Journal of Memory and Language 141 (2025) 104606

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

What's in *my* cluster? Evaluating automated clustering methods to understand idiosyncratic search behavior in verbal fluency☆

Abhilasha A. Kumar [a,*], Nancy B. Lundin [b], Michael N. Jones [c]

[a] *Bowdoin College, ME, United States*
[b] *The Ohio State University, OH, United States*
[c] *Indiana University Bloomington, IN, United States*

A B S T R A C T

Individuals routinely search through memory for concepts. This behavior is commonly studied via the verbal fluency task (VFT), where participants are typically asked to generate as many exemplars as they can from a given category (e.g., animals) or letter label (e.g., F) within a fixed amount of time. Responses in the VFT tend to be clustered in meaningful ways but individuals widely differ in the manner in which they cluster items. Despite the development of several (hand-coded and automated) methods of defining clusters and switches in the VFT, there is currently no consensus on which scoring method provides the best mechanistic account of how *individuals* search through memory in the VFT. In this work, we provide an empirical evaluation of several automated methods for defining clusters and switches in the VFT by comparing model-predicted clusters with participant-designated clusters. We find that a method that combines gradual rises and drops in a weighted composite of semantic *and* phonological similarity best predicts participant-designated cluster-switch events across three domains (*animals*, *foods*, and *occupations*). Furthermore, we propose a novel approach to understand idiosyncratic search behavior by computing a measure of discordance for each pairwise transition based on a large dataset of cluster-switch designations from independent raters ($N = 211$) for the same transitions via a pre-registered experiment. We find that transitions with high idiosyncratic scores have low lexical content (i.e., semantic and phonological similarity), and an individual's score on one domain is predictive of their score on another domain, suggesting that idiosyncratic scores may be capturing meaningful information about non-lexical sources and processes that contribute to memory search at the individual level.

Introduction

Common everyday tasks require us to actively search through our memory, such as making a grocery list, coming up with an answer to a trivia question, or trying to access a word on the tip of your tongue. Although this search process can feel effortless at times, several complex operations are involved during memory search, including but not limited to activating relevant concepts in the mental lexicon, retrieving their representations relatively quickly, and ultimately producing them to achieve the goal at hand. The Verbal Fluency Task (VFT; Bousfield and Sedgewick, 1944) is a diagnostic test widely used to study memory search, with widespread applications in both basic science as well as clinical settings. In a typical version of the task, participants are provided with a category (e.g., *Animals*) or letter label (e.g., *A*), and asked to produce as many items from that category or beginning with that letter label within a fixed amount of time.

Individuals widely vary in the manner in which they perform the VFT through the use of different strategies (Lundin et al., 2023; Unsworth et al., 2014), pauses and repetitions (Balogh et al., 2023), as well as the mental processes that contribute to patterns of clustering (i.e., producing related items such as *cat* and *dog* in close succession) and switching during the task (Hills et al., 2015). However, the study of memory search from an *individual* perspective is an ongoing challenge. While the total number of items produced (i.e., raw scores) are considered a coarse measure of performance (Henderson et al. 2023), defining what counts as a cluster within a fluency list has proven non-trivial (see Hills et al., 2015 for a discussion) and a majority of scoring approaches ignore the influence of idiosyncratic variance. Specifically, individuals report a variety of strategies used as well as experiences they bring online when performing the VFT, such as using hierarchies or categories, personal or episodic experiences, as well as imagery (Lundin et al., 2023). However, the extent to which these experiences can be incorporated into mechanistic accounts of clustering and switching within the VFT remains relatively unexplored in the literature.

Clustering and switching among individuals

The majority of the work on understanding clustering and switching

☆ This article is part of a special issue entitled: 'Individual differences in memory' published in Journal of Memory and Language.

* Corresponding author.
E-mail address: a.kumar@bowdoin.edu (A.A. Kumar).

https://doi.org/10.1016/j.jml.2024.104606
Received 5 March 2024; Received in revised form 17 December 2024; Accepted 18 December 2024
Available online 2 January 2025

behavior within the VFT has relied on hand-coded norms for the *animals* domain (Chen et al., 2020; Bose, Patra, Antoniou, Stickland, & Belke, 2022; Troyer, Moscovitch, & Winocur, 1997; Oh, Sung, Choi, & Jeong, 2019). Based on the Troyer (2000) norms, participant responses are sorted into specific subcategories (e.g. reptiles, pets, or North America for *animals*) and switch events are determined when consecutive responses belong to different subcategories. While ubiquitous, collecting and updating norms is a time and resource-intensive process. Therefore, several automated methods for determining switches and clusters within fluency lists have been proposed within the past decade. These attempts typically emphasize the *lexical content* of the specific words produced by individuals in the VFT. Several methods use distributional semantic models (DSMs; for a review, see Kumar, 2021) trained on large text corpora to derive measures of semantic similarity and define cutoffs to designate clusters and switches. For example, Hills, Jones, and Todd (2012) introduced the "similarity drop" method that estimated drops in consecutive semantic similarity from a DSM, BEAGLE (Jones and Mewhort, 2007), to assign switch events in fluency lists (see Barattieri di San Pietro, Luzzatii, Ferrari, de Girolamo, & Marelli, 2023 for similar approaches). Another related approach is to construct networks consisting of words produced by participants in the VFT, where edges between words are defined either based on some external metric of semantic similarity or derived from the behavioral data itself, such as how frequently two words were produced in succession to each other within a cohort of participants (Christensen & Kenett, 2021; Kenett, Anaki, & Faust, 2014; Sung et al., 2013; Goñi et al. 2011; Zemla & Austerweil, 2018).

Overall, automated measures have uncovered important insights into the nature of memory search in neurotypical and clinical populations. For example, while neurotypical individuals will begin the task with many switch events and over time decrease their switches, people with aphasia tend to switch at a consistent, and low, rate (Bose et al., 2022). Individuals with schizophrenia have also been observed to switch less than neurotypical participants (Lundin et al., 2020; Okruszek, Rutkowska, & Wilińska, 2013; Robert et al., 1998). Despite these insights, however, there is limited work exploring how well these methods map onto and compare against each other in predicting an *individual's* pattern of search. Indeed, the majority of the work in determining clusters and switches via automated methods focuses on validating these methods against clinical datasets for prognostic or diagnostic purposes (e.g., Bushnell et al., 2022). While this is certainly useful, focusing on diagnostic/prognostic markers in clinical populations sidesteps the critical question of *how* individuals search for concepts within memory in the first place. For instance, do individuals attend to different lexical sources (e.g., semantic and/or phonological information) to the same extent when searching through items within their lexicon? Which automated method of clustering and switching most closely aligns with how an individual conceptualizes their own search? Overall, there is a critical need to focus on psychologically motivated models of clustering and switching that shed light on memory search processes within individuals.

Most of the work on memory search within individuals has focused on exploring the validity of cluster-switch methods in diagnostic settings or at the group level, such as examining clinical populations (Zemla et al. 2020; Kenett et al., 2013), or low- and high-creative individuals (Kenett et al., 2014). Some recent work has examined individual-level variance in verbal fluency by constructing individual-level semantic networks using hierarchical Bayesian estimation (Zemla & Austerweil, 2018) or simulating random walks on individual networks based on the spreading activation mechanism (Christensen & Kenett, 2021); also see Morais et al., 2013 for a similar approach using free association data). This work has been informative in understanding how one can use fluency lists to map out the general *structure* of semantic organization for an individual. However, these methods do not provide insight into the specific lexical cues that may have been used to generate the fluency list itself, and therefore focus less on the *process* of memory search. Understanding these issues from a lexical perspective could provide more insight into the processes involved in search during the semantic version of the VFT (SFT) and ultimately aid in the development of predictive, as opposed to postdictive methods of clustering and switching (e.g., Zhang and Jones, 2022). In a recent neuroimaging study, Lundin et al. (2023) found that hippocampal and cerebellar activation was greater while switching than while clustering during VFT, and this activity increased leading up to decisions to switch, suggesting that individuals were strategically navigating their mental lexicon during the task. Critically, clusters and switches were defined based on a metric of semantic similarity from a DSM as well as participant evaluations, although the extent to which different automated methods map onto these participant evaluations remains unexplored. Understanding search from an individual perspective requires comparative work that attempts to predict and measure individual-level variance using automated methods.

Current study

The current study had two main goals. Our first goal was to examine the predictive power of a variety of hand-coded norms (based on and adapted from Troyer, 2000) and automated methods of clustering and switching by comparing model-predicted clusters to *participant-designated* and *rater-designated* clusters for a dataset of fluency lists (collected by Lundin et al., 2023; Hills et al., 2012). We were interested in evaluating which cluster-switch method would best be able to account for the broader pattern of clusters and switches identified by (a) the *same* participants who generated the fluency lists (*participant-designated*) and (b) an independent group of raters (*rater-designated*) for the same lists via a pre-registered experiment. Although there are several methods in the literature, we focused on a selection of a few psychologically motivated methods that directly map onto a search process within memory and focus on the lexical content of the items produced. We hypothesized that the methods based on hand-coded norms would generally provide a better fit to both participant and rater-designated clusters and switches, given that these norms have been curated and refined over multiple iterations, and the processes used by researchers to group items may be similar to how raters and participants group items, i.e., shared method variance may contribute to the predictive power of these norms. However, we were more interested in understanding how well the automated methods compare to each other and against this normative baseline. Taken together, these analyses could inform the broader dissemination and utility of different automated methods in the field and provide more insights into the psychological underpinnings of clustering and switching behavior from an individual perspective. A related sub-goal of our study was to evaluate the generalizability of these methods across different domains. While there has been some application of cluster and switch methods in novel domains (e.g. *items in grocery stores*; Zhao et al., 2013, *foods/fruit*, Kim et al., 2019), the vast majority of literature on the SFT has exclusively focused on the domain of *animals*. Therefore, in this study, we sought to test and extend the applicability of these methods in two other domains in addition to *animals*: *occupations* and *foods*. While category norms exist for *animals* and *foods*, to our knowledge, there are currently no validated datasets available for the *occupations* domain. Therefore, our study could be informative in understanding which method may best be able to account for variance in domains of VFT where norms are not available.

A second exploratory goal of our work was to examine a novel metric for capturing *idiosyncratic variance* in the fluency task using the rater-based and participant designations of clusters. Specifically, whether the vast array of experiences and strategies that individuals report when performing the VFT (see Lundin et al., 2023) can be contained within a single measure of idiosyncratic variance, has not been previously explored in the literature. In this work, using the designations provided by our independent raters, we computed a discordance measure between rater-designated and participant-designated clusters and switches

and assigned idiosyncrasy scores to each transition. We examined the stability of this discordance measure across domains (e.g., *animals*, *foods*, etc.) and also sought to identify the behavioral signatures of idiosyncratic transitions by investigating whether idiosyncratic transitions varied in their lexical content or correlated with measures of fluency performance, such as number of items produced, mean cluster size, and number of clusters. If idiosyncratic scores for an individual do indeed remain stable across different domains, and also systematically vary with other indicators of search and/or performance, the scores are likely capturing something specific to an individual. For instance, prior work has shown that raw scores from verbal fluency predict performance on convergent and divergent measures of creativity (Gerver et al., 2023) and clustering behavior correlates with divergent thinking (Ovando-Tellez et al., 2022). It is possible that producing more idiosyncratic responses in the VFT is linked to other measures of creativity, such as "forward flow" (Beaty et al., 2021; Gray et al., 2019), or may indicate that individuals are accessing remote/atypical connections in the lexicon which require them to inhibit semantic and/or phonologically activated or high-frequency concepts that come to mind (Gupta et al., 2012). Overall, these exploratory analyses could shed light on the processes that differentiate individuals.

Methods

Participants

Based on the pre-registered criteria (https://aspredicted.org/3X1_9FK), a total of 260 participants were recruited from the psychology participant pool at Bowdoin College ($N = 89$) for course credit and from the survey website Prolific ($N = 171$; Palan & Schitter, 2018) at the rate of $8-$11/hour. The main inclusion criterion for participation was to be a native English speaker. Out of the 246 participants who began the study, a total of 35 participants were excluded from the analysis based on pre-registered criteria (i.e., failing attention checks, incomplete experiment). The result was a final sample size of 211 participants with a mean age of 31.4 years ($SD = 14.4$ years) with an average of 14.6 years of education ($SD = 2.81$ years). The race distribution was 72.51% white, 9.95% Asian, 5.69% Black, 0.47% American Indian/Alaskan Native, and 11.37% other races/more than one race. The gender distribution was 55.45% women, 42.18% men, 0.95% nonbinary, and 0.95% transgender.

Materials

We used 60 previously generated SFT lists as our stimuli, collected by Hills et al. (2012) and Lundin et al. (2023), spanning across three domains: *animals*, *foods*, and *occupations*. Hills et al. 2012 asked participants to generate items for three minutes. The task was administered online. Lundin et al. (2023) asked participants to generate fluency lists for three minutes at a time during functional magnetic resonance imaging (fMRI), and then asked them post hoc to group items they generated based on their perceived relatedness. To limit the duration of the present independent rater task, anonymized IDs of thirty participants were randomly sampled from the Hills et al. dataset, and their fluency lists for the three domains were then combined with the lists from thirty participants from Lundin et al.'s dataset. The stimuli were then split into 10 experimental lists, such that each experimental list consisted of 6 fluency lists per domain and counterbalanced across participants.

Procedure

After consenting to participate, participants proceeded to the online experiment, programmed in jSPsych (De Leeuw, 2015) and implemented on the hosting website cognition.run. Participants were instructed that they would be grouping words across three domains (*animals*, *foods*, and *occupations*). The instructions emphasized that *how* words were grouped was entirely up to their discretion. During each experiment trial in the practice and experiment phase, participants first saw a fixation cross presented for 500 ms. They were then presented with two consecutive words from a fluency list and asked if they would "put them in the same group." Participants clicked a "yes, same group" or "no, new group" button, after which they advanced to the next experiment trial. Although participants cycled through several fluency lists for a given domain in the same sequence that the words were produced by the original participants, they were not explicitly informed that these were different "lists" generated by other participants, to limit top-down processing and demand characteristics and mimic the fluency generation process as much as possible. Instead, they were simply instructed that words may be repeated across different trials. At the end of each fluency list, participants were given an attention check, which asked them to type one of the words that they had just grouped. Therefore, each participant completed 18 total attention checks. Participants were given feedback about the accuracy of their responses during the attention check in the practice trials but not in the experimental session. At the end of the experiment, participants filled out a survey where we asked them to specify any strategies they used to group the pairs, in addition to general demographics.

Models of assigning clusters and switches

We adapted the Python package *forager* (Kumar, Apsel, Zhang, Xing, & Jones, 2023) to obtain cluster-switch designations for our three domains (*animals*, *foods*, and *occupations*), by adding new cluster-switch algorithms as well as additional functionality to process the new domains. Detailed descriptions of the different methods are provided in *forager* documentation[1] and the Github repository.[2] Below, we provide brief descriptions of the methods used in the current analyses. As discussed earlier, we focused on two normative methods and five lexical content-based automated measures to designate clusters and switches in the semantic fluency task. For all lexical content-based methods, the underlying semantic space was a large language model, the Universal Sentence Encoder (Cer et al., 2018) that can flexibly encode single (e.g., *shark*) and multi-word exemplars (e.g., *blacktip reef shark*) into a 512-dimensional vector. A total of ten responses were excluded from the *animals* domain due to being absent from *forager*'s built-in vocabulary for *animals*. These mostly included items that were either not animals (e.g., *Venus fly trap*, *house*, etc.) or fictional (e.g., *yeti*, *qilin*, etc.). For the *foods* and *occupations* domains, *forager* does not provide a pre-existing vocabulary; therefore, a new vocabulary space was constructed for these domains and all fluency responses from these domains were included. Phonological similarities were defined as the normalized edit distance between the phonetic transcriptions of words produced in the VFT (see Kumar, Lundin, and Jones, 2022) and word frequency estimates were derived from the *wordfreq* package in Python (for implementation details, please refer to the *forager* documentation).

Norm-based associative and categorical baselines (animals and foods only). We used previously validated schemes provided via the SNAFU package[3] (Zemla et al., 2020) for the *animals* and *foods* domains to assign clusters and switches. These schemes were adapted from the original Troyer (2000) subcategorization scheme and provide predefined subcategories for a list of 1227 animals and 647 foods. Within a participant-generated fluency list, based on an *associative* search model, if an item did not share any subcategory with the preceding item, it was considered a switch. If two consecutive items shared at least one subcategory, they were designated as being in the same cluster. For example, within the *animals* domain, consider the fluency list *porcupine-*

[1] https://github.com/thelexiconlab/forager.
[2] https://github.com/thelexiconlab/whats-in-my-cluster.
[3] https://github.com/AusterweilLab/snafu-py/tree/master/schemes.

peacock-chicken-turkey-owl-eagle-eel-platypus produced by a participant (50020) shown in Fig. 1, alongside their own designations for different transitions as well as the semantic and phonological similarity for each transition. Based on the associative model, the transition from *porcupine* to *peacock* is a switch (indicated by the vertical dashed yellow line) because these animals belong to different subcategories (rodents and birds, respectively) in the norms. Similarly, the transition from *eagle* to *eel* is also a switch due to no overlap in their subcategories (bird and water, respectively). Following Hills et al. (2015), we also examined a variant of the norms-based method (*categorical* search), where clusters and switches were assigned not simply based on consecutive shared category membership, but instead based on whether the current item shared a category with every other preceding item that was considered part of a cluster. In the case of our example from Fig. 1, for the transition from *owl* to *eagle*, the categorical method would not only examine whether *owl* and *eagle* share the same category, but also whether *eagle* shares the same category label as all the other items in the ongoing cluster (i.e., bird). In this case, both associative and categorical methods predict the same pattern of cluster and switches, but predictions can differ for other transitions (e.g., for *cat-dog-wolf*, *dog-wolf* would be classified as a switch because it does not belong to the ongoing cluster of pets, even though *dog* and *wolf* share the canine category).

Similarity-drop. Similarity drop is a switch heuristic proposed by Hills et al. (2012). Based on this method, a transition is classified as a switch event if there is a drop in semantic similarity between consecutive items preceded and followed by an immediate rise in semantic similarity. We considered two variants, one based purely on semantic similarity and a "multimodal" version, that incorporated consecutive semantic and phonological similarity (weighted by a parameter α via (α) [semantic similarity] + (1-α) [phonological similarity], i.e., higher α indicates more weight to semantic similarity) into the decision to assign clusters and switches, based on insights from prior work that used the same VFT data as the present study suggesting that phonological similarity may be important in understanding clustering behavior (see Kumar, Lundin, and Jones, 2022; Kumar, Zhang, Apsel, & Jones, 2023). In the example from Fig. 1, the similarity drop method assigns transitions with low semantic similarity (flanked by high semantic similarity on either end) a "switch" designation, i.e., the transitions *porcupine-peacock*, *turkey-owl*, and *eagle-eel* (indicated by vertical dashed green lines). Specifically, because the semantic similarity between *porcupine* and *peacock* is lower than the similarity between *whale-porcupine* and *peacock-chicken*, *peacock* marks the start of a new cluster, i.e., a switch event. On the other hand, due to taking phonological similarity into account, the multimodal similarity drop method (with $\alpha = 0.8$) only considers the *turkey-owl* and *eel-platypus* transitions as switches, given that *porcupine-peacock* and *eagle-eel* share high phonological similarity.

Delta-similarity. The delta similarity method (Lundin et al., 2023) is an extension of the similarity drop method that predicts switches and clusters based on whether a rise or drop in semantic similarity exceeds set thresholds. Two parameters, the "rise" and "fall" thresholds (ranging between zero and one), control *how much* the similarity needs to change to signify the start of a new cluster, i.e., a switch. For example, to evaluate whether *peacock* marks the start of a new cluster (i.e., a switch event), this method would compare the change in z-scored similarity from the preceding word to current word (e.g., *porcupine* to *peacock*) to the similarity between the current word and next word (e.g., *peacock* to *chicken*) in the list. Given that the item before *peacock*, i.e., *porcupine* marked the start of a new cluster, the algorithm would then evaluate if the observed similarity difference between *peacock-chicken* and *porcupine-peacock* is higher than a given "rise" threshold. That is, to be part of

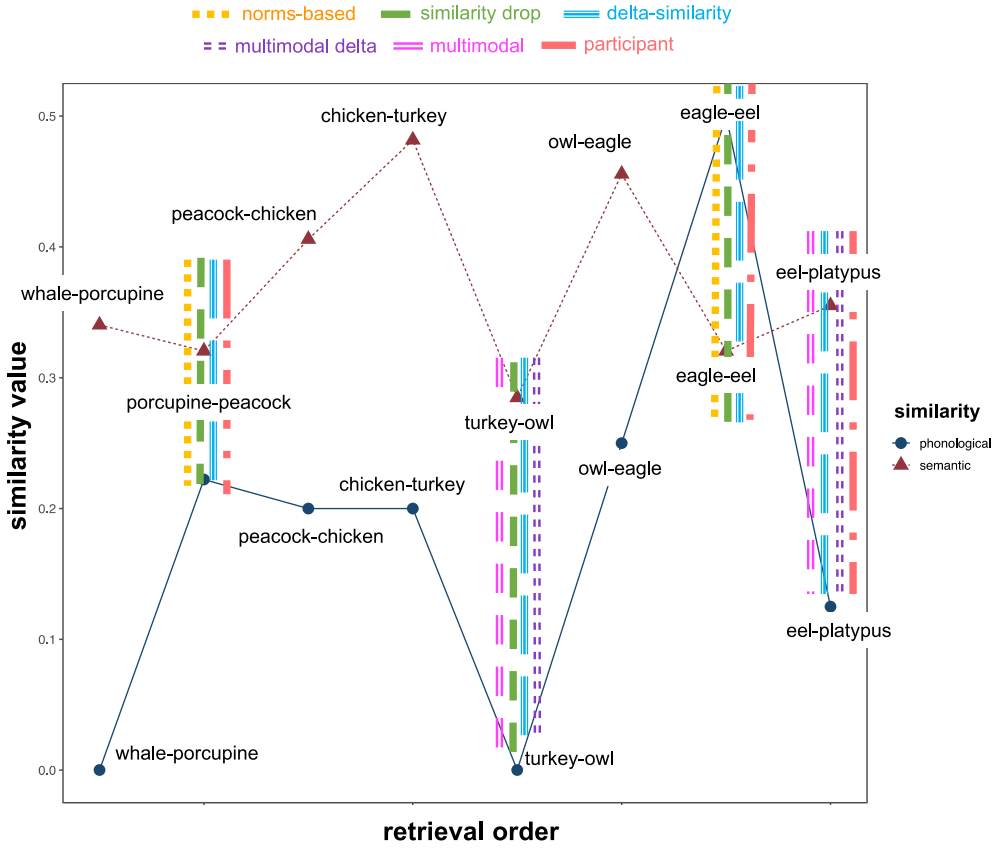

Fig. 1. Predictions of different norms-based and automated methods of clustering and switching for a series of transitions within a fluency list. Vertical lines indicate "switch" events designated by different methods. Designations for the delta method are based on rise = 0.75 and fall = 0.75. Designations for the multimodal method are based on α = 0.8. Designations for the multimodal delta method are based on α = 0.6, rise = 0.75, and fall = 0.75. Designations for both associative and categorical methods were identical and are therefore collapsed into one line.

Table 1
Model performance when predicting participant-designated clusters and switches across three verbal fluency domains using different cluster-switch methods.

Domain	Method	R^2 (conditional delta) [95 % CI]	fixed effect	AIC	BIC
animals	norm-based (associative)	0.33 [0.3,0.41]	2.61	1446.3	1462.06
	norm-based (categorical)	0.28 [0.25,0.36]	2.33	1513.14	1528.91
	multimodal delta (α = 1, rise = 0.75, fall = 0.75)	0.23 [0.2,0.3]	1.89	1633.68	1649.45
	delta (rise = 0.75, fall = 0.75)	0.23 [0.21,0.3]	1.89	1633.68	1649.45
	multimodal similarity drop (α = 0.8)	0.15 [0.14,0.23]	1.35	1725.22	1740.98
	similarity drop	0.15 [0.13,0.22]	1.33	1727.56	1743.32
	SVD (cosine = 0.9, GTOM = 0.7)	0.1 [0.09,0.17]	1.29	1806.47	1822.24
foods	norm-based (associative)	0.27 [0.24,0.35]	2.41	1716.28	1732.23
	multimodal delta (α = 0.7, rise = 1.0, fall = 0.5)	0.19 [0.17,0.27]	1.8	1816.8	1832.75
	delta (rise = 1.0, fall = 0.5)	0.18 [0.16,0.25]	1.58	1842.11	1858.06
	multimodal similarity drop (α = 1.0)	0.14 [0.12,0.21]	1.21	1874.78	1890.73
	similarity drop	0.14 [0.12,0.21]	1.21	1874.78	1890.73
	norm-based (categorical)	0.12 [0.11,0.19]	1.17	1897.14	1913.09
	SVD (cosine = 0.9, GTOM = 0.6)	0.08 [0.07,0.14]	0.96	1965.53	1981.48
occupations	multimodal delta (α = 0.9, rise = 0.75, fall = 0.25)	0.23 [0.21,0.33]	1.64	1329.53	1344.5
	delta (rise = 0.75, fall = 0.0)	0.22 [0.2,0.32]	1.62	1333.01	1347.97
	multimodal similarity drop (α = 1.0)	0.16 [0.14,0.25]	1.04	1390.32	1405.29
	similarity drop	0.16 [0.14,0.25]	1.04	1390.32	1405.29
	SVD (cosine = 0.8, GTOM = 0.3)	0.09 [0.09,0.18]	0.4	1448.87	1463.83

Note: Confidence intervals around the conditional R^2 denote 95% bootstrapped confidence intervals obtained from 1000 random samples with replacement from the empirical data.

the same cluster as *porcupine*, *peacock* should be more similar to *porcupine* than it is to *chicken* and this difference should be greater than a given threshold. In this work, we explored different thresholds for the "rise" and "fall" parameters. Unlike the similarity drop method, this method is not overly sensitive to minor changes in similarity and also allows for single-item clusters. As shown in Fig. 1, the delta similarity method's designations (indicated by vertical blue lines, with specific rise and fall thresholds) align with the similarity drop method in this case, except for the *eel-platypus* transition, where the delta method considers it a "switch" because the semantic similarity between *eel* and *platypus* is not sufficiently high to be counted as within the same cluster as *eagle* and *eel*. Additionally, we also explored a "multimodal delta" method that combined the heuristics of the delta similarity method and the multimodal method, such that "sufficiently" high *combined* semantic and phonological similarity (weighted by a parameter α) contributed to a cluster designation, and "sufficiently" low combined semantic and phonological similarity contributed to a switch designation. In Fig. 1, the multimodal delta similarity method (with certain parameter settings) only scores the *turkey-owl* and *eel-platypus* transitions as switches, due to considering relative rises and drops in *both* semantic and phonological similarity.

SVD/GTOM. This method was based on an algorithm described by Sung et al. (2013), that assigns clusters and switches based on the overall pattern of responses across all participants. The algorithm first creates a word by participant matrix by enumerating whether or not a particular word was produced by a given participant. Singular value decomposition (SVD) is then performed on this matrix and preliminary clusters are assigned based on a cosine similarity threshold between the resulting word vectors. These preliminary clusters are then further refined via a generalized topological overlap measure (GTOM), which considers the shared similarity of clusters between words above a certain threshold. The rationale behind this method is to consider words that are produced by most participants as part of the same cluster, while taking into account the shared neighborhoods of words. While we used the default number of dimensions for SVD used by Sung et al. (i.e., $n = 5$), we systematically varied the cosine and GTOM thresholds in our analyses. As shown in Fig. 1, one parameter setting of this method considers *porcupine-peacock*, *turkey-owl*, *eagle-eel*, and *eel-platypus* to be switch events, similar to the other automated methods.

Results

Predicting participant designations

We compared the norms and automated model-based designations to designations provided by the same participants who produced the fluency lists, i.e., participant designations. Specifically, in the Lundin et al. (2023) study, individuals were asked to group items together in a post-hoc manner. For instance, in Fig. 1, the participant indicated that they "switched" when going from *porcupine* to *peacock*, and then again from *eagle* to *eel*, and *eel* to *platypus* (as indicated by the vertical red lines). We evaluated how well different norms-based and automated methods were able to predict these participant-designated clusters and switches, via a generalized mixed effects model (family = binomial) with a fixed effect for the method, and a random intercept for the fluency list. Table 1 displays the overall patterns (we only report the best-performing model for methods with varying parameters). Two main findings are of note here: first, for the domains where norms were available (*animals* and *foods*), the *associative* norms-based method was the best predictor of participant-designated clusters and switches (highest R^2 and lowest BIC). Second, among the automated methods, the delta-based similarity models (multimodal and delta) best accounted for the participant-designated clusters and switches across the three domains, with phonology having more of an influence in the *foods* and *occupations* domains (as indicated by $\alpha \neq 1$, i.e., non-zero weight on phonological similarity) than the *animals* domain. Fig. 2 displays the overall performance of the multimodal delta method across different parameter settings: overall, high α, and moderate *rise* and *fall* values best predicted the participant designations across the three domains.

To understand the relative contribution of different switch methods in predicting these participant designations, we also fit a stepwise logistic mixed effects multiple regression model with each of the switch methods added as a predictor to the model. Given that several of the switch methods have different predictions based on their parameter settings, we chose the predictors based on the best-fitting parameter settings from Table 1 for all models. Overall, the model with all the switch methods provided a significantly better fit than models with fewer methods (p's $<.05$) in likelihood ratio tests across all domains), suggesting that the different methods were likely explaining unique amounts of variance. As shown in Table 2, the full model increased the explained variance in all domains relative to the single-predictor models (9 and 10 percentage points for the *animals* and *foods* domain, respectively, and 2 percentage points for *occupations*). Interestingly, the norms-

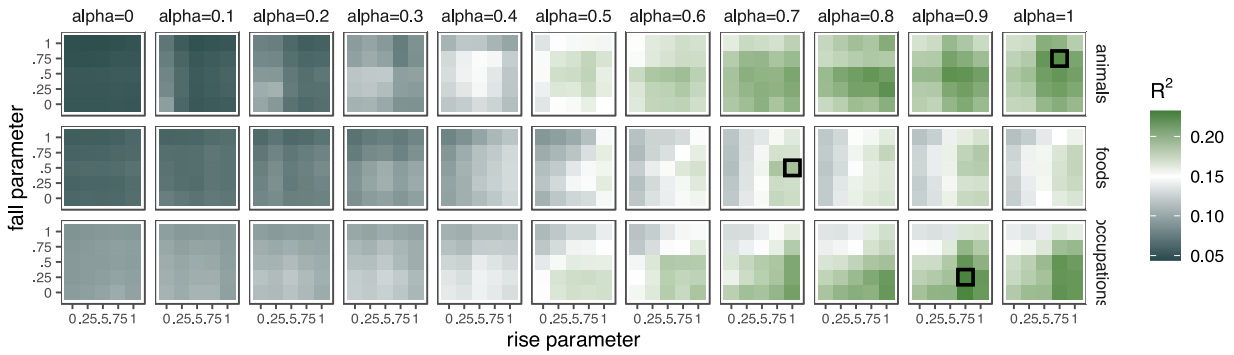

Fig. 2. Heat map of explained variance (R^2) in participant-designated clusters and switches for different parameter settings of the multimodal delta model (α denotes the weight on semantic similarity vs. phonological similarity, rise and fall thresholds denote the thresholds required to designate clusters and switches, respectively). Black outlines indicate the best-performing parameter values. Overall, high α, and moderate rise and fall values best predicted the participant designations across the three domains.

based associative method, the multimodal delta similarity method, and the SVD method were the only three predictors that remained significant across all domains, when accounting for other methods in the full model.

Predicting rater designations

Within each fluency list, we first computed an aggregate "rating" from 0 to 1 for each transition based on the data from our experiment. This rating represented the mean tendency of raters to assign a pair of words produced within a list as belonging to the same cluster vs. different clusters. A low rating closer to 0 implied that words were designated as being part of the "same" group or "cluster" (e.g., *chef-baker*, *fish-chicken*), and a high rating closer to 1 implied that words were designated as belonging to a "new group" or signifying a "switch" (e.g. *peach-pizza*, *flamingo-cow*). Fig. 3 displays the overall pattern across all words and fluency lists. As shown, raters grouped items in predictable ways. Items with high semantic similarity were rated as part of the same cluster and semantically distant items were designated as switch events, $\chi^2(1, N = 7036) = 5861.24$, $p < .001$. This pattern was steepest for *animals*, but followed the same general trend for *foods* and *occupations* as indicated by a significant interaction ($p < .001$). Interestingly, high phonological similarity was also associated with "cluster" designations, and low phonological similarity was associated with "switch" designations, $\chi^2(1, N = 7036) = 164.61$, $p < .001$, but this pattern was weaker for *animals*, relative to *foods* and *occupations*, as indicated by a significant interaction ($p < .001$). Word frequency also showed an effect, such that higher average word frequency was associated with "cluster"/ "same group" designations, $\chi^2(1, N = 7036) = 21.60$, $p < .001$.

In a second set of analyses, we explored how well the different model-based cluster-switch methods predicted these rater-designated clusters and switches. As shown in Table 3, the patterns were generally consistent with the self-designations, in that the norms-based associative model performed best overall, and the multimodal delta and delta methods were the best-performing automated methods. Therefore, among the automated methods, overall, the multimodal (and delta) methods were most predictive of participant-designated as well as rater-designated clusters and switches across domains.

Measuring idiosyncratic variance

Using the rater designations from our experiment ($N = 211$), we computed an idiosyncratic score for each transition. This idiosyncratic score measured how different a participant-designated response was from the rater-designated responses for the same transition by computing a measure of *discordance* for each transition (based on percent disagreement). For example, consider the sequence of responses shown in Table 4 for an individual (ID = 50003). When the individual designated a particular pair of consecutive responses (e.g., *duck-robin*) as related (i.e., part of a "cluster"), the idiosyncratic score for that transition was the proportion of raters who designated the same transition as a "switch" (i.e., 8 out of 20 raters). On the other hand, when the individual designated a pair of consecutive responses (e.g., *robin-woodpecker*) as

Table 2
Results from the multiple regression model predicting participant designations using multiple switch methods. Asterisk (*) indicates that the predictor was significant in the full model.

Domain	R^2	Term	Estimate	Standard error	Statistic	p value
animals	0.42	intercept	−2.98	0.25	−11.73	<.001*
		norms (associative)	1.71	0.28	6.05	<.001*
		norms (categorical)	0.47	0.27	1.73	0.084
		similarity drop	0.39	0.25	1.54	0.123
		multimodal delta (α = 1, rise = 0.75, fall = 0.75)	1.00	0.16	6.37	<.001*
		multimodal (α = 0.8)	0.35	0.25	1.41	0.159
		SVD (cosine = 0.9, GTOM = 0.7)	0.52	0.21	2.48	0.013*
foods	0.37	intercept	−3.68	0.28	−13.00	<.001*
		norms (associative)	2.10	0.22	9.41	<.001*
		norms (categorical)	−0.08	0.17	−0.45	0.651
		similarity drop	0.65	0.13	5.13	<.001*
		delta (rise = 1, fall = 0.5)	0.39	0.20	2.01	0.044*
		multimodal delta (α = 0.7, rise = 1, fall = 0.5)	0.90	0.21	4.34	<.001*
		SVD (cosine = 0.9, GTOM = 0.6)	0.61	0.20	3.08	0.002*
occupations	0.25	intercept	−1.28	0.21	−6.16	<.001*
		similarity drop	0.58	0.14	4.13	<.001*
		delta (rise = 0.75, fall = 0)	0.48	0.33	1.43	0.153
		multimodal delta (α = 0.9, rise = 0.75, fall = 0.25)	0.98	0.33	2.99	0.003*
		SVD (cosine = 0.8, GTOM = 0.3)	0.29	0.14	2.02	0.044*

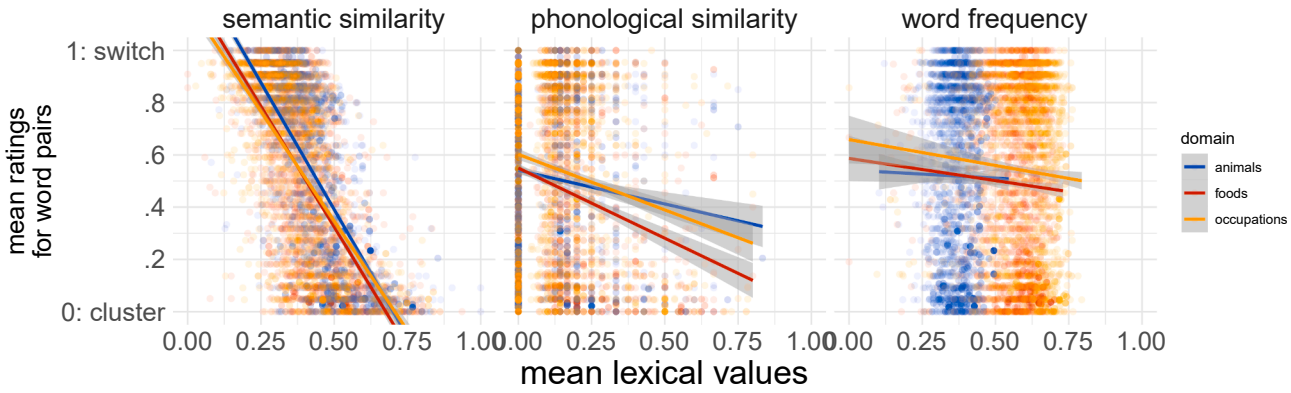

Fig. 3. Mean ratings for consecutive word pairs as a function of semantic similarity (left), phonological similarity (middle), and word frequency (right). Lower ratings indicate "cluster" responses and higher ratings indicate "switch" responses.

Table 3
Model performance when predicting rater-designated clusters and switches across three domains using different cluster-switch methods.

Domain	Model	R2 [95 % CI]	Fixed effect	AIC	BIC
animals	norm-based (associative)	0.41 [0.39,0.46]	0.43	574.09	597.39
	norm-based (categorical)	0.35 [0.33,0.4]	0.4	813.91	837.22
	delta (rise = 0.75, fall = 0.5)	0.32 [0.31,0.37]	0.38	963.53	986.83
	multimodal delta (α = 1.0, rise = 0.75, fall = 0.5)	0.32 [0.31,0.37]	0.38	963.53	986.83
	multimodal similarity drop (α = 0.8)	0.19 [0.18,0.24]	0.25	1404.93	1428.23
	similarity drop	0.19 [0.18,0.24]	0.25	1407.66	1430.96
	SVD (cosine = 0.9, GTOM = 0.8)	0.11 [0.1,0.15]	0.26	1637.88	1661.18
foods	norm-based (associative)	0.44 [0.42,0.48]	0.47	503.81	527.28
	multimodal delta (α = 0.9, rise = 0.75, fall = 0.5)	0.33 [0.31,0.38]	0.36	1071.87	1095.34
	delta (rise = 0.75, fall = 0.5)	0.32 [0.31,0.37]	0.35	1118.58	1142.06
	norm-based (categorical)	0.25 [0.24,0.3]	0.32	1254.34	1277.81
	multimodal similarity drop (α = 0.9)	0.2 [0.19,0.25]	0.23	1494.77	1518.24
	similarity drop	0.2 [0.19,0.25]	0.23	1497.81	1521.29
	SVD (cosine = 0.9, GTOM = 0.3)	0.11 [0.11,0.16]	0.16	1757.62	1781.1
occupations	delta (rise = 0.75, fall = 0.5)	0.28 [0.27,0.34]	0.32	797.13	819.39
	multimodal delta (α = 0.9, rise = 0.75, fall = 0.5)	0.28 [0.27,0.34]	0.32	795.9	818.16
	multimodal similarity drop (α = 0.9)	0.17 [0.16,0.23]	0.21	1038.06	1060.32
	similarity drop	0.17 [0.16,0.23]	0.21	1039.02	1061.28
	SVD (cosine = 0.8, GTOM = 0.3)	0.07 [0.07,0.13]	0.1	1240.8	1263.06

Note: Confidence intervals around the conditional R^2 denote 95% bootstrapped confidence intervals obtained from 1000 random samples with replacement from the empirical data.

unrelated (i.e., a switch), the idiosyncratic score was the proportion of raters who designated the same transition as a "cluster" (i.e., 20 out of 20 raters). Importantly, due to the large sample size in our behavioral experiment, each fluency list was rated on average by 21.1 raters (SD = 1.06), giving us reasonably stable estimates of idiosyncratic scores for each transition.

Do idiosyncratic transitions have any lexical markers? Fig. 4 displays the distribution of idiosyncratic scores as a function of semantic and phonological similarity, as well as word frequency estimates obtained via *forager*. Higher idiosyncratic scores were associated with low semantic, $\chi^2(1, N = 4003) = 88.51$, $p < .001$, low phonological similarity, $\chi^2(1, N = 4003) = 12.84$, $p < .001$, as well as low average word frequency, $\chi^2(1, N = 4003) = 4.38$, $p = .036$. The pattern for phonological similarity differed across domains, as indicated by a significant interaction, $\chi^2(2, N = 4003) = 9.36$; $p = .009$, such that there was no systematic relationship between idiosyncratic scores and phonological similarity for the *animals* domain, and the pattern for average frequency also appeared to be in the opposite direction for *animals* ($p = .076$ for interaction effect). Overall, idiosyncratic scores were associated with lower lexical content, i.e., lower semantic similarity, and lower frequency and phonological similarity (to an extent).

Are some individuals more idiosyncratic than others? Fig. 5 displays the distribution of idiosyncratic scores within the sample, aggregated at the individual level. There was considerable variation among individuals, where some individuals grouped their fluency items in a way that was highly consistent with rater designations and therefore had lower idiosyncratic scores, whereas others showed vast disagreement with the rater designations and therefore had higher idiosyncratic scores.

Are individual-level idiosyncratic scores stable across domains? We used simple linear regression models to predict the average idiosyncratic score for an individual on one domain (e.g., *animals*) via their score on another domain (e.g., *foods*). Then, we scrambled the scores across individuals and obtained an estimate of explained variance across 1000 permutations of the data. As shown in Fig. 6, the explained variance was significantly higher in the original data than the permuted data (p's $< .003$ for all permutation tests), suggesting that there was high systematicity in the idiosyncratic scores within an individual across all domains.

Do individual level idiosyncratic scores correlate with fluency performance? Higher idiosyncratic scores were related to fewer clusters, $F(1,84) = 14.45$; $p < .001$ (see Fig. 7), suggesting that individuals who were more idiosyncratic also switched more. Number of items produced did not have a reliable relationship with idiosyncratic scores ($p = 0.082$). Critically, these patterns did not significantly differ across domains (p's $> .80$ for interaction effects).

Participant and rater strategies

After completing a set of category and letter-based verbal fluency

Table 4
Examples of idiosyncratic transition scores in verbal fluency lists based on participant and rater-based designations.

Transition	Individual designation	Number of raters who designated this transition a "cluster"	Number of raters who designated this transition a "switch"	Idiosyncratic score
duck → robin	cluster	12	8	0.4
robin → woodpecker	switch	20	0	1
woodpecker → eagle	cluster	19	1	0.05
eagle → hawk	cluster	20	0	0
hawk → falcon	cluster	20	0	0
falcon → hippo	cluster	1	19	0.95
hippo → weasel	switch	2	18	0.1

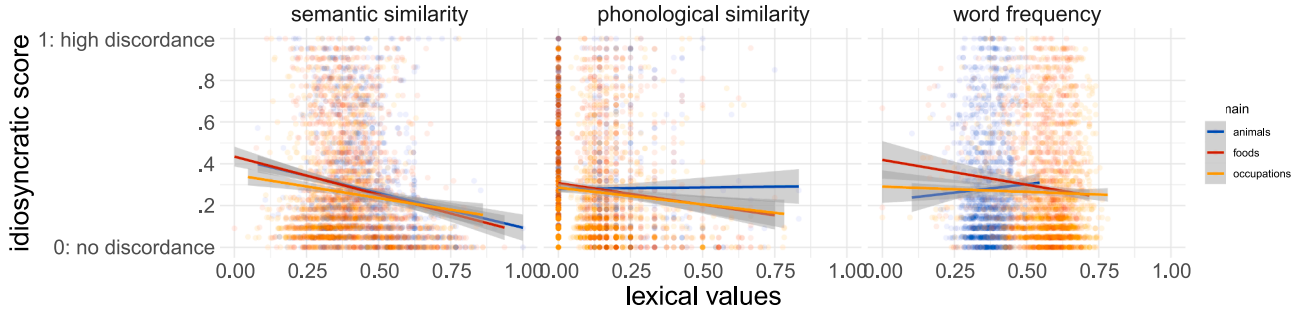

Fig. 4. Averaged idiosyncratic scores for pairwise transitions as a function of semantic similarity (left), phonological similarity (middle), and word frequency (right). Lower idiosyncratic scores indicate more agreement of the participant-designated response with the raters, whereas higher idiosyncratic scores indicate more disagreement / discordance with the raters.

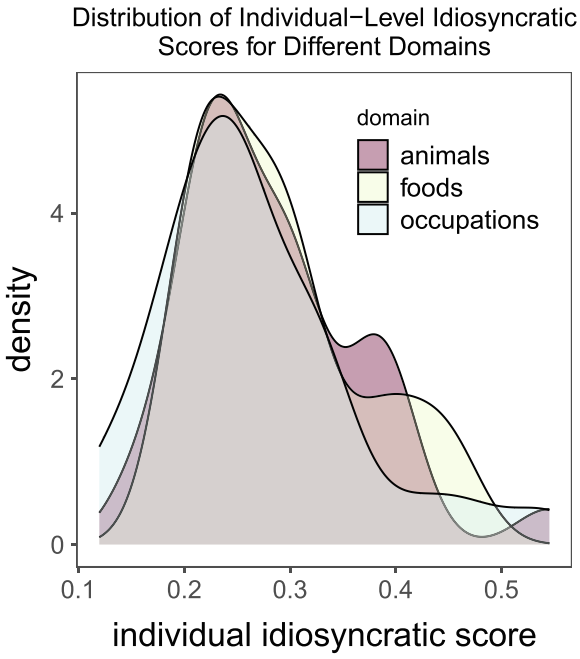

Fig. 5. Distribution of individual-level idiosyncratic scores across the three domains. Lower idiosyncratic scores indicate that the designations indicated by the participant were similar to rater designations, whereas higher idiosyncratic scores indicate more discordance between participant and rater designations for the same transitions.

tasks, Lundin et al. (2023) asked participants to report any strategies they used to generate the items and then grouped the participant-reported strategies into themes. In a similar vein, we asked the raters in the current study who classified items produced by the participants from Lundin et al.'s and Hills et al.'s studies to report the strategies they used to group the word pairs together. These strategies were then independently coded by two authors (NL and AK) using the protocol described in Lundin et al., beginning with strategy themes described by participants who generated the fluency responses, such as using semantic associations, phonetic associations, imagery, personal experiences, etc. to perform the task. Overall, there was high inter-rater reliability ($\kappa = 0.90$, $z = 39.2$, $p < .001$) across all themes as well as within each theme (mean $\kappa = 0.77$).

Table 5 displays the overall patterns of independent rater-based strategies in comparison to the participant strategies reported by Lundin et al. (2023).[4] Overall, although both raters and participants who generated the responses overwhelmingly made use of semantic characteristics to rate and generate items, there were also differences. First, less than 1% of raters used phonological or orthographic information to group items, in contrast to 63% of the participants who generated the fluency responses. This finding is unsurprising given that the Lundin et al. (2023) study included letter-based fluency tasks (in which phonetic association-based strategies are more commonly employed), whereas the current study included solely category-based fluency tasks. Interestingly, participants who generated the fluency responses were more likely than independent raters of the word pairings to report using personal life experiences and visual imagery. Finally, 7% of the raters reported strategies that were not used by the participants. These included using color as a cue (e.g., *carrot-orange*), changing strategies mid-task, using number of legs for animal groupings, etc. Therefore, raters may have been using additional cues that may not be easily available to participants during the generation process, and the participants may have used strategies to generate responses that the raters were less likely to use while judging response similarity.

Discussion

In this work, we examined the critical question of how individuals

[4] For final counts/percentages, a response was considered to be using a particular strategy if it was rated as such by at least one rater.

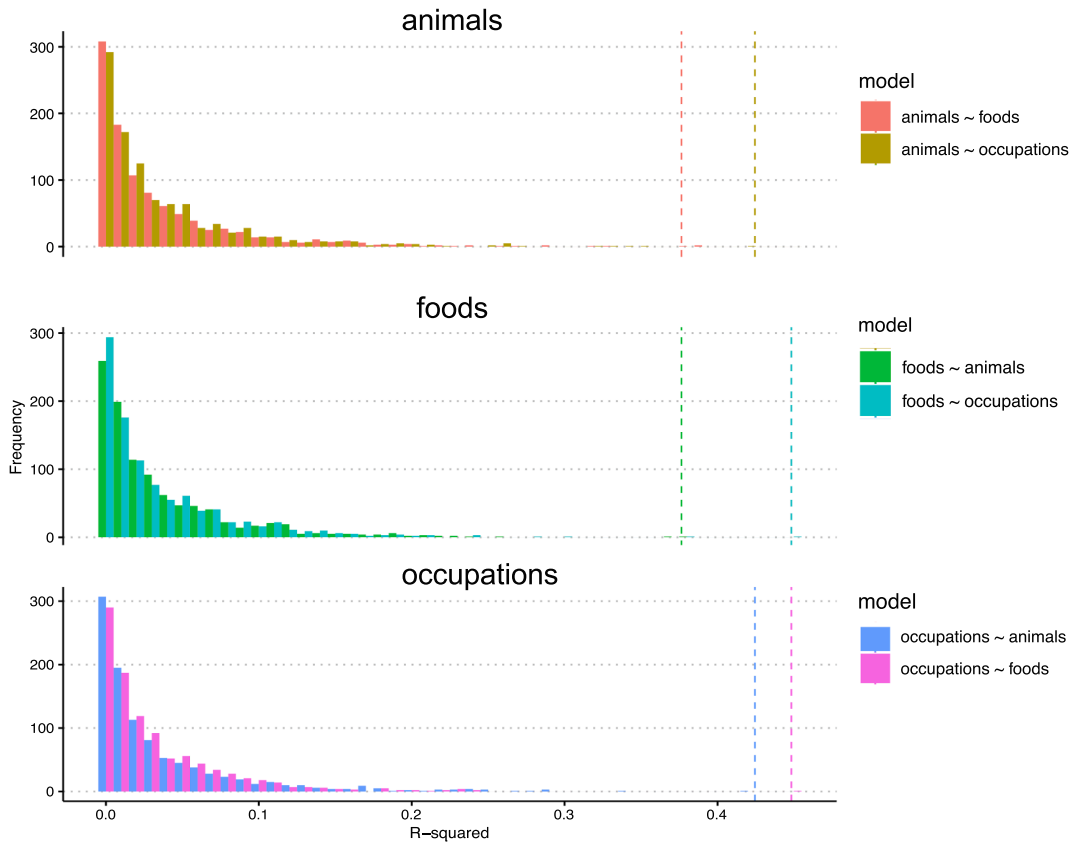

Fig. 6. Histograms of explained variance (R^2) obtained from linear models predicting idiosyncratic scores from one domain (e.g., animals) using other domains (e.g., foods, animals) via from 1000 random permutations across individuals. Dotted vertical lines denote the R^2 estimate from models based on the original data from the same individual.

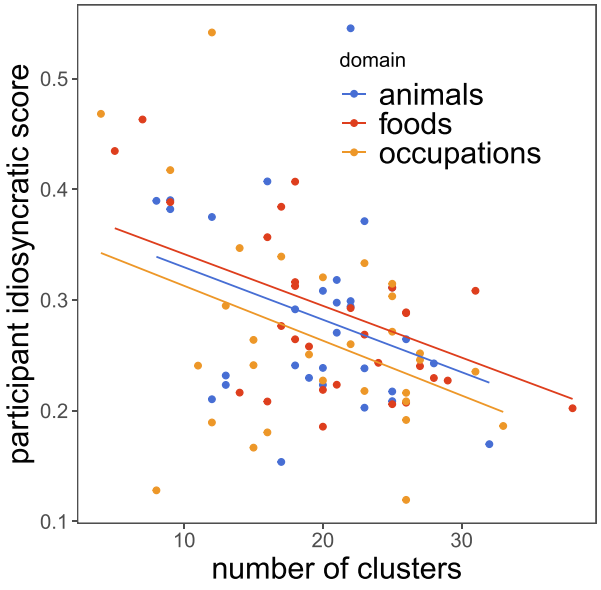

Fig. 7. Participant-level aggregated idiosyncratic scores as a function of the total number of clusters indicated by the participant.

search through memory in the verbal fluency task. We compared a variety of norm-based and automated methods of clustering and switching in their ability to predict participant-designated and rater-designated clusters and switches within fluency lists. We also explored how a novel measure of idiosyncratic variance may provide more insights into the lexical sources that contribute to searching through memory as well as overall performance in the VFT.

Predicting clustering and switching behavior

Our primary finding was that among *automated* methods of assigning clusters and switches, the "multimodal delta method" that accommodated *relative* drops in similarity via rise and fall thresholds and incorporated a weighted estimate of semantic and phonological similarity was most predictive of participant-designated clusters and switches. This finding converges with Lundin et al.'s work, where they showed that the delta method showed distinct neural activity related to switching and may be particularly indicative of critical decision-making processes implicated during search. Although they did not explore variants of the delta method that jointly examine the contribution of multiple lexical sources, prior work has suggested that search within the VFT is likely impacted by both semantic and phonological similarity (Abwender et al. 2001; Kumar, Lundin, and Jones, 2022). It is important to highlight here that the multimodal variant of the delta method explored in the current work performed on par with the delta method in the *animals* domain, and slightly outperformed the delta method in the *foods* and *occupations* domains, in predicting participant-designated clusters and switches. These differences may be suggestive of different lexical organization across domains, or may simply point to random variation across the methods, due to the smaller sample size of the Lundin et al. dataset we used in the present work. Future work could examine the predictive power of these methods on a larger scale, to fully understand domain-level differences as well as explore the relative contribution of semantic and phonological information to search processes. Our findings suggest that the contribution of phonology in semantic VFT may be small (as indicated by the low weight assigned to phonological similarity in the best-performing models, see Tables 1 and

Table 5
Distribution of strategies used by participants and raters to generate and rate fluency lists, respectively.

Strategy theme	% **participants** who used this strategy for category and letter VFT (number of participants out of N = 30)	% **raters** who used this strategy for category VFT (number of participants out of N = 211)
Finding semantically associated/related items; items within subcategories	73 % (22)	83 % (183)
Finding phonetically associated/related words, using alphabetical order or other word characteristics	63 % (19)	0.45 % (1)
Remembering or using personal life experiences	40 % (12)	3.64 % (8)
Use of visual imagery; imagined spatial navigation or other motoric behavior	23 % (7)	1.36 % (3)
Using phonetic and semantic associations	7 % (2)	0 % (0)
Unaware of strategy; listed what came to mind/ relied on instinct	7 % (2)	8.18 % (18)
Relaxing	3 % (1)	0 % (0)
Other strategies (e.g., reading words aloud; updating strategies mid-task; inhibiting particular associations; using color-based strategies; strategies based on word length/difficulty)	0 % (0)	7.7 % (17)
Vague response (missing response or unable to categorize)	0 % (0)	14.09 % (31)

3) but potentially meaningful in conceptualizing search within an integrated lexicon. Indeed, the issue of task-discrepant clustering has been acknowledged in the literature (Abwender et al., 2001; Kumar, Lundin, & Jones, 2022; Lundin et al., 2023) but model-based work that explicitly examines these questions is limited. The present work highlights the need to explore this issue from an individual perspective, especially given the differences in the strategies employed during memory search and response generation (see Table 5; Lundin et al., 2023; Unsworth et al., 2014).

Another important finding from the present study was that the norms-based methods significantly outperformed the automated methods in predicting participant-designated and rater-designated clusters and switches. While this is not surprising, given that the norms have been refined over several iterations and may reflect similar processes being used by researchers and participants, it is important to consider the relative utility of automated methods compared to simply using the norms-based methods. On one hand, norms may be most useful if the goal is to improve the diagnostic abilities of these methods (e.g., Bushnell et al., 2022; Lundin et al., 2020). It is important to highlight here, however, that even the norms-based methods only explained about 33% variance in participant designations. This may indicate a mismatch in how individuals conceptualize and reflect on their own search and how the norms (and other methods) score these lists. Critically, normative methods do not illuminate the processes that may be contributing to the formation of those clusters. Clusters may be fluid and may change based on the previous items produced (Hills et al., 2015), and normative schemes tend to oversimplify the flexible manner in which individuals cluster items from their mental lexicon as they are performing the search. As a proof of concept, we examined the predictive power of our rater-designated clusters and switches in predicting the participant-designated clusters and switches, and found that the rater designations explained nearly 52% of the total variance, in contrast to the normative methods, that captured about 33% of the variance (see Table 1). The primary differences across these two methods were that the raters designated clusters and switches for *pairwise* transitions produced by an individual and relied on a variety of strategies (as reported by the raters post-hoc, see Table 5), whereas the norms tend to group items into strict hierarchical/taxonomic categories (e.g., pets, canines, etc.), independent of the context (of other items) in which those items were produced. From this perspective, it is clear that the ways in which concepts are structured and processed do not directly map onto the norms, and devising measures that examine clusters on a continuum may be critical for advancing our understanding of how individuals search through memory.

Our additional analyses based on multiple regression models revealed that model performance improved when multiple methods were used to predict participant designations. Specifically, we found that the associative norms-based method, the multimodal delta method, and the SVD method were the only methods that significantly predicted the designations across *all* domains. This may indicate that the methods are likely capturing unique variance in the participant designations. Specifically, the norms capture hierarchical information about category membership, the multimodal delta method captures lexical similarity between concepts, and the SVD method captures response consistency across participants. All of these may be independent sources of information that individuals use to search through memory, and clustering and switching methods that combine these information sources into a composite measure are likely to be more informative and predictive of clustering and switching behavior within individuals.

Idiosyncratic variance

Another contribution of this work is the introduction of a novel measure of idiosyncratic variance in verbal fluency. Our preliminary analyses suggested that transitions that are marked by high idiosyncratic variance are also low in semantic (and phonological, to some extent) similarity. One possible explanation of this finding may be that transitions that are idiosyncratic have low lexical content and are likely driven by other non-lexical processes such as imagery, episodic/personal experiences, etc. This could suggest that when people are unable to use lexical cues to produce items in the VFT, they resort to alternative sources that could be helpful (Lundin et al., 2023). However, another interpretation of this finding may be that the raters who were classifying the transitions in the present study were more likely to rely on lexical content than the individual participants who produced the fluency lists, given that the idiosyncratic score was computed by measuring the discordance between the individual and the raters. Indeed, the delta models (that use lexical content) were better able to predict the rater designations than the participant designations (see Tables 1 and 3). Even if we consider this possibility, it appears that sources beyond the lexical content considered by the raters (and our models) may be at work when individuals designated clusters and switches for the lists that they produced in the VFT. Current models of search do not consider alternative sources and may need to account for such experiences to fully accommodate differences in memory search at the structure and process level (see Kumar, Steyvers, and Balota, 2021). Additionally, we also found important differences between the strategies employed by raters and participants, further underscoring the need to consider strategies used by participants as an individual difference marker in its own right. Indeed, some recent work on list learning suggests that spontaneous strategy employment may improve memory performance (Laine et al., 2024). Future work could examine the role of strategy in mediating memory search in a more fine-grained manner to better understand search at the individual level.

We also observed some other interesting patterns with respect to the idiosyncratic scores. Specifically, we found that we could predict an

individual's score on one domain using their score on another domain. This suggests that the idiosyncratic score is likely measuring a latent variable that is implicated across multiple domains, suggesting that an individual may have similar strategies or processes involved when performing the VFT across multiple domains. Additionally, we found that higher idiosyncratic scores were associated with fewer clusters. Although these analyses were exploratory, it is possible that navigating the search space in an idiosyncratic manner is related to wider exploration of the lexicon, and may be linked to other cognitive processes that are implicated in creative idea generation (Grever et al., 2023) or inhibitory control (Gupta et al., 2012). Future work could examine whether individuals who perform the VFT in an idiosyncratic manner may also be "different" in other cognitively meaningful ways.

Limitations and future directions

The present work focused on understanding clustering and switching behavior within the SFT using participant designations and other automated methods. Our results shed light on the lexical sources that individuals may be using to navigate their internal lexicon from one cluster to another. Of course, it is important to acknowledge here that individuals likely vary not only in the processes and information sources they use to navigate the lexicon, but also in how their specific lexicons were acquired. Indeed, previous work has used fluency lists to estimate individual semantic networks (e.g., Zemla & Austerweil, 2018), and also shown that networks estimated in this way can uncover important group-level differences (e.g., Kenett et al., 2013). Whether differences in performance among individuals arise due to *structure*-level differences in knowledge representation or *process*-level differences during tasks is an open question in the field (for reviews, see Castro & Siew; Kumar, Steyvers, & Balota, 2022). In the present work, we used the same underlying semantic representations for all individuals as a simplifying assumption and focused on the *process* of clustering and switching. Some recent work has used person-specific corpora to construct individual semantic representations (Johns, 2024), while other work has explored how different underlying semantic representations, when combined with foraging models of search, can illuminate individual differences in search behavior (Kumar et al., 2024). These approaches represent preliminary steps in tackling this issue, and ultimately, a complete account of search will involve instantiating representations and processes that *both* vary at the individual level.

Another important issue to highlight is that the current study used fluency lists and cluster-switch designations generated by participants in a post-hoc manner as a starting point for evaluating the predictive power of different automated methods. Two limitations of this approach are the nature of the sample (given that this was a fMRI study) and that individuals may have devoted varying levels of effort to this post-hoc task, and may also have different levels of insight into the cognitive processes involved in clustering and switching. Furthermore, the processes involved in generating the fluency list (i.e., engaging in search) may be different from processes involved in scoring one's own list post-hoc, or even scoring another participant's list, as the raters did in the current study. Thus, there may be some level of method variance between the original search task and the scoring tasks that the current study is unable to directly address. As an initial step in documenting and understanding individual clustering and switching behavior, these limitations underscore the need to investigate these processes in a more controlled setting in future work. The fact that these designations show reliable correlations across domains (see Figs. 4 and 5), as well as predict neural activity (Lundin et al., 2023) is a promising indicator that understanding search at the individual level is useful, but more work is needed to fully characterize the process of clustering and switching within individuals. For example, future work could investigate whether predictive (as opposed to postdictive) measures that predict clustering and switching at the item level for a given individual without having access to future items (see Zhang & Jones, 2022; Morales et al., 2024) could account for idiosyncratic variance in the fluency task. Additionally, experimental manipulations that probe an individual during the task to reflect on their search process (similar to Unsworth, 2017) may yield novel insights about the metacognitive components that contribute to and predict individual-level variance in memory search.

Conclusion

The study of differences within and between individuals is increasingly being recognized as central to accelerating scientific advances in cognitive domains (Lupyan et al., 2023). Understanding the phenomenology, or subjective experience, of individuals is critical to this effort, and recent work has begun to document differences in subjective experiences of people with respect to imagery (Bainbridge et al., 2021), inner speech (Alderson-Day & Fernyhough, 2015; Roebuck & Lupyan, 2020), and synesthesia (Jewanski, et al., 2020). Within the same vein, the current work attempted to understand differences in clustering and switching and subjective experiences during memory search, by focusing on how individuals grouped items they produced during search.

Ultimately, the subjective experience of memory search will be informative in expanding our understanding of the structural and process-level constraints that mediate mental search. However, given the practical constraints of collecting these designations, future work could investigate alternative, less cognitively demanding ways of assessing clustering and switching, in addition to the subjective experience of search, such as using gamified experimental approaches (Brändle et al., 2021) or existing individual-level corpora to extract estimates of individual-level variance (Johns, 2024). Overall, the present study investigated how several automated methods based on lexical content map onto and can predict individual designations of clustering and switching in the VFT, as well as introduced a novel measure to capture idiosyncratic behavior in the task. Our findings underscore the need to incorporate multiple lexical sources in understanding memory search as well as highlight the individual-level variance that is implicit to memory search.

CRediT authorship contribution statement

Abhilasha A. Kumar: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Nancy B. Lundin:** Writing – review & editing, Investigation, Formal analysis. **Michael N. Jones:** Writing – review & editing, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the NSF Grant (#2235362 and #2235363) "Collaborative Research: CompCog: Modeling Search within the Mental Lexicon."

Data availability

The behavioral study described in this paper was pre-registered at https://aspredicted.org/3X1_9FK. Scripts to access the data, program the experiment, and reproduce the analyses presented in this work are available at https://github.com/thelexiconlab/whats-in-my-cluster.

References

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.

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin, 141*(5), 931.

Bainbridge, W. A., Pounder, Z., Eardley, A. F., & Baker, C. I. (2021). Quantifying aphantasia through drawing: Those without visual imagery show deficits in object but not spatial memory. *Cortex, 135*, 159–172.

Balogh, R., Imre, N., Gosztolya, G., Pákáski, M., & Kálmán, J. (2023). The role of silence in verbal fluency tasks – A new approach for the detection of mild cognitive impairment. *Journal of the International Neuropsychological Society, 29*(1), 46–58.

Barattieri di San Pietro, C., Luzzatii, C., Ferrari, E., de Girolamo, G., & Marelli, M. (2023). Automated clustering and switching algorithms applied to semantic verbal fluency data in schizophrenia spectrum disorders. *Language, Cognition, and Neuroscience, 38*(7), 950–965.

Beaty, R. E., Zeitlen, D. C., Baker, B. S., & Kenett, Y. N. (2021). Forward flow and creative thought: Assessing associative cognition and its role in divergent thinking. *Thinking Skills and Creativity, 41*, 100859.

Bose, A., Patra, A., Antoniou, G. E., Stickland, R. C., & Belke, E. (2022). Verbal fluency difficulties in aphasia: A combination of lexical and executive control deficits. *International Journal of Language & Communication Disorders, 57*(3), 593–614.

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

Brändle, F., Allen, K. R., Tenenbaum, J., & Schulz, E. (2021). Using games to understand intelligence. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43, No. 43). https://escholarship.org/uc/item/17z2q92d.

Bushnell, J., Svaldi, D., Ayers, M. R., Gao, S., Unverzagt, F., Gaizo, J. D., & Clark, D. G. (2022). A comparison of techniques for deriving clustering and switching scores from verbal fluency word lists. *Frontiers in Psychology, 13*, 743557.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Kurzweil, R. (2018). Universal sentence encoder for English. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations* (pp. 169–174). https://doi.org/10.48550/arXiv.1803.11175.

Chen, L., Asgari, M., Gale, R., Wild, K., Dodge, H., & Kaye, J. (2020). Improving the assessment of mild cognitive impairment in advanced age with a novel multi-feature automated speech and language analysis of verbal fluency. *Frontiers in Psychology, 11*, 535.

Christensen, A. P., & Kenett, Y. N. (2021). Semantic network analysis (SemNA): A tutorial on preprocessing, estimating, and analyzing semantic networks. *Psychological Methods*.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47*, 1–12.

Gerver, C. R., Griffin, J. W., Dennis, N. A., & Beaty, R. E. (2023). Memory and creativity: A meta-analytic examination of the relationship between memory systems and creative cognition. *Psychonomic Bulletin & Review*, 1–39.

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

Gray, K., Anderson, S., Chen, E. E., Kelly, J. M., Christian, M. S., Patrick, J., & Lewis, K. (2019). "Forward flow": A new measure to quantify free thought and predict creativity. *American Psychologist, 74*(5), 539.

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

Henderson, S. K., Peterson, K. A., Patterson, K., Lambon Ralph, M. A., & Rowe, J. B. (2023). Verbal fluency tests assess global cognitive status but have limited diagnostic differentiation: Evidence from a large-scale examination of six neurodegenerative diseases. *Brain. Communications, 5*(2), Article fcad042.

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

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.

Jewanski, J., Simner, J., Day, S. A., Rothen, N., & Ward, J. (2020). The evolution of the concept of synesthesia in the nineteenth century as revealed through the history of its name. *Journal of the History of the Neurosciences, 29*(3), 259–285.

Johns, B. T. (2024). Determining the relativity of word meanings through the construction of individualized models of semantic memory. *Cognitive Science, 48*, Article e13413.

Kenett, Y. N., Wechsler-Kashi, D., Kenett, D. Y., Schwartz, R. G., Ben-Jacob, E., & Faust, M. (2013). Semantic organization in children with cochlear implants: Computational analysis of verbal fluency. *Frontiers in Psychology, 4*, 543.

Kenett, Y. N., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience, 8*, 407.

Kim, N., Kim, J. H., Wolters, M. K., MacPherson, S. E., & Park, J. C. (2019). Automatic scoring of semantic fluency. *Frontiers in Psychology, 10*, 1020.

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review, 28*, 40–80.

Kumar, A. A., Lundin, N. B., & Jones, M. N. (2022). Mouse-mole-vole: The inconspicuous benefit of phonology during retrieval from semantic memory. *In Proceedings of the 44th Annual meeting of the Cognitive Science Society*.

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.

Laine, M., Fellman, D., Eräste, T., Ritakallio, L., & Salmi, J. (2024). Strategy use and its evolvement in word list learning: A replication study. *Royal Society Open Science, 11*(2), 230651.

Lundin, N. B., Todd, P. M., Jones, M. N., Avery, J. E., O'Donnell, B. F., & Hetrick, W. P. (2020). Semantic search in psychosis: Modeling local exploitation and global exploration. *Schizophrenia Bulletin Open, 1*(1), Article sgaa011.

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), Article e2312462120.

Lupyan, G., Uchiyama, R., Thompson, B., & Casasanto, D. (2023). Hidden differences in phenomenal experience. *Cognitive Science, 47*(1), Article e13239.

Morais, A. S., Olsson, H., & Schooler, L. J. (2013). Mapping the structure of semantic memory. *Cognitive Science, 37*(1), 125–145.

Morales, D., Canessa, E., & Chaigneau, S. E. (2024). An Agent-Based Model of Foraging in Semantic Memory. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46). https://escholarship.org/uc/item/3vf9d7zm.

Oh, S. J., Sung, J. E., Choi, S. J., & Jeong, J. H. (2019). Clustering and switching patterns in semantic fluency and their relationship to working memory in mild cognitive impairment. *Dementia and Neurocognitive Disorders, 18*(2), 47–61.

Okruszek, Ł., Rutkowska, A., & Wilińska, J. (2013). Clustering and switching strategies during the semantic fluency task in men with frontal lobe lesions and in men with schizophrenia. *Psychology of Language and Communication, 17*(1), 93–100.

Ovando-Tellez, M., Benedek, M., Kenett, Y. N., Hills, T., Bouanane, S., Bernard, M., & Volle, E. (2022). An investigation of the cognitive and neural correlates of semantic memory search related to creative ability. *Communications Biology, 5*(1), 604.

Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27.

Robert, P. H., Lafont, V., Medecin, I., Berthet, L., Thauby, S., Baudu, C., & Darcourt, G. U. Y. (1998). Clustering and switching strategies in verbal fluency tasks: Comparison between schizophrenics and healthy adults. *Journal of the International Neuropsychological Society, 4*(6), 539–546.

Roebuck, H., & Lupyan, G. (2020). The internal representations questionnaire: Measuring modes of thinking. *Behavior Research Methods, 52*, 2053–2070.

Sung, K., Gordon, B., Vannorsdall, T. D., Ledoux, K., & Schretlen, D. J. (2013). Impaired retrieval of semantic information in bipolar disorder: A clustering analysis of category-fluency productions. *Journal of Abnormal Psychology, 122*(3), 624.

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.

Unsworth, N. (2017). Examining the dynamics of strategic search from long-term memory. *Journal of Memory and Language, 93*, 135–153.

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.

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

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.

Zhang, L., & Jones, M. N. (2022). Using "Semantic Scent" to Predict Item-Specific Clustering and Switching Patterns in Memory Search. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44). https://escholarship.org/uc/item/67m4g3d9.

Zhao, Q., Guo, Q., & Hong, Z. (2013). Clustering and switching during a semantic verbal fluency test contribute to differential diagnosis of cognitive impairment. *Neuroscience Bulletin, 29*, 75–82.