Empirical Evidence for Markov Chain Monte Carlo in Memory Search

David D. Bourgin (ddbourgin@berkeley.edu)
Joshua T. Abbott (joshua.abbott@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology
University of California Berkeley
Berkeley, CA 94720 USA

Kevin A. Smith (k2smith@ucsd.edu)
Edward Vul (evul@ucsd.edu)
Department of Psychology
University of California San Diego
La Jolla, CA 92093 USA

Abstract

Previous theoretical work has proposed the use of Markov chain Monte Carlo as a model of exploratory search in memory. In the current study we introduce such a model and evaluate it on a semantic network against human performance on the Remote Associates Test (RAT), a commonly used creativity metric. We find that a family of search models closely resembling the Metropolis-Hastings algorithm is capable of reproducing many of the response patterns evident when human participants are asked to report their intermediate guesses on a RAT problem. In particular we find that when run our model produces the same response clustering patterns, local dependencies, undirected search trajectories, and low associative hierarchies witnessed in human responses.

Keywords: Creativity; Remote Associates Test; Information retrieval; Semantic networks; Markov chain Monte Carlo.

Introduction

Exploratory search in memory is a major component of creative problem solving. Often, the demands of a task impose constraints on the solution space: when searching for a word to rhyme with A that also means B, a would-be poet is unlikely to include items that mean T or that rhyme with Q (assuming that Q does not rhyme with A and T is not a synonym of B) in her search. Similarly, an inventor would be remiss to consider inventions that do not satisfy certain requirements (e.g., novelty, usefulness) if she plans to pursue a patent. The current study proposes a formal model for the process by which people perform this type of exploratory search under multiple constraints.

Locating relevant pieces of information in memory requires a strategy for quickly traversing the space of potential solutions. Markov chain Monte Carlo (MCMC) methods have been a particularly useful in this regard, offering efficient statistical techniques for exploring spaces that would otherwise prove computationally intractable to traverse fully (Brooks, 1998). MCMC methods have been employed to model phenomena as diverse as theory change (Ullman, Goodman, & Tenenbaum, 2012), perceptual multistability (Gershman, Vul, & Tenenbaum, 2012), and conceptual fluidity (Gabora, 2000), making them a popular tool within statistical and computational cognitive modeling. MCMC methods have also been proposed within the creativity literature to model search behavior in memory (e.g., Martindale, 1995; Paulus, Levine, Brown, Minai, & Doboli, 2010), although to date there exists little empirical evidence on which to evaluate these proposals. In the current study we aim to fill this gap by evaluating a formal model of MCMC search on the Remote Associates Test (RAT; S. A. Mednick, 1962), a classic multiply-constrained search task.

Building on recent work using random walks on semantic networks to model semantic fluency (Griffiths, Steyvers, & Firl, 2007) and singly-constrained search in memory (Abbott, Austerweil, & Griffiths, 2012), we consider a range of MCMC-inspired models for search on the RAT. We find that human response patterns are well-matched by a model closely related to the Metropolis-Hastings (M-H) algorithm, providing empirical support for MCMC as a candidate for modeling creative ideation.

The plan for the paper is as follows. We first review previous work on human response patterns on the Remote Associates Test and discuss existing theories of search on this task. We then introduce Markov chain Monte Carlo methods with an emphasis on their application to creative search. We end by formally introducing the model used in the current study, and offer an analysis of its behavior in relation to human data collected by Smith, Huber, and Vul (2013).

Background

Stage models of problem solving propose that individuals first search through memory to identify potential answers to a problem and then test those candidate answers against the constraints of the problem to determine acceptability (Gruenewald & Lockhead, 1980; Raaijmakers & Shiffrin, 1981). In the current paper it is assumed that these search and test processes constitute qualitatively distinct cognitive tasks. Thus in reviewing recent work on the RAT and memory search we emphasize accounts that focus not only on the identification of correct answers but also on the path by which these answers are selected.

The Remote Associates Test

In the Remote Associates Test participants are shown a set of three cue words (e.g., 'surprise,' 'line,' 'birthday') and instructed to generate a fourth word to relate them (in this case, 'party'). Although originally introduced to test for individual differences creative ability (S. A. Mednick, 1962), early work established that performance on the RAT correlated with IQ (M. T. Mednick & Andrews, 1967) and originality during brainstorming (Forbach & Evans, 1981). More recently the RAT has also been used to measure the effects of manipulations related to creative ability, including intuition and incubation (Bowers, Regehr, Balthazard, & Parker, 1990; Vul & Pashler, 2007), the role of affect during problem solving

(Fodor, 1999; Isen, Daubman, & Nowicki, 1987), implicit learning during REM sleep (Cai, Mednick, Harrison, Kanady, & Mednick, 2009), and the relation between synesthesia and creativity (Ward, Thompson-Lake, Ely, & Kaminski, 2010). The RAT is used in the current study to investigate how people solve multiply-constrained search problems. Each RAT cue indicates a different aspect of the target solution (e.g., in the example above, 'party line' relies on a different meaning of 'party' than 'birthday party'), and there is no principled way to trade off associations between each of the three cues. Additionally, because all cues are of the same type (word-word relationships) and are designed to have a unique best solution, the RAT offers a controlled setting in which to develop a formal model of search under multiple constraints.

Human response patterns on the RAT exhibit several notable characteristics. Gupta, Jang, Mednick, and Huber (2012) found that subjects reporting low-frequency guesses within the first 30*s* of cue presentation displayed higher RAT problem accuracies overall. Harkins (2006) found that RAT problems with answers that were easily guessed when subjects were prompted by just one of the three cue words were easier to solve, suggesting that not all cues in a RAT problem are weighted equally by all problem solvers. Most recently, Smith et al. (2013) found that subjects' intermediate responses on the RAT tended to bunch around individual problem cues in a Latent Semantic Analysis (LSA) space (Carroll et al., 1997), suggesting that the RAT problem solvers were primarily using single cues to constrain their search rather than a combination of cues. The researchers also provided evidence for direct local dependencies between prior and subsequent responses, consistent with a local search strategy in which each response produced was selected from the neighborhood of a previously considered response. Responses appeared to be selected randomly from these neighborhoods, however, as the researchers found little evidence to suggest that subsequent responses become more similar to the correct answer over time.

Models of Search in Associative Memory

Several theoretical accounts of search have been proposed to account for problem-solving behavior on the Remote Associates Test. Spreading activation models (Collins & Loftus, 1975) posit a decaying activation function originating at primed concepts in a semantic network and moving outward along their edges to activate their 'close associates.' In multiply-constrained search it is assumed that solutions to a problem, as words related to multiple primed concepts (cues), will receive greater activation and thus be more likely to be produced as responses (Bolte & Kuhl, 2003). Spreading activation accounts however do not specify whether or how cues should be weighted (a relevant concern given Smith et al.'s (2013) finding that some cues tend to be used more than others when generating responses on a RAT problem), the quantitative definition of 'close associate', or a decision procedure for choosing from nodes with equal levels of activation in the network (Smith et al., 2013).

Stage models (e.g., Gruenewald & Lockhead, 1980; Raaijmakers & Shiffrin, 1981) in which a problem solver first selects a category (e.g., 'animals') and then enumerates its contents (e.g., 'cat', 'dog', 'cow', etc.) have also been proposed for free recall and singly-constrained search tasks, although it is unclear how these accounts might be expanded to account for multiply-constrained search. Finally, optimal foraging accounts (Hills, Jones, & Todd, 2012) have posited an evolutionarily conserved, area-restricted search algorithm (the Mean Value Theorem) that guides search in associative memory. Like stage models, this algorithm moves dynamically between local and global search efforts to optimize recall frequencies during semantic fluency and singly-constrained search tasks. However, recent work has demonstrated that the switching behavior that stage and foraging models have been proposed to explain can be produced by a purely local search strategy (Abbott et al., 2012). More importantly for the current study, although the above theories yield clues about how people search for answers on simple search tasks, it is not immediately apparent how they can be extended to produce testable models of multiply-constrained search.

Recent work in computational cognitive modeling and artificial intelligence has focused on using simple stochastic processes operating over large networks to model aspects of higher-level cognition. Griffiths et al. (2007) demonstrated that probabilistic methods previously used in web search produced favorable results when run on a semantic network and compared to human semantic retrieval performance. Abbott et al. (2012) subsequently reported that a random walk run on a semantic network was capable of reproducing data previously used to support a two-stage model of semantic search (Hills et al., 2012). These findings coincide with theoretical accounts proposing simulated annealing and other MCMC methods as a model for defocused search during creative problem solving (Martindale, 1995; Gabora, 2000). The success of these methods in modeling simple search tasks makes them a promising candidate for extension to more complex, multiply-constrained search. Below, we outline an MCMC-inspired account of search under multiple constraints.

Evaluating MCMC as a Search Process

Given the success of simple probabilistic search techniques in capturing human response patterns in semantic fluency tasks, we introduce and evaluate an MCMC model of multiply constrained search in memory against human data collected on the Remote Associates Test.

Human Response Data

The human data used as a baseline for our model was collected by Smith et al. (2013). In their study 56 participants completed a modified version of the Remote Associates Test in which they were instructed to report every word they considered while searching for an answer on a problem. Participants were limited to two minutes per problem and each completed 25 problems, producing on average 7.8 responses per question (though this value varied significantly with problem difficulty). Similarity ratings between responses, problem cues, and answers were calculated using a high-dimensional Latent Semantic Analysis space constructed from the TASA corpus (Zeno, Ivens, Millard, & Duvvuri, 1995).

Model Template

A simple generative model for sequential response patterns is a Markov chain (random walk) on a semantic network that at step X_0 on RAT problem i selects a starting state from the set $\mathbf{c} = [cue1, cue2, cue3]$ and then at step n randomly generates the next state $X_{n+1} = x'$ according to a probability distribution that depends only on the current state, $X_n = x$. A Markov chain of this sort was used in the current model to approximate a posterior distribution over nodes in the network, $P[x|\mathbf{c}]$, reflecting the probability of responding with concept x when presented with the cues in $\mathbf{c}$. The semantic network we ran the model on was a directed graph constructed from free response data generated from a word association task. The task asked participants to list the words that came to mind when presented with a particular one-word probe. Each response that occurred more than once during this task was subsequently used as a probe in another iteration. After repeating this task across a large number of participants, a 5018 node network was constructed containing concepts ranging from ‘a’ to ‘zucchini’ (Nelson, McEvoy, & Schreiber, 2004).[1]

In their analysis of the human response data, Smith et al. assumed that intermediate responses on the RAT problems represented incremental samples from an ongoing internal search process. To account for this behavior in the current model, a state in the chain was selected as a response if it met three conditions: (1) it was not one of the problem cues, (2) it had not already been sampled as a responses on the problem (ie., it the first time the node had been visited on the walk for that problem), and (3) at least one of the geodesic distances between the current state and the problem cues was below a threshold value, λ. All states in the walk that did not meet these criteria were not recorded as responses. Criteria (1) and (2) are consistent with the data pre-processing performed by Smith et al. on the human response data, while the third criterion was implemented as a selection metric in the model.

A method for mapping the steps in the walk to human reaction times was necessary to make an appropriate comparison to the inter-item retrieval times (IRT) collected for human responses on the RAT. Using the method introduced by Abbott et al. (2012) we denote $\tau(k)$ as the step in the walk at which the kth response was first visited and assume that the amount of time the walk spends to generate a response corresponds to the length of the response. As participants in Smith et al. typed their responses, this accounts for it taking longer for participants to type longer words. The IRT between response k and response $k-1$ on a problem was calculated using the formula

$$IRT(k) = \tau(k) - \tau(k-1) + L(k), \tag{1}$$

where $\tau(k)$ was the step in the walk on which the model first visited node k, and $L(k)$ was the length of word k. IRT values reflected the immediate search history of a particular state so that even if k were sampled as a response for multiple problems, $IRT(k)$ would differ depending on the path the model took to reach it on its walk. This calculation assumed that the time humans took to to type a letter was equivalent to the time a walker took to take a step in the network.

Model Settings

With a framework now in place we define a space of potential models by introducing three factors influencing the calculation of state transitions in our Markov chain.

Transition strategy. The first setting in the model selects the walk’s transition strategy, which can either be undirected, where the next state is sampled from a multinomial distribution over the current node’s outgoing links, or directed, where the probability of transitioning to state x' from state x is calculated using a modification of the Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970). This algorithm is a commonly used MCMC technique which approximates a target distribution, $P[x]$, using the stationary distribution of a carefully constructed random walk. At each state x in the walk a proposed transition node x' is selected from a probability distribution (the proposal distribution, $Q[x \to x']$) over the nodes incident to x in the network. The search transitions to x' from x with an acceptance probability

$$A[x \to x'] = \min\left(1,\ \frac{d(x')\, Q[x' \to x]}{d(x)\, Q[x \to x']}\right), \tag{2}$$

where the objective function, $d(x)$, is proportional to the target distribution, $P[x]$. If the transition is not accepted, the algorithm resamples a new x' from the proposal distribution and repeats until a transition is accepted. The algorithm is constructed so that after a fixed number of transitions the stationary distribution for the walk will approximate the target probability distribution, reflecting in this case the probability that a participant reports concept x when given a set of cues (the prior distribution).

The proposal distribution in our model was a multinomial distribution over the current node’s outgoing links, and the objective function, $d(x)$, was the inverse of the average geodesic distance in the network between x and the three cues in $\mathbf{c}$. Formally this is expressed by

$$\begin{gathered} A[x \to x'] = \min\left(1,\ \frac{n(x)\, d(x')}{n(x')\, d(x)}\right) \\ \text{where } d(x) = |\mathbf{c}| \times \left(\sum geo(\mathbf{c}, x)\right)^{-1}, \end{gathered} \tag{3}$$

where $geo(\mathbf{c}, x)$ is a triple containing the geodesic distances between node x and the problem cues in $\mathbf{c}$, and $n(x)$ is the

[1] Because the LSA space from Smith et al. and the semantic network used in our model were constructed from different corpora, the resulting analysis of our model is more likely to reflect general semantic properties rather than artifacts of a particular semantic space.

Table 1: Performance of the Markov chain Monte Carlo (MCMC) model.

Analysis	MCMC Model	Humans
(a) Average problem accuracy	0.27	0.42
(b) Correlation with human problem accuracies	$r(23) = 0.301$	-
(c) Average number of responses per problem	5.12	7.82
(d) Average response similarity, within vs. between cue clusters	0.304 vs. 0.177, 95% CI: $[0.120, 0.134]$	0.250 vs. 0.142, 95% CI: $[0.100, 0.115]$
(e) Permutation test to control for primary cue assignment	0.261 vs. 0.190, 95% CI: $[0.052, 0.072]$	0.209 vs. 0.160, 95% CI: $[0.032, 0.053]$
(f) Rate of change in similarity per intervening response	-0.011	-0.002
(g) Rate of change in similarity per unit of IRT	-0.084	-0.079
(h) Average response similarity for adjacent responses, within vs. between cue clusters	0.304 vs. 0.177, 95% CI: $[0.120, 0.134]$	0.247 vs. 0.139, 95% CI: $[0.101, 0.116]$
(i) Ratio of average IRT for within cluster responses to average IRT for between cluster responses	1.30	1.11
(j) Baseline vs. actual percentage of response pairs with the same primary cue	Baseline: 51.6%, Actual: 52.5%, $p < 0.05$	Baseline: 40.6%, Actual: 42.8%
(k) Between cue cluster response similarities, adjacent vs. non-adjacent responses	0.177 vs. 0.155, 95% CI: $[0.018, 0.027]$	0.139 vs. 0.113, 95% CI: $[0.021, 0.029]$
(l) Within cue cluster response similarities, adjacent vs. non-adjacent responses	0.304 vs. 0.265, 95% CI: $[0.032, 0.048]$	0.247 vs. 0.214, 95% CI: $[0.025, 0.042]$
(m) Rate of change in similarity to answer with distance to answer	-0.010	-0.004
(n) Rate of change in similarity to final response with distance to end for non-correct responses	-0.010	-0.003

Note: Confidence intervals (noted as 95% CI in the table) reflect the difference between means. All differences were statistically significant with $p < 0.001$ unless otherwise noted.

number of nodes incident to x in the network. This transition strategy captures the possibility that concepts with associations to all of the problem cues will be selected more frequently than concepts primarily associated with some subset of the cues.

Reset probabilities. The second model setting determines the probability that the search will return to one of the problem cues during a transition. This captures the notion that response generation may temporarily disrupt the direction of search in memory, 'resetting' the walk to one of the original problem cues. In the current model we used a reset condition with two components: the probability of returning the walk to one of the problem cues during a transition from an unsampled step (ρ), or transitioning from a sampled step (γ). Formally the probability of reset is

$$P[\text{reset}] = \begin{cases} \rho & \text{if } \min(geo(\mathbf{c}, x)) > \lambda \\ \gamma & \text{otherwise,} \end{cases} \tag{4}$$

where $geo(\mathbf{c}, x)$ is a triple containing the geodesic distances between the current node x and the problem cues in $\mathbf{c}$, and $0 \leq \rho, \gamma \leq 1$.

Reset cue selection. The third setting defines the way in which a reset cue is selected from $\mathbf{c}$. The reset condition can either be deterministic, in which the reset cue is designated as the cue with the smallest geodesic distance to the current state, or stochastic, in which the reset cue is selected via a multinomial distribution over the PageRank (Page, Brin, Motwani, & Winograd, 1999) scores for the cues in $\mathbf{c}$.[2]

Model Evaluation

We evaluated a Markov chain Monte Carlo search model constructed with a directed, Metropolis-Hastings transition strategy, a stochastic reset procedure, a response threshold $\lambda = 2.0$, and reset probabilities $\rho = 0.0$ and $\gamma = 0.30$ such that the walk was only liable to reset during a transition away from a *sampled* step (i.e., a step within the minimum distance lambda of one of the cues).[3] After evaluating the model's performance, we explored other setting combinations to identify elements influencing the model's fit with human data.

Performance on the Remote Associates Test

We ran 100 simulations of the MCMC model on the 25 Remote Associates Test questions from Smith et al. (2013), using a maximum duration of 1000 iterations per-problem. Consistent with the RAT literature both human and model performance fluctuated widely across problems, with average

[2] PageRank is employed here as a measure of concept fluency. For a discussion of PageRank in relation to other word fluency metrics, see Griffiths et al. (2007)

[3] These values reflect the settings with the best fit to the human data under the condition that the walk utilize the Metropolis-Hastings transition strategy

success rates ranging from 9% to 86% for humans and 1% to 88% for the model. Similarly, the number of responses generated ranged on average from 2.8 for humans and 1.7 for the model to 16.1 for humans and 10.7 for the model. The results of the MCMC model showed a medium correlation with human accuracies (Table 1: analyses *a*, *b*, and *c*).

Model performance was evaluated using the suite of analyses and LSA space developed in Smith et al. (2013). Similarity between responses was calculated as the cosine angle between each pair of words in the LSA space. Each of the model's responses was assigned a primary cue, calculated as the cue in the problem statement with the shortest cosine distance to the response in the LSA space. All adjacent responses were divided into within and across-cue pairs.

In all, the responses generated by the MCMC model reproduced the cue-response bunching patterns (Table 1: analyses *d* and *e*; responses tended to bunch around particular cues in the LSA space, suggesting they were generated primarily based on a single cue), sequential dependencies (Table 1: analyses f, g, and *h*; both the number of intervening responses and the inter-item retrieval times between responses were negatively correlated with average response-response similarity) driven by a direct association between adjacent responses (Table 1: analyses *h*, *i*, *j*, *k*; Adjacent responses within cue clusters were more similar to one another than across-cluster responses, while adjacent across-cue responses were more similar to one another than non-adjacent responses. Adjacent responses from the same cluster were also generated more quickly than adjacent responses from different clusters), focus on single cues (Table 1: analysis *l*; adjacent responses were more similar even when the responses came from the same primary cue, suggesting that sequential dependence was not exclusively the product of a single cue), and the undirected search trajectories (Table 1: analyses *m*, *n*; despite similarity to the final answer increasing significantly over the final 10 responses in a walk, the rate of increase did not differ between incorrect and correct trials) evident in human responses.

On average, our MCMC model tended to exhibit stronger response-response similarity scores than people did. This effect may generally be attributed to the way in which our semantic network was organized: because the connections between nodes were constructed using free-association norms, the network was likely to capture only the strongest and most common associations between concepts. The model output, as a function of the network, would thus contain fewer weakly-associated responses in comparison to humans despite performing similarly otherwise.

Effects of Model Assumptions

By manipulating the settings in our model we sought to identify components influencing its fit with the human data. Although an exhaustive investigation was unfeasible, several robust trends were evident. Overall, models using the Metropolis-Hastings transition strategy showed significantly higher correlations with human accuracies than did equivalent undirected models (0.243 vs. -0.076; $t(18)$=-3.89, $p = 0.001$). Interestingly, Metropolis-Hastings transitions also tended to produce higher overall problem accuracies, despite reporting on average the same number of responses per problem as undirected search models ($t(48)$=-0.44; p = 0.66), suggesting that the Metropolis-Hastings strategy may be more efficient than undirected search at exploring the network.

With a response threshold $\lambda = 2.0$ and a per-problem time limit of 1000 steps we found that both undirected and Metropolis-Hastings search strategies produced a similar number of responses to humans on average ($t(48) = -0.15$, $p = 0.88$; $t(48) = -0.44$, $p = 0.66$, respectively). Modulating the response threshold appeared to have little impact on model accuracies, likely because the models always ended their search when they reached a correct answer regardless of whether it passed the response threshold.

Problem accuracies for models using Metropolis-Hastings transitions and deterministic reset procedures had higher correlations with the human data than accuracies from models using a stochastic reset setting (0.450 vs. 0.412; $t(12)$=2.62, $p < 0.05$). This effect disappeared when using an undirected transition strategy, however ($t(10) = 0.14$, $p = 0.89$). Given that the state transitions in the Metropolis-Hastings models depended on the average geodesic distance between the current node and the problem cues, this is as expected; resetting to a random cue can significantly impact the value of $d(x)$, and thus the overall trajectory of the Metropolis-Hastings walk through the network. Because the undirected strategy did not use cue distances when selecting its transitions, it did not show this effect.

Previous work has found that participants with higher solution rates on the RAT tend to report lower-frequency intermediate responses when searching for an answer to a particular problem (Gupta et al., 2012). To investigate whether the models were capable of reproducing these patterns we compared the average typicality scores (approximated using PageRank) for correct versus incorrect response chains on each problem. The Metropolis-Hastings model reproduced this effect on 16 of the 25 RAT questions, with correct response chains showing significantly lower PageRank values than incorrect response chains. In comparison, the undirected transition strategy only reproduced this effect on six of the 25 questions.

To investigate the contribution of the network structure to our results we ran all models on a 'shuffled' network in which the edge weights between adjacent nodes were permuted. Unsurprisingly this had a disastrous effect – the models found significantly fewer answers and were far less accurate on average. Despite this poor performance, however, the Metropolis-Hastings strategy fared better in comparison to the undirected search on the shuffled network, suggesting that directed transitions conveyed a general advantage over undirected strategies regardless of network topology.

Discussion

Our results show that a Markov chain Monte Carlo-inspired search strategy in a semantic network is able to reproduce many of the human response patterns witnessed during multiply-constrained search in the Remote Associates Test. In evaluating the behavior of a Metropolis-Hastings search model we reproduced all of the results reported by Smith et al. (2013), finding evidence for response clustering around particular cues, strong local dependencies between responses, undirected search trajectories, and flat associative hierarchies for correct responses. We also found that the pattern of associative weights between concepts in our semantic network was integral to the performance of our model.

The success of the Metropolis-Hastings strategy coincides with recent MCMC models in computer vision, artificial intelligence, and computational cognitive sciences. Many of these accounts take a rationalist perspective and analyze cognition in terms of the problems it solves. Under this view, and assuming that the memory system represents a solution to the problem of information identification and retrieval (Anderson & Milson, 1989), search strategies that conserve resources by balancing computational loads between the search algorithm and the cognitive architecture will be favored over more complex, structure-independent methods. Our analyses using an MCMC model on a shuffled network, in conjunction with previous evidence using random walks in singly-constrained search, provide preliminary evidence to support this possibility. Whether or not the structure of the memory system underlies the success of these simple search strategies is a question worthy of further investigation.

Acknowledgments. This work was supported by grant number N00014-13-1-0341 from the Office of Naval Research.

References

Abbott, J., Austerweil, J., & Griffiths, T. L. (2012). Human memory search as a random walk in a semantic network. *Advances in Neural Information Processing Systems*, *25*, 3050–3058.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703–719.

Bolte, A., & Kuhl, T. G. J. (2003). Emotion and intuition: Effects of positive and negative mood on implicit judgments of semantic coherence. *Psychological Science*, *14*(5), 416–421.

Bowers, K. S., Regehr, G., Balthazard, C., & Parker, K. (1990). Intuition in the context of discovery. *Cognitive Psychology*, *22*(1), 72–110.

Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: Series D (the Statistician)*, *47*(1), 69–100.

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., & Mednick, S. C. (2009). Rem, not incubation, improves creativity by priming associative networks. *Proceedings of the National Academy of Sciences*, *106*(25), 10130–10134.

Carroll, P., Rimes, G., Kintsch, W., Mann, L., Streeter, L., & Thomas, K. (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.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Fodor, E. M. (1999). Subclinical inclination toward manic-depression and creative performance on the remote associates test. *Personality and Individual Differences*, *27*(6), 1273–1283.

Forbach, G. B., & Evans, R. G. (1981). The remote associates test as a predictor of productivity in brainstorming groups. *Applied Psychological Measurement*, *5*(3), 333–339.

Gabora, L. M. (2000). The beer can theory of creativity. In P. Bentley & D. Corne (Eds.), *Creative evolutionary systems.* San Francisco: Morgan Kauffman.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*(1), 1–24.

Griffiths, T. L., Steyvers, M. M., & Firl, A. (2007). Google and the mind predicting fluency with pagerank. *Psychological Science*, *18*(12), 1069–1076.

Gruenewald, P., & Lockhead, G. R. (1980). The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(3), 225.

Gupta, N., Jang, Y., Mednick, S. C., & Huber, D. E. (2012). The road not taken creative solutions require avoidance of high-frequency responses. *Psychological Science*, *23*(3), 288–294.

Harkins, S. G. (2006). Mere effort as the mediator of the evaluation-performance relationship. *Journal of Personality and Social psychology*, *91*(3), 436–455.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440.

Isen, A. M., Daubman, K. A., & Nowicki, G. P. (1987). Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology*, *52*(6), 1122–1131.

Martindale, C. (1995). Creativity and connectionism. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach.* Cambridge, MA: MIT Press.

Mednick, M. T., & Andrews, F. M. (1967). Creative thinking and level of intelligence. *The Journal of Creative Behavior*, *1*(4), 428–431.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

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*, *36*(3), 402–407.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web* (Tech. Rep. No. 66). Stanford, CA: Stanford University, InfoLab.

Paulus, P. B., Levine, D. S., Brown, V., Minai, A. A., & Doboli, S. (2010). Modeling ideational creativity in groups: Connecting cognitive, neural, and computational approaches. *Small Group Research*, *41*(6), 688–724.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93.

Smith, K. A., Huber, D. E., & Vul, E. (2013). Multiply-constrained semantic search in the remote associates test. *Cognition*, *128*(1), 64–75.

Ullman, T., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*.

Vul, E., & Pashler, H. (2007). Incubation benefits only after people have been misdirected. *Memory and Cognition*, *35*(4), 701–710.

Ward, J., Thompson-Lake, D., Ely, R., & Kaminski, F. (2010). Synaesthesia, creativity and art: What is the link? *British Journal of Psychology*, *99*(1), 127–141.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide* (Tech. Rep.). Brewster, NY: Touchstone Applied Science Associates.