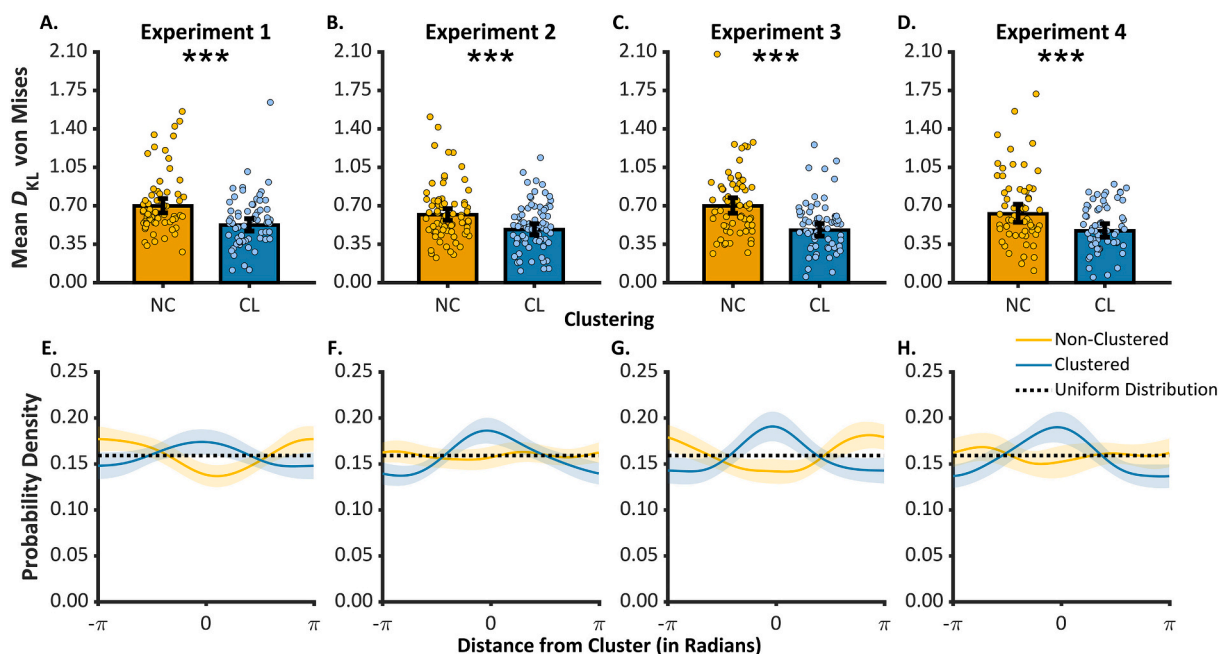

Fig. 3. Generalisation behaviour across Experiments 1–4. *A-D*: Mean divergence (D_{KL}) from the experimentally imposed von Mises distribution across Experiments 1–4 as a function of clustering (clustered and non-clustered). Individual data points represent participant scores. *E-G*: Spatial distribution of locations selected for novel words, centred to the experimentally imposed von Mises distribution, for Experiments 1–4. Error bars represent 95% confidence intervals around the mean for all plots. ***$p \leq .001$. CL = Clustered. NC = Non-Clustered.

distribution relative to the distribution for clustered words. In other words, we expected that the uniformity (or 'entropy') of responses should be greater in the non-clustered relative to the clustered condition. Inconsistent with this prediction, no difference in D_{KL} was seen between the clustered and non-clustered conditions relative to a uniform distribution, $t(124) = 0.08$, $p = .933$, $d = 0.02$, $BF_{01} = 5.24$. Indeed, the Bayes Factors indicated there was five times more evidence in favour of the null, suggesting the clustered and non-clustered condition diverged equally from the uniform distribution. Interestingly, the kernel density estimates at the centre of the experimentally imposed distribution ($\theta = 0$) show an increase for clustered responses, but a decrease for non-clustered responses (Fig. 3E). Thus, despite the two conditions having equally diverged from the uniform distribution, they may have diverged in a qualitatively distinct manner. We return to this unexpected finding following Experiment 2.

2.3. Discussion

Experiment 1 assessed how the presence of an underlying pattern (or schema) modulated memory and generalisation behaviour. Memory performance was higher for the clustered relative to the non-clustered condition (Hypothesis 1). Additionally, when presented with novel words, participants reproduced the pattern of locations presented for the clustered items, meaning they showed an ability to generalise their mnemonic information to novel, semantically related, words (Hypothesis 2). However, both conditions were equally divergent from the uniform distribution, which was not in line with expectations.

The finding that memory was benefited by the presence of a pattern is consistent with previous studies (Atienza et al., 2011; Brewer & Treyens, 1981; Greve et al., 2019). However, we note that Berens et al. (2020) did not find a difference in total information between the clustered and non-clustered conditions. They did see differences in terms of accessibility and precision, a finding we return to later (see Section 11 for analyses of accessibility and precision across Experiments 1–4).

Generalisation of the clustered items was found to be more similar to the experimentally imposed pattern than for the non-clustered items. These findings are consistent with recent evidence showing generalisation to novel instances can occur rapidly without the need for an extended period of consolidation (e.g., Sweegers & Talamini, 2014; Zeithamova et al., 2012).

Interestingly, the distribution of locations for clustered and non-clustered novel words diverged equally from a uniform distribution. Inspection of Fig. 3E suggests participants may have been less likely to place novel words in the non-clustered condition near the centre of the clustered distribution; a possible "avoidance" effect. This may suggest that the presence of a pattern (i.e., schema) in one condition influences schema-irrelevant information in the non-clustered condition. We return to this following Experiment 2.

An open question was whether generalisation behaviour would be modulated by delay. Theories of systems consolidation suggest that the extraction of schemas across a set of related experiences may take time to emerge (Kumaran & McClelland, 2012; McClelland et al., 1995), and sleep may play a critical role in this process (Inostroza & Born, 2013). Behavioural (Tompary et al., 2020) and neuroimaging (Kroes & Fernández, 2012; Wagner et al., 2015) work also suggests a time-dependent effect either in terms of the use or establishment of a schema. However, there is evidence suggesting behaviour consistent with use of a schema can remain constant (Sweegers & Talamini, 2014) or can decrease (Antony et al., 2021; Tompary et al., 2020) over time. Given that there are theoretical predictions to suggest a potential increase in generalisation over time, and that the evidence-base is mixed, a delay between Study and Test was implemented in Experiment 2. Therefore, Experiment 2 sought to replicate Experiment 1 with one change – the addition of a delay between Study and Test.

3. Experiment 2

Experiment 2 was identical to Experiment 1 with one exception – we increased the delay between Study and Test to approximately 24 h. The same preregistered hypotheses from Experiment 1 were tested. The preregistration for Experiment 2 is available here: https://osf.io/nbtm3/.

3.1. Methods

3.1.1. Participants

3.1.1.1. Power analysis. To determine the required sample size, the lowest effect size of interest for the pre-registered hypotheses was used; this was derived from the Berens et al. (2020) pilot investigation and concerned the effect of clustering on total information following a 24-h delay between Study and Test. As before, G*Power (3.1.9.2; Faul et al., 2007) was used to perform an *a priori* power analysis for a paired samples *t*-test comparing total information in the clustered and non-clustered conditions. The effect size for the analysis was estimated from the pilot investigation by Berens et al. (2020), which derived an effect size of $d = 0.31$. This effect size estimate, along with an α (one-tailed) $= 0.05$ and power $= 0.80$ were used. A suggested sample size of $N = 66$ was required.

3.1.1.2. Final sample. Eighty-six participants (77 female) were recruited for the study. The mean age was 20.63 years ($SD = 2.01$ years). Three participants did not return for the second session and 8 datasets did not converge during the mixture model and so were excluded. The final sample was 75 participants (67 female) with a mean age of 20.55 years ($SD = 1.99$ years). Similar to Experiment 1, when first analysed a lower number (66 participants) of usable datasets were present. However, when a coding issue was fixed, a sample of 75 usable datasets was obtained (hence the over-recruitment). Participants were fluent English-speakers with normal or corrected-to-normal vision and were recruited from the University of York student population and took part in exchange for course credit or cash payment.

3.1.2. Materials and procedure

The same materials and procedure were followed from Experiment 1; however, participants completed the Test phase approximately 24-h post Study, with the average delay between study and test being 23.93 h ($SD = 0.39$ h).

3.1.3. Data handling and statistical analysis

The same exclusion criteria and statistical analyses were used as in Experiment 1.

3.2. Results

3.2.1. Memory

When assessing memory performance, unlike Experiment 1, total information did not significantly differ between the two conditions, $t(148) = 0.67$, $p = .502$, $d = 0.11$, $BF_{01} = 4.63$ (Fig. 2B; Hypothesis 1). The Bayes Factor indicates four times more support in favour of the null model, suggesting no difference between conditions was present.

3.2.2. Generalisation

Fig. 3 suggests a similar pattern of results to Experiment 1. The clustered condition was significantly less divergent from the von Mises distribution than the non-clustered condition, $t(148) = 3.29$, $p = .001$, $d = 0.54$, $BF_{01} = 0.04$. In comparison to the uniform distribution, neither condition was significantly more divergent than the other, $t(148) = 0.44$, $p = .663$, $d = 0.07$, $BF_{01} = 5.22$. These results replicate Experiment 1.

3.2.3. Exploratory analyses: Comparing experiments 1 and 2

3.2.3.1. Change in generalisation.
Previous work suggests that schemas may take time to develop, with a period of sleep being an important contributor to this development (Inostroza & Born, 2013). As such, we wished to assess whether generalisation behaviour for the clustered novel items changed following a delay period. We predicted that generalisation may be greater (represented by lower D_{KL} values) following a delay period. Fig. 3A and B show the mean divergence for both conditions across experiments.

To assess a change over time, we generated a GLME using the same general parameters described above. However, instead of the effect of clustering, we assessed whether there was an effect of Delay (0 = Immediate Test, 1 = Delayed Test) on the divergence between the experimentally imposed von Mises distribution and responses to clustered novel words. No random slopes were present in the model. It was found that there was no significant evidence of a change over time, $t(136) = 1.08$, $p = .280$, $d = 0.19$, $BF_{01} = 3.20$. This suggests that following a period of sleep, participants adherence to the von Mises distribution for clustered novel items did not change.

3.2.3.2. Avoidance behaviour.
In Fig. 3E and F, there was possible evidence for a lack of uniformity in the distribution of locations for novel words in the non-clustered condition. Participants appear to show avoidance of the centre of the cluster for novel non-clustered words (though visual inspection suggests this effect is perhaps greater in Experiment 1 than 2). To assess this possible avoidance more formally, we compared non-clustered kernel density estimates at the centre of the cluster to the density expected if the responses were uniformly distributed ($2\pi^{-1}$). If participants were actively avoiding the centre of the cluster, their kernel density for non-clustered items at this location will be significantly lower than the uniform value. A GLME was computed using the same parameters as previously described, but without the clustering variable. Instead, a fixed effect of Delay (0 = Immediate Test, 1 = Delayed Test) was added, with random intercepts for each participant. No random slopes were specified in the model.

The kernel density plots for all experiments are shown in Fig. 4. For the analysis, we first compared immediate test and delayed test to the uniform value, separately. At immediate test, there was significantly reduced probability density compared to the uniform value, $t(136) = 4.06$, $p < .001$, $d = 0.35$, $BF_{01} = 0.003$; this suggests participants actively avoided locations at the centre of the cluster for novel non-clustered items. In contrast, there was no significant reduction in probability

density during Delayed Test, $t(136) = 1.38$, $p = .169$, $d = 0.10$, $BF_{01} = 4.65$. Along with this, a significant effect of delay was observed, $t(136) = 2.06$, $p = .041$, $d = 0.35$, $BF_{01} = 0.78$. Here, there was a decrease in the avoidance effect in Experiment 2 relative to Experiment 1 (i.e., distributions of novel words were more uniform following a delay). However, the Bayes Factor was anecdotal.

3.3. Discussion

Experiment 2 replicated the key generalisation finding from Experiment 1 – participants' distributions of locations were more similar to the underlying pattern in the clustered condition compared to the non-clustered condition (Hypothesis 2). However, we did not replicate the difference in overall memory performance (Hypothesis 1). The lack of difference in total information instead agrees with the results of a previous registered report using a similar paradigm (Berens et al., 2020). Additionally, we found that, following a delay period, participants adherence to the underlying pattern did not change for clustered items. This is contrary to some lines of evidence suggesting schematic extraction may take time to develop and therefore an increased capacity to generalise should be observed (e.g., Inostroza & Born, 2013; Kumaran & McClelland, 2012; McClelland et al., 1995). However, this finding is in line with other studies reporting generalisation based on an underlying pattern remains relatively stable over time (Sweegers & Talamini, 2014).

We also saw an "avoidance effect" in the non-clustered condition, where participants avoided placing novel words in the non-clustered condition at the centre of the cluster. Despite old non-clustered words being drawn from a uniform distribution, and being from a separate semantic category to the clustered words, participants were biased away from the clustered location. This avoidance effect was present in Experiment 1 and decreased in Experiment 2 where it was no longer present. Thus, this avoidance effect appears immediately but possibly decreases over a 24 delay (though see results of Experiments 3–4). Given this avoidance effect was not predicted, we performed two further experiments with pre-registered analyses to replicate this effect.

4. Experiment 3

Experiment 3 aimed to replicate Experiment 1 with the Test phase immediately following the Study phase. Experiment 4 aimed to replicate Experiment 2, with a 24-h delay between Study and Test. Both experiments were run online, rather than in person, due to coronavirus
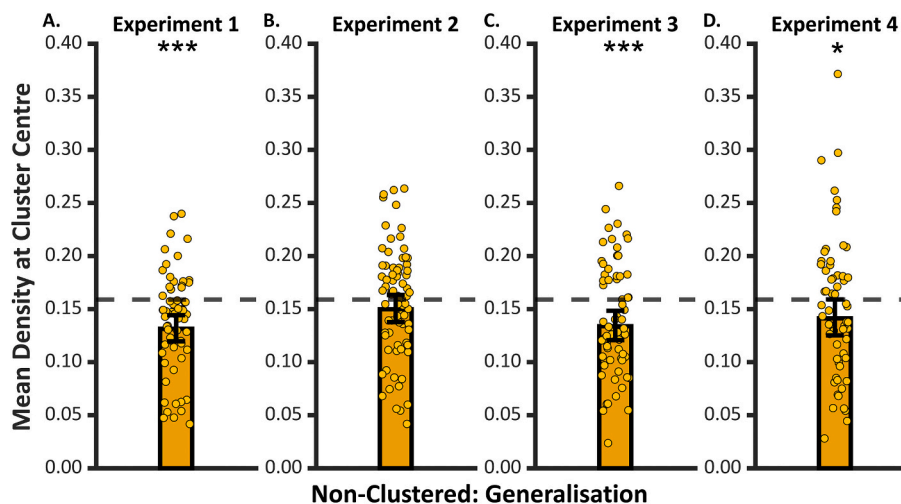

Fig. 4. Kernel density estimates for the non-clustered condition, centred to the experimentally imposed cluster, across experiments. Experiments 1, 3 and 4 all show evidence of an avoidance effect, with significantly lower kernel density at the centre of the cluster. Error bars are 95% confidence intervals around the mean estimate. Individual data points represent participant scores. The dashed line represents the uniform probability value (0.159). *$p < .050$, ***$p < .001$.

restrictions. The hypotheses from Experiments 1 and 2 were repeated in Experiments 3 and 4, with the exclusion of the comparison of the generalisation trial distributions with a uniform distribution (given this comparison was not informative in Experiments 1–2). Critically an additional preregistered analysis was included concerning the avoidance effect in the non-clustered generalisation condition; this was the same as the exploratory analysis of Experiments 1 and 2 (see Section 4.1.4.2, below). We predicted that participants would show a significant reduction in probability density for non-clustered novel words at the centre of the cluster (relative to a uniform distribution, as in Experiment 1). The preregistration for Experiment 3 is available here: https://osf. io/2wsn8/.

4.1. Methods

4.1.1. Participants

4.1.1.1. Power analysis. To determine the required sample, we assessed the range of effect sizes from Experiment 1 ($d = 0.35$–0.77) and set a minimum effect size of theoretical interest (i.e., Hypothesis 3, $d = 0.35$). Using G*Power (3.1.9, Faul et al., 2007) and estimating the sample size required for a one-sample *t*-test with this effect size, $\alpha = 0.05$ and power $= 0.80$. A sample size of 52 usable datasets was needed. However, given this estimate, and the power analysis previously conducted for Experiment 1, a final sample size of 60 usable datasets was set.

4.1.1.2. Final sample. Eighty-nine participants (35 female) with a mean age of 24.91 years (*SD* = 4.99 years) were recruited for the experiment. Three participants left before the study phase, 10 were excluded during Study and 3 were excluded at Test due to inattention. One did not complete the test phase despite completing the study phase. Two attempted the study phase twice and so were excluded. This left 70 participants that passed the initial checks. Of those, 4 did not provide a response on 20 or more memory trials and 3 datasets did not converge during the mixture model. Therefore, the final sample was 63 participants (25 female) with a mean age of 25.27 years (*SD* = 4.99 years). All participants were fluent English-speakers with normal or corrected-to-normal vision and were recruited through Prolific.co and received monetary compensation for their time.

4.1.2. Materials

The same word lists and Introspection Questionnaire were used from Experiments 1 and 2. However, rather than using four sub-lists for each category (i.e., human-made and natural) as in Experiments 1–2, 30 words from each category were randomly selected for each participant and assigned to the generalisation condition. The remaining 90 words were assigned to the memory condition. This was done due to practical constraints when coding the online experiment.

4.1.3. Procedure

The same general procedure was followed as in Experiment 1, but through the online platform Prolific. Participants recruited from Prolific were directed to a secure website hosting the online experiment. Participants were only able to use a laptop or desktop computer to run the task, with handheld devices (e.g., smartphone, tablet) being excluded. Before the start of the Study Phase, participants watched a short introductory video about how the session progressed and how to respond. A PDF document of written instructions was also provided (https://osf. io/qxfuj/). The instructions emphasised the need to visualise the object related to the cue word appearing at the cued location before responding on each study trial and how participants were to be asked to recall these locations at test. The video instructions replaced the practise trials used in-lab, as using instructions in this format online produced similar results for memory trials (Berens et al., 2020).

4.1.3.1. Study phase. The Study Phase was identical to Experiments 1 and 2. Once completed, participants moved immediately onto the Test Phase.

4.1.3.2. Test phase. One minor change was made to the Test phase. In Experiments 1 and 2, participants were presented with a fixation cross (1 s) followed by the word alone (2 s) and then the opportunity to reposition the marker to the remembered or generalised location (10s). In Experiments 3–4, the word was not shown alone for 2 s. In-lab, participants provided a response on average within 2.38 s of being able to replace the marker, with almost all responses collected within 7.23 s. As such, the additional 2 s of the word alone was removed given the 10s time window for responding. Following Test, participants were asked to complete the Introspection Questionnaire.

4.1.3.3. Introspection questionnaire. The same questions as Experiments 1 and 2 were used online. We also included an additional question about whether the participant had help completing the task; this was to be used as an exclusion criterion (though not pre-registered) had participants reported they did have help completing the task. No such report was given.

4.1.4. Data handling and statistical analysis

4.1.4.1. Exclusion criteria. All exclusions from Experiments 1 and 2 were used in Experiments 3 and 4. However, participants could also be excluded during Study or Test for not following task instructions; this was quantified as having reaction times of <2 s across a total of 70 trials. Specifically, participants would receive a warning message through their browser should the number of trials with reaction times <2 s hit 10, 30, 45 and 60 trials. This message asked participants to either: slow down and ensure they imagined the object appearing at each location (Study Phase) or encouraged them to try to remember the location for each word (Test Phase). This was an exclusion that was not pre-registered, but used previously (see Berens et al., 2020) as a way of maximising participant performance when conducting the experiment online.

4.1.4.2. Statistical analysis. The same statistical analyses were used for Hypotheses 1 (Total Information) and 2 (D_{KL} von Mises). For Hypothesis 3, we compared the probability density estimates at the centre of participants experimentally imposed cluster for non-clustered novel words to the density for a uniform distribution. If participants were actively avoiding the centre of the cluster, then they will not be distributing locations randomly (or uniformly) and so their kernel density at this location should be significantly below that of a uniform value. To test this, a GLME was fit using a log link function and a gamma distribution to model the spread of the data, estimated using the maximum likelihood fitting method within the MATLAB Statistics and Machine Learning Toolbox. This was an intercept only model with random intercepts for each participant. The derived model was then used to conduct a one-sample *t*-test comparing the beta of the intercept model to the log of the uniform kernel density value. A one-tailed test was used for this analysis as a directional effect was predicted. Note, this analysis was almost identical to the exploratory analysis performed across Experiments 1 and 2, but without the inclusion of any fixed effects. Cohen's *d* and Bayes Factors are reported and use the same parameters as described previously.

4.2. Results

4.2.1. Memory

Total Information was not significantly different between the clustered and non-clustered conditions, $t(124) = 0.73$, $p = .466$, $d = 0.13$, $BF_{01} = 4.13$ (Fig. 2C). There was four times more evidence in favour of

the null model, suggesting that an underlying pattern does not benefit overall memory performance. This result is consistent with Experiment 2, and Berens et al. (2020), but contrary to Experiment 1.

4.2.2. Generalisation

As in Experiment 1, participants showed an ability to generalise, with the clustered condition being significantly less divergent from the von Mises distribution than the non-clustered, $t(124) = 5.00$, $p < .001$, $d = 0.89$, $BF_{01} = 4.66 \times 10^{-5}$ (Fig. 3C). These results replicate Experiments 1 and 2, showing participants can generalise from old words to novel words in the same semantic category.

4.2.3. Avoidance

The next analysis tested whether the avoidance effect observed in Experiment 1 would replicate. As shown in Fig. 3G, participants do show some evidence of avoidance behaviour in their location selection. This was confirmed by a significant reduction in the probability density for non-clustered items at the centre of the cluster, $t(62) = 3.35$, $p = .001$ (one-tailed), $d = 0.42$, $BF_{01} = 0.03$ (Fig. 4C). This replicates the avoidance effect found in Experiment 1. Specifically, participants actively avoid placing the locations of novel non-clustered words at the centre of the cluster.

4.3. Discussion

The aim of Experiment 3 was to replicate the findings of Experiment 1, particularly the evidence of an avoidance effect. First, there was no significant benefit to overall mnemonic information available in the clustered compared to the non-clustered condition (as in Experiment 2, but not 1). Second, we replicated the generalisation behaviour seen in Experiment 1. The distribution of locations for novel clustered words was more similar to the underlying von Mises distribution than for novel non-clustered words (Hypothesis 2). Finally, we replicated the exploratory analysis of Experiment 1, showing that participants were less likely to position novel non-clustered words in the centre of the cluster (Hypothesis 3). Experiment 4 aimed to replicate the lack of avoidance following a delay period, as in Experiment 2.

5. Experiment 4

Experiment 4 was identical to Experiment 3, apart from the inclusion of a 24-h delay between Study and Test (as in Experiment 2). We had the same three key hypotheses as in Experiment 3. The preregistration for Experiment 4 is available here: https://osf.io/fjze8/.

5.1. Methods

5.1.1. Participants

5.1.1.1. Power analysis.
As before, the required sample size was determined based on the smallest effect size of interest. The minimum effect size of interest (taken across all previous experiments) was $d = 0.43$ for the total information effect. Using G*Power (3.1.9, Faul et al., 2007) estimating the sample size for a paired-samples t-test that for this effect size, $\alpha = 0.05$ and power = 0.80, a sample size of 45 usable datasets was required. However, to ensure similar power to previous experiments, a final sample size of 60 usable datasets was set.

5.1.1.2. Final sample.
A total of 79 participants (32 female) with a mean age of 24.15 years ($SD = 4.54$ years) were recruited for the study. Of those, 3 participants failed attentional checks during Study, 3 failed to return for the Test phase, 5 did not provide a sufficient number of responses, and 8 datasets failed to converge during mixture modelling. The final sample was 60 participants (25 female) with a mean age of 24.07 years ($SD = 4.43$ years). All participants were fluent English-

speakers with normal or corrected-to-normal vision, were recruited through Prolific.co, and received monetary compensation for their time.

5.1.2. Materials and procedure

The experiment was identical to Experiment 3, except for two features. First, a delay between Study and Test was introduced, similar to Experiment 2. Participants completed the Study Phase and then 24-h later completed the Test Phase. The average delay was 23.74 h ($SD = 0.25$ h). Additionally, participants watched two separate instruction videos, one at the beginning of the Study Phase and another at the beginning of the Test Phase. Written instructions were also provided (https://osf.io/bxru4/).

5.1.3. Data handling and statistical analysis

Data handling, exclusion, and statistical analyses were identical to Experiment 3.

5.2. Results

5.2.1. Memory

There was no difference between the clustered and non-clustered conditions in terms of total information, $t(118) = 1.04$, $p = .299$, $d = 0.19$, $BF_{01} = 3.15$ (Fig. 2D). As in Experiments 2 and 3, support for the null hypothesis was found.

5.2.2. Generalisation

Fig. 3H shows the pattern of locations selected by participants for novel items in this experiment. It was found that clustered items were significantly less divergent from the von Mises distribution than the non-clustered items, $t(118) = 3.85$, $p < .001$, $d = 0.70$, $BF_{01} = 0.01$. These results replicate all previous experiments.

5.2.3. Avoidance

Participant's non-clustered kernel density estimates at the centre of the cluster were compared to a uniform distribution. We found significant evidence of an avoidance effect, $t(59) = 2.00$, $p = .025$ (one-tailed), $d = 0.26$, $BF_{01} = 1.07$. This replicates the findings of Experiments 1 and 3, but not Experiment 2 (where no avoidance effect was present following a 24-h delay). We return to the possible effect of delay on this avoidance effect in the across-experiment exploratory analyses below.

5.3. Discussion

Experiment 4 replicated previous experiments. We found no evidence for a difference in total information between the clustered and non-clustered conditions (as seen in Experiments 2 and 3, but not 1). We showed that the distribution of novel clustered words was more similar to the underlying distribution than for novel non-clustered words (as in Experiments 1–3). Further, we found evidence that participants were less likely to place novel words in the non-clustered condition near the centre of the cluster (as in Experiments 1 and 3). This was contrary to predictions given the finding of Experiment 2, which found the avoidance effect was no longer apparent following a delay period. To assess this further, we performed an exploratory analysis of the change in avoidance behaviour as a function of time.

6. Exploratory analysis: Across-experiment comparisons

Across four experiments we provide evidence for (1) no difference in overall memory performance (total information) for old words in the clustered relative to the non-clustered condition, (2) less divergence between the experimentally imposed pattern (von Mises distribution) and the novel word responses in the clustered condition relative to the non-clustered condition, and (3) avoidance of the centre of the clustered pattern for non-clustered novel words. We next carried out a set of across-experiment analyses to compare these effects across (1) delay and

(2) setting, to ensure the effects are robust to these changes. Further, we present new analyses assessing: (1) the components making up total information (accessibility and precision), (2) an avoidance effect for non-clustered old words (i.e., memory trials) and (3) evidence for greater density of locations for clustered novel words (i.e., generalisation trials) at the cluster centre.

Each analysis used a similar GLME structure, assessing whether the metric of interest was affected by clustering (0 = Non-Clustered, 1 = Clustered), delay (0 = Immediate Test or 1 = Delay Test) or setting (0 = In-lab, 1 = Online) along with their interactions. All models, unless otherwise specified, had two random effects per participant. The first was random intercepts per subject, and the other was random slopes for the effect of clustering (if the effect of clustering was included). When no effect of clustering was included, no random slopes were present in the model.

For old words (memory trials), we assessed (1) total information, accessibility and precision for the clustered relative to non-clustered condition. Given the evidence for avoidance for non-clustered novel words, we also assessed (2) evidence for avoidance behaviour for old non-clustered words. For new words (generalisation trials), we assessed (1) D_{KL} (relative to the experimentally imposed von Mises distribution) for clustered relative to non-clustered new words, and (2) probability density estimates at the centre of the von Mises distribution for non-clustered new words. We also assessed the (3) probability density estimates for clustered new words to further assess generalisation behaviour across experiments.

6.1. Memory

6.1.1. Total information

For total information, there was a main effect of delay, $F(1,514) = 48.63$, $p < .001$, $d = 0.39$, $BF_{01} = 5.66 \times 10^{-10}$, with total information decreasing across time. All other main effects and interactions were non-significant ($p \geq .114$, $d \leq 0.12$, $BF_{01} \geq 3.50$). To explore the lack of total information effect further, we analysed the metrics that make up total information: accessibility and precision.

6.1.2. Accessibility

Fig. 5 shows the effect of clustering on accessibility, with greater accessibility in the clustered relative to the non-clustered condition. This was confirmed statistically with a main effect of clustering, $F(1,514) = 14.24$, $p < .001$, $d = 0.16$, $BF_{01} = 0.02$. Participants showed greater accessibility in the clustered relative to the non-clustered condition. There was also a main effect of delay, $F(1,514) = 32.50$, $p < .001$, $d = 0.33$, $BF_{01} = 1.63 \times 10^{-6}$. Here, accessibility decreased from immediate ($M = 0.61$, $SE = 0.03$) to delayed ($M = 0.46$, $SE = 0.03$) test. No other significant effects were observed ($p \geq .217$, $d \leq 0.08$, $BF_{01} \geq 6.57$).

6.1.3. Precision

For precision, there was a significant effect of clustering, $F(1,514) = 6.76$, $p = .010$, $d = 0.11$, $BF_{01} = 0.73$. As shown in Fig. 5, precision was greater in the non-clustered compared to the clustered condition. A main effect of delay was also present, $F(1,514) = 34.20$, $p < .001$, $d = 0.27$, $BF_{01} = 8.28 \times 10^{-7}$, with precision decreasing from immediate ($M = 1.40$, $SE = 0.02$) to delayed ($M = 1.14$, $SE = 0.02$) test. A significant
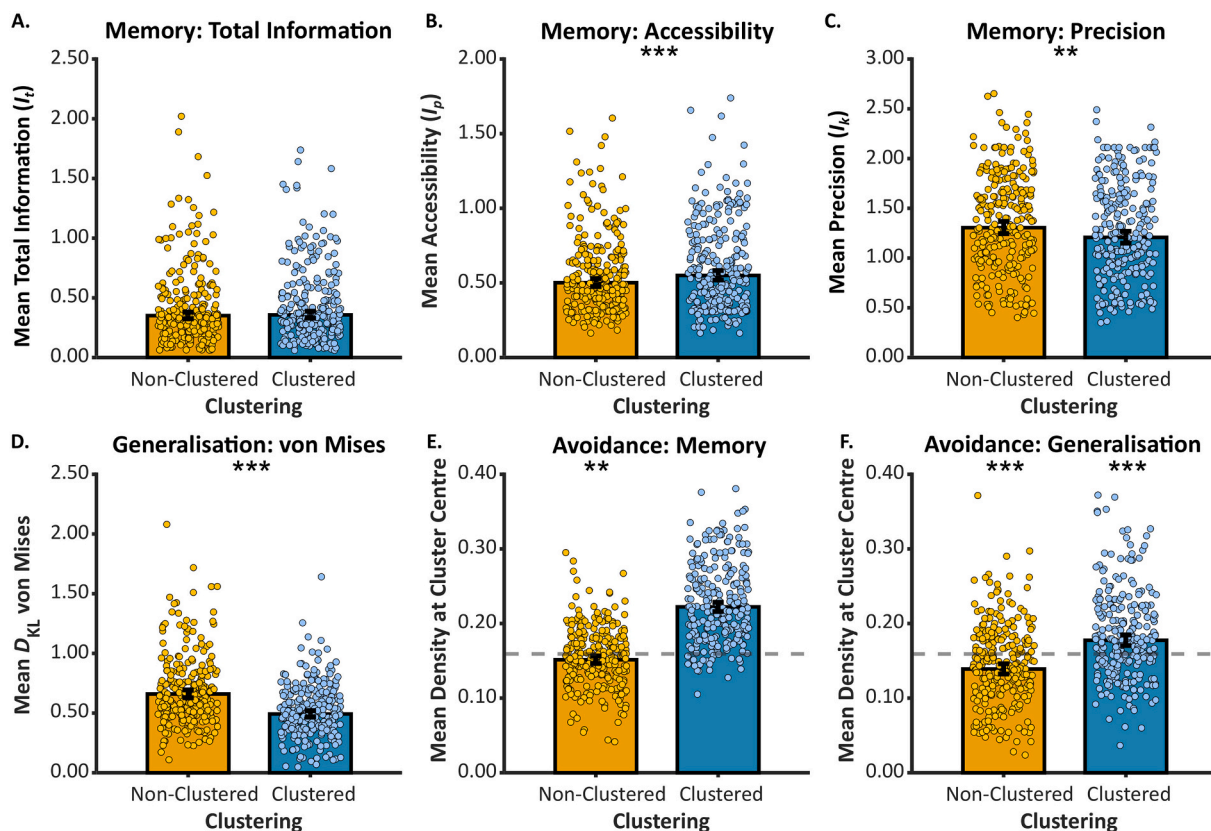

Fig. 5. Across experiment analyses. A final summary of the overall results reported within the paper. All plots show the effect of clustering collapsed across delay and setting. (A-C) Memory Metrics: The effect of clustering on each memory metric: (A) total information, (B) accessibility, and (C) precision. (D) D_{KL} von Mises Generalisation: The effect of clustering on generalisation to the von Mises distribution. (E-F) Probability Density at Cluster Centre: Plotting both (E) memory and (F) generalisation densities at the cluster's centre to assess differences from a uniform density. Though not tested, for completeness the clustered memory data is also plotted. Individual data points represent participant scores. The dashed line represents the uniform probability value (0.159). Error bars represent 95% confidence intervals around the mean for all plots. **$p < .01$, ***$p < .001$.

interaction between clustering and setting was also observed, $F(1,514) = 7.35$, $p = .007$, $d = 0.20$, $BF_{01} = 0.33$. Post-hoc comparisons revealed that the clustering effect was more apparent when the study was conducted online compared to in-lab. To control for familywise error, the Bonferroni-Holm correction was applied and reported as p_{BH}. There was significantly greater precision in the non-clustered condition when tested online ($M = 1.40$, $SE = 0.03$) compared to the clustered conditions both in-lab ($M = 1.24$, $SE = 0.03$; $p_{BH} = .031$, $d = 0.13$, $BF_{01} = 0.52$) and online ($M = 1.18$, $SE = 0.04$; $p_{BH} = .002$, $d = 0.17$, $BF_{01} = 0.03$). Additionally, the non-clustered condition showed significantly greater precision when performed online than in-lab ($M = 1.23$, $SE = 0.03$; $p_{BH} = .024$, $d = 0.18$, $BF_{01} = 0.28$). All other post-hoc tests were not significant ($p \geq .362$, $d \leq 0.04$, $BF_{01} \geq 12.32$). No other main effects and interactions were significant ($p \geq .215$, $d \leq 0.06$, $BF_{01} \geq 9.33$).

6.1.4. Avoidance

One question we asked was whether a similar avoidance effect (as seen for non-clustered new words) was also present for non-clustered old words. In short, is memory for schema-irrelevant information also biased? We compared the probability density of old non-clustered words, relative to a uniform distribution. Two main effects were modelled: delay and setting, with no random slopes. Fig. 5 shows the probability density of locations selected for both clustered and non-clustered memory trials when centred to the cluster. We found that the probability density for non-clustered old words was lower than predicted by a uniform distribution, $F(1,257) = 8.18$, $p = .005$, $d = 0.09$, $BF_{01} = 0.48$, suggesting similar avoidance behaviour for non-clustered old words was present in our data. No effect of delay ($F(1,257) = 0.01$, $p = .911$, $d = 0.01$, $BF_{01} = 9.74$) or setting ($F(1,257) = 0.37$, $p = .542$, $d = 0.05$, $BF_{01} = 8.38$) was observed. However, there was an interaction between the two, $F(1,257) = 4.03$, $p = .046$, $d = 0.30$, $BF_{01} = 0.96$. Exploration of the post-hoc effects found no significant effects even before correction ($p \geq .069$, $d \leq 0.23$, $BF_{01} \geq 1.52$). Notably, the Bayes Factor for the interaction was anecdotal. Based on these exploratory results, we performed a pre-registered secondary analysis on the data from Berens et al. (2020), where a similar avoidance effect in the non-clustered condition for memory trials was seen (see Supplementary Material). We therefore provide evidence that schematic information affects memory behaviour for schema-irrelevant information.

6.2. Generalisation

6.2.1. D_{KL} von Mises

Next, we wished to assess whether the effects of clustering, delay or setting influenced D_{KL} values for the distribution of locations for new words relative to the experimentally imposed von Mises distribution. For this model, we found that only clustering was significant, $F(1,514) = 65.20$, $p < .001$, $d = 0.34$, $BF_{01} = 1.80 \times 10^{-13}$. Specifically, the distribution of clustered new words was less divergent from the von Mises than the distribution for non-clustered new words. All other main effects and interactions were non-significant ($p \geq .051$, $d \leq 0.09$, $BF_{01} \geq 2.88$). Generalisation behaviour was therefore consistent over delay and setting.

6.2.2. Kernel density

Examination of the probability density estimates at the centre of the cluster was then undertaken for both the clustered and non-clustered conditions, separately. In both instances, no random slopes were included in the model. For the non-clustered condition, there was significant evidence of avoidance with reduced kernel density at the cluster centre, $F(1,257) = 28.94$, $p < .001$, $d = 0.17$, $BF_{01} = 1.53 \times 10^{-5}$. No main effects of delay, setting, or an interaction were observed ($p \geq .066$, $d \leq 0.17$, $BF_{01} \geq 1.87$). The avoidance effect of non-clustered new words was therefore consistent across delay and setting.

Finally, we assessed the probability density at the centre of the cluster for clustered novel responses. There was significantly greater

density relative to a uniform distribution, $F(1,257) = 28.29$, $p < .001$, $d = 0.17$, $BF_{01} = 2.12 \times 10^{-5}$. No main effect of delay, setting, or an interaction were observed ($p \geq .209$, $d \leq 0.12$, $BF_{01} \geq 4.58$). Therefore, participants were more likely to place new clustered words near the centre of the cluster, and this was not modulated by delay or setting.

7. General discussion

We assessed whether participants use patterns (schematic information) to guide memory and generalisation behaviour using a precision long-term memory paradigm. Across four experiments and a secondary analysis of published data (Berens et al., 2020), we found that schematic information modulated both memory and generalisation behaviour. Critically, we found schematic information in one condition (the clustered condition) modulated memory and generalisation behaviour for an unrelated condition (the non-clustered condition). Participants were less likely to place both old (memory trials) and new (generalisation trials) words in the non-clustered condition near the centre of the clustered pattern.

This avoidance behaviour was seen for both studied words (memory trials) and semantically-related unstudied words (generalisation trials) in exploratory analyses in Experiment 1, preregistered analyses in Experiment 3–4, and a secondary analysis of the data from Berens et al. (2020). Further, we found no evidence for an effect of delay (though we note the Bayesian analyses were inconclusive). Therefore, we find consistent evidence that schematic information influences memory and generalisation behaviour for schema-irrelevant information. We first focus on this key finding before discussing the memory and generalisation results and their implications for schema theory.

7.1. Schema-irrelevant information and avoidance

We saw clear evidence that the presence of a pattern influenced memory and generalisation behaviour in the non-clustered condition. Participants avoided placing non-clustered items at the location of the cluster for both memory and generalisation trials. Therefore, the presence of schematic information biases memory and generalisation behaviour for schema-irrelevant information.

Previous work has shown that schemas can negatively bias recall of information (Lew & Howe, 2017; Roediger & McDermott, 1995; Warren, Jones, Duff, & Tranel, 2014). For example, Bartlett (1932) demonstrated that retrieval for events in a narrative were biased by a participant's existing knowledge of the world. Warren et al. (2014) showed that, while healthy controls display relatively high levels of false recall in the presence of a schema, patients with vmPFC damage show relatively fewer errors. The Deese-Roediger-McDermott (DRM) false memory effect can also be interpreted as a memory bias (increase in false alarms) in the presence of a schema (Cann, McRae, & Katz, 2011). These studies have predominantly focussed on binary measures of memory, demonstrating increased false alarms or errors in the presence of a schema. Here our focus was on information irrelevant to the schematic information being learnt, rather than false memory or biases for schema-related information.

Studies have used schema-irrelevant information as a control condition to compare to schema-congruent and incongruent conditions (e.g., Frank et al., 2018; Greve et al., 2019). These studies show that memory performance is enhanced for schema-congruent and -incongruent information relative to schema-irrelevant information. The present experiments suggest that the presence of schematic information can bias memory for these irrelevant items. Changes in performance for schema-irrelevant information may have been previously missed due to a lack of appropriate control comparison. For example, in Greve et al. (2019), retrieval of schema-irrelevant items may have been reduced by the presence of schematic information, resulting in what appears to be a schema benefit. Instead, the results may be caused by the presence of a schema biasing (i.e., hindering) the retrieval of less relevant

information. As we compared behaviour in our non-clustered condition to that expected of a uniform distribution (representing the distribution of locations expected if no biases were present), we demonstrated biases for schema-irrelevant information that may have been missed in previous studies. Therefore, future studies should be aware that a schema-irrelevant control condition may not be an appropriate baseline given our results.

Concerning previous precision long-term memory studies, results such as those from Tompary et al. (2020) may have masked this avoidance effect. Tompary et al. (2020) used two clustered conditions on opposite sides of a circle (180° apart), meaning the effects on schema-relevant information will have overshadowed any effects of schema-irrelevant information. It was only with the inclusion of a non-clustered condition, where word-locations were drawn from a uniform distribution, that we revealed an effect of the clustered pattern on the semantically distinct non-clustered words.

What produces the avoidance behaviour in the non-clustered condition? One possibility is that the avoidance effect is driven by a "mutual exclusivity" bias (Clark, 1988; Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992). This bias is often studied in language learning and refers to the tendency to assign only one label to an object. For example, if children are presented with two objects, one familiar and one novel, and asked to identify what object is being referred to when a novel word is presented, they typically select the novel object (Markman & Wachtel, 1988). This suggests they are less willing to assign more than one label to a given object, even though several labels may encompass the same object (e.g., a cat is both a mammal and an animal). Though much of the work on mutual exclusivity has focused on children, recent work examining adult word learning has suggested that the bias helps with generalising to novel words (Lake, Linzen, & Baroni, 2019).

Though this bias is often thought to help guide language development, a similar bias could drive our avoidance effect. As participants identified that semantically related words (e.g., natural words) were associated with a general location (e.g., top-right quadrant), they might have been more inclined to group words from the other category (e.g., human-made) on the opposite side of the circle. In short, they attributed the top-right quadrant of the circle as "natural only", despite human-made words also appearing in this area. The simplest version of a mutual exclusivity bias might be an explicit process at retrieval where, when a non-clustered word is presented, participants actively retrieve a schema related to the clustered condition and use an *"if not in the clustered category, place on the opposite side of the circle to the schema"* strategy. Although possible, analysis of the post-retrieval debrief suggests very few participants were using explicit strategies such as these. However, a mutual exclusivity bias could explain the avoidance effect either explicitly or implicitly, under the assumption that participants are either explicitly or implicitly categorising the words at the level of "human-made" and "natural". Although we cannot rule out implicit categorisation at this superordinate level, the debrief suggested that few participants (4 of 261) spontaneously referred to these semantic groups. Instead, participants were more likely to (explicitly) categorise words into subordinate categories (e.g., "household objects, animals and fruit", "fruit and vegetables, household items, mammals", "fruit, technology… cars… exotic animals, weather", "planets, animals, food").

If a mutual exclusivity bias were driving the avoidance effect, one critical question is what is driving this bias? One explanation would be that the non-clustered condition is unlike most groupings found in the real world. Returning to the earlier example, if you have a "bathroom" schema, it is probably less likely that you will find non-bathroom related items in this location relative to elsewhere in the house. In short, the "bathroom" schema does not tell you where the microwave will be, but it likely provides information about where it is unlikely to be. Although real-world examples of more uniformly distributed items may exist, for example, house plants throughout a home, they may be rare, meaning we have little experience with them. Participants may apply this real-world sampling to the present experiments, presuming non-clustered words are less likely to be located in the clustered area of the circle.

Another possible explanation for the avoidance effect in the non-clustered condition is that participants' behaviour was guided by relative probabilities representing the likelihood of having studied a particular type of word at each location. This contrasts with making location responses based on absolute probabilities representing the overall 'density' of different types of words at each location. Non-clustered word locations were drawn from a uniform distribution, such that the absolute probability of encountering a non-clustered word was close to uniform around the circle. However, the relative probability of encountering a non-clustered relative to clustered word differed around the circle – the relative probability was lower in the clustered area of the circle relative to the other side of the circle. If participants' location responses were influenced by assessing the *relative* probability of having studied a word-location association from a given semantic category, we would expect to observe an avoidance effect in the non-clustered condition.

This base-rate neglect proposal (ignoring the absolute density of words in a given location) is a well-documented bias in the literature (Hawkins, Hayes, Donkin, Pasqualino, & Newell, 2015; Welsh & Navarro, 2012; Wolfe, 2007). It could potentially drive the avoidance effect if either implicit or explicit categorisation of the words were occurring (as would be necessary for the mutual exclusivity bias) or if participants were not categorising but were sensitive to the semantic distances between individual words. As such, it could explain the avoidance effect without categorisation at the superordinate level (i.e., human-made and natural). Given that research has suggested that base-rate neglect is driven by explicit processes (Lovett & Schunn, 1999; c.f. Bohil & Wismer, 2015; Wismer & Bohil, 2017), it is likely that the effect seen here (if base-rate neglect is the correct explanation) would be sensitive to whether participants are learning word-location associations under conditions that preclude explicit awareness.

A further alternative is that the avoidance effect could be driven by proactive or retroactive interference (at encoding or retrieval) between word-location associations (Anderson & Neely, 1996; Baddeley & Hitch, 1977; Barnes & Underwood, 1959; Jenkins & Dallenbach, 1924; Kliegl, Pastötter, & Bäuml, 2015; Sadeh, Ozubko, Winocur, & Moscovitch, 2016; Underwood, 1957; Wixted, 2004). Specifically, dense clustering of word-location associations in one part of the circle may interfere with those specific associations (location-based interference). This interference would apply irrespective of semantic category, resulting in worse memory performance for word-location associations in the clustered area relative to locations on the other side of the circle. This explains our avoidance effect for non-clustered memory trials – participants were poorer at remembering locations near the cluster, decreasing the probability of placing old words in this area of the circle.

For the clustered condition, interference would also apply. However, there are more words located in that area of the circle belonging to the clustered condition, and all those items are tested at retrieval. Therefore, there is a sampling bias at retrieval in the clustered condition. This sampling bias may outweigh any potential interference effect, such that no avoidance behaviour is found. One way to assess this in the future would be to only test a subset of clustered items at retrieval such that the true locations of those items were distributed uniformly.

Performance for generalisation trials should follow that seen in the memory trials. Irrespective of whether generalisation in the paradigm is driven by an encoding-based or retrieval-based model of schema processing (see below), any schematic representation is likely to be built on the word-location associations that are strongly encoded and/or more easily retrieved. This would naturally produce an avoidance effect for the non-clustered new words, given that the memory bias is already present for non-clustered old words.

Although our word lists came from two superordinate semantic categories and were therefore categorically distinct, there is some semantic overlap between lists (as seen in the word2vec semantic distance measure we used to select the words). While the proposed mechanisms

above do not rely on semantic overlap between the two categories, such overlap may modulate the extent of avoidance in the non-clustered condition. Indeed, the proposed mechanisms are primarily driven by a lack of similarity between the two categories. In the case of the mutual exclusivity bias and base rate neglect account, these rely on semantic distance, rather than overlap, between the two categories. Evidence suggests that the mutual exclusivity bias works under a wide range of circumstances, even when the items themselves show little semantic overlap (Markman, Wasow, & Hansen, 2003). In the case of location-based interference, we stated this would occur regardless of semantic category, so overlap is again not a necessary precondition for this explanation. Further experiments directly manipulating overlap would clarify this possible relationship. We would predict that separating the words further may lead to exacerbating the avoidance effect as classification of items would become easier for participants.

Finally, two exploratory analyses of the data were conducted to potentially delineate between the above proposals (see Supplementary Materials). The first analysis attempted to better understand the shape of the distribution for the avoidance effect. Specifically, whether avoidance appeared as a separate cluster on the opposite side of the circle (as might be predicted by the mutual exclusivity proposal) or as a relatively uniform distribution with a reduction in density just around the clustered area (as might be predicted by interference or base rate neglect). This analysis found that the latter pattern provided a better fit for participant data, which provides tentative support either the interference or base rate neglect proposal, but not mutual exclusivity. Additionally, an analysis of possible proactive and retroactive interference was undertaken. Specifically, an examination of how avoidance changed over the course of testing was undertaken by splitting data into three sections: beginning, middle and end. Trials were then ordered based on the order items were presented at study or at test. It was found that avoidance was relatively consistent regardless of when the items appeared at Study or Test, though only strong evidence for avoidance was seen for items presented at the middle or end of the study and tests phases. These two analyses perhaps provide more support for the base rate neglect and interference proposals than the mutual exclusivity proposal. Nonetheless, formal modelling is required to better understand the shape of the distribution and possible mechanisms that may drive this behaviour.

In sum, the avoidance effect may be driven by (1) a mutual exclusivity bias, (2) a base-rate neglect bias, and (3) location-based interference. We return to these three accounts following a discussion of the broader generalisation and memory results.

7.2. Generalisation

Across experiments, participants could use the underlying pattern to make informed decisions about where to locate novel semantically-related words. Distributions across new words in the clustered condition were more similar to the underlying pattern (von Mises distribution) than new words in the non-clustered condition (as measured by D_{KL}). We also saw evidence for greater peak kernel density at the centre of the pattern for clustered new words relative to a uniform distribution. This finding suggests that participants were more likely to place new words in the clustered condition towards the centre of the pattern than if the words had been placed randomly.

The evidence for generalisation presented here, both immediately and following a delay, is in line with previous research (Berens & Bird, 2021; Djonlagic et al., 2009; Durrant, Taylor, Cairney, & Lewis, 2011; Ellenbogen, Hu, Payne, Titone, & Walker, 2007; Graves, Antony, & Turk-Browne, 2020; Mirković, Vinals, & Gaskell, 2019; Sweegers & Talamini, 2014; Tompary et al., 2020). Though, there is conflicting evidence concerning whether generalisation performance increases, decreases, or remains constant over longer delays (Sweegers & Talamini, 2014; Tompary et al., 2020). We saw no clear evidence for a change in generalisation behaviour over a 24-h delay, suggesting relative stability

over 1 day (which included one night of sleep). Longer delays using a similar experimental approach would be needed to draw definitive conclusions about generalisation behaviour over extended timescales.

The finding of immediate generalisation performance, if such behaviour is based on a schematic representation, is at odds with standard models of systems consolidation (e.g., McClelland et al., 1995). Here, new schematic representations are thought to be formed as a function of hippocampal to neocortical transfer over (at a minimum) several hours, and sleep is thought to play a crucial role in this systems consolidation process (see Rasch & Born, 2013). Although novel information can be rapidly integrated into an existing schema (Fernández & Morris, 2018; Kumaran, Hassabis, & McClelland, 2016; van Buuren et al., 2014), this rapid transfer is not thought to occur when establishing new schemas as is the case here, where no location-based schema for a semantic grouping of words should exist before the experiment.

Updated models that incorporate a retrieval-based generalisation mechanism, such as the REMERGE model (Kumaran & McClelland, 2012), more readily accommodate our findings of immediate generalisation. During immediate generalisation, where systems consolidation would not have had chance to take place, participants will rely more on retrieval-based mechanisms. Over time, as systems consolidation occurs, there will be a move to more encoding-based mechanisms supported by a generalised neocortical-based schema.

It is plausible that there is a shift from retrieval-based to encoding-based generalisation over time in our experiments, but that both mechanisms support similar generalisation behaviour. However, recent research suggests generalisation behaviour might decrease over time when using the precision long-term memory paradigm, which would be inconsistent with the extraction of a stable schematic representation. In their study, Tompary et al. (2020) showed that schematic representations may decline over time alongside memory for individual word-location associations. Antony et al. (2021) found a similar pattern of results using a spatial navigation object-location task. As participants' memory performance declined for individual object-location associations over time, so did their adherence to the pattern of locations. Finally, although they did not assess generalisation to new words, Berens et al. (2020) showed that the distribution of remembered word locations decreased in similarity to the underlying pattern over 4 days and this decrease correlated with memory accessibility (i.e., the proportion of word-location associations retrieved). Given generalisation behaviour is seen immediately following encoding, and that it appears to decline over time as memory for individual item-locations declines, generalisation in these paradigms may be more driven by retrieval-based mechanisms rather than be supported by stable schematic representations. Whether more stable long-term schematic representations emerge over longer timescales, with multiple encoding sessions, in these paradigms (as has been shown in rodents; Richards et al., 2014) remains an open question.

7.3. Memory

The presence of schematic information in the clustered condition modulated memory-guided behaviour in both the clustered and non-clustered conditions. First, in an exploratory analysis, we replicated the results of Berens et al. (2020), showing the presence of a pattern increased accessibility (proportion remembered) but decreased precision (the angle of error for word-location associations that were remembered). Our pre-registered analyses comparing Total Information (the product of accessibility and precision, divided by a constant) in the clustered relative to non-clustered condition showed no overall boost in memory performance between conditions (though a small but significant difference was seen in Experiment 1). This lack of an increase in overall memory performance again replicates the results of Berens et al. (2020).

Previous studies have shown an overall benefit to memory for schematic vs non-schematic information (Atienza et al., 2011; Brewer &

Treyens, 1981; Frank et al., 2018; Greve et al., 2019). The present findings might appear to contradict these studies. However, most previous analyses have used binary measures of memory (correct vs incorrect) that are conceptually similar to the accessibility measure used in the present studies. Thus, our increase in accessibility in the clustered relative to non-clustered condition is consistent with previous findings.

Importantly, our ability to assess accessibility and precision suggests this increase in accessibility comes at a cost – a corresponding decrease in precision. This lack of precision is similar to previous findings suggesting that the presence of a schema leads to the loss of more fine-grained detail information but enhanced memory for face-location associations that had a schematic element (Sweegers, Coleman, van Poppel, Cox, & Talamini, 2015). Other studies have reported similar memory biases as a consequence of schematic information (Berens et al., 2020; Mäntylä & Bäckman, 1992; Pezdek, Whetstone, Reynolds, Askari, & Dougherty, 1989; Richter et al., 2019; Tompary et al., 2020; Tompary & Thompson-Schill, 2021; Zeng, Tompary, Schapiro, & Thompson-Schill, 2021), as well as increases in false positives to novel items that are related to an underlying schema (Neuschatz et al., 2002). Therefore, our results are consistent with previous studies that schematic information can increase performance on certain memory measures but decrease performance on others.

Further, our findings concerning accessibility and precision suggest the increase in "information" in terms of accessibility is equivalent to the decrease in terms of precision (hence the lack of difference in Total Information), such that schematic information in this paradigm does not increase overall memory performance. Although we cannot yet generalise beyond the present experimental approach, one possibility is that this accessibility versus precision trade-off (or the trade-off between hits and false-alarms in other experiments) might result in no net memory benefit in the presence of a schema. In short, schematic information alters memory behaviour, but our results question whether it benefits overall memory performance.

7.4. Conclusion

Across four experiments, we provide evidence for memory and generalisation effects for both schema-relevant and -irrelevant information. Critically, we have shown that memory and generalisation behaviour is biased away from a schematic location for schema-irrelevant information. These effects appear immediately after encoding and appear relatively stable over a 24-h period. We have outlined three broad explanations for this behaviour: (1) a mutual exclusivity bias account, (2) a base-rate neglect account and (3) a location-based interference account. Whereas the mutual exclusivity bias account would likely require implicit or explicit categorisation of the words as human-made or natural, the latter two accounts may not require such categorisation.

Given that these effects emerge immediately after encoding, with evidence of decline over longer delays in other experiments (e.g., Antony et al., 2021; Berens et al., 2020; Tompary et al., 2020), the generalisation behaviour is likely driven by a retrieval-based mechanism that relies on memory for the individual episodes to infer a location for novel items by relying on close semantic neighbours. In this way, categorisation at the superordinate level is unnecessary so long as the participant is sensitive to the semantic relatedness among individual items. Using either the base rate neglect or interference mechanism (regardless of semantic category), both of which may rely on a retrieval-based approach, could explain our effects in a parsimonious manner without the need for explicit strategies or semantic categorisation of the words.

Formal modelling is likely to provide further theoretical insight. For example, accessibility and precision measures have recently been suggested to emerge from a single d-prime measure in a signal detection framework (Schurgin, Wixted, & Brady, 2020). Careful analysis of the shapes of the distributions produced by these models compared to our experimental data may help delineate between models. Incorporating

both location-based interference and semantic relatedness in such a framework may be able to accommodate our findings without the need for schematic representations or semantic categorisation. Indeed, the presence of both location- and semantic-based interference may explain our generalisation effects and memory effects seen in the current study and Berens et al. (2020); for example, the differences in accessibility and precision between the clustered and non-clustered condition.

Regardless of the exact mechanism, our results highlight that the presence of schematic information can affect memory and generalisation behaviour for schema-relevant and -irrelevant information. Experimentally, these results have implications for future studies that use schema-irrelevant information as a control condition, where behaviour is assumed not to be affected by the presence of a schema. Theoretically, the results provide insight into schema processing. They suggest that schematic information affects memory and generalisation behaviour immediately after encoding for both schema-relevant and -irrelevant information in a manner that is not clearly predicted by existing schema theories.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105203.

References

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. L. Bjork, & R. A. Bjork (Eds.), *Memory* (pp. 237–313). Academic Press. https://doi.org/10.1016/B978-012102570-0/50010-0.

Anderson, R. C. (1984). Some reflections on the Acquisition of Knowledge. *Educational Researcher, 13*(9), 5–10. https://doi.org/10.3102/0013189X013009005

Antony, J. W., Stiver, C. A., Graves, K. N., Osborne, J., Turk-Browne, N. B., & Bennion, K. A. (2021). Spatial gist extraction during human memory consolidation. *Journal of Experimental Psychology. Learning, Memory, and Cognition.* https://doi.org/10.1037/xlm0000894

Atienza, M., Crespo-Garcia, M., & Cantero, J. L. (2011). Semantic congruence enhances memory of episodic associations: Role of theta oscillations. *Journal of Cognitive Neuroscience, 23*(1), 75–90. https://doi.org/10.1162/jocn.2009.21358

Baddeley, A., & Hitch, G. J. (1977). Recency re-examined. In S. Dornic (Ed.), *Attention and performance VI* (pp. 647–667). Lawrence Erlbaum Associates.

Barnes, J. M., & Underwood, B. J. (1959). 'Fate' of first-list associations in transfer theory. *Journal of Experimental Psychology, 58*(2), 97–105. https://doi.org/10.1037/h0047507

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge University Press.

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9*(10), 1–11. https://doi.org/10.1167/9.10.7

Berens, S. C., & Bird, C. M. (2021). *Medial temporal and prefrontal cortices encode structural task representations at different levels of abstraction.* https://doi.org/10.1101/2021.08.11.455977

Berens, S. C., Richards, B. A., & Horner, A. J. (2020). Dissociating memory accessibility and precision in forgetting. *Nature Human Behaviour, 4*(8), 866–877. https://doi.org/10.1038/s41562-020-0888-8

Bohil, C. J., & Wismer, A. J. (2015). Implicit learning mediates base rate acquisition in perceptual categorization. *Psychonomic Bulletin & Review, 22*(2), 586–593. https://doi.org/10.3758/s13423-014-0694-2

Born, J., & Wilhelm, I. (2012). System consolidation of memory during sleep. *Psychological Research, 76*(2), 192–203. https://doi.org/10.1007/s00426-011-0335-6

Brady, T., Schacter, D. L., & Alvarez, G. (2018). The adaptive nature of false memories is revealed by gist-based distortion of true memories. *PsyArXiv.*. https://doi.org/10.31234/osf.io/zeg95

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology, 13*(2), 207–230. https://doi.org/10.1016/0010-0285(81)90008-6

van Buuren, M., Kroes, M. C. W., Wagner, I. C., Genzel, L., Morris, R. G. M., & Fernández, G. (2014). Initial investigation of the effects of an experimentally learned schema on spatial associative memory in humans. *Journal of Neuroscience, 34*(50), 16662–16670. https://doi.org/10.1523/JNEUROSCI.2365-14.2014

Cann, D. R., McRae, K., & Katz, A. N. (2011). False recall in the Deese–Roediger–McDermott paradigm: The roles of gist and associative strength. *Quarterly Journal of Experimental Psychology, 64*(8), 1515–1542. https://doi.org/10.1080/17470218.2011.560272

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language, 15*(2), 317–335. https://doi.org/10.1017/S0305000900012393

Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., & Damasio, A. R. (1996). A neural basis for lexical retrieval. *Nature, 380*(6574), 499–505. https://doi.org/10.1038/380499a0

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 364*(1536), 3773–3800. https://doi.org/10.1098/rstb.2009.0111

Djonlagic, I., Rosenfeld, A., Shohamy, D., Myers, C., Gluck, M., & Stickgold, R. (2009). Sleep enhances category learning. *Learning & Memory, 16*(12), 751–755. https://doi.org/10.1101/lm.1634509

Dudai, Y. (2012). The restless engram: Consolidations never end. *Annual Review of Neuroscience, 35*, 227–247. https://doi.org/10.1146/annurev-neuro-062111-150500

Durrant, S. J., Taylor, C., Cairney, S., & Lewis, P. A. (2011). Sleep-dependent consolidation of statistical learning. *Neuropsychologia, 49*(5), 1322–1331. https://doi.org/10.1016/j.neuropsychologia.2011.02.015

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., & Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences, 104*(18), 7723–7728. https://doi.org/10.1073/pnas.0700094104

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fernández, G., & Morris, R. G. M. (2018). Memory, novelty and prior knowledge. *Trends in Neurosciences, 41*(10), 654–659. https://doi.org/10.1016/j.tins.2018.08.006

Frank, D., Montaldi, D., Wittmann, B., & Talmi, D. (2018). Beneficial and detrimental effects of schema incongruence on memory for contextual events. *Learning & Memory, 25*(8), 352–360. https://doi.org/10.1101/lm.047738.118

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia, 53*, 104–114. https://doi.org/10.1016/j.neuropsychologia.2013.11.010

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology, 28*(1), 99–108. https://doi.org/10.1037/0012-1649.28.1.99

Graves, K. N., Antony, J. W., & Turk-Browne, N. B. (2020). Finding the pattern: On-line extraction of spatial structure during virtual navigation. *Psychological Science, 31*(9), 1183–1190. https://doi.org/10.1177/0956797620948828

Greve, A., Cooper, E., Tibon, R., & Henson, R. N. (2019). Knowledge is power: Prior knowledge aids memory for both congruent and incongruent events, but in different ways. *Journal of Experimental Psychology: General, 148*(2), 325–341. https://doi.org/10.1037/xge0000498

Harlow, I. M., & Donaldson, D. I. (2013). Source accuracy data reveal the thresholded nature of human episodic memory. *Psychonomic Bulletin & Review, 20*(2), 318–325. https://doi.org/10.3758/s13423-012-0340-9

Hawkins, G. E., Hayes, B. K., Donkin, C., Pasqualino, M., & Newell, B. R. (2015). A Bayesian latent-mixture model analysis shows that informative samples reduce base-rate neglect. *Decision, 2*(4), 306–318. https://doi.org/10.1037/dec0000024

Head, H., & Holmes, G. (1911). Sensory disturbances from cerebral lesions. *Brain, 34* (2–3), 102–254. https://doi.org/10.1093/brain/34.2-3.102

Henson, R. N., & Gagnepain, P. (2010). Predictive, interactive multiple memory systems. *Hippocampus, 20*(11), 1315–1326. https://doi.org/10.1002/hipo.20857

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Hintzman, D. L. (1986). 'Schema abstraction' in a multiple-trace memory model. *Psychological Review, 93*(4), 411–428. https://doi.org/10.1037/0033-295X.93.4.411

Hunt, R. R., & Worthen, J. B. (2006). *Distinctiveness and memory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195169669.001.0001

Inostroza, M., & Born, J. (2013). Sleep for preserving and transforming episodic memory. *Annual Review of Neuroscience, 36*(1), 79–102. https://doi.org/10.1146/annurev-neuro-062012-170429

Jenkins, J. G., & Dallenbach, K. M. (1924). Obliviscence during sleep and waking. *The American Journal of Psychology, 35*(4), 605–612. https://doi.org/10.2307/1414040

van Kesteren, M. T. R., Fernández, G., Norris, D. G., & Hermans, E. J. (2010). Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proceedings of the National Academy of Sciences, 107*(16), 7550–7555. https://doi.org/10.1073/pnas.0914892107

van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences, 35*(4), 211–219. https://doi.org/10.1016/j.tins.2012.02.001

Kleider, H. M., Pezdek, K., Goldinger, S. D., & Kirk, A. (2008). Schema-driven source misattribution errors: Remembering the expected from a witnessed event. *Applied Cognitive Psychology, 22*(1), 1–20. https://doi.org/10.1002/acp.1361

Kliegl, O., Pastötter, B., & Bäuml, K.-H. T. (2015). The contribution of encoding and retrieval processes to proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(6), 1778–1789. https://doi.org/10.1037/xlm0000096

Korkki, S. M., Richter, F. R., Jeyarathnarajah, P., & Simons, J. S. (2020). Healthy ageing reduces the precision of episodic memory retrieval. *Psychology and Aging, 35*(1), 124–142. https://doi.org/10.1037/pag0000432

Kroes, M. C. W., & Fernández, G. (2012). Dynamic neural systems enable adaptive, flexible memories. *Neuroscience & Biobehavioral Reviews, 36*(7), 1646–1666. https://doi.org/10.1016/j.neubiorev.2012.02.014

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences, 20*(7), 512–534. https://doi.org/10.1016/j.tics.2016.05.004

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories. *Psychological Review, 119*(3), 573–616. https://doi.org/10.1037/a0028681

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the emergence of conceptual knowledge during human decision making. *Neuron, 63*(6), 889–901. https://doi.org/10.1016/j.neuron.2009.07.030

Lake, B. M., Linzen, T., & Baroni, M. (2019). Human few-shot learning of compositional instructions. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 611–617). Montreal, QB: Cognitive Science Society. https://doi.org/10.48550/arXiv.1901.04587.

Lew, A. R., & Howe, M. L. (2017). Out of place, out of mind: Schema-driven false memory effects for object-location bindings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(3), 404–421. https://doi.org/10.1037/xlm0000317

Lovett, M. C., & Schunn, C. D. (1999). Task representations, strategy variability, and base-rate neglect. *Journal of Experimental Psychology: General, 128*(2), 107–130. https://doi.org/10.1037/0096-3445.128.2.107

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences, 17*(8), 391–400. https://doi.org/10.1016/j.tics.2013.06.006

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology, 9*(1), 111–151. https://doi.org/10.1016/0010-0285(77)90006-8

Mäntylä, T., & Bäckman, L. (1992). Aging and memory for expected and unexpected objects in real-world settings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(6), 1298–1309. https://doi.org/10.1037/0278-7393.18.6.1298

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20*(2), 121–157. https://doi.org/10.1016/0010-0285(88)90017-5

Markman, E. M., Wasow, J. L., & Hansen, M. B. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology, 47*(3), 241–275. https://doi.org/10.1016/S0010-0285(03)00034-3

MATLAB. (2019). *version 9.7.0 (R2019).* Natick, Massachusetts: The MathWorks Inc.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *26. Advances in Neural Information Processing Systems* (pp. 3111–3119). Curran Associates, Inc.

Mirković, J., & Gaskell, M. G. (2016). Does sleep improve your grammar? Preferential consolidation of arbitrary components of new linguistic knowledge. *PLoS One, 11*(4). https://doi.org/10.1371/journal.pone.0152489

Mirković, J., Vinals, L., & Gaskell, M. G. (2019). The role of complementary learning systems in learning and consolidation in a quasi-regular domain. *Cortex, 116*, 228–249. https://doi.org/10.1016/j.cortex.2018.07.015

Neuschatz, J. S., Lampinen, J. M., Preston, E. L., Hawkins, E. R., & Toglia, M. P. (2002). The effect of memory schemata on memory and the phenomenological experience of

naturalistic situations. *Applied Cognitive Psychology, 16*(6), 687–708. https://doi.org/10.1002/acp.824

Nilakantan, A. S., Bridge, D. J., VanHaerents, S., & Voss, J. L. (2018). Distinguishing the precision of spatial recollection from its success: Evidence from healthy aging and unilateral mesial temporal lobe resection. *Neuropsychologia, 119*, 101–106. https://doi.org/10.1016/j.neuropsychologia.2018.07.035

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Peich, M.-C., Husain, M., & Bays, P. M. (2013). Age-related decline of precision and binding in visual working memory. *Psychology and Aging, 28*(3), 729–743. https://doi.org/10.1037/a0033236

Pezdek, K., Whetstone, T., Reynolds, K., Askari, N., & Dougherty, T. (1989). Memory for real-world scenes: The role of consistency with schema expectation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(4), 587–595. https://doi.org/10.1037/0278-7393.15.4.587

Piaget, J. (1926). Psychology. *The Monist, 36*(3), 430–455.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3, Pt.1), 353–363. https://doi.org/10.1037/h0025953

Preston, A. R., & Eichenbaum, H. (2013). Interplay of Hippocampus and prefrontal cortex in memory. *Current Biology, 23*(17), 764–773. https://doi.org/10.1016/j.cub.2013.05.041

Quent, J. A., Henson, R. N., & Greve, A. (2021). A predictive account of how novelty influences declarative memory. *Neurobiology of Learning and Memory, 179*. https://doi.org/10.1016/j.nlm.2021.107382

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological Reviews, 93*(2), 681–766. https://doi.org/10.1152/physrev.00032.2012

Richards, B. A., Xia, F., Santoro, A., Husse, J., Woodin, M. A., Josselyn, S. A., & Frankland, P. W. (2014). Patterns across multiple memories are identified over time. *Nature Neuroscience, 17*(7), 981–986. https://doi.org/10.1038/nn.3736

Richter, F. R., Bays, P. M., Jeyarathnarajah, P., & Simons, J. S. (2019). Flexible updating of dynamic knowledge structures. *Scientific Reports, 9*(1), Article 1. https://doi.org/10.1038/s41598-019-39468-9

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Rojahn, K., & Pettigrew, T. F. (1992). Memory for schema-relevant information: A meta-analytic resolution. *British Journal of Social Psychology, 31*(2), 81–109. https://doi.org/10.1111/j.2044-8309.1992.tb00958.x

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology, 4*(3), 328–350. https://doi.org/10.1016/0010-0285(73)90017-0

Sadeh, T., Ozubko, J. D., Winocur, G., & Moscovitch, M. (2016). Forgetting patterns differentiate between two forms of memory representation. *Psychological Science, 27*(6), 810–820. https://doi.org/10.1177/0956797616638307

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General, 133*(4), 534–553. https://doi.org/10.1037/0096-3445.133.4.534

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 372*(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour, 4*(11), 1156–1172. https://doi.org/10.1038/s41562-020-00938-0

Smith, D. J., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(1), 3–27. https://doi.org/10.1037/0278-7393.26.1.3

Sun, S. Z., Fidalgo, C., Barense, M. D., Lee, A. C. H., Cant, J. S., & Ferber, S. (2017). Erasing and blurring memories: The differential impact of interference on separate aspects of forgetting. *Journal of Experimental Psychology: General, 146*(11), 1606–1630. https://doi.org/10.1037/xge0000359

Sweegers, C. C. G., Coleman, G. A., van Poppel, E.a. M., Cox, R., & Talamini, L. M. (2015). Mental schemas hamper memory storage of goal-irrelevant information. *Frontiers in Human Neuroscience, 9*(629). https://doi.org/10.3389/fnhum.2015.00629

Sweegers, C. C. G., & Talamini, L. M. (2014). Generalization from episodic memories across time: A route for semantic knowledge acquisition. *Cortex, 59*, 49–61. https://doi.org/10.1016/j.cortex.2014.07.006

Tamminen, J., Davis, M. H., Merkx, M., & Rastle, K. (2012). The role of memory consolidation in generalisation of new linguistic information. *Cognition, 125*(1), 107–112. https://doi.org/10.1016/j.cognition.2012.06.014

Tompary, A., & Thompson-Schill, S. L. (2021). Semantic influences on episodic memory distortions. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001017

Tompary, A., Zhou, W., & Davachi, L. (2020). Schematic memories develop quickly, but are not expressed unless necessary. *Scientific Reports, 10*(1). https://doi.org/10.1038/s41598-020-73952-x

Tulving, E., & Kroll, N. (1995). Novelty assessment in the brain and long-term memory encoding. *Psychonomic Bulletin & Review, 2*(3), 387–390. https://doi.org/10.3758/BF03210977

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review, 64*(1), 49–60. https://doi.org/10.1037/h0044616

Wagner, I. C., van Buuren, M., Kroes, M. C., Gutteling, T. P., van der Linden, M., Morris, R. G., & Fernández, G. (2015). Schematic memory components converge within angular gyrus during retrieval. *ELife, 4*. https://doi.org/10.7554/eLife.09668

Warren, D. E., Jones, S. H., Duff, M. C., & Tranel, D. (2014). False recall is reduced by damage to the ventromedial prefrontal cortex: Implications for understanding the neural correlates of schematic memory. *Journal of Neuroscience, 34*(22), 7677–7682. https://doi.org/10.1523/JNEUROSCI.0119-14.2014

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain: A Journal of Neurology, 107*(3), 829–854. https://doi.org/10.1093/brain/107.3.829

Welsh, M. B., & Navarro, D. J. (2012). Seeing is believing: Priors, trust, and base rate neglect. *Organizational Behavior and Human Decision Processes, 119*(1), 1–14. https://doi.org/10.1016/j.obhdp.2012.04.001

Wismer, A. J., & Bohil, C. J. (2017). Base-rate sensitivity through implicit learning. *PLoS One, 12*(6), Article e0179256. https://doi.org/10.1371/journal.pone.0179256

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology, 55*(1), 235–269. https://doi.org/10.1146/annurev.psych.55.090902.141555

Wolfe, C. R. (2007). Adaptive redundancy, denominator neglect, and the base-rate fallacy. *Behavioral and Brain Sciences, 30*(3), 286–287. https://doi.org/10.1017/S0140525X07001951

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron, 75*(1), 168–179. https://doi.org/10.1016/j.neuron.2012.05.010

Zeng, T., Tompary, A., Schapiro, A. C., & Thompson-Schill, S. L. (2021). Tracking the relation between gist and item memory over the course of long-term memory consolidation. *ELife, 10*, Article e65588. https://doi.org/10.7554/eLife.65588

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*, 233–235. https://doi.org/10.1038/nature06860

Zola-Morgan, S. M., & Squire, L. R. (1990). The primate hippocampal formation: Evidence for a time-limited role in memory storage. *Science, 250*(4978), 288–290. https://doi.org/10.1126/science.2218534