Journal of Experimental Psychology: General

Limitations to Optimal Search in Naturalistic Active Learning

Lisheng He, Russell Richie, and Sudeep Bhatia

Online First Publication, March 28, 2024. https://dx.doi.org/10.1037/xge0001558

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

https://doi.org/10.1037/xge0001558

Limitations to Optimal Search in Naturalistic Active Learning

Lisheng He[1], Russell Richie[2], and Sudeep Bhatia[3]
[1] SILC Business School, Shanghai University
[2] MindCORE and Program in Cognitive Science, University of Pennsylvania
[3] Department of Psychology, University of Pennsylvania

Optimality in active learning is under intense debate in numerous disciplines. We introduce a new empirical paradigm for studying naturalistic active learning, as well as new computational tools for jointly modeling algorithmic and rational theories of information search. Participants in our task can ask questions and learn about hundreds of everyday items but must retrieve queried items from memory. To maximize information gain, participants need to retrieve sequences of dissimilar items. In eight experiments ($N = 795$), we find that participants are unable to do this. Instead, associative memory mechanisms lead to the successive retrieval of similar items, an established memory effect known as semantic congruence. The extent of semantic congruence (and thus suboptimality in question asking) is unaffected by task instructions and incentives, though participants can identify efficient query sequences when given a choice between query sequences. Overall, our results indicate that participants can distinguish between optimal and suboptimal search if explicitly asked to do so, but have difficulty implementing optimal search from memory. We conclude that associative memory processes may place critical restrictions on people's ability to ask good questions in naturalistic active learning tasks.

Public Significance Statement

The ability to guide our own learning is one of the greatest developmental tools we have. A common theory of such self-guided or "active" learning is that people tend to ask questions that gather the most information. However, most work supporting this theory is based on research with simple domains where questions need not be generated from memory. We show that in more complex domains, where more realistic questions must be recalled, subjects are far less able to generate optimal queries. This suggests that optimal question asking may be more limited in naturalistic, everyday domains than previously thought.

Keywords: active learning, optimal search, semantic congruence, word embeddings, computational modeling

People often choose what information they want to gather. A child can choose which toy to explore; a student can choose what textbook to read; and a scientist can choose what experiments to run. This kind of learning is known as active learning and has been the subject of intense study in recent years in several fields, including developmental and cognitive psychology (Coenen et al., 2019; Gopnik, 1996), education (Carr et al., 2015), neuroscience (Friston, 2009), and machine learning (Settles, 2009). Although there are many questions to ask about active learning, perhaps the most pressing question about active learning is this: how and why do people seek the particular information they seek?

Theories of rational cognition provide an increasingly popular answer to this question (Anderson, 1990; Griffiths et al., 2010). These theories propose that people search for information optimally;

Lisheng He https://orcid.org/0000-0003-4857-601X

Some of the data and ideas in the article were presented at the 2022 Conference of the Cognitive Science Society, the 2022 National Science Foundation Augmented Intelligence workshop, and the 2022 Chinese Decision Psychology Young Scholar Forum. Lisheng He was supported by the National Natural Science Foundation of China (72101156). Sudeep Bhatia was supported by the National Science Foundation (SES-1847794). The funding sources have no involvement in the study design, data analysis, and article preparation. All experimental materials, data, and code can be found at Open Science Framework: https://osf.io/5e6tk/.

Lisheng He and Russell Richie contributed equally to this work. Lisheng He served as lead for data curation, formal analysis, investigation, methodology, visualization, and writing–original draft and contributed equally to funding acquisition and validation. Russell Richie served as lead for investigation, methodology, project administration, and writing–original draft, contributed equally to data curation and formal analysis, and served in a supporting role for visualization. Sudeep Bhatia served as lead for conceptualization, funding acquisition, methodology, supervision, and writing–review and editing, contributed equally to writing–original draft, and served in a supporting role for investigation. Lisheng He and Russell Richie contributed equally to conceptualization and writing–review and editing.

Correspondence concerning this article should be addressed to Lisheng He, SILC Business School, Shanghai University, Shanghai, China. Email: hlisheng@shu.edu.cn

that is, they generate queries that provide the most information possible. The rational account of active learning has been successfully tested in many domains in psychology (Coenen et al., 2019; Meder et al., 2021), including causal learning (Bramley et al., 2015), categorization (Markant & Gureckis, 2014), spatial search (Gureckis & Markant, 2009), rule learning (J. D. Nelson et al., 2001), and function learning (A. Jones et al., 2018; Wu et al., 2018). Rational theories have also benefited from research on optimal experimental design in statistics, which provides a mathematically rigorous framework for specifying query optimality (A. Atkinson et al., 2007; Cavagnaro et al., 2011; Kiefer, 1959; Lindley, 1956; Myung & Pitt, 2009; Myung et al., 2013; Ouyang et al., 2016).

Of course, the rational account is not the only theory of active learning. One of the earliest experimental investigations of active learning was conducted by Wason (1966, 1968). Using a card selection task, Wason showed that participants seldom asked the optimal question and instead showed a tendency for confirmatory search (Wason, 1966). There are compelling rational accounts of Wason's findings (Klayman & Ha, 1987; Oaksford & Chater, 1994, 2007), but some researchers have also argued that confirmatory search is a direct product of fundamental cognitive mechanisms, including associative mechanisms implicated in memory (Bhatia, 2016; Glöckner & Betsch, 2008; Holyoak & Simon, 1999).

It would not be surprising if the associative structure of memory were to constrain the types of questions people generate in active learning. After all, associative memory has been shown to play a crucial role in closely related tasks, such as categorization (e.g., Nosofsky & Palmeri, 1997), judgment (e.g., Juslin et al., 2008), and decision making (e.g., Aka & Bhatia, 2021). That said, there have been few attempts to explore memory processes in naturalistic active learning, as most studies are conducted using artificial stimuli, varying on a very small number of dimensions, that are unlike many of the real entities in the world that people must learn about. For example, there are very few instances in real life where people have to learn the property values of various triangles and squares with different types of shadings and patterns, a common experimental task in the psychology literature. By contrast, many real-world property learning tasks involve natural objects, like foods or animals, for which individuals have rich existing representations. Additionally, most experiments on active learning ask subjects to choose among a relatively small number of experimenter-generated queries, reducing the role of memory processes, which would be expected to influence subjects' retrieval of queries (but see Rothe et al., 2018; Wilke et al., 2009 for exceptions). Overall, it is not clear, based on prior work, whether or not people behave optimally when they must retrieve sequences of (complex, naturalistic) questions from memory.

One important source of conflict between theories of rational search and theories of memory search involves the role of similarity. Optimal search often requires asking questions that are dissimilar to each other, as asking the same (or a similar) question repeatedly will usually provide the same (or similar) answers/information. Consider, for example, a task in which the learner has to determine how much of a new nutrient there is in different food items. The learner can ask questions about each item sequentially (how much of the nutrient is there in a bagel? how much in pita? how much in an egg?) and must retrieve each item (bagel, pita, egg) from memory before the query. As similar items usually have similar properties, for the questions to be maximally informative, the queried items must be as different from each other as possible. It is much better to follow up a query about bagel with a query about egg than a query about pita.

This optimal search strategy is the opposite of what researchers have observed in most recall tasks. Typically, when asked to retrieve items from memory, people generate sequences of semantically similar items, an effect known as semantic congruence. The semantic-congruence effect is remarkably robust and emerges across a variety of tasks including free association (De Deyne et al., 2019; D. L. Nelson et al., 2004), free recall from lists (Howard & Kahana, 2002; Romney et al., 1993), semantic memory search (Bousfield & Sedgewick, 1944; Hills et al., 2012), and memory-based decision making (Aka & Bhatia, 2021; Bhatia, 2019; Z. H. Zhang et al., 2021). This is due to the associative structure of memory (Anderson & Bower, 2014; R. C. Atkinson & Shiffrin, 1968). Retrieved items cue successive items based on their strength of association. Items that are similar are more associated with each other, which is why the retrieval of bagel is more likely to cue pita than egg.

How is this conflict resolved in naturalistic active learning tasks? Are people able to search optimally and retrieve sequences of dissimilar items, or are they fundamentally constrained by the associative memory processes that lead to semantic congruence in other recall tasks? As discussed above, most studies on active learning are conducted under rarefied conditions that do not require memory search. This is largely due to the difficulty in modeling naturalistic active learning, in which people can search over and ask questions about thousands of common items and entities. Such items do not always have easily quantifiable representations, and researchers are thus unable to specify formal models of memory search that operate over these representations. Yet without such models, and the quantitative representations that enable them, researchers can say little about the memory processes that underly naturalistic question asking and whether these processes are actually optimal.

Overview of Approach

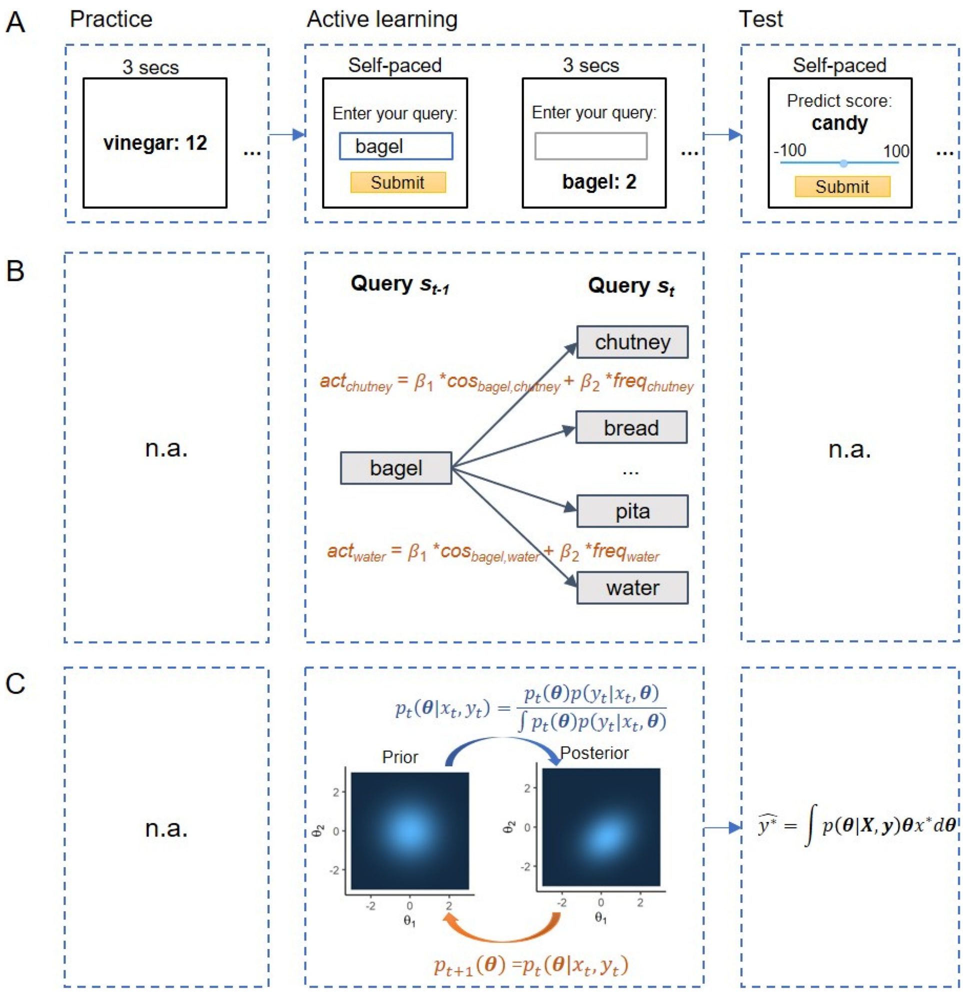
Fortunately, recent work has shown the promise of natural language processing models, trained on large text corpora, for tackling the problem of representation (Bhatia & Aka, 2022). For example, distributed semantic models (DSMs) use patterns of word–word or word–document co-occurrence in very large collections of texts, to build real-valued, high-dimensional vector representations of thousands or even millions of real words and phrases (Griffiths et al., 2007; M. N. Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov et al., 2013; Pennington et al., 2014). As DSMs are trained on vast amounts of text corpora, these representations are very rich and can capture a lot of what people know and associate with words (and their corresponding objects and concepts). Crucially, objects that are semantically similar, like bagel and pita, tend to have similar word distributions in text, and therefore end up with vector representations that are close to each other by metrics like cosine similarity or Euclidean distance. For this reason, similarity measurements in DSM vector spaces can be used to describe many psychological phenomena related to semantic similarity, and more generally semantic representation and memory retrieval (Bhatia et al., 2019; Günther et al., 2019), including semantic-congruence effects in free association, free recall from lists, semantic memory search, and memory-based decision making (Aka & Bhatia, 2021; Bhatia, 2019;

Griffiths et al., 2007; Hills et al., 2012; Howard & Kahana, 2002; M. N. Jones & Mewhort, 2007; Mandera et al., 2017).

We use DSMs to model memory search in a new naturalistic active learning task (Figure 1A). In the task, participants (a) learn a novel property by querying different entities in a category and getting feedback on those entities' property scores, and then (b) in the test phase predict the scores of a fixed set of test items. Property scores for the entities are constructed by prespecified linear functions on their DSM vectors, giving similar items similar property scores. Experiments 1a, 2, and 3 implement this task with 1,594 food items, while Experiment 1b implements it with 1,734 animals. Additionally, Experiment 2 compares the queries in the active learning task with recall in a standard semantic memory search task. Experiment 3 provides detailed coaching on how to do well in the active learning task. Experiments 4a, 4b, 5a, and 5b do not directly use the task but instead ask participants to judge the optimality of

Figure 1

Experimental and Modeling Setup



Note. (A) The practice, active learning, and test phases of the naturalistic active learning task. In the practice phase, the participants were given five pairs of items and scores, one at a time. The practice phase was designed to provide the participants with a broad understanding of the task. In the active learning phase, participants generated their own queries, submitted the queries, and were shown the queried items' scores. Items were queried successively. In the test phase, participants were presented with 20 preselected items one at a time and asked to predict the items' scores based on what they had learned in the active learning phase. (B) In the memory model, activation of candidate queries at time t is linearly dependent on a query's similarity to the query at time $t-1$, as well as a query's frequency. Activation is subsequently passed through a softmax function to obtain probabilities for sampling every possible query given the previous query, $\Pr[s_t|s_{t-1}]$. (C) In the Bayesian learning model, the learner comes into the active learning phase with a prior belief of the property, which we set at a standard multivariate normal distribution over $\boldsymbol{\theta}$. After receiving the feedback upon query at t, the learner updates the belief of $\boldsymbol{\theta}$ using the Bayes rule (the arrow and equation on top). The posterior belief after query t becomes the prior belief for the subsequent query at $t+1$ (the arrow and equation at the bottom). In the test phase, the learner makes predictions for test item x^* based on the posterior belief of $\boldsymbol{\theta}$ after the full active learning phase. See the online article for the color version of this figure.

individual searches or search sequences in the task. All experiments are incentivized, and participants can earn more money with better performance at test. Experiments 2 and 3 are preregistered.

Our use of linear functions on DSM vector spaces is driven by our prior work which has found that human judgments of the real-world properties of common objects (such as the healthiness of foods, the riskiness of various technologies, the leadership qualities of famous individuals, and the similarities of objects like birds and vegetables) can be accurately approximated using a linear function on the objects' DSM vectors (Bhatia, 2019; Bhatia & Stewart, 2018; Bhatia et al., 2022; Gandhi et al., 2022; Richie & Bhatia, 2021; Zou & Bhatia, 2021a; see Richie et al., 2019 for a comprehensive analysis). In some of this work, we have also found that more flexible nonlinear functions do not add predictive value in describing human judgments. The reason for this is likely the richness of the word vectors—many complex real-world properties of objects can be represented using linear combinations of 300+ dimensions. We were also motivated by our recent work on (passive) category learning that finds that linear functions on DSM vectors are much easier to learn than nonlinear functions on these vectors (Zou & Bhatia, 2021b). This work was motivated by the classical findings of Shepard et al. (1961), and essentially replicated these findings using DSM vectors as the underlying attribute structures for objects. Ultimately, linear functions on DSM vectors accurately specify human judgments of real-world properties and are also the easiest to learn, which is why they are a good starting point for our analysis.

Participants' queries allow us to measure the extent of semantic congruence or incongruence in search behavior. We specify semantic congruence as the averaged cosine similarity between the DSM vectors of each pair of two adjacent items in the query sequence. To model the dynamics that underlie query generation, we specify memory search as a Markov random walk over the hundreds or thousands of items in the category (foods or animals; Abbott et al., 2015; De Deyne et al., 2019). Our memory model predicts the probability of retrieving an item as a (positive or negative) function of the semantic similarity with the previously retrieved item, as well as the item's word frequency (Figure 1B). By fitting this model on participants' query sequences from our experiments, we can quantify the effect of semantic similarity (and dissimilarity) on memory processes in naturalistic question asking.

Our memory model also allows us to make predictions about participant responses in the test phase of our task, and how these responses depend on underlying memory dynamics. In particular, by assuming that participants optimally extrapolate from the training data to the test data (i.e., that they are ideal Bayesian learners) we can formally specify the relationship between test accuracy (the ability of our participant to learn the underlying property function) and retrieval strategy (the tendency of participants to display semantic congruence or incongruence in search; Figure 1C). Additionally, we measure the efficiency of our participant's queries using Bayesian D-optimality, a well-known efficiency criterion in research on optimal experimental design (Kiefer & Wolfowitz, 1959; Myung & Pitt, 2009). D-optimality allows us to evaluate the quality of participant queries, and by doing so, relate the mechanisms revealed by our algorithmic memory model to the computational goals of the task. As the function that participants are required to learn is linear in the DSM vector space, we use linear D-optimality. However, for robustness, we also examine how the linear D-optimality relates to the measurement of optimality for Gaussian process function learning, the latter of which allows for nonlinear mapping between word vectors and property scores (Hayes et al., 2019; Lucas et al., 2015; Wu et al., 2018).

Experiments 1a and 1b

The goal of these experiments was to investigate naturalistic active learning in a new experimental task involving thousands of foods (Experiment 1a) and animals (Experiment 1b). Participants were asked to learn a hypothetical property possessed by foods or animals and were allowed to ask about nearly any food or animal in the training phase of the task. We used a variety of computational techniques to analyze query generation, its relationship with memory retrieval, and its effect on learning performance.

Method

Participants

Participants were 396 U.S. residents who spoke fluent English recruited from Prolific Academic (https://www.prolific.ac; Experiment 1a: $M_{\text{age}} = 34$ years, $SD_{\text{age}} = 12$ years, $N = 198$, 71% female, 27% male, 0.5% other/prefer not to say; Experiment 1b: $M_{\text{age}} = 32$ years, $SD_{\text{age}} = 12$ years, $N = 198$, 62% female, 36% male, 1.5% other/prefer not to say). In all experiments in this article, no other demographic information besides age and gender was collected, and gender was collected as a three-way choice between the categories above (although in subsequent experiments, we changed "other" to "nonbinary" in response to participant feedback). For data quality control, we only recruited the participants with an approval rate of over 80% on Prolific.

Stimuli

All food and animal stimuli were collected via the Natural Language Toolkit interface to WordNet (Bird et al., 2009; Miller, 1995). We extracted all nouns that were descendants (hyponyms, hyponyms of hyponyms, etc.) of the first and second synsets of the word food (Experiment 1a) and the first synset of the word animal (Experiment 1b), and then filtered out ambiguous and nonfood/nonanimal items and those items that did not have vectors in a standard word vector model, word2vec trained on Google News (Mikolov et al., 2013). This led to 1,594 usable food items and 1,734 usable animal items, each of which was represented by a 300-dimensional vector.

Our use of the Google News word2vec model was driven by prior work, which has found that distances in this vector space are good predictors of human similarity judgment (Mikolov et al., 2013; Pereira et al., 2016) and that linear functions of this space can be used to accurately predict both similarity judgments (Richie & Bhatia, 2021) and judgments of other properties of objects (Bhatia, 2019; Bhatia & Stewart, 2018; Bhatia et al., 2022; Gandhi et al., 2022; Richie & Bhatia, 2021; Zou & Bhatia, 2021a; see Richie et al., 2019 for a comprehensive analysis). Importantly, this work shows that the accuracy of the word2vec model persists for animals and foods (stimuli in our experiments) despite the fact that the underlying corpus contains news articles. Most of this work also finds that there are few differences between this model and other popular models like GloVe (Pennington et al., 2014), though word2vec does have the benefit of having a larger vocabulary with multiword phrases (e.g., words like "polar_bear"). Indeed, we have found that only 60% of the animal words and 61% of the food words in our stimuli sets are in the vocabulary of a popular GloVe model trained on the Common Crawl corpus. Out of the words that are in both vocabularies, there is a Pearson correlation of $r = .66$, 95% CI [0.66, 0.66] in the cosine similarities of animals and $r = .64$, 95% CI [0.64, 0.64] in the cosine similarities of foods, suggesting our upcoming results would be similar if we had used GloVe instead of word2vec. Of course, there are also more recent models, like Bidirectional Encoder Representations from Transformers, or BERT, which specify vectors for sentences. We decided not to use these models in our article as prior work has shown that their vectors are not good for modeling similarity without further fine tuning (Reimers & Gurevych, 2019).

Prior research suggests that the 10-dimensional principal components of the original 300-dimensional word2vec vectors in one domain can predict item properties nearly as well as the full 300-dimensional vectors do (Richie et al., 2019). Thus, to generate coherent and learnable properties, we subjected all 1,594 food vectors and all 1,734 animal vectors to (separate) principal component analysis (PCA) and used the first 10 principal components (which explained 30% and 32% of the variation in foods and animals, respectively). We further quantified word frequency using the Google Books NGram corpus (Michel et al., 2011; Trenkmann, 2016).

We applied random linear functions on the word vector principal components to generate artificial property scores for all items. The generated weight vectors were 10-dimensional, corresponding to the 10-dimensional word vector principal components. Thus, for an item i, described by a 10-dimensional (decomposed) word vector $\boldsymbol{x}_i$, the property score was $y_i = \sum \boldsymbol{x}_i \cdot \boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^*$ was the random weight vector. The weight vectors were all sampled from a multivariate standard normal distribution and different experiments used different random samples. In Experiment 1a, only one randomly generated weight vector was used. In Experiment 1b, we generated four random weight vectors and each participant was assigned to one of the weight vectors (see Tables A1 and A2 for example items with high and low scores in different experiments). To avoid unnecessary cognitive load for participants, we scaled all property scores such that the values ranged from −100 to +100 and rounded scores to the nearest integer for display. As these properties were generated using random functions on the word vectors, they do not necessarily have concise descriptions consisting of one word (e.g., tasty) or even a short phrase (e.g., American dinner foods).

The reason we used random linear functions on the decomposed word vectors was to ensure that our artificial properties tracked important and discriminating dimensions of foods and animals. We had initially tried to specify random functions on the unreduced 300-dimensional vectors but found that they did not generate properties on which our objects varied greatly. By contrast, the principal components decomposition extracts the dimensions on which the objects in a given domain (foods or animals) have the greatest variance, and random functions on these dimensions subsequently generate properties with systematic and interpretable differences across different objects in the domain. It is worth noting that a linear function on the PCA-reduced vectors is also a linear function on the original unreduced word vector space.

Procedures

Our task consists of three phases (Figure 1A). Before the active learning task, participants participated in a practice phase, where they were presented five items and the corresponding property scores, one at a time, in a random order. Each item-score pair appeared onscreen for 3 s. To aid participant learning and encourage sampling of a diverse range of items, we randomly sampled five items 100,000 times and selected the set with the highest Bayesian D-optimality (see our discussion of D-optimality below). This ensured that the practice items would expose participants to a reasonable span of the food space or animal space. To ensure that the practice items were sufficiently familiar to participants, before sampling practice items, we removed items occurring less than 40 times in the SUBTLEX-US corpus of American English subtitles (Brysbaert & New, 2009).

In the active learning phase, the participants' task was to enter the food or animal items of their own choice in a textbox. The entered word/phrase was passed to the predefined property dictionary, and the corresponding score was returned and appeared onscreen for 3 s. In case the entry was not in our dictionary, we encouraged participants to check for spelling errors or alternate spelling forms and reenter the word/phrase, or enter another item until the item was accepted. Participants were asked to enter a total of 20 items.

Following the active learning phase, we tested participants' learning performance with 20 preselected food or animal items, presented in a random order. We selected 20 test foods and 20 test animals using the same procedure as for the practice items. For each item, participants were asked to indicate their best guess of its property score using a slider from −100 to +100. To aid participants' test performance, we displayed the scores of the 20 items they entered in the active learning phase throughout the test phase. No feedback was provided in the test phase. Practice and test items can be found in Table 1.

After completing the experiment, participants were given a base payment of $2 and were given a bonus of $1.00 if their test performance (measured by root-mean-square error [RMSE]) was in the top 10%, and $0.50 if they were in the top 50%.

Table 1
Practice Items and Test Items, for Experiments 1a, 1b, 2, and 3

Experiments 1a, 2, and 3 (food)		Experiment 1b (animals)	
Practice items	Test items	Practice items	Test items
Vinegar	Candy	Turkey	Mammoth
Margarita	Cream	Wolf	Trout
Trout	Macaroni	Camel	Sheep
Muffin	Caramel	Moth	Buzzard
Noodle	Tomato	Alligator	Ox
	Dumplings		Mare
	Oregano		Chicken
	Fish		Parrot
	Coconut		Crocodile
	Tongue		Pony
	Chardonnay		Badger
	Guacamole		Snake
	Sugar		Coyote
	Egg		Sardine
	Eggplant		Rhino
	Alcohol		Insect
	Grape		Doggie
	Pumpkin		Lamb
	Corn		Shellfish
	Bumbo		Frog

Memory Search Model

We modeled the memory search as a Markov information acquisition process, where the probabilities of searching each of the candidate queries were only dependent on the similarity between the candidate queries and the most recently searched query, as well as the word frequency of the candidate queries themselves. In line with classic memory retrieval models, we assumed that participants searched for a word among all candidate queries in the memory space S. Assuming the Markov property, the model predicted the switch from one query s_{t-1} to another query s_t using transition probabilities $\Pr[s_t|s_{t-1}]$, where $s_{t-1}, s_t \in S$ and $t \in \{2, 3, \ldots, T\}$ are the time steps. We allowed $\Pr[s_t|s_{t-1}]$ to be a function of two key cognitive mechanisms—semantic congruence and word frequency—giving us:

$$\Pr[s_t|s_{t-1}] = \sigma(\beta_1 \mathrm{sim}_{s_{t-1},s_t} + \beta_2 \mathrm{freq}_{s_t}), \tag{1}$$

where $\mathrm{sim}_{s_{t-1},s_t}$ is the cosine similarity between s_{t-1} and s_t, freq_{s_t} represents the frequency (log-transformed) of candidate query s_t, and $\sigma(\cdot)$ is the softmax function that sets $\sum_{s_t \in S} \Pr[s_t|s_{t-1}] = 1$.

We fit the memory search models under a hierarchical Bayesian framework, which provided both group- and individual-level estimation of β_1 and β_2. The hierarchical Bayesian model fitting was carried out in *rstan*, an R implementation of Stan (Stan Development Team, 2021). The group-level grand means μ_k were set at the standard normal distribution, and the individual-level degree of deviation from the grand mean σ_k was set to follow a half-Cauchy distribution (with location = 0 and scale = 5). On the individual level, the model allowed each participant's parameters to deviate from the grand mean with different size $\delta_{k,j}$ (drawn from a prior standard normal distribution), resulting in an individual-level parameter $\beta_{k,j} = \mu_k + \sigma_k \delta_{k,j}$, where j indexes participant ID and k denotes different parameters in the memory search models. All group- and individual-level parameters were estimated simultaneously via fitting the individual-level memory search data. To check the convergence of Markov chain Monte Carlo (MCMC), we ran five independent chains for each fit and estimated an R-hat for the convergence check (Gelman & Rubin, 1992). Each of the five chains contained 2,000 iterations after 1,000 warmup samples, totaling 10,000 formal samples for each fit. All $\hat{R}$ values were below 1.05, indicating excellent convergence of the MCMC simulations.

Ideal Bayesian Learning Model

Participants learned about the target property from the scores given as feedback to queries. To formally capture this dynamic learning process, we assumed that the participants were ideal Bayesian learners who took as input the scores y_t of their query (quantified by a word vector $\boldsymbol{x}_t$) and learned a linear mapping between them (Figure 1C). Participants updated their belief of weights $\boldsymbol{\theta}$ that determined the linear mapping after observing each pair of $\boldsymbol{x}_t$ and y_t using the Bayes rule: $p_t(\boldsymbol{\theta}|\boldsymbol{x}_t, y_t) = \frac{p_t(\boldsymbol{\theta})p(y_t|\boldsymbol{x}_t, \boldsymbol{\theta})}{\int p_t(\boldsymbol{\theta})p(y_t|\boldsymbol{x}_t, \boldsymbol{\theta})}$, where $\boldsymbol{x}_t$ is an 11-length vector, a concatenation of a 10-dimensional word vector principal components and the constant 1. As in a typical (Bayesian) linear regression, we allowed for a nonzero intercept in the Bayesian learning model and thus $\boldsymbol{\theta}$ was an 11-dimensional vector, that is, a concatenation of the 10-dimensional weight (on the word vectors) and an intercept (note that our algorithm for generating the property scores assumed an intercept of zero, but that was unknown to the participants). At onset, the prior belief of $\boldsymbol{\theta}$ was set as a Gaussian distribution centered at 0 with unit standard deviation.

In the test phase, we assumed that participants made predictions on the test items based on the updated posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})$, where $\boldsymbol{X}$ is the design matrix corresponding to their 20 queries in the active learning phase and $\boldsymbol{y}$ are the corresponding scores. The predicted scores at test can be written as $\widehat{y^*} = \int p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})\boldsymbol{\theta}\boldsymbol{x}^* d\boldsymbol{\theta}$, where $\boldsymbol{x}^*$ is the vector corresponding to the test item. A nice property of the Bayesian learning model's predictions at test was that its predictions were not sensitive to the order in which queries were generated.

To be consistent with the score generation process, we used the original scores (rather than the rescaled scores in the range [−100, 100] that were displayed to participants) in the Bayesian learning model. Accordingly, in evaluating participants' predictive accuracy using Pearson's R and RMSE, we back-transformed their prediction scores at test using the same set of scaling factors, to be compatible with the original scores.

Bayesian D-Optimality of Queries

The ideal Bayesian learning model also allowed us to evaluate query efficiency with Bayesian D-optimality, one of the most often used optimality criteria (Chaloner & Verdinelli, 1995; Kiefer & Wolfowitz, 1959; Myung & Pitt, 2009). Mathematically, Bayesian D-optimality of the full set of queried items is the determinant of the Fisher information matrix $D = \det\{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \Sigma^{-1}\}$, where $\boldsymbol{X}$ is the 11×20 design matrix corresponding to the 20 queried items, and Σ is the 11×11 covariance matrix before querying the items. At the beginning of the experiment, Σ is set as an identity matrix, corresponding to the standard multivariate normal prior distribution on $\boldsymbol{\theta}$. Intuitively, if the queried items are sparsely distributed in the space, the design matrix typically has a high Bayesian D-optimality. By contrast, if the queried items are close to one another, the design is likely to have a low Bayesian D-optimality.

Note that it is also possible to evaluate single-query informativeness at each time step in the learning phase using Bayesian D-optimality. This measure could be included in the memory search model, in addition to the two predictors in Equation 1. However, the more parsimonious memory search model in Equation 1 outperformed an extended model with Bayesian D-optimality as an additional predictor in memory with decisive Bayes factors (BF; Experiment 1a: BF = 5.58×10^8; Experiment 1b: BF = 8.89×10^8), due to the collinearity between Bayesian D-optimality and semantic congruence. Therefore, we report the main memory search model in Equation 1.

We also attempted to compare our D-optimality measure with a measure of optimality for Gaussian process function learning, the latter of which allows very flexible nonlinear relationships between predictor and response variables (Hayes et al., 2019; Lucas et al., 2015; Wu et al., 2018). Gaussian process function learning does not assume any specific functional form. Instead, it merely assumes that items nearby in the word vector space (i.e., X) should score similarly on the target property (i.e., y). Fortunately, information-theoretic tools allowed us to measure the query optimality for Gaussian process function learning despite its flexibility. Specifically, we measured the design entropy for Gaussian process function learning (see Gramacy, 2020, p. 224). The higher the design entropy, the more optimal the query sequence for Gaussian process function learning. We found that D-optimality and Gaussian process design entropy were highly correlated in our datasets, with Spearman correlations above

Table 2
Summary of Hierarchical Bayesian Estimation of Memory Model Parameters

Experiment	Group-level *M* and 95% CI		Percentage of individual-level 95% CI beyond 0	
	Semantic congruence (β_1)	Word frequency (β_2)	Semantic congruence (β_1)	Word frequency (β_2)
Experiment 1a	0.62 [0.57, 0.67]	0.97 [0.93, 1.02]	93%	99%
Experiment 1b	0.71 [0.66, 0.76]	0.93 [0.90, 0.97]	98%	100%
Experiment 2	0.68 [0.63, 0.74]	0.86 [0.80, 0.91]	100%	100%
Experiment 3	0.42 [0.36, 0.50]	0.87 [0.80, 0.94]	67%	98%

Note. CI = confidence interval.

.8 for all experiments (see Figure A1). The implication is that in our data, queries that are suboptimal with D-optimality are also suboptimal with Gaussian process entropy and that replacing D-optimality with Gaussian process entropy in our models is unlikely to alter any results. This shows that our basic findings are largely robust to attributing different inductive biases to participants.

Transparency and Openness

All experimental materials, data, and code can be found at the Open Science Framework (OSF; He et al., 2023).

Results

Search Sequences Display Semantic Congruence

Participants tended to query items that were semantically related to previously recalled items. Across Experiments 1a and 1b, *t* tests revealed that participants' average cosine similarities between consecutive queries exceeded the mean similarity among all possible transitions (all $t > 59$ and $p < 10^{-10}$ for foods and animals). At the individual level, all of the 198 participants in Experiment 1a and the 198 participants in Experiment 1b displayed semantic congruence higher than the random level, and 94% and 99% of participants in Experiments 1a and 1b, respectively, reached the conventional significance threshold of $p < .05$.

Participants also tended to query items that were encountered and discussed frequently in the world, like apple or egg, more often than rare (at least to the typical participant in our samples of U.S. residents) items like medlar or pawpaw (both types of fruit). We compared the average (log) word frequency of each participant's queries to that of all possible foods (Experiment 1a) or animals (Experiment 1b). All of the 198 participants in Experiment 1a and the 198 participants in Experiment 1b in Experiment 1b displayed the tendency of querying high-frequency items, and 99% and 100% of participants in Experiments 1a and 1b, respectively, reached the significance threshold of $p < .05$.

Memory Search Model Fits

As mentioned above, we used a computational memory search model to formally capture memory retrieval dynamics in the active learning phase. Hierarchical Bayesian model fitting of the model provided both group- and individual-level estimation of β_1 and β_2, allowing us to test the extent to which the memory-based active learning process was determined by similarity with the previous item and word frequency (Table 2). On the group level, cosine similarity had a strong effect on sequential memory search for food items in Experiment 1a and for animals in Experiment 1b, as indicated by a positive value of β_1. Likewise, frequent words/phrases were much more likely to be queried than infrequent ones for both foods in Experiment 1a and animals in Experiment 1b, as indicated by a positive value of β_2. The individual-level estimation also suggested that a majority of our participants displayed these tendencies.

Individual-level parameters correlated in the expected ways with the model-free measures of memory properties: β_1 almost perfectly correlated with adjacent semantic similarity (the average cosine similarity of successively retrieved items; Figure 2A), and β_2 perfectly correlated with (log-transformed) word frequency (Figure 2B). These results serve as a useful sanity check and suggest that the underlying cognitive processes in our naturalistic active learning task resembled those underlying a typical memory task, processes that likely lead to suboptimal queries.

Bayesian Learning Model Predicts Test Performance

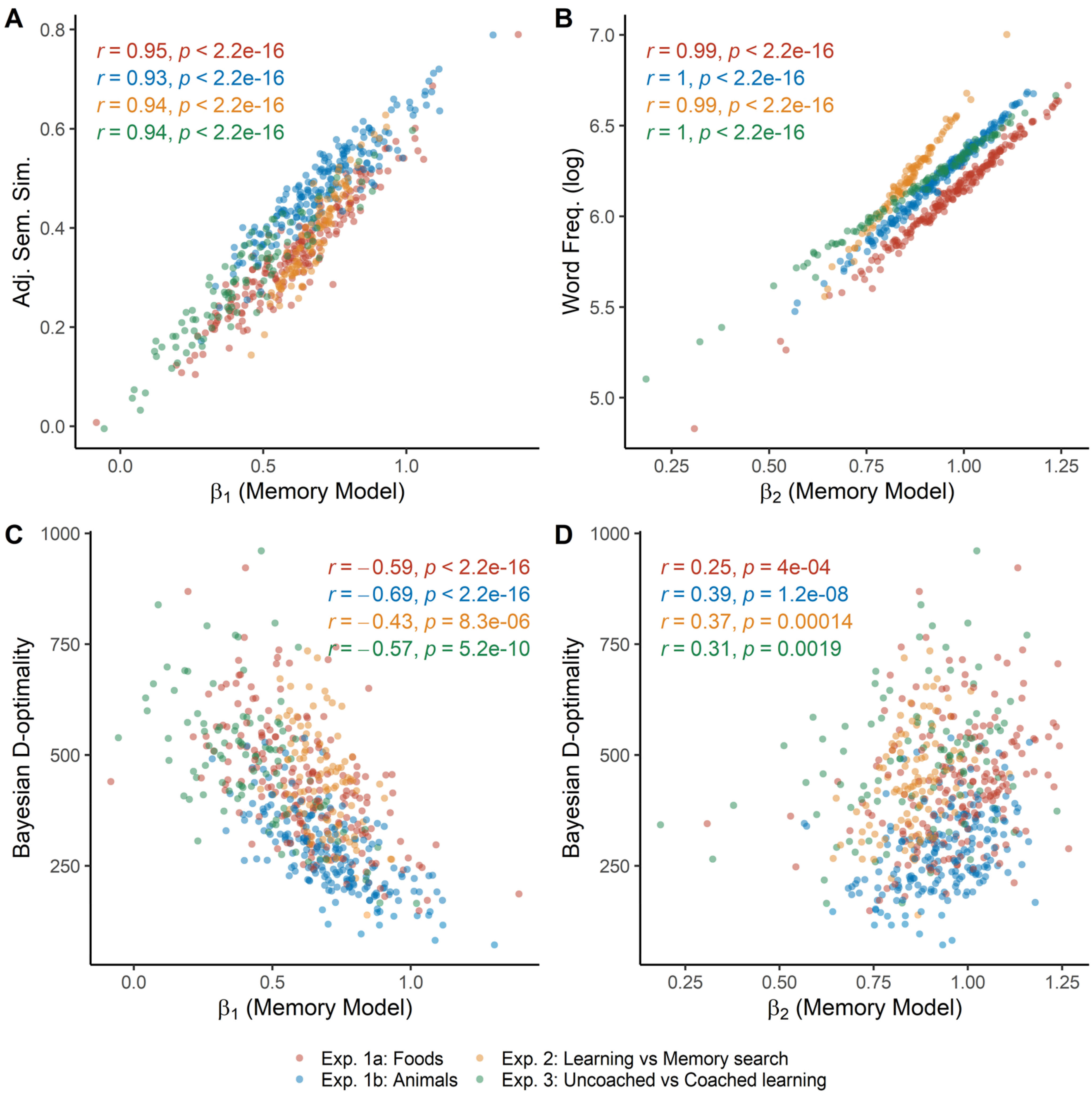
The ideal Bayesian learning model captured the between-participant variation in test performance (Figure 3). We found that the predicted performance at test by the Bayesian learning model trained on each participant's queries was correlated with the actual performance of the participants measured by both Pearson's *R* (Experiment 1a: $r = .418$, $p < 10^{-9}$, 95% CI [0.295, 0.526]; Experiment 1b: $r = .236$, $p < .001$, 95% CI [0.100, 0.363]) and RMSE (Experiment 1a: $r = .219$, $p = .002$, 95% CI [0.082, 0.358]; Experiment 1b: $r = .166$, $p = .019$, 95% CI [0.028, 0.299]). The results suggest that the ideal Bayesian learning model was able to at least partly capture the heterogeneity in test performance across participants.

Note that participants' actual test performance was significantly better than the baseline model that assigned random scores at test (baseline predictions were generated by randomly shuffling participant responses at test, and comparing these shuffled responses with the true test scores 100 times per participant). This indicates that despite their suboptimality in query generation, participants were able to use the observed data in the training phase to predict property scores at test. However, actual test performance did not reach the accuracy levels predicted by the ideal Bayesian learning model ($ps < 10^{-12}$ in all experiments, Figure 3). This could be because of additional sources of noise during the test phase. Although the ideal Bayesian learner is noiseless, participants themselves could have made random errors when giving ratings.

Semantic Congruence Hinders Optimal Search and Subsequent Learning

We correlated the Bayesian D-optimality of each participant's query sequence with the estimated degree of semantic congruence, β_1, in the

Figure 2
Memory Retrieval in Naturalistic Active Learning



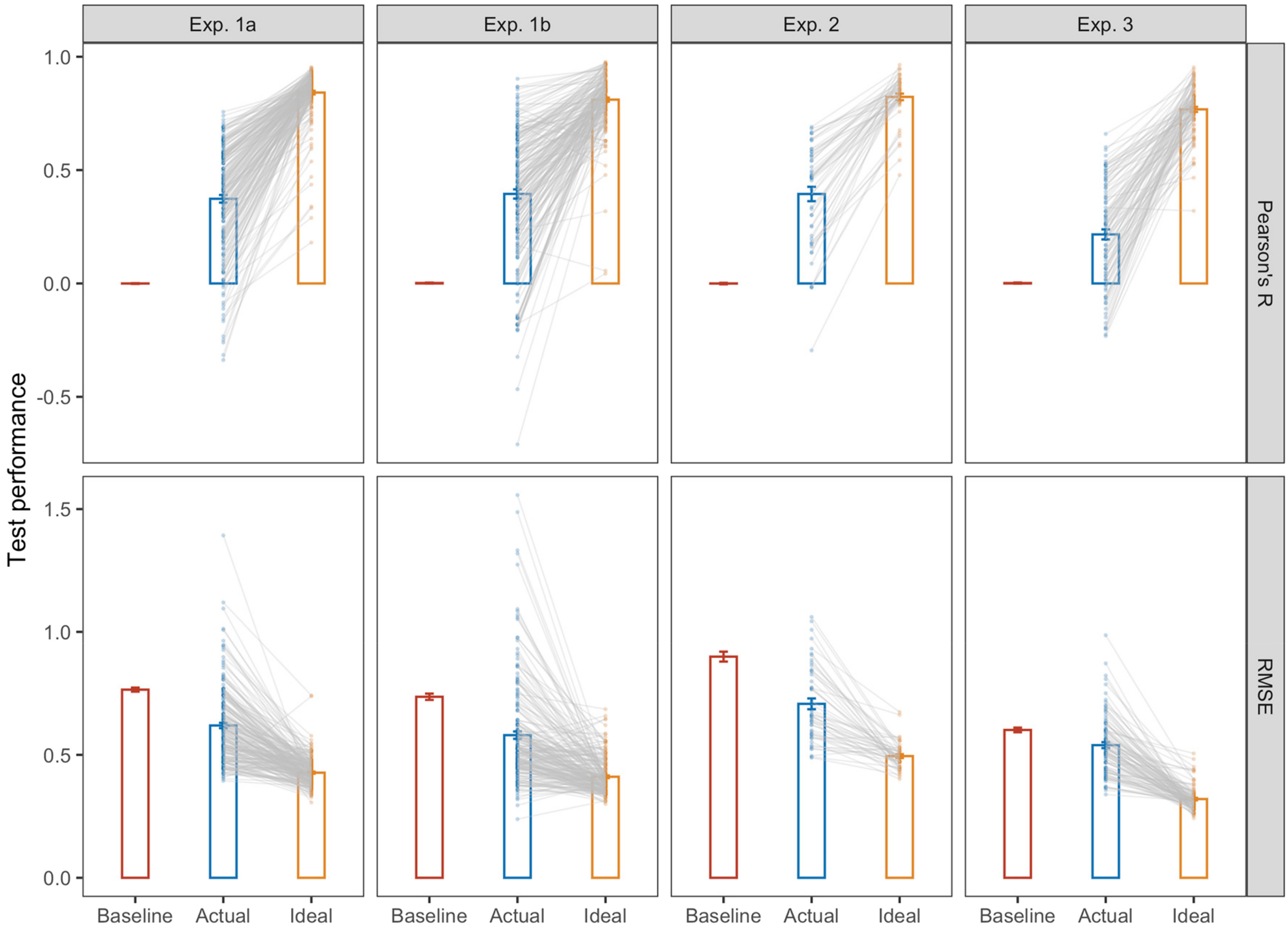
Note. (A) Correlations between adjacent semantic similarity and β_1 in the memory model. (B) Correlations between the (log-transformed) word frequency of retrieved items and β_2 in the model. (C) Correlations between β_1 and Bayesian D-optimality. (D) Correlations between β_2 and Bayesian D-optimality. This analysis collapsed all the participant-generated query sequences from different experimental conditions in Experiment 2 ($N = 102$) and Experiment 3 ($N = 100$). See the online article for the color version of this figure.

memory search model (Figure 2C). As the two variables entail a strong tradeoff, participants whose queries are more positively influenced by the similarity with previous items have lower levels of Bayesian D-optimality; these correlations are expectedly negative. Interestingly, Bayesian D-optimality was positively correlated with β_2 (Figure 2D). The reason for this is that words with higher frequencies have larger vector norms in many DSMs (Kintsch, 2014; Wilson & Schakel, 2015). Querying words with large vector norms, in turn, tends to increase the volume of the space spanned by the set of queries, and volume is very closely related to Bayesian D-optimality.

Semantic incongruence was also associated with better test performance as predicted by the ideal Bayesian learning model. We found

Figure 3

Baseline, Actual and Ideal Test Performance Across Experiments, Measured With Pearson's R and RMSE, Respectively



Note. Each point in the graphs corresponds to a single participant, and connected actual and ideal points show the relationship between participants' actual performance and the performance of a Bayesian learner trained on their queries. The baseline performance is that of a model that guesses test scores randomly. Experiment 2 used the active learning condition in this analysis ($N = 50$). Experiment 3 used both the coached learning and uncoached learning conditions ($N = 100$). RMSE = root-mean-square error. See the online article for the color version of this figure.

that β_1 was positively correlated with the ideal Bayesian learning model's predicted RMSE at test (Experiment 1a: $r = .318$, $p < 10^{-5}$, 95% CI [0.187, 0.438]; Experiment 1b: $r = .381$, $p < 10^{-7}$, 95% CI [0.255, 0.494]) and negatively with its predicted Pearson's R at test (Experiment 1a: $r = -.115$, $p = .105$, 95% CI [−0.251, 0.024]; Experiment 1b: $r = -.222$, $p = .002$, 95% CI [−0.350, −0.085]), though the Pearson's R correlation for Experiment 1a does not cross the threshold for significance.

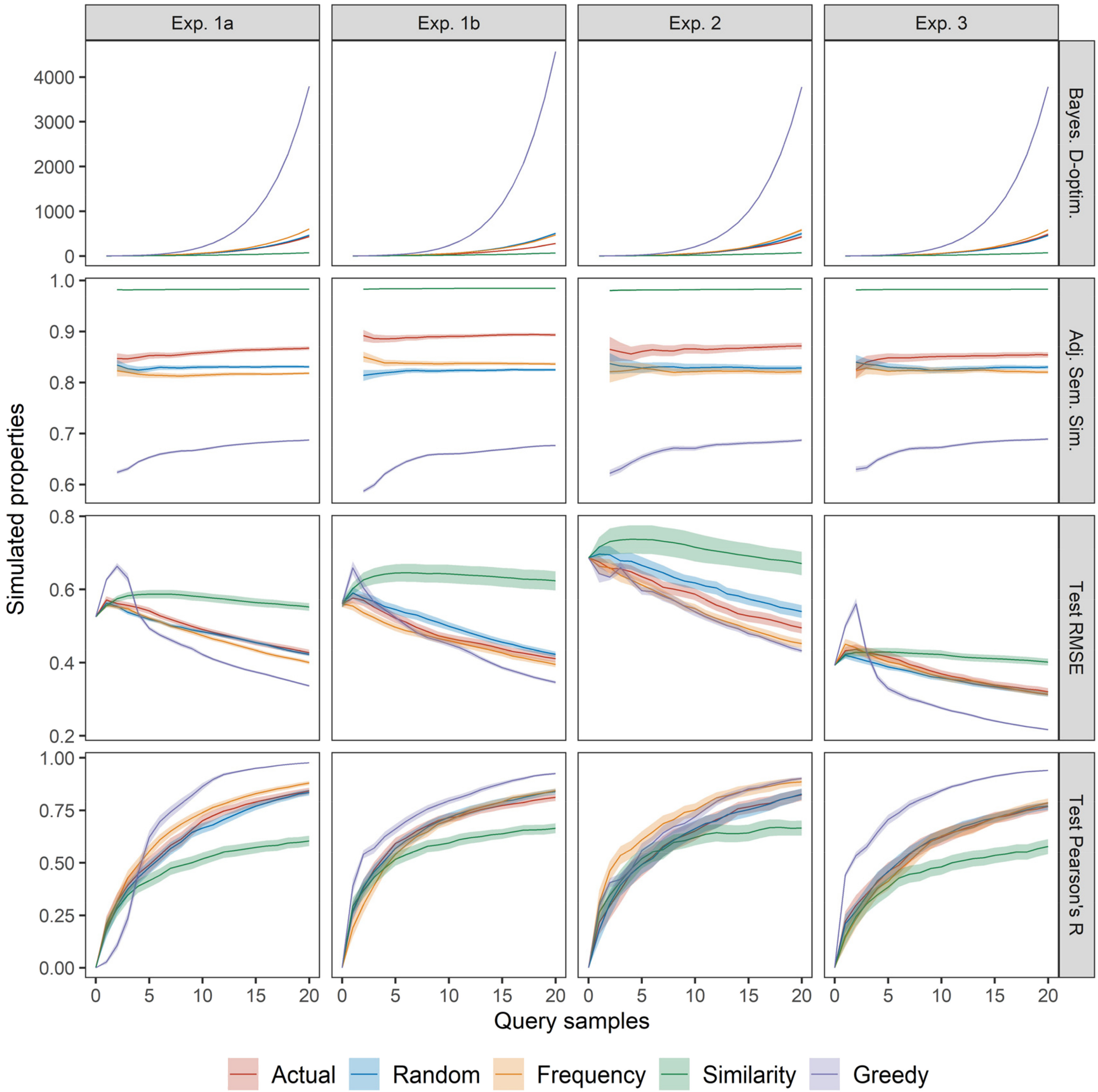
Additionally, participants who displayed more semantic incongruence also performed better at test. β_1 was positively correlated with participants' actual RMSE at test in both Experiment 1a ($r = .155$, $p = .029$, 95% CI [0.016, 0.288]) and Experiment 1b ($r = .137$, $p = .054$, 95% CI [−0.002, 0.271]. Likewise, β_1 was negatively correlated with participants' actual Pearson's R at test in both Experiment 1a ($r = -.078$, $p = .276$, 95% CI [−0.215, 0.062]) and Experiment 1b ($r = -.182$, $p = .010$, 95% CI [−0.314, −0.044]). Again, the RMSE correlation for Experiment 1b and the Pearson's R correlation for Experiment 1a do not cross the threshold for significance.

Simulated Query Strategies and Test Performance

To more systematically examine the tradeoff between semantic congruence in memory retrieval and optimal search in active learning, we simulated a semantic similarity-based retrieval strategy and compared its Bayesian D-optimality, as well as the predicted test performance by the ideal Bayesian learning model, with that of a number of other retrieval strategies (Figure 4). The semantic similarity-based strategy produced the lowest query Bayesian D-optimality and achieved the worst test performance among all retrieval strategies. In stark contrast, the D-optimality Greedy strategy that kept selecting from the most informative queries according to the Bayesian D-optimality criterion always produced the lowest semantic congruence in queries and the best performance at test. Retrieval strategies that were searched based on word frequency or based on random sampling achieved intermediate D-optimality and accuracy rates at test.

The participants' actual query sequences were far from optimal, as compared with the D-optimality Greedy strategy (Figure 4). They

Figure 4
Properties of Actual and Simulated Memory Query Strategies



Note. Actual query uses the participants' actual sequences of queries. Random query randomly selects from all the items with no replacement. Frequency-based query randomly selects from the top-100 most frequent items with no replacement (being equivalent to selecting one from five high-frequency items at each query step). Similarity-based query keeps randomly selecting one from the top five most cosine-similar items that has not been queried before. Greedy query keeps randomly selecting one from the top five items with the highest Bayesian D-optimality. The top two panels indicate the Bayesian D-optimality and adjacent semantic similarity of simulated queries, whereas the bottom two panels indicate the test performance (RMSE and Pearson's R) of an ideal Bayesian learning model that makes predictions based on simulated queries. The simulation sample size equals the sample size in each experiment, respectively ($N = 198, 198, 50, 100$ for Experiments 1a, 1b, 2, and 3, respectively). The shaded bands are 95% confidence intervals. RMSE = root-mean-square error. See the online article for the color version of this figure.

also displayed more semantic congruence and achieved lower Bayesian D-optimality, than the random or frequency-based queries. Overall, these results, once again, suggest that participants were unable to ask the most informative questions when the questions had to be generated from memory.

Discussion

We examined naturalistic active learning for foods (Experiment 1a) and animals (Experiment 1b). Our paradigm allowed us to observe and model participant query generation over a vast domain involving over a thousand potential items. In line with prior research on memory search, we found strong evidence for semantic congruence, that is, retrieved items tended to cue the subsequent retrieval of semantically similar (rather than dissimilar) items. This result was further validated by the fits of our Markov memory model, which showed a strong positive effect of semantic similarity (as well as an effect of word frequency) on recall. Furthermore, as expected, participant-level semantic-congruence effects were negatively correlated with Bayesian D-optimality scores, as well as performance at test. Finally, a formal comparison of various memory search strategies showed that observed search processes were much worse than optimal.

Experiment 2

Experiment 1 showed that participants display semantic congruence (rather than semantic incongruence) in naturalistic active learning, which hinders learning. In Experiment 2, we wished to investigate the magnitude of this effect by comparing the extent of semantic congruence in our task with a standard semantic memory search task in which participants were asked to retrieve all the items they could from a category without any overarching learning goal. This allowed us to examine the degree to which participants were able to modulate memory search and prioritize the retrieval of dissimilar items when given an active learning task. Experiment 2 also used a different set of property scores to Experiment 1 and thus tested the robustness of the findings in Experiment 1. This experiment was preregistered on AsPredicted.org: https://aspredicted.org/IVD_VYF. In our preregistration, we planned to include effects of both semantic congruence and D-optimality in our memory model and hypothesized that both conditions would show effects of semantic congruence, while only the active learning condition would show an effect of D-optimality. However, due to high collinearity between semantic congruence and D-optimality (Figure 2C), we ultimately dropped D-optimality from the memory model. This omission should not affect our ability to test Experiment 2's core hypothesis, that participants would be able to modulate search when retrieving items from memory for the purpose of (efficient) active learning, as compared to memory retrieval without a goal.

Method

Participants

One hundred and two U.S. residents fluent in English were recruited from Prolific Academic in the same way as in previous experiments ($M_{\text{age}} = 32$ years, $SD_{\text{age}} = 13$ years; 60% female, 40% male). They were randomly assigned to an active learning condition ($N = 50$) or to a semantic memory search ($N = 52$) condition.

Stimuli

The stimuli were the same as in Experiment 1a. However, a new random linear function was used to derive property scores. The list of high- and low-score items can be found in Table A1.

Procedures

The procedure and instructions for the active learning condition were identical to those of Experiment 1a, except that there was no practice phase. As in Experiments 1a and 1b, we displayed the 20 items they entered in the active learning phase and their corresponding scores throughout the test phase to avoid the potential memory confound in the test phase. In the semantic memory search condition, participants were simply instructed to recall 20 foods, with no testing phase (and, again, no practice phase).

Memory Search Model

We used the same memory search model as in Experiments 1a and 1b to fit the memory retrieval data. The hierarchical Bayesian model fitting collapsed the query sequences in both conditions.

Ideal Bayesian Learning Model

In the active learning condition, we applied the same ideal Bayesian learning model as in Experiment 1a to predict the participants' reported property scores at test, based on their queries at the active learning phase.

Results

Replicating Experiment 1a, participants in Experiment 2 displayed strong semantic congruence in sequential search from memory (see Table 2 for a summary). On the individual level, our memory search model was able to capture the effects almost perfectly (Figure 2A), and semantic congruence in memory search was strongly correlated with the Bayesian D-optimality of the queries (Figure 2C). Here, we analyzed and collapsed the query sequences in both the active learning and semantic memory search conditions. Although there was no learning taking place in the semantic memory search condition, we calculated the Bayesian D-optimality of query sequences in that condition as if participants were to learn a property as those in the active learning condition. This "as-if" Bayesian D-optimality served as a benchmark for evaluating the query optimality in the active learning condition.

In the active learning condition, the Bayesian learning model trained on the participants' memory queries was able to capture the individual differences in test performance measured either by Pearson's R ($r = .528$, $p < .0001$, 95% CI [0.293, 0.703]) or RMSE ($r = .157$, $p = .277$, 95% CI [−0.081, 0.302]) although the RMSE measure did not reach the conventional significance threshold. The degree of semantic incongruence during item querying from memory, β_1, was associated with the test performance in the predicted directions (i.e., negatively correlated with test Pearson's R, and positively correlated with test RMSE), although statistical significance did not reach the conventional threshold (all ps $> .1$). These nonsignificant results were likely due to relatively small sample size (with $N = 50$).

In a direct comparison with the simulated D-optimality Greedy retrieval strategy that kept selecting from the most informative queries, Figure 4 shows that the actual query sequences in Experiment 2 were far from optimal. Consistent with Experiments 1a and 1b, participants in Experiment 2 also displayed more semantic congruence and achieved lower Bayesian D-optimality, than the random or frequency-based queries.

Our main interest in Experiment 2 was to evaluate and compare query properties in the active learning and the semantic memory search conditions. In our preregistration, we predicted that semantic congruence would be lower in the active learning condition; however, we found that the two conditions differed neither on semantic congruence (i.e., β_1), $t(96.7) = 0.818$, $p = .416$, 95% CI [−0.052, 0.022], nor on Bayesian D-optimality of the queries, $t(99.8) = 0.719$, $p = .474$, 95% CI [−62.2, 29.1] (see Figure 5). These findings suggest that semantically congruent memory search was so strong that it cannot be moved by the demand of efficient learning. Neither did participants in the two conditions differ on the search of high-frequency items (i.e., β_2), $t(99.1) = 0.034$, $p = .973$, 95% CI [−0.032, 0.031].

Discussion

Experiment 2 replicated the core results of Experiment 1a. As in Experiment 1a, participants displayed semantic congruence in retrieval, which hindered their ability to learn the property scores in our task. More importantly, contrary to our preregistered hypotheses, Experiment 2 showed that the extent of semantic congruence in our active learning task was identical to that in a simple semantic memory search task without any learning objective. This indicates that associative biases in memory are so strong that they completely hinder participants' ability to actively learn. In other words, retrieval strategies are not only suboptimal, they are no better than retrieval strategies in the absence of any learning goal.

Experiment 3

Why do participants generate suboptimal queries? Is that they are unable to modulate associative memory mechanisms, which prioritize the retrieval of semantically similar, rather than dissimilar, items? Or, do they just not know how to query efficiently in our active learning task? In Experiment 3, we examined the boundaries of the above findings by testing whether explicit task instructions could improve search efficiency for active learning. In particular, we contrasted our standard active learning condition with a new active learning condition in which we coached participants to query more optimally by sampling dissimilar items. This experiment was preregistered on AsPredicted.org: https://aspredicted.org/blind.php?x=9nr57q. Similar to Experiment 2, our preregistration included a plan to include effects of both semantic congruence and D-optimality in our memory model and hypothesized that participants in the coached condition would display weaker effects of semantic congruence, and possibly stronger effects of D-optimality. Again, we dropped D-optimality from modeling due to collinearity with semantic congruence, but this does not alter our ability to test the hypothesis that coaching should modulate search strategy in a way that is beneficial for learning, relative to the standard active learning condition without such coaching. Our preregistration also included the hypothesis that the coached condition would perform better in the test phase than the standard active learning condition.

Method

Participants

One hundred participants were recruited in the same way as in previous experiments ($M_{\text{age}} = 33$ years, $SD_{\text{age}} = 11$ years; 67% female, 31% male, 2% other/prefer not to say). They were randomly assigned to an uncoached active learning condition ($N = 58$) or to a coached active learning ($N = 42$) condition.

Stimuli

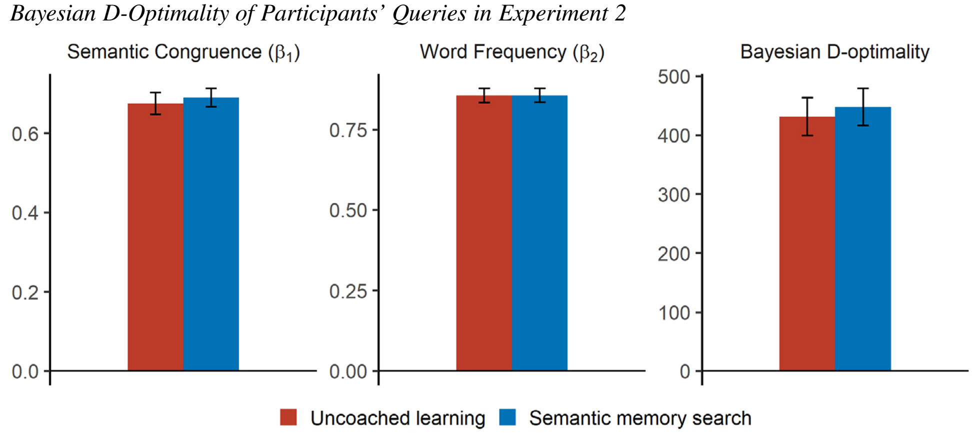
The stimuli were the same as in Experiment 1a, except that a new random linear function was used to derive property scores. The list of high- and low-score items can be found in Table A1.

Procedures

The procedures and instructions for the uncoached condition were identical to those of Experiment 1a. In the coached condition,

Figure 5

Semantic Congruence, Word Frequency (Estimated Individual-Level β_1 and β_2 in Memory Model), and Bayesian D-Optimality of Participants' Queries in Experiment 2



Note. Error bars represent 95% confidence intervals. See the online article for the color version of this figure.

participants were additionally given instructions on how to query efficiently by retrieving dissimilar items. The additional instructions read:

> Please think carefully about the "usefulness" of a particular query before you submit it. Generally speaking, asking about items very similar to items you've already asked about will not tell you much about the property. For example, if you've already learned that "surgeon" has a high score, then if you ask about "doctor," it will probably also have a high score, and you won't have learned much about what the property is. (Or take the extreme case where you ask about "surgeon" twice—you learn nothing the second time you ask about it!) If you already asked about "surgeon," you will learn more if you ask about an item rather different from your previous queries, like "plumber" or "soldier."

Memory Search Model

We used the same memory search model as in Experiments 1a, 1b, and 2 to fit the memory retrieval data in both conditions.

Ideal Bayesian Learning Model

The ideal Bayesian learning model was the same as in Experiments 1a, 1b, and 2.

Results

Replicating Experiment 1a, participants in Experiment 3 displayed strong semantic-congruence effects in search from memory (Table 2), which were captured by our memory model (Figure 2A). Collapsing across conditions, individual-level semantic congruence in memory search was strongly (negatively) correlated with the Bayesian D-optimality of the queries (Figure 2C). Participants also displayed more semantic congruence and achieved lower Bayesian D-optimality than the random or frequency-based queries (Figure 4). Additionally, semantic congruence in memory querying (i.e., β_1) was associated with test performance as predicted by the ideal Bayesian learning model, as well as the actual test performance. Collapsing across conditions, β_1 was negatively correlated with predicted test Pearson's R ($r = -.233$, $p = .020$, 95% CI [−0.410, −0.038]), and positively correlated with predicted test RMSE ($r = .302$, $p = .002$, 95% CI [0.112, 0.470]). Across participants, β_1 was also negatively correlated with actual test Pearson's R ($r = -.264$, $p = .007$, 95% CI [−0.438, −0.071]) and positively correlated with actual test RMSE ($r = .297$, $p = .003$, 95% CI [0.106, 0.466]). Overall, semantic incongruence in memory querying led to better performance at test.

Our primary goal was to test our hypothesis that the "usefulness" instruction in the coached learning condition should help overcome the semantic-congruence bottleneck and improve query informativeness, compared with the uncoached learning condition. As Figure 6 shows, the uncoached learning and coached learning conditions differed neither on semantic congruence (i.e., β_1), $t(79.9) = 1.214$, $p = .229$, 95% CI [−0.136, 0.033], nor on word frequency (i.e., β_2), $t(86.4) = 0.323$, $p = .747$, 95% CI [−0.063, 0.087]. Overall, the key query properties did not differ in the two conditions. Taking the learning task into consideration, the two conditions did not differ on Bayesian D-optimality, $t(93.0) = 1.018$, $p = .311$, 95% CI [−88.6, 28.6]. Neither did they differ on the actual performance in the test phase, test Pearson's R: $t(88.9) = 1.049$, $p = .297$, 95% CI [−0.042, 0.136]; test RMSE: $t(77.1) = 0.221$, $p = .826$, 95% CI [−0.056, 0.045]. This contradicted our preregistered predictions.

Discussion

Experiment 3 aimed to examine the boundaries of the effect documented in this article. In particular, it tested whether participants could be coached to alter their retrieval strategies, thereby improving performance at test. Although Experiment 3 replicated our findings in Experiment 1, contrary to our preregistered hypotheses, we found that coaching did not change the degree of semantic congruence in search and likewise did not have an effect on test scores. It is important to note that our uncoached condition in Experiment 3 is nearly identical to the active learning condition in Experiment 2 (except for the underlying function that needs to be learned), which is directly contrasted with a semantic search condition in that experiment. Putting together these two experiments, we can conclude that the degree of semantic clustering in a simple semantic search task is similar to that in an uncoached active learning task which is similar to that in a coached active learning task. In other words, coaching participants in an active learning task does not alter their behavior relative to a semantic memory search task.

One possible conclusion from this result is that the associative memory biases that are responsible for the effects shown in this article are remarkably stubborn and difficult to modulate. That said, it could also be the case that the search strategies in our coached condition (in which participants are encouraged to search for dissimilar items) are identical to the ones they would have used in the uncoached condition and in the pure semantic memory search condition of Experiment 2. This implies that the null effect of coaching could be because of the specific strategies that were coached, and that other types of coaching or experience could alter search strategies and change the results of our experiment.

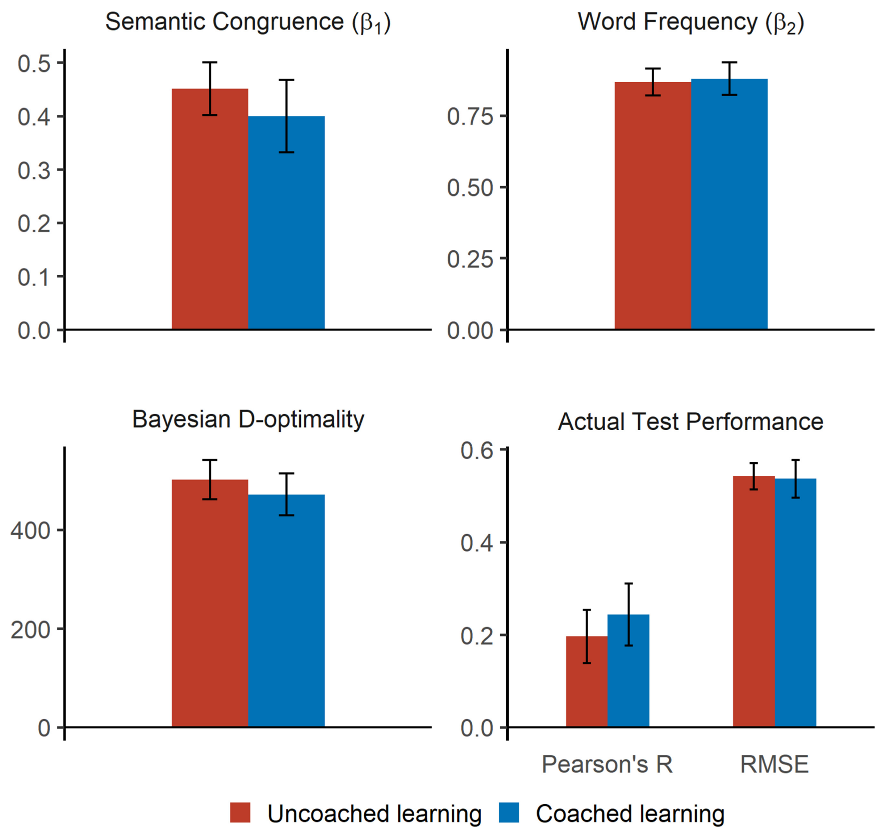
Overview of Experiments 4 and 5

In Experiment 1, we showed that associative memory biases place critical restrictions on participants' ability to search optimally in our naturalistic active learning task. Experiments 2 and 3 further demonstrated that search strategies are insensitive to task goals and do not change with explicit instruction. These results raise an important question: Even though participants may not be able to implement optimal search, can they nonetheless distinguish optimal and suboptimal queries from each other?

Experiments 4 and 5 were designed to answer this question. In Experiment 4, subjects were told to assume they have already queried a particular item (e.g., margarita) and asked which of two other items (e.g., pea or martini) is the optimal query. Thus, Experiment 4 approximated the item-by-item querying structure of Experiments 1–3, but relieved the demands on participants to generate queries from memory, which is subject to semantic congruence. In Experiment 4a, we merely explained the active learning task to subjects before presenting the binary choice task (for foods and animals); in Experiment 4b, subjects actually completed the active learning task (for animals) before completing the binary choice task (for foods). To preview, in both Experiments 4a and 4b, we found subjects were still biased toward selecting suboptimal items, despite the minimal demands on memory search. Thus, we carried

Figure 6

Semantic Congruence, Word Frequency (Estimated Individual-Level β_1 and β_2 in Memory Model) and Bayesian D-Optimality of Participants' Queries and Actual Test Performance in Experiment 3



Note. Error bars represent 95% confidence intervals. See the online article for the color version of this figure.

out Experiment 5, which asked subjects to choose the more optimal of two query sets (of 10 or 20 items). By presenting a choice between two complete query sets that differed greatly in optimality (cf., a choice between two items), we heightened the difference in optimality between choices. As in Experiment 4, participants in Experiment 5a did not complete the active learning task before choosing between query sets; participants in Experiment 5b did. Moreover, Experiment 5a used empirically observed query sets from participants in Experiment 1a, while Experiment 5b used query sets simulated with the similarity-based and optimality-based retrieval strategies described in the Simulated query strategies and test performance section. In Experiments 5a and 5b, subjects could in fact reliably choose the optimal query set. Participant-level accuracy rates across Experiments 4a, 4b, 5a, and 5b are shown in Figure 7. We explain these experiments in more detail now.

Experiment 4a

Method

Participants

Forty-nine participants were recruited from Prolific with the same restrictions as previous experiments ($M_{\text{age}} = 33$ years, $SD_{\text{age}} = 12$ years; 53% female, 40% male, 6% other/prefer not to say).

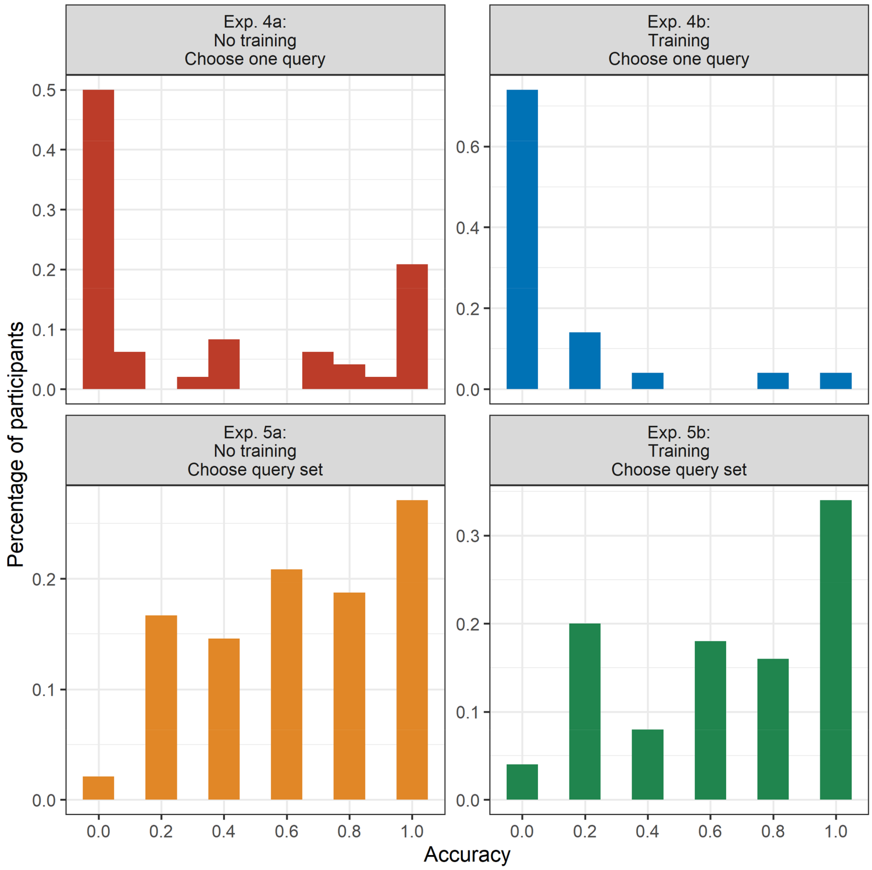
Stimuli

For each category (foods and animals), we first selected the five practice items of Experiments 1–3 (foods: vinegar, margarita, trout, muffin, and noodle; animals: turkey, wolf, camel, moth, alligator). For each practice item, we then found the five most similar and five most dissimilar foods (animals) according to cosine similarity in the 10-dimensional word2vec principal components and selected one highly similar and one highly dissimilar item from among these 10 items (see Table A3 for all [practice item, similar, dissimilar] triples). Item selection was subject to the same frequency threshold used to generate items for the test phase of the active learning conditions in Experiments 1–3 (SUBTLEX-US frequency ≥40). Choice trials were then presented as follows:

> If you knew the property score of margarita and no other foods, which of the following foods and its score would help you learn the property more effectively?
>
> 1. pea and its property score
> 2. martini and its property score

Figure 7

Distribution of Participant-Level Accuracy Rates in Experiments 4a, 4b, 5a, and 5b



Note. See the online article for the color version of this figure.

Procedures

Participants were asked to imagine that they were participating in our active learning task, which we explained in detail using language adapted from Experiment 1a, and told that how well one could learn a property would depend on what items were queried. Participants were asked to judge, given that a particular item was already queried, which of two possible queries would give more information about the property, although they were not told anything about what made a query set more informative. Participants made a choice for a practice pair (occupations), followed by five choice pairs for foods, and five choice pairs for animals. Order of these two groups, order of pairs within a group, and position of the correct answer on the bottom or top were randomized for each participant. Participants were given a base payment of $2.00 and were given a bonus of $0.50 if their performance was in the top 50%.

Results

The top left panel of Figure 7 shows the distribution of participant-level accuracy rates in Experiment 4a. As can be seen, 50% of subjects always chose the suboptimal query, 20% of subjects always chose the optimal query, and the remainder sometimes chose the optimal query and sometimes chose the suboptimal query.

To test whether this preference for suboptimal queries was reliable, we used the R package *rstanarm* (Goodrich et al., 2022) to conduct a Bayesian logistic regression. In particular, we regressed choice (optimal = 1, suboptimal = 0) onto a fixed intercept, and random intercepts for subject and trial, that is, `choice ~ 1 + (1|Subject) + (1|Trial)`. We ran four chains, with 5,000 samples per chain, with the first 2,500 samples of each chain treated as warmup. R-hats for all parameters were close to 1.0, and the number of effective samples was over 1,000 for all parameters.

The effect of interest is the fixed intercept. The 95% credible interval for this effect was below 0 (−4.3, −0.2; $M = -2.2$), suggesting that subjects reliably chose the suboptimal query.

Experiment 4b

Method

Participants

Fifty participants were recruited from Prolific with the same restrictions as previous experiments ($M_{\text{age}} = 38$ years, $SD_{\text{age}} = 15$ years; 58% female, 40% male, 2% nonbinary/prefer not to say).

Stimuli

Active learning stimuli were identical to Experiment 1b (animals). Binary choice stimuli were identical to Experiment 4a (foods).

Procedures

The procedure was identical to Experiment 4a, except that participants first performed the active learning task of Experiment 1b (animals), and then performed the binary choice task of Experiment 4a for just foods. Participants were given a base payment of \$3.00 and were given a bonus of \$1 if their performance in the active learning test phase was in the top 10%, and \$0.50 if their performance was in the top 50%. They received an additional bonus of \$0.50 if their performance was in the top 50% of the binary choice phase.

Results

The top right panel of Figure 7 shows the distribution of participant-level accuracy rates in Experiment 4b. We conducted the same Bayesian logistic regression as in Experiment 4a. R-hat and number of effective samples were similarly satisfactory. Once again, the 95% credible interval for the fixed-effect intercept was below 0 (-5.5, -2.6; $M = -3.9$), suggesting that subjects reliably chose the suboptimal query.

Contrary to what was expected, relative to Experiment 4a, the distribution of accuracy rates in Experiment 4b shifted to the left, reflecting an increased preference for suboptimal queries. This was confirmed by a direct statistical comparison between the two experiments: `choice ~ 1 + experiment + (1|trial) + (1|subject)`, where `experiment = 1` indicates Experiment 4b and `experiment = 0` indicates Experiment 4a. The probability of direction (see Makowski et al., 2019), the percentage of the fixed-effect intercept posterior below or above 0 (whichever is larger than 50%), of `experiment` is 100%, and the 95% credible interval is (-6, -1.2), strong evidence that participants in Experiment 4b made more suboptimal choices than those in Experiment 4a did. Experiment 4b had an active learning task before doing the binary choice task, whereas Experiment 4a did not. It is possible that the extracognitive load or distraction due to the training had led to worse performance in Experiment 4b. It is also possible that the active learning task made the participants accustomed to choosing the most similar items and that tendency carried forward into the binary choice task. Nonetheless, participants in both Experiments 4a and 4b failed to choose the more optimal query in an ideal Bayesian learning manner.

Experiment 5a

Method

Participants

Forty-eight participants were recruited from Prolific with the same restrictions as previous experiments ($M_{\text{age}} = 29$ years, $SD_{\text{age}} = 9$ years; 63% female, 37% male).

Stimuli

The stimuli in Experiment 5a were five pairs of query sets sampled from the participant queries generated in Experiment 1a. Specifically, we selected the five query sets with the highest Bayesian D-optimality, and the five query sets with the lowest Bayesian D-optimality. Each highly optimal query set was paired with one of the least optimal query sets. These pairs are shown in Table A4.

Procedures

Participants were instructed on the nature of the active learning task, as they were in Experiment 4a. Participants were then asked to judge, in a pair of two query sets, which set would give more information about the property, although they were not told anything about what made a query set more informative. To start, participants made a choice for a practice pair, followed by five pairs of query sets sampled from the participant queries generated in Experiment 1a. Participants then made choices for all five pairs of query sets. The position of the correct answer on the left or right was randomized once such that, for a given choice pair, the correct answer was on the same side for all participants. The order of choice pairs was randomized for each participant. Participants were given a base payment of \$1.35 and were given a bonus of \$0.50 if their performance was in the top 50%.

Results

The bottom left panel of Figure 7 shows the distribution of participant-level accuracy rates in Experiment 5a, which skew to the right, tentatively reflecting the overall preference for the optimal query sets. We conducted the same Bayesian logistic regression as in Experiments 4a and 4b. R-hat and the number of effective samples were similarly satisfactory. The probability of direction was 93% (95% credible interval: -0.4, 2.2; $M = 0.9$), notable evidence that subjects chose the optimal query set more often than the suboptimal query set.

Experiment 5b

Method

Participants

Fifty participants were recruited from Prolific with the same restrictions as previous experiments ($M_{\text{age}} = 34$ years, $SD_{\text{age}} = 11$ years; 62% female, 34% male, 4% nonbinary/prefer not to say).

Stimuli

Active learning stimuli were identical to those of Experiment 1b (animals). Stimuli for the binary choice phase were five choice pairs of two simulated 10-item query sets. Within a choice pair, the foods at the beginning of the sequence were identical and were randomly selected. However, for one sequence, subsequent items were chosen by the D-optimality Greedy strategy in simulated query strategies and test performance. For the other sequence, the semantic similarity-based strategy in simulated query strategies and test performance was used. Both strategies searched in the 10-dimensional principal component space, and item selection was subject to the same frequency threshold used to generate test items for the test phase of the active learning conditions of Experiments 1–3 (SUBTLEX-US frequency ≥ 40). These choice pairs are shown in Table A5.

Procedures

Participants first performed the active learning experiment of Experiment 1b (animals) and then performed the binary choice experiment for just foods. Within the binary choice experiment, each pair of 10-item sequences was presented side by side in a table format. The position of the optimal query sequence on the left or right was randomized once such that, for a given choice pair, a sequence was on the same side for all participants. Participants indicated their choices on radio buttons listed vertically. The order of choice pairs was randomized for each participant, as was the position of the radio button for the correct answer on the top or bottom. Participants were given a base payment of $3.00 and were given a bonus of $1 if their performance in the active learning test phase was in the top 10%, and $0.50 if their performance was in the top 50%. They received an additional bonus of $0.50 if their performance was in the top 50% of the binary choice phase.

Results

The bottom right panel of Figure 7 shows the distribution of participant-level accuracy rates in Experiment 5b. We conducted the same Bayesian logistic regression as in Experiments 4a, 4b, and 5a. The R-hat and number of effective samples were similarly satisfactory. The probability of direction was 99% (95% credible interval: [0.2, 1.8]; $M = 1.0$), suggesting that subjects reliably chose the optimal query. In short, Experiments 5a and 5b showed that participants were able to identify the more efficient set of queries when the complete sets of queries, rather than single queries, were presented. That being said, there was a lot of variability across participants (Figure 7). Although the modal participant consistently made optimal choices, a nonnegligible proportion actually chose the suboptimal options more often than the optimal ones.

It can also be seen that, relative to Experiment 5a, participants' preferences have shifted to the right in Experiment 5b, reflecting increased preference for optimal queries. This shift of preferences may be due to a few different reasons. It could be because of the additional training in Experiment 5b, as noted in Figure 7. The simulated query sets may also widen the gap in optimality between the optimal and suboptimal query sets in Experiment 5b, relative to the participant-generated query sets in Experiment 5a. The different domains under examination (foods in Experiment 5a vs. animals in Experiment 5b) may also play a role.

Discussion

Participants in Experiments 1–3, which required generating queries from memory, overwhelmingly made suboptimal queries. Participants in Experiment 4, which presented choices between two experimenter-provided queries, still preferred suboptimal queries, despite reduced demands on memory. Only in Experiment 5, which presented choices between two complete query sets, did participants make optimal choices. We interpret this pattern of findings more in the General Discussion, to which we now turn.

General Discussion

We found that participants failed to generate optimal behavior in our experiments, behavior that maximized information gain by querying sequences of dissimilar items. Rather, associative memory mechanisms led to the successive retrieval of similar items (Experiments 1a and 1b). Additional preregistered experiments showed that, contrary to hypotheses, participants' querying behavior was no more optimal—or less similarity-driven—in our active learning task than a traditional semantic search task (Experiment 2), and no more optimal or less similarity-driven when directly told to query more optimally by querying dissimilar items (Experiment 3), suggesting that optimal active learning may be at the mercy of extremely stubborn memory constraints, which are difficult to alleviate by task instructions. A final set of experiments showed that participants had difficulty correctly distinguishing the efficiency of individual pairs of queries (Experiments 4a and 4b). Instead, they displayed an obvious tendency to regard the suboptimal items as more informative than the optimal items. However, they were able to distinguish the efficiency of full query sets (Experiments 5a and 5b). This indicates that people can, in the right settings, understand what optimality entails (see also Rothe et al., 2018 for similar results in a related memory-based task).

However, the fact that biases persisted when people were asked to choose between experimenter-generated pairs of stimuli indicates that our biases may not be limited to only memory. Of course, memory may still be an important determinant of these biases. For example, it could be that people evaluate experimenter-generated queries based on which ones seem more accessible in their minds, that is, the item in the prompt primes the selection of the more similar option. This could explain why people are closer to optimality when given full query sets (in which there is no item in the prompt). Of course, it is also possible that people believe that an appropriate amount of confirmatory search by querying similar items is an efficient strategy for the property learning task in the experiments.

Our results stand in stark contrast with theoretical positions that propose that people search in optimal ways. Although such theories have been successful in explaining human inquiry in domains ranging from causal learning (Bramley et al., 2015) to spatial search (Gureckis & Markant, 2009), prior work has documented important limitations to optimal search. For example, it seems that people are only able to partially represent the hypothesis space (Bramley et al., 2017, 2018; Markant et al., 2016), which leads them to select queries that are informative with respect to an approximate representation but suboptimal with respect to an ideal learner. It is also the case that in search from memory, people's decisions to switch queries appear to be independent of the efficiency of the queries (Wilke et al., 2009). Other work has argued that people show a tendency for confirmatory search (Wason, 1966), which is a byproduct of associative memory mechanisms like the ones documented in this article (Bhatia, 2016; Glöckner & Betsch, 2008; Holyoak & Simon, 1999). We suspect that any setting in which participants must formulate sequences of queries in natural language will probably be constrained by memory processes, particularly the similarity-driven associative memory search. For example, in Rothe et al.'s (2018) experiments, participants were asked to generate a single question to gather information about the layout of enemy battleships in the game Battleship. These participants already struggled to generate the most informative questions. If they were asked to generate additional questions, we suspect subsequent questions would be similar to previous ones, leading to suboptimal query sequences.

Although associative memory processes curtail optimal active learning, that does not mean that people's memory processes are inherently flawed. Rather, memory serves multiple cognitive functions and the associative biases documented in this article may reflect optimal tradeoffs between diverging task demands. Indeed, many researchers have argued that association or similarity-driven memory search is part of an optimal system for semantic memory retrieval (Hills et al., 2012). Related work has shown that associative memory processes implicated in judgment and decision biases are adaptive in that they often lead to accurate inference and generalization with minimal cognitive cost (Bhatia, 2017; Tenenbaum & Griffiths, 2001). Regulating these processes in active learning tasks may be too effortful, and people may be optimally trading off the performance with the cognitive cost required to succeed in our task (Lieder & Griffiths, 2020). This theory predicts that even though we were unable to reduce semantic congruence and increase optimal search through coaching, performance may improve with higher incentives or practice.

One important limitation of our work is the fact that we specified our property scores using a linear function on the DSM space. This choice was motivated by considerable prior work that has found that many real-world properties can be approximated as linear functions on DSM spaces (Bhatia, 2019; Bhatia & Stewart, 2018; Bhatia et al., 2022; Gandhi et al., 2022; Richie & Bhatia, 2021; Zou & Bhatia, 2021a; see Richie et al., 2019 for a comprehensive analysis) and that it is easier for people to learn linear functions than nonlinear functions on underlying attribute spaces (Shepard et al., 1961; Zou & Bhatia, 2021b). Thus, we believe that to the extent that people are using inductive biases, these biases likely favor linear functions. However, it is nonetheless possible that participants entered our task with inductive biases favoring nonlinear functions, and optimally generated search queries given these nonlinear inductive biases. To test this possibility, we replicated our measurements of linear D-optimality with the design entropy for Gaussian process function learning and found that the two measures were highly correlated, indicating that search queries that were suboptimal under our linear measure would also be suboptimal under a more flexible and potentially nonlinear mapping between word vectors and property scores. In fact, it has even been argued that these two types of function learning may be two views of the same solution to function learning (Lucas et al., 2015). Of course, it could be the case that the similarity-based search strategies documented in this article are optimal in nonlinear settings, which is why future work should attempt to explicitly design stimuli sets in which these metrics diverge. Future work should also explore how behavior changes as participant beliefs are refined over the course of the task. It could, again, be the case that the similarity-based search processes documented in this article become more efficient once weights have been learned and the decision maker becomes highly confident about the underlying function.

Other future directions include the refinement of our memory and learning models. For example, participants in our study learned about novel target properties. Yet they came into the experiments with idiosyncratic knowledge about food items or animals. Thus, it is likely they held different prior beliefs about the novel target properties. As prior belief is not the focus of this article, we assumed all participants held the same prior belief in the experiments. In future work, the shape of prior belief can be set as free parameters and the same framework can be used to derive the prior representation of target properties in a given domain (H. Zhang et al., 2015). Individual differences in this regard can be revealed.

The Bayesian learning model also assumes that participants maintain a distribution of belief over multiple hypotheses (possible coefficients on the latent high-dimensional representations). However, other research suggests that in a closely related—and simpler—active category learning setting, participants maintain a single hypothesis at a time (Markant & Gureckis, 2014) and acquire easily interpretable queries (Cheyette et al., 2023). Previous research also reveals other simple heuristics, such as the split-half heuristic (Navarro & Perfors, 2011) and the likelihood difference heuristic (J. D. Nelson, 2005), in active learning tasks. More recently, researchers have found support for Gaussian processes in function learning, especially in explaining how people use similarity-based generalization to guide search (Schulz et al., 2019; Wu et al., 2018). Other work has shown that uncertainty plays a crucial role in guiding search behavior and resolving explore-exploit tradeoffs (Schulz & Gershman, 2019; Speekenbrink, 2022). It is possible that such processes play a role in the query search in our active learning tasks and subsequently explain some of the departures from optimality. If people only hold one hypothesis at a time, obtaining the information of a semantically similar item is likely to be more interpretable than a distant item, while not necessarily being less informative. Future work should consider explicitly modeling these processes in memory-based active learning.

Therefore, we suggest that suboptimality may emerge from both associative retrieval and from misconception about what is most informative in the naturalistic active learning settings. These results add to our understanding of the complexity of and the limitations to optimal search for naturalistic active learning. On top of the associative memory processes that prevent people from accessing more informative queries (in Experiments 1–3), the preference for simple hypotheses and easily interpretable data may be other important driving forces. Teasing apart these factors, as well as their interactions, in naturalistic active learning remains a promising direction for future research.

Our work contributes to the emerging body of research that offers researchers a naturalistic search domain to study active learning (Bramley et al., 2018; Hornsby & Love, 2022; Liefgreen et al., 2020; Z. H. Zhang et al., 2021). Additionally, our computational models integrate insights from several fields and are able to jointly describe both algorithmic memory search processes (which we have specified using a Markov random walk model) as well as the optimality or suboptimality of these search processes for active learning (which we have specified using an ideal Bayesian learner and the Bayesian D-optimality metric). In this way, our article presents a powerful new research paradigm for naturalistic active learning. There has been an increasing interest in porting computational cognitive models beyond abstract lab stimuli, to attempt to describe everyday cognition. This has been driven by the availability of new machine learning models that offer quantitative representations for natural entities (in the form of word vector representations (Bhatia, 2019; Bhatia & Aka, 2022; Bhatia & Stewart, 2018; Bhatia et al., 2019; Gandhi et al., 2022; Lu et al., 2019; Zou & Bhatia, 2021a, 2021b), or image vectors (Hebart et al., 2020; Peterson et al., 2018; Trueblood et al., 2021), as well as the growing demand from policy makers and practitioners for theory-driven behavioral and cognitive insights. Our research is part of this trend, and we look forward to future work that applies established algorithmic and rational theories of cognition to rich stimuli sets to better understand human cognition and behavior in the wild.

Constraints on Generality

Although our participants were adults from western cultures, we feel confident that the results of our article reflect general cognitive tendencies that would replicate across cultures and languages. That said, it would not be surprising if the underlying semantic representations and prior beliefs over these representations vary across cultures. Future work could try to use our paradigm to formally model cultural differences in memory-based search for active learning.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review*, *122*(3), 558–569. https://doi.org/10.1037/a0038693

Aka, A., & Bhatia, S. (2021). What I like is what I remember: Memory modulation and preferential choice. *Journal of Experimental Psychology: General*, *150*(10), 2175–2184. https://doi.org/10.1037/xge0001034

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Anderson, J. R., & Bower, G. H. (2014). *Human associative memory*. Psychology Press.

Atkinson, A., Donev, A., & Tobias, R. (2007). *Optimum experimental designs, with SAS* (Vol. 34). Oxford University Press.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *Psychology of learning and motivation* (Vol. 2, pp. 89–195). Academic Press.

Bhatia, S. (2016). The dynamics of bidirectional thought. *Thinking & Reasoning*, *22*(4), 397–442. https://doi.org/10.1080/13546783.2016.1187205

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20. https://doi.org/10.1037/rev0000047

Bhatia, S. (2019). Predicting risk perception: New insights from data science. *Management Science*, *65*(8), 3800–3823. https://doi.org/10.1287/mnsc.2018.3121

Bhatia, S., & Aka, A. (2022). Cognitive modeling with representations from large-scale digital data. *Current Directions in Psychological Science*, *31*(3), 207–214. https://doi.org/10.1177/09637214211068113

Bhatia, S., Olivola, C., Bhatia, N., & Ameen, A. (2022). Predicting leadership perception with large-scale natural language data. *Leadership Quarterly*, *33*(5), Article 101535. https://doi.org/10.1016/j.leaqua.2021.101535

Bhatia, S., Richie, R., & Zou, W. L. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31–36. https://doi.org/10.1016/j.cobeha.2019.01.020

Bhatia, S., & Stewart, N. (2018). Naturalistic multiattribute choice. *Cognition*, *179*, 71–88. https://doi.org/10.1016/j.cognition.2018.05.025

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media.

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *The Journal of General Psychology*, *30*(2), 149–165. https://doi.org/10.1080/00221309.1944.10544467

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338. https://doi.org/10.1037/rev0000061

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *105*, 9–38. https://doi.org/10.1016/j.cogpsych.2018.05.001

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731. https://doi.org/10.1037/xlm0000061

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Carr, R., Palmer, S., & Hagel, P. (2015). Active learning: The importance of developing a comprehensive measure. *Active Learning in Higher Education*, *16*(3), 173–186. https://doi.org/10.1177/1469787415589529

Cavagnaro, D. R., Pitt, M. A., & Myung, J. I. (2011). Model discrimination through adaptive experimentation. *Psychonomic Bulletin & Review*, *18*(1), 204–210. https://doi.org/10.3758/s13423-010-0030-4

Chaloner, K., & Verdinelli, I. (1995). Bayesian Experimental design: A review. *Statistical Science*, *10*(3), 273–304. https://doi.org/10.1214/ss/1177009939

Cheyette, S. J., Callaway, F., Bramley, N. R., Nelson, J., & Tenenbaum, J. (2023). People seek easily interpretable information. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45). https://escholarship.org/uc/item/5sm2b484

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2019). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, *26*(5), 1548–1587. https://doi.org/10.3758/s13423-018-1470-5

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006. https://doi.org/10.3758/s13428-018-1115-7

Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

Gandhi, N., Zou, W., Meyer, C., Bhatia, S., & Walasek, L. (2022). Computational methods for predicting and understanding food judgment. *Psychological Science*, *33*(4), 579–594. https://doi.org/10.1177/09567976211043426

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Glöckner, A., & Betsch, T. (2008). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making*, *3*(3), 215–228. https://doi.org/10.1017/S1930297500002424

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2022). "*rstanarm: Bayesian applied regression modeling via Stan*." R package (Version 2.21.3) [Computer software]. https://mc-stan.org/rstanarm/

Gopnik, A. (1996). The scientist as child. *Philosophy of Science*, *63*(4), 485–514. https://doi.org/10.1086/289970

Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244. https://doi.org/10.1037/0033-295x.114.2.211

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033. https://doi.org/10.1177/1745691619861372

Gureckis, T., & Markant, D. (2009, July 29–August 1). *Active learning strategies in a spatial concept learning game*. Proceedings of the Annual Meeting of the Cognitive Science Society, Amsterdam, Netherlands. https://escholarship.org/uc/item/55b5m116

Hayes, B. K., Banner, S., Forrester, S., & Navarro, D. J. (2019). Selective sampling and inductive inference: Drawing inferences based on observed and missing evidence. *Cognitive Psychology*, *113*, Article 101221. https://doi.org/10.1016/j.cogpsych.2019.05.003

He, L., Richie, R., & Bhatia, S. (2023). *Limitations to optimal search in naturalistic active learning*. https://osf.io/5e6tk

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185. https://doi.org/10.1038/s41562-020-00951-3

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440. https://doi.org/10.1037/a0027373

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General*, *128*(1), 3–31. https://doi.org/10.1037/0096-3445.128.1.3

Hornsby, A. N., & Love, B. C. (2022). Sequential consumer choice as multi-cued retrieval. *Science Advances*, *8*(8), Article eabl9754. https://doi.org/10.1126/sciadv.abl9754

Howard, M. W., & Kahana, M. J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, *46*(1), 85–98. https://doi.org/10.1006/jmla.2001.2798

Jones, A., Schulz, E., Meder, B., & Ruggeri, A. (2018, July 25–28). *Active function learning*. Proceedings of the 40th Annual Meeting of the Cognitive Science Society, Madison, USA. https://portal.fis.tum.de/en/publications/active-function-learning

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37. https://doi.org/10.1037/0033-295x.114.1.1

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*(1), 259–298. https://doi.org/10.1016/j.cognition.2007.02.003

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, *21*(2), 272–304. https://doi.org/10.1111/j.2517-6161.1959.tb00338.x

Kiefer, J., & Wolfowitz, J. (1959). Optimum designs in regression problems. *The Annals of Mathematical Statistics*, *30*(2), 271–294. https://doi.org/10.1214/aoms/1177706252

Kintsch, W. (2014). Similarity as a function of semantic distance and amount of knowledge. *Psychological Review*, *121*(3), 559–561. https://doi.org/10.1037/a0037017

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211–228. https://doi.org/10.1037/0033-295x.94.2.211

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295x.104.2.211

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, Article e1. https://doi.org/10.1017/S0140525X1900061X

Liefgreen, A., Pilditch, T., & Lagnado, D. (2020). Strategies for selecting and evaluating information. *Cognitive Psychology*, *123*, Article 101332. https://doi.org/10.1016/j.cogpsych.2020.101332

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005. https://doi.org/10.1214/aoms/1177728069

Lu, H. J., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10), 4176–4181. https://doi.org/10.1073/pnas.1814779116

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215. https://doi.org/10.3758/s13423-015-0808-5

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, *10*, Article 2767. https://doi.org/10.3389/fpsyg.2019.02767

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94–122. https://doi.org/10.1037/a0032108

Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-directed learning favors local, rather than global, uncertainty. *Cognitive Science*, *40*(1), 100–120. https://doi.org/10.1111/cogs.12220

Meder, B., Crupi, V., & Nelson, J. D. (2021). What makes a good query? Prospects for a comprehensive theory of human information acquisition. In C. Dezza, E. Schulz, & C. M. Wu (Eds.), *The drive for knowledge: The science of human information-seeking* (pp. 101–123). Cambridge University Press.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., Aiden, E. L., & Team, G. B. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013, December 5–10). *Distributed representations of words and phrases and their compositionality*. Advances in neural information processing systems, Lake Tahoe, NV, USA (pp. 3111–3119). https://proceedings.neurips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Miller, G. A. (1995). Wordnet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, *57*(3–4), 53–67. https://doi.org/10.1016/j.jmp.2013.05.005

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*(3), 499–518. https://doi.org/10.1037/a0016104

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120–134. https://doi.org/10.1037/a0021110

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. https://doi.org/10.3758/bf03195588

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979–999. https://doi.org/10.1037/0033-295x.112.4.979

Nelson, J. D., Tenenbaum, J. B., & Movellan, J. R. (2001, August 1–4). *Active inference in concept learning*. Proceedings of the 23rd conference of the Cognitive Science Society, Edinburgh, UK. https://escholarship.org/uc/item/90m7n1xf

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300. https://doi.org/10.1037/0033-295X.104.2.266

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631. https://doi.org/10.1037/0033-295x.101.4.608

Oaksford, M., & Chater, N. (2007). *Bayesian Rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Ouyang, L., Tessler, M. H., Ly, D., & Goodman, N. (2016). *Practical optimal experiment design with probabilistic programs*. arXiv preprint. https://arxiv.org/abs/1608.05046

Pennington, J., Socher, R., & Manning, C. D. (2014, October 25–29). *Glove: Global vectors for word representation*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar. https://aclanthology.org/D14-1162/

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3–4), 175–190. https://doi.org/10.1080/02643294.2016.1176907

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669. https://doi.org/10.1111/cogs.12670

Reimers, N., & Gurevych, I. (2019, November 3–7). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong (pp. 3982–3992). https://aclanthology.org/D19-1410.pdf

Richie, R., & Bhatia, S. (2021). Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive Science*, *45*(8), Article 13030. https://doi.org/10.1111/cogs.13030

Richie, R., Zou, W., Bhatia, S., & Vazire, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology*, *5*(1), Article 50. https://doi.org/10.1525/collabra.282

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. *Psychological Science*, *4*(1), 28–34. https://doi.org/10.1111/j.1467-9280.1993.tb00552.x

Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*(1), 69–89. https://doi.org/10.1007/s42113-018-0005-5

Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., & Gershman, S. J. (2019). Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences*, *116*(28), 13903–13908. https://doi.org/10.1073/pnas.1821028116

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7–14. https://doi.org/10.1016/j.conb.2018.11.003

Settles, B. (2009). *Active learning literature survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42. https://doi.org/10.1037/h0093825

Speekenbrink, M. (2022). Chasing unknown bandits: Uncertainty guidance in learning and decision making. *Current Directions in Psychological Science*, *31*(5), 419–427. https://doi.org/10.1177/09637214221105051

Stan Development Team. (2021). *RStan: The R interface to Stan* [Computer software]. https://mc-stan.org/

Tenenbaum, J. B., & Griffiths, T. L. (2001, August 1–4). *The rational basis of representativeness*. Proceedings of the 23rd Annual Conference of the Cognitive Science Society, Edinburgh, UK.

Trenkmann, M. (2016). *PhraseFinder—Search millions of books for language use*. https://phrasefinder.io

Trueblood, J. S., Eichbaum, Q., Seegmiller, A. C., Stratton, C., O'Daniels, P., & Holmes, W. R. (2021). Disentangling prevalence induced biases in medical image decision-making. *Cognition*, *212*, Article 104713. https://doi.org/10.1016/j.cognition.2021.104713

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology* (pp. 135–151). Penguin.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161

Wilke, A., Hutchinson, J., Todd, P. M., & Czienskowski, U. (2009). Fishing for the right words: Decision rules for human foraging behavior in internal search tasks. *Cognitive Science*, *33*(3), 497–529. https://doi.org/10.1111/j.1551-6709.2009.01020.x

Wilson, B. J., & Schakel, A. M. (2015). *Controlled experiments for word embeddings*. arXiv preprint. https://arxiv.org/abs/1510.02675

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924. https://doi.org/10.1038/s41562-018-0467-4

Zhang, H., Daw, N. D., & Maloney, L. T. (2015). Human representation of visuo-motor uncertainty as mixtures of orthogonal basis distributions. *Nature Neuroscience*, *18*(8), 1152–1158. https://doi.org/10.1038/nn.4055

Zhang, Z. H., Wang, S. C., Good, M., Hristova, S., Kayser, A. S., & Hsu, M. (2021). Retrieval-constrained valuation: Toward prediction of open-ended decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(20), Article e2022685118. https://doi.org/10.1073/pnas.2022685118

Zou, W., & Bhatia, S. (2021a). Judgment errors in naturalistic numerical estimation. *Cognition*, *211*, Article 104647. https://doi.org/10.1016/j.cognition.2021.104647

Zou, W., & Bhatia, S. (2021b, July 26–29). *Learning new categories for natural objects*. Proceedings of the Annual Meeting of the Cognitive Science Society, Virtual meeting. https://escholarship.org/uc/item/89b6z6ns

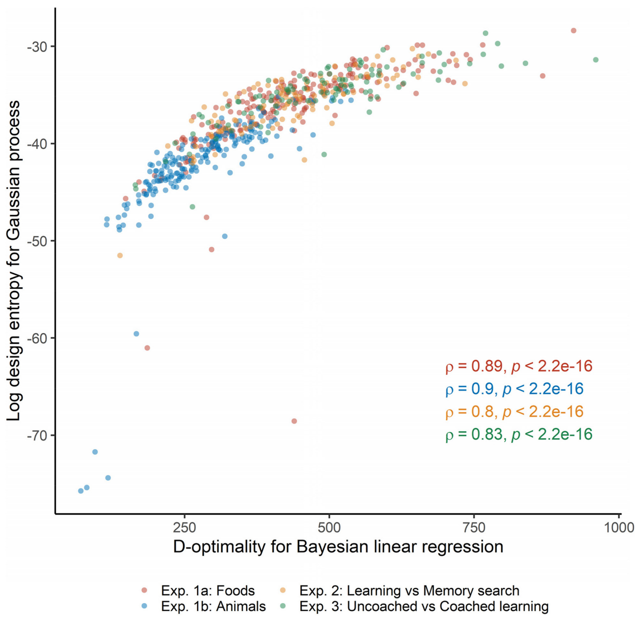
(*Appendix follows*)

Appendix

Experimental Materials and Supporting Results

Figure A1

The Correspondence Between D-Optimality for Bayesian Linear Regression and Log Design Entropy for Gaussian Process Function Learning in Experiments 1a, 1b, 2, and 3



Note. The two hyperparameters for the Gaussian process were set at length scale = 1 and nugget = 0.01. Of course, the choice of the hyperparameters (especially length scale) may have an impact on the measurement of log design entropy. A larger length scale makes the log design entropy more closely aligned to the Bayesian D-optimality, whereas a smaller length scale (which corresponds to a more complex nonlinear relationship) makes the former deviate from the latter. See the online article for the color version of this figure.

Table A1

Food Items With the Highest and Lowest Scores for Each of the Random Linear Functions Used in Experiments 1a, 2, and 3

Experiment 1a	Experiment 2	Experiment 3
Highest scores		
Apple	Pumpkin	Soft drink
Turnip	Apple	Scotch whiskey
Medlar	Cherry tomato	Malt whiskey
Pawpaw	Strawberry	Alcoholic beverage
Crab apple	Candy cane	Tequila
Macoun	Cupcake	Liquor
Potato	Navel orange	Beer
Damson	Potato	Wine
Casaba	Blueberry	Coke
Sugar beet	Tomato	Vino
Lowest scores		
Skim milk	Arak	Sour cream
Coconut milk	Poteen	Cornbread
Lingcod	Ouzo	Garlic salt
Broth	Camomile tea	Orzo
Cocktail sauce	Broth	Caster sugar
Spanish mackerel	Brandy	Pie crust
Albacore	Pastis	Streusel
Lemon juice	Curacao	Cornmeal
Chicken broth	Rotgut	Casserole
Lime juice	Firewater	Matzo meal

(*Appendix continues*)

Table A2
Animals With the Highest and Lowest Scores for Each of the Four Random Linear Functions Experiment 1b

Function 1	Function 2	Function 3	Function 4
Highest scores			
Mastodon	Cowry	Basset hound	Striper
Stegosaurus	Horsehead	Golden retriever	Gopher
Plesiosaur	Giant	Poodle	Snook
Bernese mountain Dog	Pterodactyl	Collie	Crawdad
Ichthyosaur	Blowfish	Puppy	Shad
Apatosaurus	Dragon	Pooch	Walleye
Malamute	Serpent	Beagle	Tarpon
Tyrannosaurus rex	Kine	Dalmatian	Smallmouth
Shetland sheepdog	Medusa	Cocker spaniel	Crappie
Golden retriever	Manta	Dog	Redfish
Lowest scores			
Swiftlet	Springer spaniel	Tunny	Bos indicus
Skylark	Cocker spaniel	Tuna	Gorilla gorilla
Mealybug	Miniature poodle	Threadfin	Serow
Greenfly	Weimaraner	Dolphinfish	Banteng
Spider mite	Siberian husky	Jack mackerel	Canis familiaris
Tobacco budworm	Labrador retriever	Skipjack tuna	Pan troglodytes
Armyworm	Shetland sheepdog	Hogfish	Bos taurus
Tsetse	Great pyrenees	Yellowfin	Pongo pygmaeus
Bollworm	Yorkshire terrier	Striped marlin	Malayan tapir
Mealy bug	Golden retriever	Yellowfin tuna	Giraffa camelopardalis

Table A3
Choice Pairs for Experiments 4a (Animals and Foods) and 4b (Foods)

Domain	Previous query	Optimal	Suboptimal
Animals	Turkey	Goose	Hippo
	Wolf	Moose	Maggot
	Camel	Donkey	Jellyfish
	Moth	Wasp	Yak
	Alligator	Rattlesnake	Dodo
Foods	Vinegar	Buttermilk	Meatball
	Margarita	Martini	Pea
	Trout	Herring	Toast
	Muffin	Cake	Shellfish
	Noodle	Dumpling	Cranberry

Note. Each choice pair between an optimal and suboptimal item (goose and hippo, respectively) is conditioned on having already chosen the previous query (e.g., turkey).

(*Appendix continues*)

Table A4
Choice Pairs for Experiment 5a

Optimal	Suboptimal
Pair 1	
Noodle, trout, fish, apple, pear, water, beer, wine, cheese, burger, ham, milk, lettuce, tomato, ketchup, strawberry, plum, grape, beef, pork	Apple, pineapple, candy, strawberry, ketchup, avocado, banana, pomegranate, apple, pear, potato, carrot, grape, bean, corn, apple, apple, apple, apple, apple
Pair 2	
Broccoli, pizza, sushi, mackerel, cake, rice, apple, chocolate, beer, wine, chocolate, pea, butter, potato, salmon, yam, parfait, tuna, pudding, juice	Chicken, pineapple, apple, banana, turnip, carrot, mango, strawberry, blueberry, radish, lettuce, cabbage, corn, kale, broccoli, cucumber, parsnip, potato, kiwi fruit, persimmon
Pair 3	
Vinegar, noodle, muffin, trout, margarita, biscuit, steak, fries, wine, fish, gammon, salad, chicken, broccoli, potato, lettuce, beer, lager, salad, cider	Vinegar, salt, sugar, pepper, cinnamon, nutmeg, chili, cayenne, rosemary, basil, tomato, onion, courgette, mushroom, carrot, potato, aubergine, lettuce, cabbage, radish
Pair 4	
Pizza, chocolate, beer, water, salmon, lobster, broccoli, corn, potato, whiskey, sausage, watermelon, wine, salad, grape, salt, sugar, banana, apple, lemon	Vinegar, salt, pepper, garlic, paprika, potato, carrot, chips, vegetable, broccoli, cauliflower, corn, pea, banana, spinach, tomato, onion, lettuce, beetroot, pumpkin
Pair 5	
Fish, alcohol, beef, chicken, lamb, pork, burger, beer, wine, cereal, chocolate, water, lettuce, radish, apple, cherry, pear, trout, cabbage, cola	Bagel, toast, baguette, bread, sandwich, donut, croissant, dough, cake, pastry, cheesecake, dessert, fruitcake, chocolate cake, chocolate, vanilla, blueberry, cranberry, brownie, cookie

Note. The more optimal set is on the left in each pair.

Table A5
Choice Pairs for Experiment 5b

Optimal	Suboptimal
Pair 1	
Apple, flounder, soda, cilantro, cognac, barbeque, wheat, limpet, casserole, coho	Apple, strawberry, grapefruit, pineapple, avocado, celery, onion, romaine, asparagus, chard
Pair 2	
Soda, flounder, cauliflower, cognac, barbecue, flour, broth, grape, lolly, halibut	Soda, cola, juice, coca_cola, coke, beer, alcohol, rum, gin, vodka
Pair 3	
Bread, coral, booze, cilantro, grape, fish, soda, filet, marzipan, wheat	Bread, loaf, waffle, cake, muffin, marshmallow, birthday_cake, pie, flapjack, kibble
Pair 4	
Tagliatelle, coho, coke, corn, barbeque, cognac, lollipop, vinegar, papaya, fish	Tagliatelle, gorgonzola, risotto, cannelloni, rigatoni, potato_pancake, quiche, ravioli, penne, linguine
Pair 5	
Cheese, flounder, booze, grape, soda, broth, coho, barbecue, lollipop, rice	Cheese, english_muffin, apple_sauce, oatmeal, cottage_cheese, granola, muffin, pretzel, waffle, loaf

Received May 20, 2022
Revision received December 17, 2023
Accepted January 18, 2024 ■