Investigating Category Transitions and Interresponse Times in Fluency Tasks

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Dear Reviewers,

Thank you for the opportunity to submit our manuscript, *"Investigating Category Transitions and Interresponse Times in Fluency Tasks."* This work was reviewed previously. Follow-up analyses from reviewer suggestions took several weeks computationally. The editor requested that we resubmit as a new submission with a cover letter describing the original responses to reviewers, which is below. We are grateful for your thoughtful and constructive feedback, which has substantially improved the quality and clarity of our work. The final version of the manuscript incorporates all of the requested changes, and we are confident that the revisions offer a stronger theoretical contribution and address each concern in full.

Summary of Key Revisions and Final Updates:

- **Clarification of IRT Calculation (p. 5):**
  We clarified how interresponse times (IRTs) were computed—from the final keypress ending one response to the first keypress of the subsequent response—and explained how this method may slightly differ from IRTs in spoken fluency tasks. To contextualize this difference, we now include a comparison to mean IRTs reported in a spoken fluency study (Orologa et al., 2024), noting that our values are slightly elevated due to typing delays but do not affect the pattern of post-transition slowdowns.

- **Finalization of Transition Point Analysis (pp. 14–15, 24–26, 31):**
  Following reviewer feedback and additional analysis, we updated the manuscript to reflect that across **all methods of defining transitions**—including participant similarity ratings, delta-similarity, norms-based categorical and associative metrics—the **longest IRT consistently occurred at the post-transition point (t+1)**. This contradicts prior expectations that the longest pause would occur **during** the semantic switch (t0) and suggests a fundamental shift in how memory foraging should be modeled.

- **Addressing Reviewer Concern on t0 Significance (p. 15):**
  We explicitly note that t0 was also significantly elevated compared to most other time points in Experiment 1, though it did not exceed t+1. This helps clarify that transitions incur a retrieval cost, but the maximal delay appears during initiation into the new category, not the switch itself.

- **Robustness Across Algorithmic and Behavioral Measures (pp. 14–15, 24–26, 30–31):**
  We now emphasize that the t+1 peak is **robust across all similarity metrics**, including those that introduce a greater number of "micro-switches" within a broad category (e.g., eagle to pigeon). This strengthens our conclusion that the

observed pattern is not a methodological artifact, but a stable characteristic of cognitive search.

- **Theoretical Integration of Fluid vs. Static Category Boundaries (pp. 31–32):** We now offer a clearer discussion of the implications of our findings for theories of memory organization, suggesting that category transitions may not reflect fixed boundaries in a semantic space but instead emerge dynamically in real time. We interpret the t+1 slowdown as a consequence of reorienting within a newly activated semantic patch, rather than traversing between topologically distant locations.

- **Expansion of Experiment 2 and General Discussion (pp. 24–26, 30–32):** We revised and expanded several sections of the Experiment 2 Discussion and General Discussion to reflect new findings and theoretical developments. These changes improve coherence across sections and directly address concerns about the consistency and implications of the IRT patterns observed.

We believe the current version of the manuscript meaningfully advances our understanding of memory foraging by demonstrating a consistent post-transition slowdown that holds across different similarity metrics and task types. The findings have implications for how we model semantic search and challenge previous assumptions about where cognitive "travel costs" are incurred.

Thank you for your consideration. We look forward to your feedback.

Sincerely,
Corey M. Magaldino and Eric L. Amazeen
Arizona State University
cmagaldi@asu.edu

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Investigating Category Transitions and Interresponse Times in Fluency Tasks

Corey M. Magaldino and Eric L. Amazeen

Arizona State University, Department of Psychology, Tempe, Arizona, US

Email Correspondence: cmagaldi@asu.edu

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Abstract

Semantic fluency tasks involve recalling items from a given category (e.g., animals). It is well

documented that these tasks produce heavy-tailed distributions of interresponse times (IRTs).

Heavy-tailed distributions have been observed in a variety of contexts promoting efficient

search. The current work investigates the role of categorical transitions within a single semantic

category, multiple semantic categories, and non-semantic categories (e.g., letter categories).

Counterintuitively, findings suggest the longer IRTs requisite for producing heavy-tails did not

occur at the categorical transitions. Rather, the longest IRTs occurred immediately after

switching categories. This work highlights similarities in foraging patterns across different

domains from the physical and spatial to the cognitive and abstract.

Keywords: heavy-tailed distributions, semantic fluency, memory foraging, free recall

Investigating Category Transitions and Interresponse Times in Fluency Tasks

Humans can typically retrieve several items from memory quickly and without much effort. For instance, if asked to name items from a grocery store or shopping mart, people generally can begin producing responses almost instantaneously. Phenomenologically, the experience starts effortlessly, but retrieval of novel items becomes more difficult as time progresses. Cognitive psychologists have been trying to better understand this retrieval process for the better part of a century (Bousfield & Sedgewick, 1944; Williams & Hollan, 1981; Tombaugh et al., 1999; Henry et al., 2004; Mioshi et al., 2006; Maseda et al., 2014).

The verbal fluency task has been historically used to probe the memory retrieval process (Bousfield & Sedgewick, 1944). Verbal fluency tasks have two prototypical paradigms: semantic fluency and letter fluency (Zemla, Cao, Mueller, & Austerweil, 2020). The difference between the two variants is whether participants generate responses from a semantic category (e.g., animals) or a letter category (e.g., words that start with the letter 'S'). In this experimental paradigm, participants are typically asked to produce as many responses as possible from a particular semantic category in a fixed amount of time (i.e., "name as many animals as you can in 10 minutes"). The verbal fluency task has been used to identify schizophrenia (Sung et al., 2012; Sung et al., 2016), Alzheimer's Disease (Monsch et al., 1992; Locascio et al., 1995; Salmon et al., 2002; Zemla & Austerveil, 2019), as well as assess general cognitive performance (Abwender et al., 2001; Hills, Mata, Wilke, & Samanez-Larkin, 2013; Lezak, 1995; Troyer et al., 1997; Troyer, 2000).

Early work in semantic fluency tasks argues that human memory retrieval is a process of sampling with replacement (Bousfield & Sedgewick, 1944). In this description, responses are selected at random and the characteristic 'slowing down' in number of responses and

interresponse times (IRTs) is explained by reselecting already sampled responses, which need to be discarded. In this explanation, as the time within a given trial increases, the IRT increases as a function of resampling the same items more frequently. Modern modeling approaches support the sampling-with-replacement interpretation and suggest an exponential decay function adequately describes the retrieval process from semantic memory (Wixted & Rohrer, 1994).

Heavy-tailed Distributions

However, more recently, Rhodes at Turvey (2007) adopted an approach that primarily focused on the time intervals between retrievals rather than the exponentially decaying rate of recall. Rhodes and Tuvey (2007) reported that distributions of IRTs conformed to a Lévy distribution – a specific type of heavy-tailed distribution. The Lévy distribution is characterized by a high peak and a heavy-tail when compared to the typical normal distribution (see Figure 1). In general, the heavy-tailed description is caused by a higher frequency of extreme events (e.g., extremely long pauses between responses in this case). In addition to the heavy-tails, these distributions are characterized by frequent, short bursts of activity which, in turn, produce the prominent sharp peak (see Figure 1). In the context of free recall tasks, short bursts would correspond to several items being recalled in very rapid succession.

This class of distributions is not only known for their high peaks and heavy-tails, these distributions are commonly observed in animal foraging behavior. Examples of species that have shown heavy-tailed distributions in their movement trajectories while foraging include but are not limited to: honeybees (Reynolds, Smith, Reynolds, Carreck, & Osborne, 2007), fruit flies (Reynolds & Frye, 2007), spider monkeys (Boyer et al., 2006), the albatross (Viswanathan et al., 1996), the Cory's shearwater (Focardi & Cecere, 2013), the Great White shark (Sims, Humphries, Bradford, & Bruce, 2011), and the Bigeye tuna (Humphries et al., 2016). Moreover,

the movement patterns of humans in contexts ranging from hunter-gatherer tribes (the Ju/'hoansi; Brown, Liebovitch, & Glendon, 2006; Lee, 1979, 1993; Lee and DeVore, 1976; the Hadza group of Tanzania; Raichlen et al., 2013; the Me`Phaa of Mexico, the Brazilian Cariri farmers, and Amazonian farmers; Reynolds, 2018) to modern everyday human behaviors and movements (Rachev & Mittnik, 2000; Voit, 2005; Jia et al., 2012; Rhee et al., 2011). For example, humans typically make few large movements throughout the day (e.g., commuting from home to work) accompanied by much more frequent small bursts of activity (e.g. small movements around the workplace).

In animal ecology, the observed heavy-tailed distributions are created by coupling of short bursts at the mean with a higher-than-normal probability for the extreme values at the tails. Spatially, this is observed as many small distance movements and a higher-than-normal probability for long distance movements to occur. In short, the shape of the distribution relies on these long-distance movements to produce the hallmark heavy-tails. In cognitive search, the shape of the distributions relies on the existence of long IRTs to produce the heavy-tails. It follows that these long IRTs should be occurring between semantically dissimilar items. If semantic memory is organized based on patches of semantically related items, the long IRTs should correspond to traveling between distant semantic patches. In other words, the long IRTs required to produce a heavy-tailed distribution should be occurring at the transition between categories or subcategories. These categorical transitions may be the key to understanding the relationship between semantic recall and the prevalence of heavy-tailed distributions in fluency tasks.

Optimality

Although it is apparent that heavy-tailed distributions are observed in a variety of contexts, findings suggest that this pattern is seen in so many contexts because species evolved to select the most efficient food search strategies (Wosniack et al., 2017; Humphries et al., 2010). In other words, these distributions are so prevalent because they are typically the product of an optimal strategy and the literature supports this claim. Empirically, organisms that rely on a heavy-tailed foraging strategy show the lowest rates of starvation (Humphries & Sims, 2014). Furthermore, actual and simulated foraging trajectories show a heavy-tailed process as being close to the optimal strategy (Viswanathan et al., 1996, 2002; Reynolds & Frye, 2007; Reynolds et al., 2007). Heavy-tailed distributions have been observed in the most ancient forms of foraging and have been shown to be within 2% of the optimal route in some contexts (Reynolds, 2011a; Sims et al., 2014).

The ubiquity of these distributions in foraging paired with its presence in free recall as reported by Rhodes and Turvey (2007) has direct implications for the memory retrieval process. It could be the case that retrieving words from a particular semantic category is a similar process to foraging for food in nature, or the two are at least governed by a similar dynamic. Albeit speculative, it could be the case that a heavy-tailed process developed in such distinct contexts because it is related to efficiency and optimality – a valuable trait for memory retrieval considering there is no known limit on human's long-term memory storage (Cowan, 2008).

Although Rhodes and Turvey (2007) were the first to report a heavy-tailed distribution in human memory through the retrieval of animal names, recent research has demonstrated their utility in the cognitive space; namely, heavy-tailed distributions have been observed in the free recall of country names (Patten et al., 2020) as well as the movements people use to identify objects they cannot see (Magaldino & Amazeen, 2021). Proponents of these heavy-tailed

explanations argue that the presence of heavy-tailed distributions occurs as a product of highly

efficient, near-optimal patterns of behavior to forage for resources in a patchy environment.

Other work has characterized memory retrieval as analogous to optimal foraging for food

when resources are distributed in patches (Szary, Dale, Kello, & Rhodes, 2015; Hills, Jones, &

Todd, 2009; 2012; Hills, Todd, & Jones, 2015). There is consensus in the literature that semantic

search is optimal through the lens of classic optimal foraging theory (Abbot, Austerweil, &

Griffith, 2015; Charnov, 1976; Hills, Jones, Todd, 2012). Classically, optimal foraging theory

argues that a foraging strategy is optimal if an agent leaves a current resource patch when the

rate of food within the current patch falls below the overall average rate of food acquisition (also

known as the Marginal Value Theorem; see Charnov, 1976). Routinely, results demonstrate

similarities between spatial foraging and memory retrieval. That is, results indicate that

individuals search locally through memory based on similarity until producing another response

within a similar cluster is more difficult than jumping to a new cluster (Abbot, Austerweil, &

Griffith, 2015; Hills, Jones, Todd, 2009; 2012). The conclusion drawn from this work is that

humans are foraging through semantic space and when a given region is depleting, they move to

a new region of the semantic space (Hills, Todd, & Jones, 2015). This local-to-global patterning

is precisely the type of process to produce heavy-tailed distributions seen more recently in the

cognitive domain (Szary et al., 2015; Rhodes & Turvey, 2007; Patten et al., 2020; Magaldino &

Amazeen, 2021).

Semantic Distance

Given that so many independent studies have come to similar descriptions comparing

memory retrieval to spatial foraging, it is critical to understand the semantic topology an

individual is traversing. Rhodes and Turvey (2007) measured IRTs as a proxy for measuring the

distance traversed in semantic space. More recently, Patten et al. (2020) extended this work and

measured the distance between responses in two ways. First, Patten et al. (2020) had participants

retrieve countries and used the geographical distance between the actual responses as a proxy for

semantic distance. Second, Patten et al. (2020) used multidimensional scaling based on

participant-generated similarity ratings to create a semantic topography. Multidimensional

scaling collapses unmeaningful dimensions as a means to represent the level of similarity

between items in a dataset (Shephard, 1980). In this context, multidimensional scaling is similar

to constructing a semantic network for each individual participant (Patten et al., 2020).

Another common approach to approximating the semantic distance participants are

traversing involves using a standard vector-based model (such as HAL or word2vec). In these

approaches, responses are vectorized and distance is calculated as cosine similarity between

vectors (Johns, Mewhort, & Jones, 2019). For instance, Hills, Todd, & Jones (2015) used

semantic representations produced from a BEAGLE vector-based model (Jones, Kintsch, &

Mewhort, 2006; Jones & Mewhort, 2007). Similar to many linguistic machine-learning

algorithms, BEAGLE was trained on a Wikipedia corpus with millions of entries (Willits,

D'Mello, Duran, & Olney, 2007). Once trained, the BEAGLE algorithm can produce

measurements of pairwise similarity between words based on their co-occurrence and relatedness

in the corpus (Jones et al., 2006). These vector-based machine learning algorithms calculate

similarities based on translating words to vectors that represent meaning. The pairwise

similarities offer a proxy for the structure of the semantic space where participants are believed

to be foraging (Hills, Todd, & Jones, 2015).

Alternatively, natural language processing tools, like WordNet, offer ways to

approximate the semantic space participants are traversing. WordNet is a human-curated lexical

database that organizes the structure of concepts based on ontological relationships rather than

co-occurrence in a given corpus (see Miller, 1995; Fellbaum, 1998). In WordNet, words are

grouped into sets of synonyms, known as synsets. Each synset represents a distinct concept and

synsets are interlinked through semantic and lexical relations, resulting in a full network of

related words and concepts. The WordNet database is inspectable and navigable by users, which

allows for the construction of semantic distance between ontological elements. One such metric,

the Wu-Palmer Similarity Measure, is commonly used to model the relatedness between words

in natural language processing (Banches, 2013; Malik, Subramaniam, & Kaushik, 2013;

Oussalah & Mohamed, 2021). The Wu-Palmer Similarity measure is an edge-counting method

based on the distance between two ontological nodes and the distance that separates the two

nodes from their closest common ancestor (Wu & Palmer, 1994).

Each of these semantic methods makes assumptions about how humans represent words

and concepts (Kumar, 2021; Brychcin & Konopik, 2014). Whether using a lexical database or a

vectorized word embedding model, there are still concerns about accurately capturing the

nuances of lexical semantics. For instance, edge counting methods may equate a twin sibling and

a long-distant uncle as equally similar because they both exist one step away from the twins'

parents. On the other hand, vectorized models may be biased against infrequently occurring

similarities; for instance, accurately identifying the similarity between falcons and dolphins, not

based on their animal meaning but as being related to football teams from the southeast, is

difficult. The current study opted to use similarity ratings provided by participants to ground

responses in the participants' own native space. This was true for Experiment 1; however,

Experiments 2 and 3 took advantage of objective category transitions defined by the task. Given

shortcomings with other methods, we believed allowing a participant to simply report how

similar they believed their responses were would be the most accurate illustration of what they

believe to be similar. Similarity ratings will serve as a proxy for the semantic distance between

participant responses.

Current Study

First, the current study aims to replicate prior findings from Rhodes and Turvey (2007)

and Patten et al. (2020) that IRTs from free recall during a semantic fluency task adhere to a

heavy-tailed distribution. Furthermore, we predict that similarity ratings associated with

responses will also adhere to a heavy-tailed distribution, suggesting that the IRTs and semantic

distances are correlated and represent a trajectory traversed during free recall. Second, the

current study will investigate whether the long IRTs requisite for a heavy-tailed distribution are

occurring at the transition point between categories or subcategories. We predict long IRTs to be

coupled with dissimilar responses if the IRT is a product of traversing semantic space. If results

demonstrate that long IRTs are occurring at category switches, this finding will support a heavy-

tailed explanation that participants are making semantic leaps from one resource-abundant zone

(e.g., the subcategory of reptiles) to another resource-abundant zone (e.g., the subcategory of

mammals). However, if findings demonstrate that long IRTs are a product of exhausting a

particular category, this finding will support that our memory retrieval process more closely

adheres to sampling-with-replacement, where longer IRTs are caused by repeatedly resampling

and discarding previous responses.

To clarify, the 'exhaustion' hypothesis posits that as participants repeatedly sample items

from the same category, they increasingly re-encounter already-supplied items in their internal

search, discarding them until a novel candidate emerges. If this were the driving mechanism, we

would predict a more gradual buildup of IRTs over a 'run' of items within that category—

culminating in a distinct peak when only a final (or near-final) item can be retrieved. In other words, each new item from the same category should take longer than the last, because the chance of 'resampling' previously used items increases, and so does the difficulty of finding fresh exemplars.

By contrast, a classic foraging account would suggest that once the local 'patch' of items becomes less efficient, the participant is likely to incur a large 'travel time' (a longer IRT) precisely at the moment of switching to a new category. If the exhaustion hypothesis fully explained our data, we might see the longest IRT at or near the end of a long category run, immediately before a switch occurs.

Ultimately, it is well-documented that participants switch between clusters or semantically related categories during semantic fluency tasks (Bousfield & Sedgewick, 1944; Troyer et al.,1997; Hills et al., 2012). The goal of the present research is to better understand the relationship between IRTs and semantic similarity when participants are put in circumstances that require several subcategory switches (e.g., free recalling items from a given category for 10 minutes). The general theme is that by probing the patterns during the retrieval process, we may gain insight into the underlying cognitive processes. The present experiments explore these processes by examining category transitions both within a single domain (Experiment 1) and across multiple domains (Experiments 2 and 3), highlighting how the similarity measures and transitions are operationalized.

Experiment 1

Experiment 1 required participants to freely recall items from a single semantic category (animals) to probe the foraging behavior through subcategory clusters. In Experiment 1, we tested whether the distribution of IRTs and similarity ratings adheres to a heavy-tailed

distribution as predicted by several studies in both the cognitive search and animal foraging literature. If participants were foraging through semantic space during free recall, both the IRT distribution and the similarity rating distribution should be heavy tailed. Furthermore, participants' IRTs and similarity ratings should be negatively correlated, such that a long period of time between responses should correspond to semantically dissimilar responses. Long IRTs should correspond to category switches, defined as response pairs associated with low self-report similarity ratings. That is, the longest IRTs and lowest similarity should be related to global transitioning from one patch (e.g., reptiles) to another patch (e.g., mammals). If results fail to corroborate these predictions, the pattern of results may conflict with the explanation that participants are navigating a semantic topology during these cognitive search tasks.

Method

Participants

A total of 50 students enrolled in introductory psychology courses at Arizona State University were recruited to voluntarily participate in this study in exchange for partial course credit. A priori power analyses revealed approximately 45 participants were required for the experiment to be adequately powered to detect effects reported in Patten et al. (2020). Although Patten et al. (2020) report multiple effects, the medium effect size was selected as the most conservative estimate. Each participant provided informed consent prior to participation.

Materials

The experiment was run on a Dell Workstation computer with Microsoft Windows XP. The experiment was conducted, and data was collected through PsychoPy v. 2021.2.3 (Pierce, Gray, Simpson, MacAskill, Hochenberger, Sogo, Kastman, & Lindelov, 2019).

Procedure

Participants were asked to list as many animals as they could for 10 minutes by typing their responses into a word processing program made with PsychoPy v. 2021.2.3. Participants were instructed to avoid repeating responses; however, participants were told not to worry if they realized an inadvertent repetition in hindsight. Participants were instructed to prioritize the speed at which they entered each response and that they should not be concerned with spelling. Participants were only told the identity of the relevant category when they were ready to start to avoid preemptively storing multiple responses. After recalling animals for 10 minutes, the fluency portion of the experiment was complete. Inter-response times (IRTs) were computed as the elapsed time between a response and its subsequent response—that is, from the final keypress ending one word to the first keypress of the next word. This measure excludes the actual production time required to type a word, though it may still capture residual planning or editing delays inherent to typed responses.

Immediately following, participants were prompted to rate, on a 0–10 scale, how similar they perceived each adjacent pair of responses to be. We instructed them: 'A 0 means the two items are not similar at all. A 10 means they are extremely similar in any way you see fit—for instance, because they belong to the same subcategory of animal, share a habitat, or any other similarity you notice.' We did not further constrain or define 'similarity.' As a result, many participants produced ratings close to extremes (0 or 10), consistent with a dichotomous view of similarity.

To identify a subcategory transition, we labeled any pair with a rating below 5 as a 'low-similarity' transition. Although this threshold was partly arbitrary, it captured the stark rating drop participants used to indicate a switch (e.g., going from 'whale' to 'lion'). For future studies, we note that a 'delta-similarity' method (Hills et al., 2012; Lundin et al., 2023) or a vector-based

CATEGORY TRANSITIONS AND IRTS 14

similarity approach (Kumar et al., 2023) might yield more nuanced identification of transitions, including rapid subcategory switching. Participants' own pairwise responses were displayed on the screen sequentially, meaning participants were asked to rate the similarity between their first and second response, their second and third response, and so on sequentially until they had rated all of the responses they provided. Entries were later corrected for any misspellings.

Results

Distribution Estimation

Participant IRTs and similarity ratings were first modeled at the global level as histograms to inspect the general distribution shape observed in all the Experiment 1 IRT data (see Figures 2 and 3). Visually, the histogram of IRT has the predicted high peak and heavy-tail; however, similarity ratings produced a bimodal distribution. Participants seemed to anchor at the ends of the similarity scale, essentially treating it as a categorical variable with responses clustering at extremely dissimilar and extremely similar (see Figure 3). Because the general shape of the histogram does not conform to any of the distributions of interest, similarity ratings were no longer used in future experiments.

Three distributions were fit to the empirical IRT data using maximum likelihood estimation (MLE; see Figure 4). The distributions being compared were the normal distribution and two classes of distributions that accommodate heavy tails (the power law distribution and the lognormal distribution). We compared the empirical distribution of IRTs with 1,000 simulated iterations for each of the distributions listed above (normal, power law, and lognormal) and tested for goodness of fit using a Kolmogorov-Smirnov (KS) test. At the global level, simulations revealed that 0 iterations conformed to a normal distribution in Experiment 1. In contrast to Patten et al. (2020) and Szary, Kello, and Dale (2015), we found 0 iterations

conformed to a power law distribution in Experiment 1 at the global level. The best fitting

distribution for the global data in Experiment 1 was the lognormal distribution with 736 of 1,000

iterations failing the KS test – meaning they failed to reject the null hypothesis that the simulated

data and empirical data were from the same continuous distribution (see Table 1 for simulation

results).

Following the global inspection of the IRT distribution shape, the distribution shape was

tested at the individual participant level. The same simulation method was employed where each

individual participant's empirical distribution was compared to 1,000 simulated iterations for the

distributions of interest (normal, power law, and lognormal). At the individual level, results

mirrored those at the global level. Few iterations conformed to a normal distribution in

Experiment 1. On average, 67.24 ($SD = 205.48$) of 1,000 normal distribution iterations failed to

reject the null hypothesis of the KS test. Similar to Patten et al. (2020), we found on average

651.48 ($SD = 424.64$) of 1,000 power law distribution iterations failed to reject the null

hypothesis of the KS test. Lastly, our data suggests that the lognormal distribution fit best with

959.96 ($SD = 84.37$) iterations failing to reject the null hypothesis of the KS test on average. The

pattern of results from the simulation analysis provides support that heavy-tailed distributions

(both the power law and lognormal) fit the empirical distribution of IRTs substantially better

than a normal distribution.

Relationship between IRT and Similarity Ratings

We predicted a strong, negative correlation between participant IRTs and their

corresponding similarity ratings. If participants were foraging a semantic landscape while

performing a free recall task, it follows that pauses between provided responses should have a

direct relationship with the semantic similarity between responses. We predicted that the longer

IRTs would be associated with dissimilar responses. Participants IRT and similarity ratings were weakly negatively correlated at the global level, $r(3271) = -0.09$, $p < .001$. At the individual level, only 2 of 50 participants had a significant correlation between IRT and similarity ratings. Participants generally treated similarity ratings as a categorical variable with a dichotomous outcome at either end of the scale (similar or dissimilar) rather than a ratio scale of similarity, which we believe dampened the strength of this relationship.

In response to reviewer feedback, we performed additional analyses to see if alternative similarity metrics revealed a stronger relationship with IRT. Specifically, we computed a "semantic similarity" measure using the forager algorithm (see Kumar et al., 2023). Across these methods, we found that the algorithmic semantic similarity reliably produced a stronger negative correlation with IRT than the original behavioral similarity ratings. The correlation remained significant ($r(3278) = -0.25$, $p < .001$), reaffirming that participants' typed responses and broad 0–10 rating scale likely introduced variability that weakened the linkage between IRT and semantic distance. We therefore retained the main findings reported here while acknowledging that future studies might optimize similarity measures or use vector-based algorithms (Kumar et al., 2023) to capture finer-grained distance metrics.

Transition Identification

We aimed to test if the long IRTs associated with a heavy-tailed response distribution were occurring at transition points between subcategories of animals. Participants' similarity ratings were used to identify the point at which transitions between categories occurred. When participants provided a low similarity rating (below 5), it was assumed that they were moving between categories. Items within subcategories generally had a high degree of similarity. For instance, if a participant was traversing through the 'aquatic animal' subcategory, we would

expect things like orca and whale to be similar; however, when they decide to leave the 'aquatic animal' subcategory and move to a new subcategory, the similarity ratings between those items were typically much lower (if the participant went from 'whale' to 'lion' for example). After identifying all of the subcategory transition points, we aimed to test if the average transition IRT was significantly longer than the surrounding average IRTs. To avoid ambiguity, transitions were defined from an participant entry and the subsequent entry. For each transition, we gathered the previous 3 responses and the following 3 responses to construct a 7-point pre/post transition series (see Figure 9). It is worth noting that multiple transitions could occur within a short span; however, excluding rapid, back-to-back transitions did not change the core outcome.

Furthermore, to explore whether alternative transition definitions would yield different results, we repeated these transition analyses using the additional transition methods. We computed transitions using (a) a "delta similarity" measure, which looks at the relative drop in similarity compared to adjacent pairs (Hills et al., 2012; Lundin et al., 2023), (b) a norms-based categorical measure, and (c) norms-based associated measure (based on Troyer et al., 1997 and extended by Zemla et al., 2020). In the delta-similarity approach, a transition was flagged whenever the similarity score showed a stark relative drop (rather than an absolute rating threshold). In the norms-categorical approach, we identified whether two consecutive items belonged to distinct subcategories (e.g., mammal vs. bird) based on standardized norms. Lastly, in the norms-associative approach, we considered transitions when two items had a low normative association strength. Across each of these alternative similarity measures, the pattern of average IRTs remained consistent. Using several algorithmic approaches (delta-similarity, norms-categorical, norms-associative) the heightened IRT once again appeared at the post-transition position ($t+1$), matching the behavioral similarity rating result. This convergence

shows that the *t+1* slowdown is robust to how switches are detected and is therefore unlikely to be an artefact of any single similarity metric. (see Figures 9-12 for transition plots for each transition method; see Tables 2 and 3 for switch comparisons across transition methods).

A repeated measure analysis of variance revealed global differences in average IRT based on behavioral transition point, $F(6,294) = 14.02$, $\eta^2_{partial} = 0.22$, $p < .001$. Least significant difference (LSD) post-hoc pairwise comparisons revealed average IRT for the *t+1* transition significantly differed from all other transition points, all *ps* < .05. Additionally, LSD post-hoc pairwise comparisons revealed average IRT for the *t0* transition point significantly differed from all other transition points (all *ps* < .05) except *t+2* ($p = .081$).   This result suggests that the longest IRT did not occur immediately at the transition point. Instead, the longest IRT was observed immediately following the transition.

Discussion/Significance

Experiment 1 aimed to test the presence of heavy-tailed distributions in both IRTs and similarity ratings. Results suggest the presence of a heavy-tail in the distribution of IRTs. Simulation work demonstrated that the lognormal distribution fits the present data rather well at both the global level and the individual level when compared to a normal non-heavy-tailed distribution. Simulations provided mixed results for the power law distribution. At the individual level, the power law distribution fit approximately 65% of iterations; however, at the global level, there was no evidence that a power law distribution fit the shape of the empirical distribution observed from participant IRTs. Similarity ratings did not conform to any distribution tested due to the bimodal nature of participant responses. Essentially, participants gravitated toward the anchor points at each end of the spectrum and did not fully utilize the range

of the similarity scale. Similarity ratings were employed as a proxy for semantic distances; however, future work is needed to better capture the relatedness between participant responses.

Furthermore, Experiment 1 aimed to test if there is a relationship between IRT and similarity ratings, suggesting that longer IRTs correspond to traversing between dissimilar items. Although there was a significant relationship between IRT and similarity ratings at the global level, the strength of the relationship was almost nonexistent. Given the shortcoming of similarity ratings, there is little evidence of a link between animal foraging models and human memory models whereby humans are foraging through a semantic landscape during free recall. There is reason to believe a similar dynamic system would evolve for organisms in environments where resources are distributed into patchy clusters, especially if that system is optimal at obtaining resources; however, the failure of similarity ratings to capture inter-word relatedness made detecting this relationship difficult.

Interestingly, we predicted that the longest IRTs would occur at transition points. Results indicate that, on average, the longest IRTs were occurring immediately following a transition. To put this into context, imagine a participant is recalling animals and they are in the 'aquatic animal' subcategory and switching to the 'farm animal' subcategory. Sample responses may be: 'orca', 'whale', 'cow', 'goat.' In this scenario, our results portray the longest IRT occurring between 'cow' and 'goat' and not at the transition from 'whale' to 'cow,' as we predicted. This finding suggests that the heavy tails routinely observed may not be caused by the IRTs at the transitions themselves, but instead, the heavy tail of the distribution may be caused as participants begin to generate responses in a new subcategory.

Prior work has demonstrated heavy-tailedness in cognitive search (Magaldino & Amazeen, 2021; Patten et al., 2020; Rhodes & Turvey, 2007). Experiment 1 replicated heavy-

tailed patterning in free recall observed in the literature; however, there is considerably less work investigating subcategory transition points. The current work aimed to study these transition points. Results from Experiment 1 demonstrate that the longest IRTs on average occurred after a subcategory transition between dissimilar responses. This finding suggests that there may not be a massive, static semantic topology that participants are traversing – where all categories and subcategories are not presently waiting for a would-be forager to select their respective responses. Instead, results align more closely with the notion that categories emerge and disappear dynamically based on task demands. We conclude Experiment 1 by arguing that the pattern of results reported here suggests participants spontaneously acquire a new item from another category or subcategory with no significant delay from the previous response from a different category. The increase in response time we observe is time for the new category to self-organize and emerge once transitioned.

Finally, our analyses using delta similarity, norms-based categories, and norms-based association reveal that this $t+1$ effect (the highest IRT appearing after the transition) is not specific to only the behavioral similarity ratings. When we define transitions algorithmically (e.g., using delta similarity or standardized category norms), the maximum IRT frequently aligns with the post transition point ($t+1$). Regardless of whether we used behavioral ratings or algorithmic methods to identify transitions, the highest IRT consistently occurred immediately after the switch ($t+1$). This convergence of findings- despite the algorithmic methods introducing more intra-category "micro-switches" - suggests the $t+1$ effect is a robust feature of recall dynamics. For example, transitions such as "eagle" to "pigeon" might be grouped together by a norms-based similarity metric due to both being birds, but participants may view them as

semantically distinct based on personal criteria (e.g., predator vs scavenger). This reinforces the idea that semantic similarity is not one-dimensional.

Experiment 2

Although the first experiment probed local-to-global switches between categories, all responses were constrained to be within the single semantic category associated with animals; therefore, category switches were switches between subcategories of animals in actuality. Experiment 2 was less constrained and allowed for an additional semantic category for participants to sample – essentially widening the semantic space participants are operating within. Experiment 2 asked participants to freely recall either animals or vegetables in any fashion they would like. By providing two global semantic categories, we can test whether participants exhaust one global category before making a long-distance leap to another global category. Although prior work has demonstrated heavy-tailed distributions in free recall IRTs, more work is required to better understand what produces the reported heavy-tails. Experiment 1 found evidence that the heavy-tails are produced by the response immediately following a transition. Experiment 2 tests if that finding persists with separate distinct categories (animals, vegetables) rather than subcategories of a single semantic category (e.g., animals). Furthermore, Experiment 1 relied on similarity ratings to identify when a subcategory transition occurred. An advantage of Experiment 2 is that there are objective category switches depending on whether the previous response was an animal or a vegetable. Results from Experiment 2, in comparison with Experiment 1, provide evidence about the nature of global category switches. Due to the dichotomous fashion participants adopted when reporting similarity ratings, further analysis of similarity ratings was not conducted.

Method

The same general method used in Experiment 1 was adopted for Experiment 2 with the only change being the addition of the second semantic category (vegetables).

Participants

A total of 46 students enrolled in introductory psychology courses at Arizona State University were recruited to voluntarily participate in this study in exchange for partial course credit. A priori power analyses revealed that 45 participants were required for the experiment to be adequately powered to detect effects reported in Patten et al. (2020). Although Patten et al. (2020) report multiple effects, the medium effect size was selected as the most conservative estimate. Each participant provided informed consent prior to participation.

Materials

The experiment was run on a Dell Workstation computer with Microsoft Windows XP. The experiment was conducted, and data was collected through PsychoPy v. 2021.2.3 (Pierce, Gray, Simpson, MacAskill, Hochenberger, Sogo, Kastman, & Lindelov, 2019).

Procedure

The procedure for Experiment 2 followed the same procedure as Experiment 1. The key difference between the two experiments is in the instructions prompted to participants. Participants were asked to list as many animals or vegetables as they could for 10 minutes by typing their responses into a word processing program made with PyschoPy v. 2021.2.3. The order of presentation for "animals" and "vegetables" in the instructions to participants was counterbalanced such that half of the participants saw "…list as many animals or vegetables…" and the other half of the participants saw "…list as many vegetables or animals…" Participants were free to enter either animals or vegetables at any time throughout the experiment.

Counterbalancing was done to minimize any bias the order of categories may have on participants' response patterns.

Results

The same general analysis plan as in Experiment 1 was applied to Experiment 2, excluding similarity ratings as noted above. Analysis first investigated the shape of the IRT distribution. Simulations were used to assess the goodness of fit between the observed IRT distribution and data simulated from known distributions. The distributions tested were two classes of heavy tailed distributions (the lognormal distribution and the power law distribution) and the normal distribution. Lastly, IRTs at transition points were compared to neighboring IRTs to determine if the longest IRTs were occurring at inter-category switches.

Distribution Estimation

Participant IRTs were first modeled at the global level as a histogram to inspect the general shape of the distribution observed in the Experiment 2 IRT data (see Figure 5). Upon initial inspection, the profile of the histogram of IRT conformed to the high peak and heavy-tail. Similar to Experiment 1, two heavy tailed distributions and the normal distribution were fit to the empirical data using MLE (see Figure 6). The empirical distribution of IRTs was compared with 1,000 simulated iterations for each of the distributions being tested (normal, power law, and lognormal). Goodness of fit was tested using KS tests in the same global fashion as Experiment 1. KS tests comparing the observed IRT distribution to each of the simulated distributions corroborated results from Experiment 1. Of 1,000 simulations each, 0 iterations conformed to data simulated from the normal distribution. We found 0 of 1,000 instances of simulated power-law data fit the empirical distribution. However, there was evidence that data produced from a lognormal distribution fit the global IRT distribution. In total, 705 of 1,000 simulated lognormal

datasets failed to reject the null hypothesis of the KS test (see Table 1 for simulation results).

Ultimately, data simulation results from Experiment 2 mirror results from Experiment 1.

Notably, at the global level, there was no evidence of distribution fit for the normal distribution

nor the power law distribution. Our results suggest that the lognormal distribution is the best

approximation of the observed IRT data at the global level.

Distribution shape was tested at the individual participant level. For each participant,

observed IRT data was tested against 1,000 iterations of simulated data derived from either a

normal distribution, a lognormal distribution, or the power law distribution. The normal

distribution generally did not fit individual IRT data with only 84.04 ($SD = 234.31$) of 1,000

iterations failing to reject the null hypothesis of the KS test. We found on average 567.37 ($SD =$

452.03) of 1,000 power law distribution iterations failed to reject the null hypothesis of the KS

test. Results from individual IRT distribution fitting suggested that the lognormal distribution fit

the empirical IRT data best with 999.91 ($SD = 0.46$) iterations failing to reject the null hypothesis

of the KS test on average. Results from distribution fitting in Experiment 2 follow the same

general pattern as results from Experiment 1. That is, there is little to no evidence that IRT data

is normally distributed. There is mixed evidence that IRT data is fit by a heavy-tailed

distribution for individual participant data. Approximately half (567.37 of 1,000, on average) of

simulated iterations fit the power law distribution at the individual level. Convincingly, nearly all

(999.91, on average) simulated lognormal iterations failed to statistically differ from the

individual IRT distributions.

Relationship between IRT and Similarity Ratings

We predicted a strong, negative correlation between participant IRTs and their

corresponding similarity ratings. However, given issues with the behavioral similarity ratings

observed in Experiment 1, we opted to use algorithmic methods of determining semantic

similarity. We performed additional analyses to see if alternative similarity metrics revealed a

stronger relationship with IRT. Specifically, we computed a "semantic similarity" measure using

the forager algorithm (see Kumar et al., 2023). Across these methods, we found that the

algorithmic semantic similarity reliably produced a stronger negative correlation with IRT than

the original behavioral similarity ratings. The correlation remained significant ($r(3023) = -0.29$,

$p < .001$), reaffirming that participants' typed responses and broad 0–10 rating scale likely

introduced variability that weakened the linkage between IRT and semantic distance. We

therefore retained the main findings reported here while acknowledging that future studies might

optimize similarity measures or use vector-based algorithms (Kumar et al., 2023) to capture

finer-grained distance metrics.

Transition Identification

Given evidence that IRT data is distributed with heavy-tails, we aimed to test if those

heavy-tails observed in response distributions were caused by local-to-global transitions between

categories. In Experiment 1, we relied on similarity ratings to determine whether or not a distant

transition occurred. A key advantage of Experiment 2 is that there were objective category

switches that were easily identifiable based on participant responses. In Experiment 2, transition

points occurred whenever a participant changed categories from "vegetable" to "animal" or vice

versa. Once all inter-category transition points were identified, we tested if the average IRT at

the transition point between categories was significantly longer than the neighboring response

IRTs. Using the same procedure as Experiment 1, we gathered the previous 3 responses and the

following 3 responses from the transition point to construct a 7-point series centered around the

transition point (see Figure 9).

Additionally, to explore whether alternative transition definitions would yield different results, we repeated these transition analyses using the additional transition methods. We computed transitions using (a) a "delta similarity" measure, (b) a norms-based categorical measure, and (c) norms-based associated measure. Across each of these alternative similarity measures, the pattern of average IRTs remained consistent. Using each of these transition methods, the heightened IRT again appears at *t+1*, mirroring the results when using objective category switches and the results from similarity ratings in Experiment 1 (see Figures 9-12 for transition plots for each transition method; see Tables 2 and 3 for switch comparisons across transition methods).

A repeated measure analysis of variance revealed global differences in average IRT based on transition point, $F(6, 294) = 11.83$, $\eta^2_{partial} = 0.19$, $p < .001$. Least significant difference (LSD) post-hoc pairwise comparisons revealed average IRT for the *t+1* transition significantly differed from all other transition points, all *ps* < .05. Although we predicted the longest IRT would occur at the inter-category transition, this result suggests that, on average, the longest IRT occurred immediately after the inter-category transition.

Discussion / Significance

The findings of Experiment 2 build upon those in Experiment 1, supporting the idea that participant IRTs are distributed in a heavy-tailed fashion. The rationale behind Experiment 2 was to test if the predictions from Experiment 1 persisted when the task was less semantically constrained. That is, Experiment 1 tested if our data produced heavy-tailed IRT distributions as observed in Rhodes and Turvey (2007) and Patten et al. (2020) for a single semantic category. By adding a second semantic category, Experiment 2 increased the space by which participants were free to forage. At the global level, the shape of the IRT distribution was analyzed using

simulations to compare the observed data with three candidate distributions: the normal

distribution, the power law distribution, and the lognormal distribution. Similar to Experiment 1,

the lognormal distribution emerged as the best approximation of the observed IRT data.

Distribution fitting at the individual participant level revealed consistent results with Experiment

1. The lognormal distribution demonstrated a superior fit compared to the normal and power law

distribution for individual IRT data.

An advantage of Experiment 2 was the objective identification of category switches

within participant responses. In Experiment 1, transition points occurred based on similarity

ratings, but, in Experiment 2, transitions were identified based solely on responses category

changing from "animals" to "vegetables" or vice versa. This eliminated any potential error as to

where the category transition occurred.

Interestingly, our results were similar regardless of which transition-defining method we

used. Under the behaviorally anchored definition in Experiment 1(i.e., identifying transitions

strictly by the participant's self-reported category), we found that the longest IRT occurred

immediately after the inter-category transition ($t+1$) In Experiment 2, when we implemented

objective category boundaries to identify switches (e.g., an item was either a vegetable or an

animal, objectively), we produced an even more pronounced IRT following an inter-category

transition ($t+1$). Furthermore, when we employed algorithmic transition definitions (e.g., delta

similarity or norms-based methods), the highest IRT still emerged at the post-transition position

($t+1$). Thus, the $t+1$ effect is robust across every method of defining a switch.

Surprisingly, the longest IRT occurred immediately after the inter-category transition

($t+1$), rather than at the transition point itself ($t0$). This pattern held not only for transitions

defined behaviorally, but also for those identified algorithmically using delta similarity, norms-

categorical, and norms-associative methods. Prior work in semantic foraging (e.g., Patten et al., 2020) suggested that the greatest retrieval delay would occur during the traversal between distant items - implying a peak IRT at $t0$. However, the consistent emergence of a $t+1$ peak across all transition methods challenges this view. It suggests that the major retrieval cost does not stem from the transition itself, but rather from reorienting or initiating a new search within the next semantic patch. This finding aligns more closely with the idea that categories and subcategories dynamically self-organize in response to task demands, rather than being statically represented within a fixed semantic topology.

Importantly, even though algorithmic methods inevitably introduce additional "micro-switches" within broader categories—such as between "farm animals" and "wild animals" or "leafy greens" and "root vegetables"—the same post-transition IRT peak persists. This robustness strengthens the interpretation that the t+1 effect reflects a fundamental aspect of cognitive search. Rather than being an artefact of a particular switch definition, the delay appears to reflect the cognitive reconfiguration required to activate new semantic regions. Future work might extend this framework by manipulating the strength of associative threads between responses or by comparing typed and spoken fluency tasks to assess the relative contribution of motor production time. However, current results suggest that regardless of how transitions are detected, the cognitive slowdown reliably follows—not coincides with—the switch.

Experiment 3

Similar to Experiment 2, Experiment 3 tested two categories; however, the categories were not semantic categories. Instead, participants were instructed to select responses that began with the letter S or the letter T. The rationale was to further expand the semantic space from Experiment 2. With letter categories, there was no barrier limiting responses to a certain number

of provided semantic categories; letter categories could traverse an infinite number of semantic categories. Experiment 3 allowed us to further probe global transitions in free recall; for instance, it was unclear if heavy-tailed patterning would emerge without a forced semantic constraint. The heavy-tailed explanations of cognitive search makes sense if there is a semantic space to traverse, but do heavy-tailed patterning emerge in letter fluency as well? It could be the case that participants move sporadically to unrelated items in semantic space when constrained to a letter category (e.g., space, seaweed, scepter, scholastic). Alternatively, participant responses may still have been semantically related, and therefore responses may still be distributed with heavy-tails (e.g., sailboat, sea, surfer, starfish). The latter case would have supported a more general semantic landscape for memory rather than a specific landscape within a given category.

Method

The same general method used in Experiment 1 and Experiment 2 was adopted for Experiment 3, with the only significant change being that participant responses could be from two letter categories – either words that began with the letter *S* or words that began with the letter *T*.

Participants

A total of 43 students enrolled in introductory psychology courses at Arizona State University were recruited to voluntarily participate in this study in exchange for partial course credit. A priori power analyses revealed that 45 participants were required for the experiment to be adequately powered to detect effects reported in Patten et al. (2020). Although Patten et al. (2020) report multiple effects, the medium effect size was selected as the most conservative estimate. Each participant provided informed consent prior to participation.

Materials

The experiment was run on a Dell Workstation computer with Microsoft Windows XP. The experiment was conducted, and data was collected through PsychoPy v. 2021.2.3 (Pierce, Gray, Simpson, MacAskill, Hochenberger, Sogo, Kastman, & Lindelov, 2019).

Procedure

The procedure for Experiment 3 followed the same procedure as Experiments 1 and 2. The key difference between Experiment 3 and Experiment 2 is in the instructions prompted to participants. Participants were asked to list as many *S*-words or *T*-words as they could for 10 minutes by typing their responses into a word processing program made with PyschoPy v. 2021.2.3. The order of presentation for "*S*-words" and "*T*-words" in the instructions to participants was counterbalanced such that half of the participants saw "…list as many *S*-words or *T*-words…" and the other half of the participants saw "…list as many *T*-words or *S*-words…" Participants were free to enter either *S*-words or *T*-words at any time throughout the experiment. Counterbalancing was done to minimize any bias the order of categories may have on participants' response patterns.

Results

The same general analysis plan as in Experiment 1 and Experiment 2 was applied to Experiment 3. Analysis began by first examining the shape of the IRT distribution. After inspecting the profile of the IRT distribution, the empirical distribution was compared with 1,000 simulated iterations of each of the distributions tested (normal, power law, and lognormal). Goodness of fit was tested using KS tests in the same fashion as Experiments 1 and 2. Distributions were tested at both the global and individual levels. Participant IRTs at the category transition (from *S*-words to *T*-words or vice versa) were compared to neighboring responses to test if the longest IRTs were occurring at the transition point.

Distribution Estimation

Participant IRTs at the global level were initially inspected as a histogram to observe the general shape of the distribution in Experiment 3 IRT data (see Figure 7). As seen in previous experiments, the general shape of the distribution displays the characteristic high peak and corresponding heavy-tail. The normal, lognormal, and power law distributions were all fit to IRT data from Experiment 3 using MLE (see Figure 8). Participant IRTs observed in Experiment 3 were compared with 1,000 simulated iterations for each of the distributions tested (normal, power law, lognormal). Goodness of fit was assessed using KS tests, following the same simulation approach as in Experiment 1 and Experiment 2. At the global level, there was no evidence of the normal distribution or the power law distribution fitting the empirical distribution of participant IRTs. Out of 1,000 simulations, zero iterations conformed to the normal distribution and the power law distribution. However, 597 of 1,000 simulations of lognormal data failed to reject the null hypothesis of the KS test (refer to Table 1). The results of distribution estimation and fitting in Experiment 3 are similar to those in Experiment 1 and Experiment 2. At the global level, our results consistently suggest that the lognormal distribution serves as the most accurate approximation of the observed IRT distribution.

As in prior experiments, we extended the evaluation of distribution fit to the individual participant level. For each participant, their observed IRT data was tested against 1,000 iterations of simulated data derived from a normal distribution, a lognormal distribution, and a power law distribution. The normal distribution exhibited poor fit to the individual IRT data with 187.56 ($SD = 328.02$) out of 1,000 iterations failed to reject the null hypothesis of the KS test on average. The power law distribution also exhibited a poor fit to the individual IRT data in Experiment 3. An average of 183.02 ($SD = 281.34$) out of 1,000 power law distribution

CATEGORY TRANSITIONS AND IRTS   32

iterations failed to reject the null hypothesis. This result contrasts prior experiments where

approximately half of the iterations conformed to a power law distribution. Results from IRT

distribution fitting at the individual level consistently indicated that the lognormal distribution

provided the best fit to the empirical IRT data. On average, 978.60 ($SD = 101.64$) iterations out

of 1,000 simulated lognormal datasets failed to statistically differ from the empirical IRT

distributions. Results from distribution analysis mirrored the general pattern of results observed

in Experiment 1 and 2. Our results demonstrate little to no evidence supporting participant IRTs

are normally distributed. There is mixed evidence for a heavy-tailed distribution fitting the

empirical data. We found little support for the power law distribution in Experiment 3; however,

another class of heavy-tailed distribution, the lognormal distribution fits the data well.

Transition Identification

Given the consistent observation of heavy-tailed distributions in IRT data, our objective

was to investigate whether these heavy tails were a consequence of local-to-global transitions

between categories. As with Experiment 2, Experiment 3 had objective category boundaries that

were readily identifiable. Simply put, category switches occurred when a participant switched

from an *S*-word response to a *T*-word response or vice versa. Once all transition points were

identified, we examined whether the average IRT at the transition point between categories was

significantly longer than the neighboring response IRTs. Employing the same methodology as in

prior experiments, we constructed a 7-point series centered around the transition point and

extended three responses in either direction from the transition point (see Figure 9).

Unfortunately, Experiment 3 results were not subjected to the forager algorithm to determine

semantic similarity nor other transition methods since the lexical database required to include all

possible S and T words would exceed our computing power.

A repeated measure analysis of variance was conducted comparing average IRT across the 7-point series to test if the longest IRTs were occurring at the transition point. Significant global differences in average IRT were detected based on transition point, $F(6, 252) = 5.15$, $\eta^2_{partial} = 0.11$, $p < .001$. LSD post-hoc pairwise comparisons revealed the average IRT for transition $t+1$ significantly differed from all other transition points, all $ps < .05$. As observed in prior experiments, the longest IRT did not occur at the transition point. Our results indicated that the longest IRT occurred in the response immediately following a category transition.

Discussion / Significance

Experiment 3 extended our exploration into the nature of memory recall for a letter fluency task, where the semantic constraints were removed. Similar to Experiment 1 and 2, Experiment 3 tested two categories; in this case, the categories were not semantic but instead participants were instructed to generate words beginning with the letter *S* or the letter *T*. The motivation behind this experiment was to further expand the semantic search space and investigate whether the heavy-tailed patterns observed in Experiments 1 and 2 persisted without the forced semantic constraint. This approach aimed to test the generality of the heavy-tailed patterning observed in free recall of semantic categories when the semantic constraints were relaxed.

The findings of Experiment 3 provide insight into the structure of free recall during letter fluency tasks. Despite the lack of semantic constraints, evidence of heavy-tailed distributions persisted in IRT data at both the global and individual levels, aligning with heavy-tailed foraging explanations observed in previous experiments. Rhodes and Turvey (2007) and Patten et al. (2020) demonstrated the presence of heavy-tailed distributions in IRT data for free recall within a single category. Experiment 3 extended this finding to free recall of multiple non-semantic

categories. The lognormal distribution consistently emerged as the best-fitting model for the observed IRT data at both the global and individual participant levels.

Experiment 3 had objective category boundaries similar to Experiment 2. Category transitions were identified based on the response category changing from an *S*-word to a *T*-word or vice versa. We predicted that the longest IRT would occur during the inter-category transition. As reported in Experiment 2, the longest IRT occurred at the response following an inter-category transition in Experiment 3. The persistence of this effect for both the semantic fluency task and the letter fluency task suggests the finding may be general to cognitive search. We posit that responses from new categories and subcategories unfold gradually throughout a trial, rather than being locations on a semantic landscape.

General Results

We were interested in how the change in task constraints may have affected the outcomes for all experiments. Each experiment loosened the semantic constraints associated with a free recall task. The first experiment was constrained to a single semantic category (animals). The second experiment expanded the search space by including a second semantic category (vegetables) that participants were free to access at any point during an experimental trial. The third experiment continued to expand the search space by removing semantic constraints. In Experiment 3, constraints were relaxed and the search space was drastically increased. Participants were instructed to respond based on the first letter of their response, rather than any particular semantic category.

Average IRT across experiments and the average length of response list were examined to test if Experiment 3 produced a qualitatively different experience from the semantically constrained Experiment 1 and Experiment 2. Differences in average IRT were detected based on

Experiment, $F(2, 138) = 12.99$, $\eta^2 = 0.16$, $p < .001$. Post-hoc pairwise comparisons revealed

average IRTs in Experiment 3 were significantly faster ($M = 6.36$; $SD = 2.31$) than average IRTs

in Experiment 1 ($M = 9.08$; $SD = 3.16$) and Experiment 2 ($M = 9.22$; $SD = 3.32$; all ps < .001).

There were differences detected in average response list length across experiments, $F(2, 138) =$

30.99, $\eta^2 = 0.31$, $p < .001$. Post-hoc pairwise comparisons were conducted to test which

experiment produced the longest response list. Results indicated that the average response list for

Experiment 3 ($M = 101.49$; $SD = 32.98$) was significantly longer than the average response lists

in Experiment 1 ($M = 65.42$; $SD = 19.56$) and Experiment 2 ($M = 66.78$; $SD = 19.67$; all ps <

.001). On average, intervals between responses were much faster and the length of the list of

responses produced was much longer. This pattern of results suggests that the letter fluency task

in Experiment 3 was qualitatively different than the semantic fluency tasks used in Experiment 1

and Experiment 2. Namely, the results support the interpretation that it was easier to provide

responses in the less constrained Experiment 3.

Although it was clear that participants produced more responses and were faster between

responses, we were interested in how participant responses slowed down throughout

experimentation. In the semantic fluency task of Experiment 1 and Experiment 2, we expected

that participants would slow down as they exhausted their vocabulary of responses within the

given semantic category; however, in the letter fluency task of Experiment 3, we expected

participants to have a lower rate of slowing down due to the abundance of potential words

available. We divided each participant's data into 10 bins that were each approximately 10% of

their total responses to test the rate at which participants slowed down. We then averaged across

participants to produce a series where each point corresponded to the average IRT for each 10%

bin for all participants (see Figures 13). The average IRT for the first 10% bin and the last 10%

bin for Experiment 1(M_{first} = 3.05, SD_{first} = 0.13; M_{last} = 13.33, SD_{last} = 0.92) and Experiment 2

(M_{first} = 3.80, SD_{first} = 0.24; M_{last} = 13.68, SD_{last} = 1.36) were much lower than in Experiment 3

(M_{first} = 3.97, SD_{first} = 0.20; M_{last} = 7.58, SD_{last} = 0.45). Notably, the difference between the first

10% bin and the last 10% bin was more pronounced in Experiment 1 (M_{diff} = 10.28) and

Experiment 2 (M_{diff} = 9.88) than in Experiment 3 (M_{diff} = 3.61), indicating participants slowed

down much less in Experiment 3 than the other two experiments. This pattern of slowing down

throughout a trial being dampened in Experiment 3 is easily identifiable when inspecting the

slopes observed in Figures 13.

　　　We explored the degree to which the tailedness of the distribution changed as a result of

experiment. Global data for each experiment was fit using MLE for a stable distribution. Stable

distributions are a class of distributions that include both heavy-tailed distributions and the

normal distribution as subclasses within the stable distribution family. Each distribution was fit

using a four-moment parameterization (alpha, beta, gamma, and delta). Using this stable

distribution parameterization, alpha reflects the general shape of the tailedness of the

distribution, beta reflects the skewness, gamma reflects the variation around the mean, and delta

reflects the mean or location of the distribution (Kateregga, Mataramyura, & Taylor, 2017). Of

these parameters, we were interested in the alpha parameter as it corresponds to the tailedness of

the distribution, where lower values indicate more data distributed in the tails. For each

experiment, we found the participant-level alpha parameter and averaged it across all

participants. Differences in average alpha parameters were detected based on experiment, F(2,

138) = 10.44, η^2 = 0.13, p < .001. Post-hoc pairwise comparisons revealed that the average alpha

parameter for participants was significantly higher in Experiment 3 (M = 1.12; SD = 0.14) than

in Experiment 1 (M = 0.93; SD = 0.25) and Experiment 2 (M = 0.95, SD = 0.22, all ps < .001).

Ultimately, this finding suggests that Experiment 1 and Experiment 2 had heavier tails than Experiment 3. In short, there is evidence that the shape of the distribution was sensitive to changes in task constraints with the least constrained producing the least heavy tail.

General Discussion

In general, this area of research examines the patterns of memory retrieval to elucidate the underlying cognitive mechanisms. Early explanations argued that the increases in response time throughout the task were a product of sampling with replacement (Bousfield & Sedgewick, 1944), and that exponential functions could model the increasing recall times (see review by Wixted & Rohrer, 1994). More recently, researchers recognized that responses do not simply decay as trial time increases – bursts of responses occur late in a trial, with clusters of related items recalled rapidly and long pauses reflecting transitions to distant ideas (Rhodes & Turvey, 2007). This bursty and clustered behavior typically produces heavy-tailed distributions: a high frequency of bursts near the mean and long-range movements between resource-dense clusters yield the heavy tails.

Heavy-tailed distributions appear in various foraging and movement contexts (Humphries et al., 2016; Sims et al., 2014; Reynolds, 2018) and, increasingly, in the cognitive domain as well (Patten et al., 2020; Magaldino & Amazeen, 2021; Szary et al., 2015). If our results adhere to predictions, they provide evidence that common foraging mechanisms might govern both physical resource allocation and abstract cognitive search. Given that optimal heavy-tailed search strategies are advantageous in patchy environments (Hills et al., 2012; Humphries & Sims, 2014; Reynolds et al., 2007; Sims et al., 2014), it is reasonable to expect that such dynamics would evolve in divergent domains (Todd & Hills, 2020). The present study further links these systems

by testing whether free recall exhibits the kind of heavy-tailed behavior predicted by foraging

models.

The experiments reported here offer insights into the nature of free recall and semantic

fluency. Across three experiments, we examined IRT distributions under varying constraints.

Experiment 1 focused on a single semantic category (animals), Experiment 2 expanded the

search space by including a second category (vegetables), and Experiment 3 further relaxed

constraints by using a letter-fluency task. In all cases, the lognormal distribution emerged as the

best-fitting model, reinforcing the view that free recall is characterized by heavy-tailed search

patterns.

A central aim of this work was to determine where the long IRTs—the "heavy tails"—are

generated within a trial. If humans navigate a semantic landscape, one might expect the longest

IRTs to occur precisely at the transition between semantically unrelated items (e.g., switching

from animals to vegetables). However, our data show that the longest IRT consistently appears

immediately following the transition ($t+1$) - whether the switch is defined by participants'

similarity ratings or by any of three algorithmic similarity metrics (delta-similarity, norms-

categorical, or norms-associative). The convergence across methods strengthens the evidence

that the primary slowdown comes after the categorical leap. This was consistent not only for

semantic categories but also for the letter-based fluency task in Experiment 3.

Algorithmic similarity measures inevitably identify additional "micro-switches" inside a

broad category. For instance, in Experiment 2, participants averaged approximately 8 switches

when we used objective categories; however, when using algorithmic or behavioral methods, that

number grew to around 40 (see Table 3 comparing the number of switches of algorithmic

methods to those of the objective category in Experiment). These algorithmic similarity methods,

including delta similarity and category norms, inevitably identify more micro-switches within broad categories - such as distinguishing between farm animals and wild animals - than objective category boundaries allow. For example, "cow" to "goat" would not register as a switch under objective criteria but would under algorithmic similarity if their contextual linkage is weak. That extra noise increases the number of detected transitions, yet the $t+1$ slowdown still survives. The persistence of the effect in the presence of these intra-category switches underlines its robustness.

Importantly, supplementary analyses revealed that the IRT peak consistently occurred at t+1 across all definitions of transition—whether behavioral or algorithmic. When transitions are determined using our behavioral method—relying on a 0–10 rating scale with a cutoff (e.g., ratings below 5 indicating a switch)—the maximum IRT emerges at $t+1$ This observed effect— longer IRTs immediately following (rather than at) a category transition—differs from classic foraging theory predictions (Abbott et al., 2015) and from prior semantic distance–based findings (e.g., Montez, Thompson, & Kello, 2015). Alternatively, participants may partially pre-activate a new semantic patch, rapidly produce the new category item, and only then incur a temporal cost when they begin exploring that patch. This possibility suggests memory foraging could be more dynamic and partially parallel than previously assumed.

This robust finding provides a fresh perspective on memory foraging. Previous studies (e.g., Montez et al., 2015) have shown that larger semantic leaps correlate with longer response pauses, and models like those proposed by Abbott et al. (2015) suggest that a substantial travel cost should occur at the boundary between semantic clusters. Our data, however, reveal a peak delay occurring after the switch, which could reflect either a lag in participants' conscious categorization or a cognitive reorganization period as the new category self-organizes.

A potential concern is whether our use of typed responses may have artificially elongated IRTs, particularly at transitions. However, this effect should be consistent across all points in the series, not selectively at *t+1*. Prior fluency studies using spoken responses have reported mean IRTs around 2-4 seconds (e.g., Orologa et al., 2024), while our mean IRTs were ~8 seconds. This inflation likely reflects the added motor demands of typing as well as longer trials, but there is no reason to believe it would uniquely affect post-transition IRTs. The persistence of the *t+1* effect, even under algorithmic methods and despite increased intra-category transitions, suggests a robust underlying cognitive mechanism.

Additionally, when examined across experiments, we observed differences in response dynamics. Participants in Experiment 3 exhibited faster average IRTs and produced longer response lists than those in the semantic fluency tasks of Experiments 1 and 2, implying a larger search space and less pronounced slowing. Despite these quantitative differences, the heavy-tailed nature of IRT distributions persisted regardless of semantic constraint, supporting the generality of foraging-like mechanisms across both semantic and non-semantic domains.

In sum, our findings contribute to the broader understanding of cognitive search by demonstrating that heavy-tailed IRT distributions emerge even when semantic constraints are varied. The present experiments provide converging evidence that inter-response times in free recall follow heavy-tailed distributions, regardless of whether participants search within semantic or non-semantic categories. Across all tasks, the lognormal distribution emerged as the most consistent and best-fitting model, reinforcing the idea that memory search exhibits the same bursty, clustered dynamics seen in foraging and movement across physical space. These findings support the broader theoretical claim that evolved systems—including human cognition—rely on similar search dynamics when navigating patchy environments, whether those environments are

spatial, perceptual, or conceptual. Taken together, this work strengthens the bridge between

ecological theories of foraging and the cognitive architecture of memory, highlighting how

adaptive search strategies may be a unifying principle across diverse behavioral domains. These

results invite further investigation into the real-time dynamics of category switching and the

development of more nuanced measures of semantic similarity to fully capture the complexity of

cognitive foraging.

Declarations

Open Practices Statement

The data for all experiments are available at: https://osf.io/pbc7m/?view_only=82725788933449d79f3c4ed6f0f90b1c. None of the experiments were preregistered.

Data Availability Statement

The data sets generated during and analyzed during the current study are availability in an OSF repository, https://osf.io/pbc7m/?view_only=82725788933449d79f3c4ed6f0f90b1c.

Funding

No funds, grants, or other support was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval

This study was performed in line with the principle of the Declaration of Helsinki. Approval was granted by the Institutional Review Board of Arizona State University (STUDY00015260, 9/29/2022).

Consent to Participate

Informed consent was obtained from all individual participants included in the study prior to beginning the experiment.

Consent for Publication

The authors affirm that participants provided informed consent for publication of data.

Code Availability

The code used for analysis is available from the corresponding author upon reasonable request. All data are available in an OSF repository, https://osf.io/pbc7m/?view_only=82725788933449d79f3c4ed6f0f90b1c.

References

Abbott, J., Austerweil, J., & Griffiths, T. (2015). Random walks on semantic networks can

resemble optimal foraging. Psychological Review, 122(3), 558–569.

Abwender, D.A., Swan, J.G., Bowerman, J.T., & Connolly, S.W. (2001). Qualitative analysis of

verbal fluency output: Review and comparison of several scoring methods. Assessment,

8(3), 323–338.

Banches, R.E. (2013). *Text Mining With Matlab*. Springer, NY.

Boyer D., Ramos-Fernandez, G., Miramontes, O., Mateos, J. L., Cocho, G., Larralde, H., Ramos,

H., Rojas, F. (2006). Scale-free foraging by primates emerges from their interaction with

a complex environment. *Proceedings of the Royal Society B: Biological Sciences, 273*,

1743-1750. doi:10.1098.rspb.2005.3462

Brown, C. T., Liebovitch, L. S., & Glendon, R. (2006). Lévy flights in Dobe Ju/\hoansi Foraging

Patterns. *Human Ecology, 35*(1), 129-138. doi:10.1007/s10745-006-9083-4

Brychcin, T., & Konopik M. (2014). Semantic spaces for improving language modeling.

Computer Speech & Language, 28(1), 192-209.

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.

Charnov, E.L. (1976). Optimal foraging: The marginal value theorem. *Theoretical Population*

Biology, 9, 129-136.

Cowan, N. (2008). What are the differences between long-term, short-term, and working

memory? *Progress in Brain Research,* 169, 323-338.

Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Focardi, S., & Cecere, J. G. (2013). The Lévy flight foraging hypothesis in a pelagic seabird. *Journal of Animal Ecology, 83,* 353-364. doi:10.1111/1365-2656.12147

Goñi, J., Arrondo, G., Sepulcre, J., Martincorena, I., de Mendizábal, N.V., Corominas-Murtra, B., ..., et al. (2011). The semantic organization of the animal category: Evidence from semantic verbal fluency and network theory. Cognitive Processing, 12(2), 183–196.

Henry, J.D., Crawford, J.R., & Phillips, L.H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. *Neuropsychologia*, 42, 1212-1222.

Hills, T.T., Jones, M.N., & Todd, P.M. (2009). Optimal foraging in semantic memory. In: Proceedings of the annual meeting of the Cognitive Science Society, pp 620–625.

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

Hills, T.T., Mata, R., Wilke, A., & Samanez-Larkin, G.R. (2013). Mechanisms of age-related decline in memory search across the adult life span. Developmental Psychology, 49(12), 2396–2404.

Hills, T.T., Todd, P.M., & Jones, M.N. (2015). Foraging in semantic fields: How we search through memory. Topics in Cognitive Science, 7(3), 513–534.

Humphries, N. E., Queiroz, N., Dyer, J. R. M., Pade1, N. G., Musy, M. K., Schaefer, K. M., Fuller, D. W., Brunnschweiler, J. M., Doyle, T. K., Houghton, J. D. R., Hays, G. C., Jones, C. S., Noble, L. R., Wearmouth, V. J., Southall, E. J., & Sims, D. W. (2010). Environmental context explains Lévy and Brownian movement patterns of marine predators. *Nature*, 465, 1066– 1069.

Humphries, N. E. & Sims, D. W. (2014). Optimal foraging strategies: Levy walks balance

searching and patch exploitation under a very broad range of conditions. *Journal of*

Theoretical Biology, 358, 179– 193.

Humphries, N.E., Schaefer, K. M., Fuller, D. W., Phillips, G. E., Wilding, C., & Sims, D. W.

(2016). Scale-dependent to scale-free: Daily behavioural switching and optimized

searching in a marine predator. *Animal Behavior, 113,* 189-201.

doi:10.1016/j.anbehav.2015.12.029

Jia, T., Jiang, B., Carling, K., Bolin, M., & Ban, Y. (2012). An empirical study on human

mobility and its agent-based modeling. *Journal of Statistical Mechanics: Theory and*

Experiment, 11, doi:10.1088/1742-5468/2012/11/p11024

Jones, M. N., Kintsch, W., & Mewhort, D.J.K. (2006). High-dimensional semantic space

accounts of priming. *Journal of Memory and Language*, *55*(4), 534-552.

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.

Jun, K.S., Zhu, X., Rogers, T.T., Yang, Z., & et al (2015). Human memory search as initial-visit

emitting random walk. *Advances in Neural Information Processing Systems*, 1072–1080.

Kateregga, M., Mataramvura, S., & Taylor, D. (2017). Parameter estimation for stable

distributions with application to commodity future log-returns. *Cogent Economics and*

Finance, 5. doi:10.1080/23322039.2017.1318813

Kim, N., Kim, J.H., Wolters, M.K., MacPherson, S.E., & Park, J.C. (2019). Automatic scoring of

semantic fluency. *Frontiers in Psychology, 10*, 1–16.

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges.

Psychonomic Bulletin & Review, *28*, 40-80.

Kumar, A.A., Apsel, M., Zhang, L., Xing, N., Jones. M.N. (2023). forager: A Python package

  and web interface for modeling mental search. *Behavior Research Methods*, 1-17

Lee, R. B. (1979). *The !Kung San: men, women, and work in a foraging society*, Cambridge

  University Press, Cambridge.

Lee, R. B. (1993). *The Dobe Ju/'hoansi, harcourt brace*, New York.

Lee, R. B., and DeVore, I. (Eds.) (1976). *Kalahari hunter–gatherers: Studies of the !Kung San

  and their neighbors*, Harvard University Press, Cambridge.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York, NY: Oxford

  University Press.

Linz, N., Tröger, J, Alexandersson, J., & Konig, A. (2017). Using neural word embeddings in the

  analysis of the clinical semantic verbal fluency task. In IWCS 2017 - 12th international

  conference on computational semantics, Montpellier, France (pp. 1–7).

Locascio, J.J., Growdon, J.H., & Corkin, S. (1995). Cognitive test performance in detecting,

  staging, and tracking Alzheimer's disease. *Archives of Neurology*, 52, 1087-1099.

Lundin, N. B., Brown, J. W., Johns, B. T., Jones, M. N., Purcell, J. R., Hetrick, W. P., ... & Todd,

  P. M. (2023). Neural evidence of switch processes during semantic and phonetic foraging

  in human memory. *Proceedings of the National Academy of Sciences*, 120(42),

  e2312462120.

Malik, R., Subramaniam, L. V., & Kaushik, S. (2007). Automatically selecting answer templates

  to respond to customer emails. In Proceeding of the IJCAI, 7, 1659-1664.

Maseda, A., Lodeiro-Fernandez, L., Lorenzo-Lopez, L., Nunez-Naveria, L., Balo, A., & Millan-

  Calenti, J.C. (2014). Verbal fluency, naming and verbal comprehension: three aspects of

  language as predictors of cognitive impairment. *Aging Mental Health*, 18, 1037-1045.

Miller. G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*. 38(11), 39-41.

Mioshi, E., Dawnson, K., Mitchell, J., Arnold, R., & Hodges, J.R. (2006). The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening. *International Journal of Geriatric Psychiatry*, 21, 1078-1085.

Monsch, A.U., Bondi, M.W., Butters, N., Salmon, D.P., Katzman, R., & Thal, L.J. (1992). Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. *Archives of Neurology*, 49(2), 1253-1258.

Montez, P., Thompson, G., & Kello, C. T. (2015). The role of semantic clustering in optimal memory foraging. *Cognitive Science, 39*(8), 1925-1939.

Mueller, K.D., Koscik, R.L., LaRue, A., Clark, L.R., Hermann, B., Johnson, S.C., & Sager, M.A. (2015). Verbal fluency and early memory decline: Results from the Wisconsin Registry for Alzheimer's Prevention. Archives of Clinical Neuropsychology, 30(5), 448–457.

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.

Nolan, J. P. (1997). Numerical calculation of stable densities and distribution functions. *Communications in Statistics: Stochastic Models, 13*(4), 759-774.

Nolan, J. P. (2021). *Univariate stable distributions: Models for heavy tailed data*. Springer Nature. doi:10.1007/978-3-030-52915-4

Orologa, E., Chatzopoulos, G., Nikolaidis, D., Kosmidis, M. H., & Proios, H. (2024). The importance of response times measures in verbal fluency tasks. *Psychology: The Journal*

of the Hellenic Psychological Society, 29(1), 71–81.

https://doi.org/10.12681/psy_hps.34095

Oussalah, M., & Mohamed, M. (2021). Knowledge-based sentence semantic similarity: algebraical properties. *Progress in Artificial Intelligence*. doi: 10.1007/s13748-021-00248-0

Patten, K.J., Greer, K., Likens, A.D., Amazeen, E. L., & Amazeen, P.G. (2020). The trajectory of thought: Heavy-tailed distributions in memory foraging promote efficiency. *Memory & Cognition,* 48, 772-787.

Pierce, J.W., Gray J.R., Simpson, S., MacAskill, M.R., Hochenberger, R., Sogo, H., Kastman, E., & Lindelov, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods.* doi:10.3758/s13428-018-01193-y

Rachev, S. T., & Mittnik, S. (2000). *Stable Paretian Models in Finance*. Wiley.

Raichlen, D.A., Wood, B.M., Gordon, A.D., Mabulla, A.Z.P., Marlowe, F.W. & Pontzer, H. (2014). Evidence of Lévy walk foraging patterns in human hunter-gatherers. *Proceedings of the National Academy of Science USA*, 111, 728–733.

Raoux, N., Amieva, H., Le Goff, M., Auriacombe, S., Carcaillon, L., Letenneur, L., & Dartigues, J.F. (2008). Clustering and switching processes in semantic verbal fluency in the course of Alzheimer's disease subjects: Results from the PAQUID Longitudinal Study. Cortex, 44(9), 1188–1196.

Reynolds, A. M. (2011a). Chemotaxis can provide biological organisms with good solutions to the traveling salesman problem. *Physical Review E,* 83.

Reynolds, A. M. (2011b). On the origin of bursts and heavy tails in animal dynamics. *Physica A‑Statistical Mechanics and Its Applications*, 390, 245– 249.

Reynolds, A. M. (2018). Lévy foraging patterns of rural humans. *PLoS One,* 13.

doi:10.1371/journal.pone.0199099

Reynolds, A.M. & Frye, M.A. (2007). Free-flight odor tracking in *Drosophila* is consistent with

an optimal intermittent scale-free search. *PLoS One*, 2, 1– 9.

Reynolds, A.M., Smith, A.D., Reynolds, D.R., Carreck, N.L. & Osborne, J.L. (2007). Honeybees

perform optimal scale-free searching flights when attempting to locate a food

source. *Journal of Experimental Biology*, 210, 3763– 3770.

Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S.J., & Chong, S. (2011). On the Levy-walk nature of

human mobility. *IEEE‐ACM Transactions on Networking*, 19, 630– 643.

Rhodes, T., & Turvey, M. T. (2007). Human memory retrieval as Lévy foraging. *Physica A:*

Statistical Mechanics and Its Applications, 385(1), 255-260.

doi:10.1016/j.physa.2007.07.001

Ross, T.P., Calhoun, E., Cox, T., Wenner, C., Kono, W., & Pleasant, M. (2007). The reliability

and validity of qualitative scores for the controlled oral word association test. Archives of

Clinical Neuropsychology, 22 (4), 475–488.

Salmon, D.P., Thomas, R., Pay, M., Booth, A., Hofstetter, C., Thal, L., … et al. (2002).

Alzheimer's disease can be accurately diagnosed in very mildly impaired individuals.

Neurology, 59, 1022-1028.

Sims, D.W., Humphries, N.E., Bradford, R.W. & Bruce, B.D. (2011). Levy flight and Brownian

search patterns of a free-ranging predator reflect different prey field

characteristics. *Journal of Animal Ecology*, 81, 432– 442.

Sims. D. W., Reynolds, A. M., Humphries, N. E., Southall, E. J., Wearmouth, V. J., Metcalfe, B.,

& Twitchett, R. J. (2014). Hierarchical random walks in trace fossils and the origin of

optimal search behavior. *Proceedings of the National Academy of Science, 111*, 11073-

11078. doi:10.1073/pnas.1405966111

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468),

390-398.

Sung, K., Gordon, B., Vannorsdall, T.D., Ledoux, K., Pickett, E.J., Pearlson, G.D., & Schretlen,

D.J. (2012). Semantic clustering of category fluency in schizophrenia examined with

singular value decomposition. Journal of the International Neuropsychological Society,

18(3), 565–575.

Sung, K., Gordon, B., & Schretlen, D.J. (2016). Semantic structure can be inferred from category

fluency tasks via clustering analyses: Reply to Voorspoels others.(2014). Cortex; A

journal devoted to the study of the nervous system and behavior, 75, 249–254.

Szary, J., Dale, R., Kello, C. T., Rhodes, T. (2015). Patterns of interaction-dominant dynamics in

individual versus collaborative memory foraging. *Cognitive Processes*, *16*(4), 389-99.

doi: 10.1007/s10339-015-0731-8

Todd, P. M., & Hills, T. T. (2020). Foraging in mind. *Current Directions in Psychological

Science, 29*(3), 309-315.

Tombaugh, T.N., Kozak, J., & Rees, L. (1999). Normative data stratified by age and education

for two measures of verbal fluency: Fas and animal naming. *Archives of Clinical

Neuropsychology,* 14, 167-177.

Troyer, A.K. (2000). Normative data for clustering and switching on verbal fluency tasks.

Journal of Clinical and Experimental Neuropsychology, 22(3), 370–378.

Troyer, A.K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two

    components of verbal fluency: Evidence from younger and older healthy adults.

    Neuropsychology, 11(1), 138–146.

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2014). Strategic search from long-term memory:

    An examination of semantic and autobiographical recall. *Memory,* 22 (6), 687-699.

Viswanathan, G.M., Afanasyev, V., Buldyrev, S.V., Murphy, E.J., Prince, P.A. & Stanley,

    H.E. (1996). Lévy flight search patterns of wandering albatrosses. *Nature*, 381, 413.

Viswanathan, G. M., Bartumeus, F., Buldyrev, S. V., Catalan, J., Fulco, U. L., Havlin, S., da

    Luz, M. G. E., Lyra, M. L., Raposo, E. P., & Stanley, H. E. (2002). Lévy flight random

    searches in biological phenomena. *Physica A: Statistical Mechanics and its Applications,*

    314, 208-213. doi:10.1016/S0376-4371(02)01157-3

Voit, J. (2004). The Statistical Mechanics of Financial Markets. In Balian, R., Beiglbock, W.,

    Grosse, H. & Tirring, W. (Eds.), *Springer Texts and Monographs in Physics*. Springer.

    doi:10.1007/b137351

Voorspoels, W., Storms, G., Longenecker, J., Verheyen, S., Weinberger, D.R., & Elevåg, B.

    (2014). Deriving semantic structure from category fluency: Clustering techniques and

    their pitfalls. Cortex, 55, 130-147.

Weron, R. (2001). Levy-stable distributions revisited: Tail index > 2 does not exclude the levy-

    stable regime. *International Journal of Modern Physics C, 12*(02), 209-223.

Williams. M. D., & Hollan, J. D. (1981). The process of retrieval from very long-term memory.

    Cognitive Science, 5 (2), 87-119.

Willits J. A., D'Mello, S. K., Duran, N. D., & Olney, A. (2007). Distributional statistics and

    thematic role relationships. In D. S. McNamara & J. G. Trafton (Eds.), *Proceeding of the*

29th Annual Conference of the Cognitive Science Society (pp. 707-712). Austin, TX: Cognitive Science Society.

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review, 1*(1), 89-106

Woods, D.L., Wyma, J.M., Herron, T.J., & Yund, E.W. (2016). Computerized analysis of verbal fluency: Normative data and the effects of repeated testing, simulated malingering, and traumatic brain injury. PloS one, 11 (12), 439.

Wosniack, M.E., Santos, M.C., Raposo, E.P., Viswanathan, G.M., da Luz, M.G. (2017). The evolutionary origins of Lévy walk foraging. *PLOS Computational Biology, 13*(10). doi.org/10.1371/journal.pcbi.1005774

Wu, Z., & Palmer, M. (1994) Verb semantics and lexical selection. *In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* 133-139, Las Cruces, NM.

Zemla, J.C., & Austerweil, J.L. (2018). Estimating semantic networks of groups and individuals from fluency data. *Computational Brain & Behavior*, 1(1), 36–58.

Zemla, J.C., & Austerweil, J.L. (2019). Analyzing knowledge retrieval impairments associated with Alzheimer's disease using network analyses. Complexity 2019.

Zemla, J.C., Cao, K., Mueller, K. D., Austerweil, J. L. (2020). SNAFU: The Semantic Network and Fluency Utility. *Behavior Research Methods*, 52, 1681-1699.

Table 1

Simulated Distribution Fits

		Normal Distribution	Power Law Distribution	Lognormal Distribution
Global	Experiment 1	0	0	736
	Experiment 2	0	0	705
	Experiment 3	0	0	587
Individual	Experiment 1	67.24	651.48	983.06
	Experiment 2	84.04	567.37	999.91
	Experiment 3	187.56	183.02	978.60

Note. The above table demonstrated the simulated distribution fits for all experiments at the global and individual levels. Empirical data was compared to simulated distributions ($n = 1,000$). Distribution fit was assessed using a two-sample Kolmogorov-Smirnov (KS) test at a 5% significance level. Table values indicate the number of simulations where the KS test failed to reject the null hypothesis that the simulated data and empirical data were from the same continuous distribution.

Table 2

Average IRTs and Semantic Similarity

	Switching Method	Switch IRT	Nonswitch IRT	Switch Sem. Sim.	Nonswitch Sem. Sim.
	Sim. Ratings	8.82 (11.06)	7.35 (9.38)	0.41 (0.14)	0.41 (0.14)
	Associative	10.42(12.04)	5.39 (6.94)	0.36 (0.10)	0.49 (0.13)
Experiment 1	Categorical	10.28(11.92)	5.58(7.33)	0.36 (0.10)	0.48 (0.13)
	Delta Sim.	10.15(11.70)	6.23(8.49)	0.34 (0.08)	0.49 (0.12)
	Associative	9.98 (11.34)	5.23 (6.44)	0.36 (0.11)	0.48 (0.14)
Experiment 2	Categorical	9.34 (10.42)	6.72 (9.46)	0.37 (0.11)	0.44 (0.15)
	Delta Sim.	10.16 (11.57)	6.25 (7.79)	0.37 (0.09)	0.44 (0.12)

Note. The above table shows the average IRT in seconds and average semantic similarity for

Experiments 1 and 2 for different switching methods (similarity ratings, associative, categorical,

and delta similarity).

Table 3

Average Number of Switches by Switching Method and Experiment

	Similarity Ratings	Objective Categories	Associative	Categorical	Delta Similarity
Experiment 1	36.46 (11.19)	N/A	39.00 (9.79)	38.92 (9.96)	35.55 (10.26)
Experiment 2	35.20 (13.36)	8.48 (3.85)	44.65 (11.95)	41.35 (11.68)	36.46 (10.43)
Experiment 3	64.53 (20.34)	25.91 (13.36)	N/A	N/A	N/A

Note. The above table shows the average number of switches by switching method and experiment. Switching methods include participants self-report of similarity, objective categories defined by the task, associative, categorical, and delta similarity. Parentheses indicate the standard deviation for the given average.
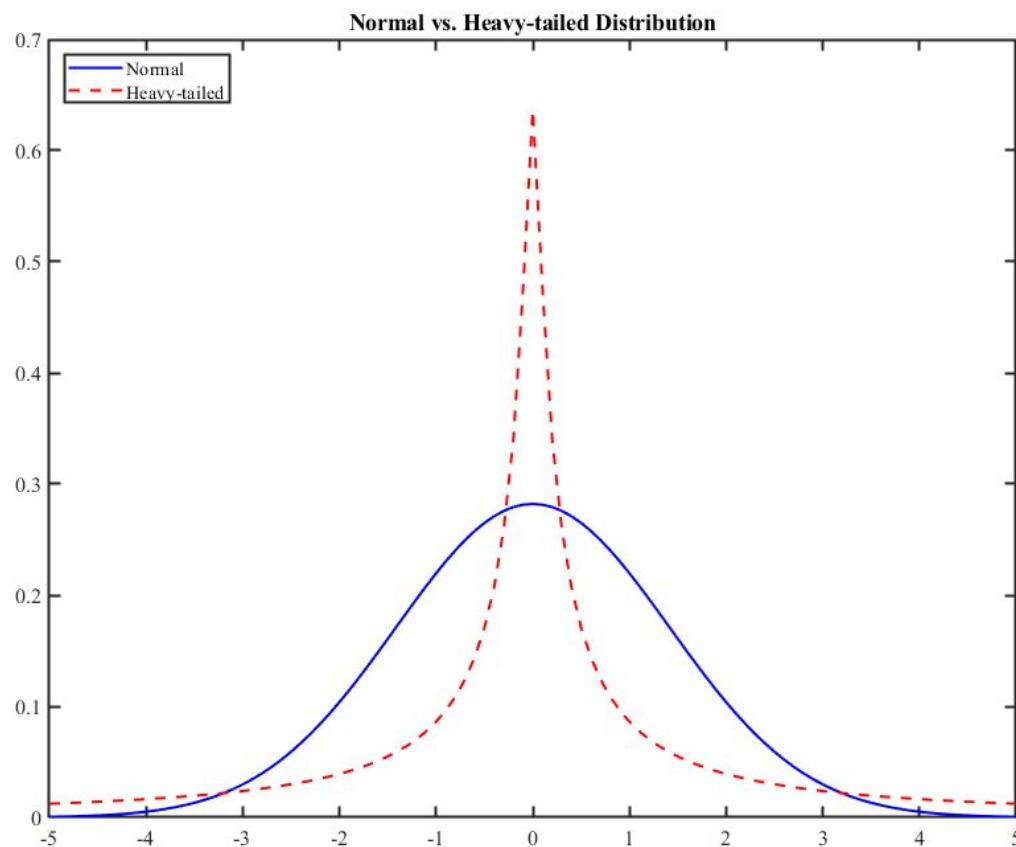
Figure 1. Illustration of the difference between the normal distribution and a class of heavy-tailed distribution (Lévy distribution as reported in Rhodes and Turvey, 2007). Adapted from Magaldino & Amazeen 2021.
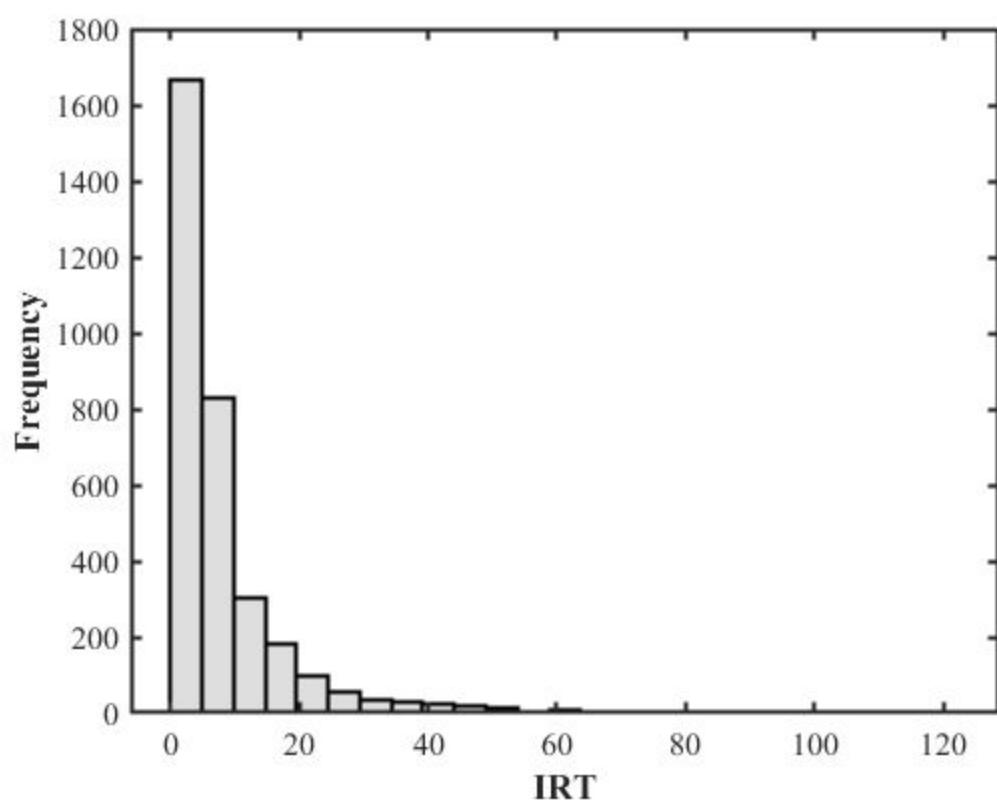
*Figure 2*. Histogram showing the frequency of all participant IRTs from Experiment 1.
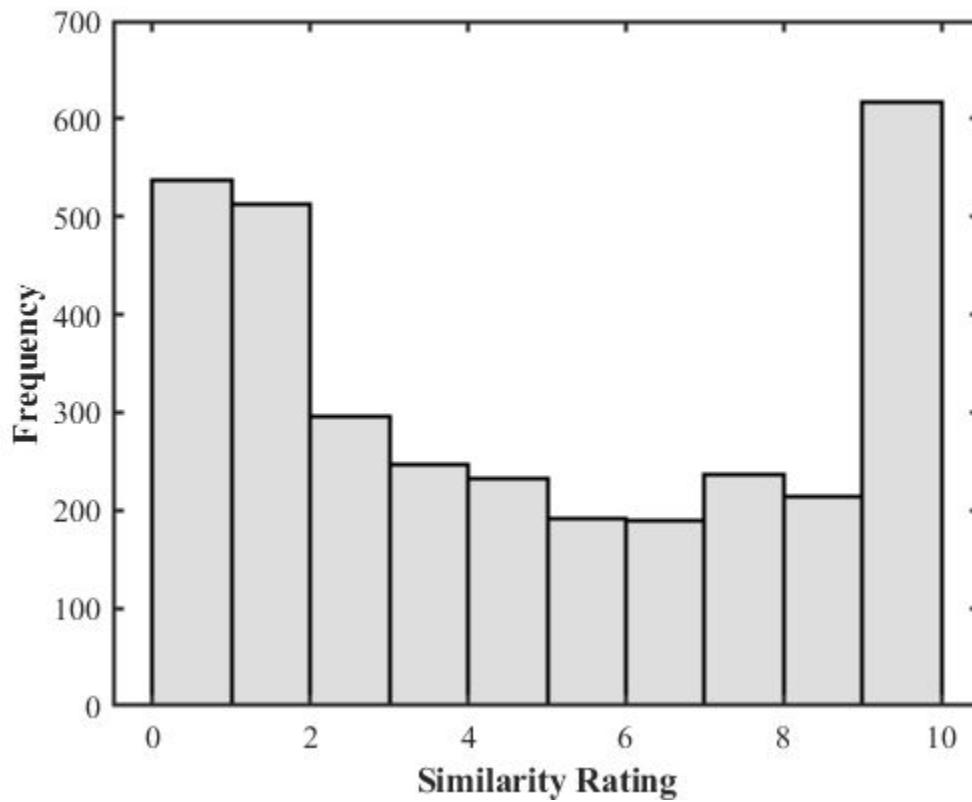
Figure 3. Histogram showing the frequency of all participant similarity ratings from Experiment 1.
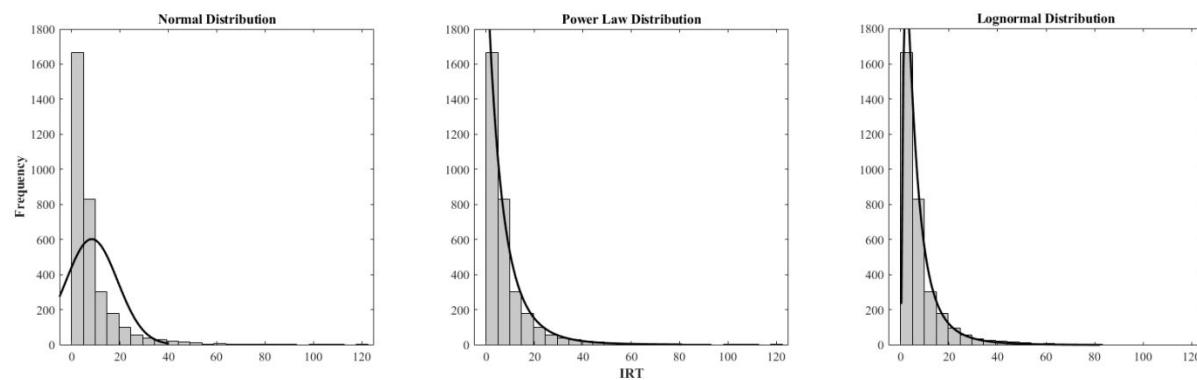
Figure 4. Distribution fit of the total IRT data for Experiment 1 for the normal distribution (left),

the power law distribution (middle), and the stable distribution (right). Distribution parameters

were estimated using maximum likelihood estimation.

60

Figure 5. Histogram showing the frequency of all participant IRTs from Experiment 2.
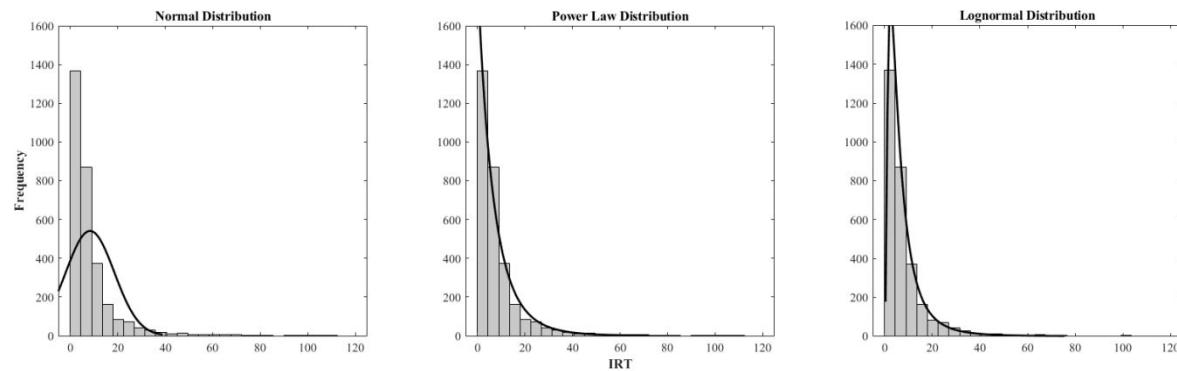
Figure 6. Distribution fit of the total IRT data for Experiment 2 for the normal distribution (left), the power law distribution (middle), and the stable distribution (right). Distribution parameters were estimated using maximum likelihood estimation.

CATEGORY TRANSITIONS AND IRTS   62
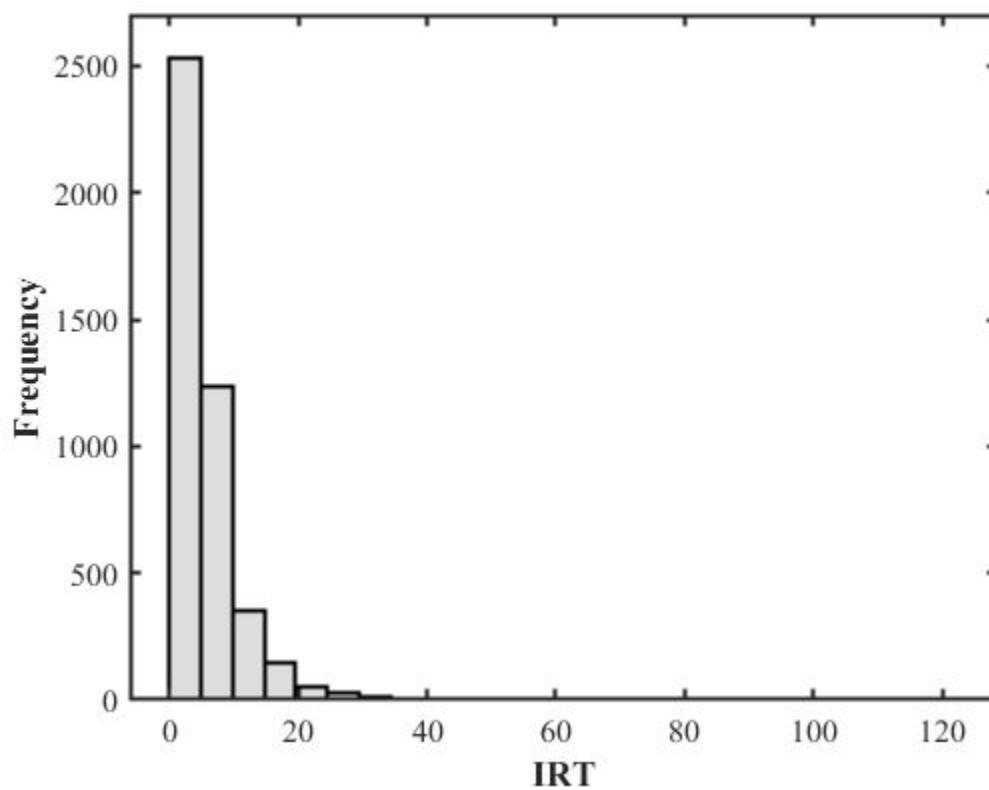
Figure 7. Histogram showing the frequency of all participant IRTs from Experiment 3.
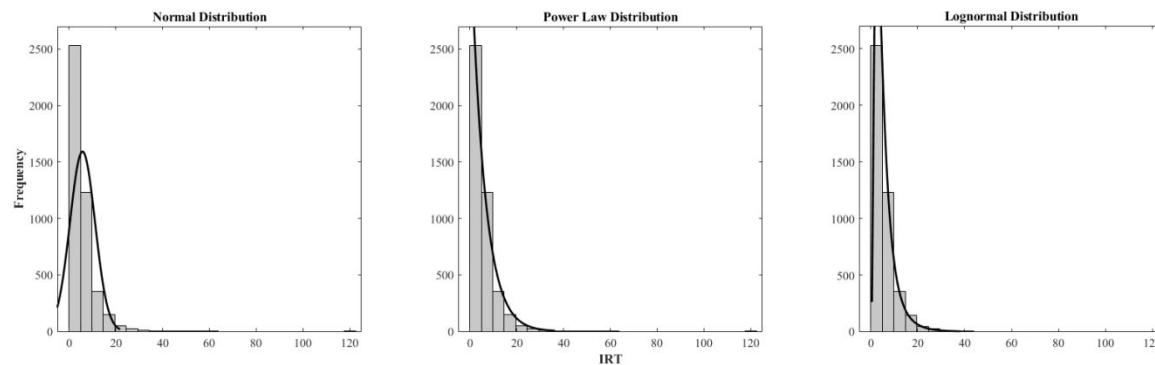
Figure 8. Distribution fit of the total IRT data for Experiment 3 for the normal distribution (left),

the power law distribution (middle), and the stable distribution (right). Distribution parameters

were estimated using maximum likelihood estimation.
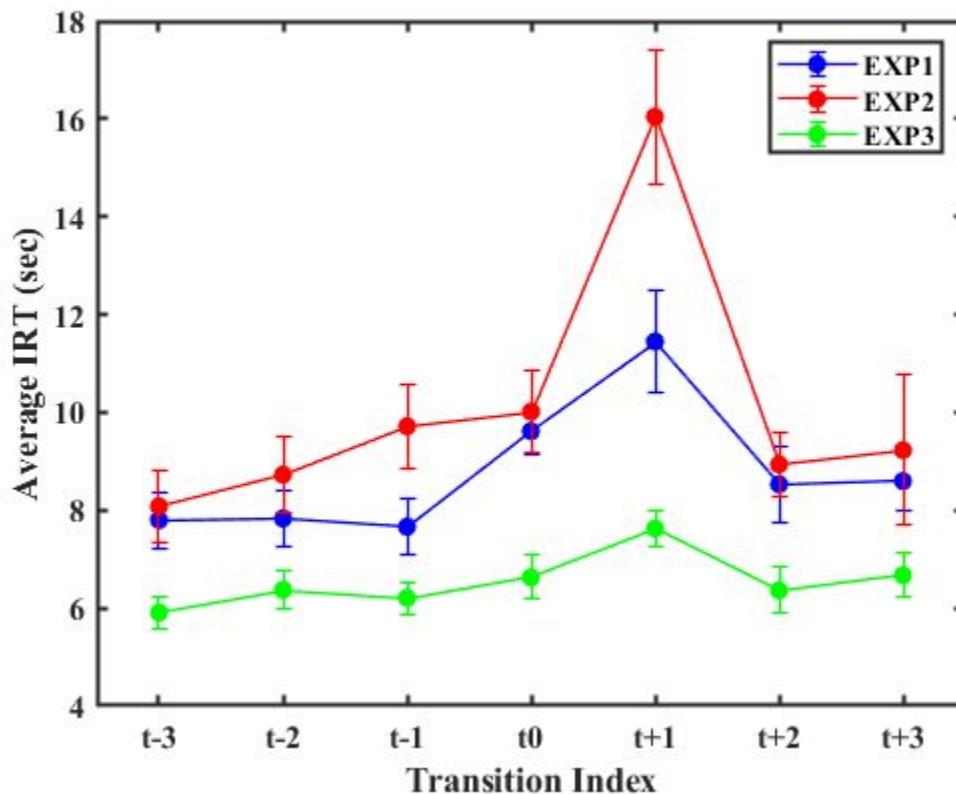
Figure 9. Line graph depicting average IRTs around a transition point (*t0*) for all experiments.

Transition points were identified using participants' similarity ratings in Experiment 1.

Experiment 2 and 3 used objective category switches (either animal/vegetable or S/T-words).

The *x*-axis illustrates average IRTS for ± 3 responses from the transition point.
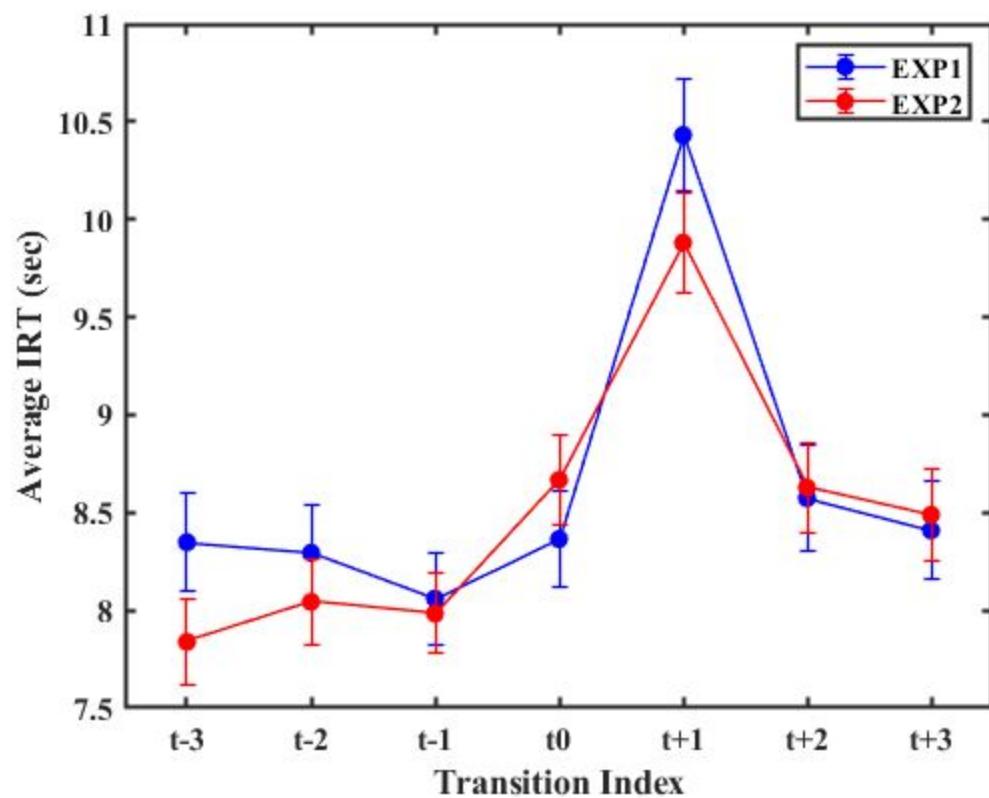
Figure 10. Line graph depicting average IRTs around a transition point (*t0*) for Experiment 1 and

2. Transition points were identified using norm-based associative search. The *x*-axis illustrates

average IRTS for ± 3 responses from the transition point.
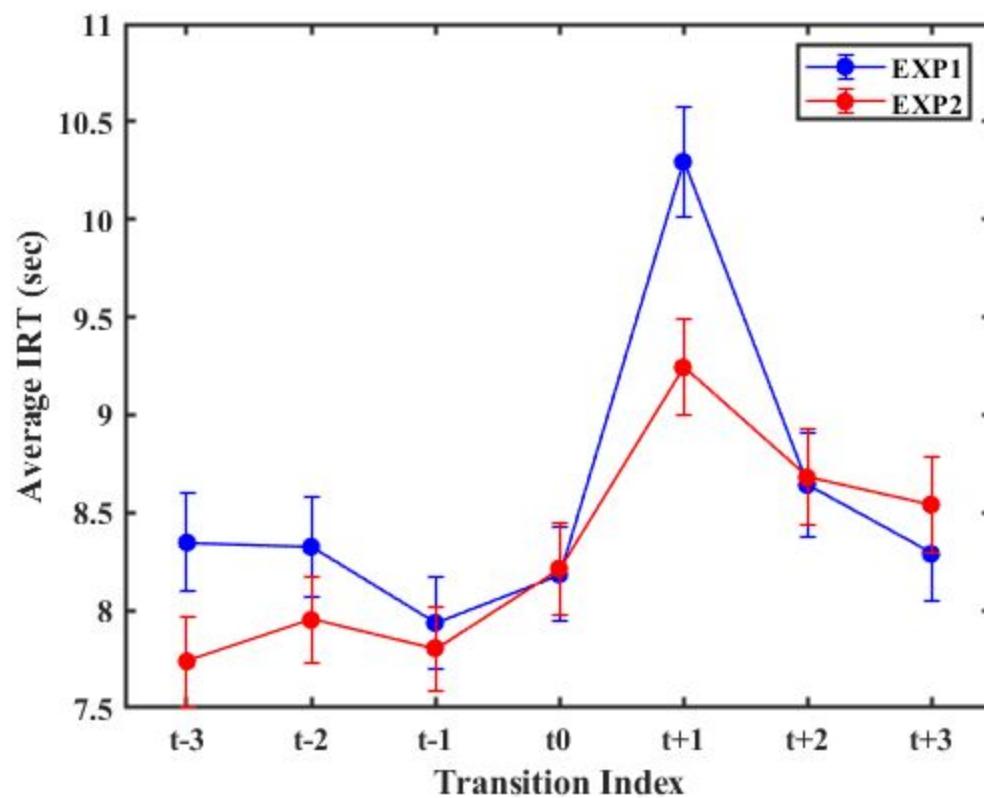
Figure 11. Line graph depicting average IRTs around a transition point (*t0*) for Experiment 1 and

2. Transition points were identified using norm-based categorical search. The *x*-axis illustrates

average IRTS for ± 3 responses from the transition point.
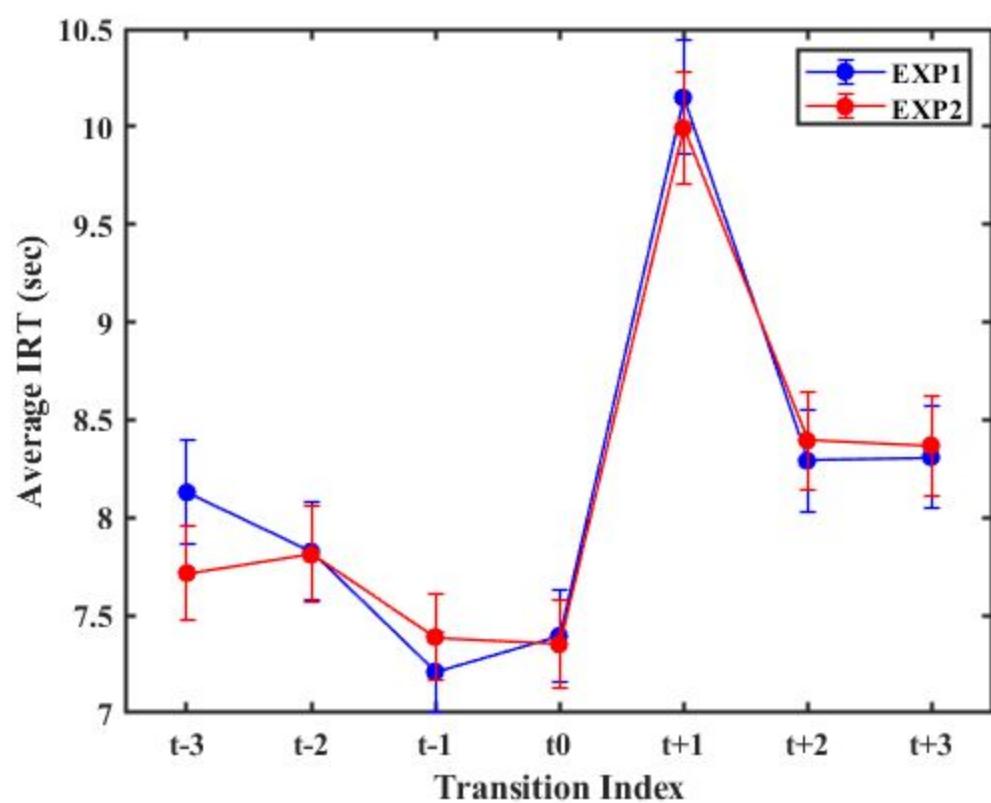
Figure 12. Line graph depicting average IRTs around a transition point (*t0*) for Experiment 1 and

2. Transition points were identified using delta similarity with a 0.5 rise and fall criteria. The *x*-

axis illustrates average IRTS for ± 3 responses from the transition point.
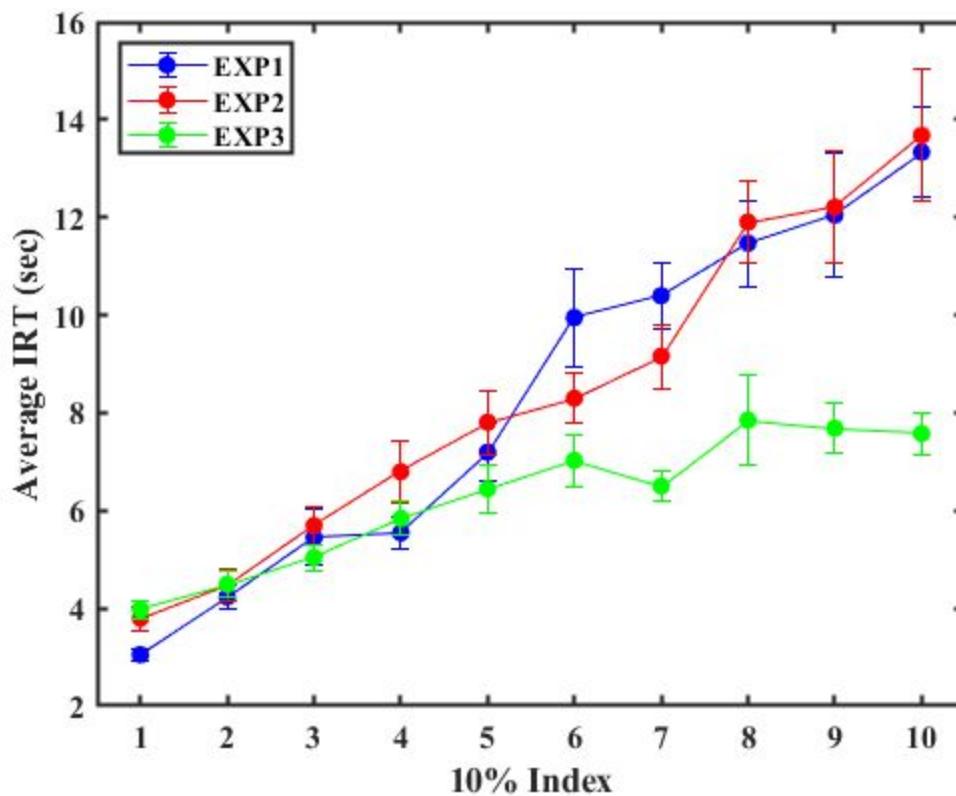
Figure 13. Average IRT across the course of Experiment 1-3. Participant responses were

segmented into bins corresponding to 10% of their total responses. Each point on the *x*-axis

represents the average IRT for all 10 of those segments. Experiments are separated by color.