



# Unique signatures in verbal fluency task performance in schizophrenia and depression<sup>☆</sup>

Sunghye Cho<sup>a,\*</sup>, Yan Cong<sup>b</sup>, Aarush Mehta<sup>c,d</sup>, Amir H. Nikzad<sup>c,d,e</sup>, Sarah A. Berretta<sup>c,d,f</sup>, Leily M. Behbehani<sup>c,d,g</sup>, Mark Liberman<sup>a</sup>, Sunny X. Tang<sup>a,c,d,e</sup>

<sup>a</sup> Linguistic Data Consortium & Linguistics Department, University of Pennsylvania, Philadelphia, PA, USA

<sup>b</sup> School of Languages and Cultures, Purdue University, West Lafayette, IN, USA

<sup>c</sup> Zucker Hillside Hospital, Northwell Health, New Hyde Park, NY, USA

<sup>d</sup> Institute of Behavioral Science, Feinstein Institutes for Medical Research, Manhasset, NY, USA

<sup>e</sup> Department of Psychiatry, Donald and Barbara Zucker School of Medicine, Hempstead, NY, USA

<sup>f</sup> School of Behavioral and Brain Sciences, University of Texas at Dallas, Richardson, TX, USA

<sup>g</sup> Department of Psychology, Yale University, New Haven, CT, USA

## ARTICLE INFO

### Keywords:

Schizophrenia  
Depression  
Verbal fluency  
Automated speech analysis  
Semantic knowledge

## ABSTRACT

**Background:** Verbal fluency tasks reveal consistent impairments among individuals with schizophrenia (SSD) and shed valuable insight on semantic and cognitive processing. However, comparisons with other psychiatric groups remain limited. Doing so may help clarify whether characteristics are unique to SSD versus shared across diagnoses. This study compared verbal fluency responses among participants with SSD, unipolar depression, and healthy volunteers (HV).

**Methods:** Participants with SSD ( $n = 64$ ), unipolar depression ( $n = 70$ ), and HVs ( $n = 37$ ) completed letter- and category-guided verbal fluency tasks. Using automated pipelines, we extracted total correct responses, pause durations, first response latency, semantic and phonemic distances, and the number and size of semantic/phonemic clusters. Associations between fluency metrics and clinical ratings of speech disturbances (e.g., incoherence, inefficiency) were examined within patient groups.

**Results:** Groups differed significantly in total fluency scores, particularly on the category fluency task, with the greatest impairment in SSD and intermediate effects in depression. Significant pairwise group differences emerged in the number of semantic and phonemic clusters in the category task, while cluster size showed fewer differences. In HVs, both semantic and phonemic distances increased with longer pauses, suggesting strategic cluster switching. This pattern was absent in SSD and depression groups. SSD exhibited longer first latency and pauses relative to HVs; participants with depression showed no change in latency, but similarly prolonged pauses.

**Conclusion:** Individuals with SSD and depression exhibit distinct patterns in verbal fluency, with category tasks revealing more robust group differences. Automated fluency analysis offers a scalable approach for differentially detecting cognitive-linguistic impairments across psychiatric populations.

## 1. Introduction

Verbal fluency tasks represent a concise method for eliciting clinically-salient information that taps into linguistic/semantic networks, executive functioning, working memory, and lexical access. During a category-guided fluency task, participants are asked to list as many items as possible within one category (e.g., animals, vegetables).

The letter-guided fluency task requires participants to name as many words as possible starting with a certain letter (e.g., “f,” “s”). The tasks usually last for one minute and participants are asked to avoid giving proper nouns (e.g., “Bob,” “Boston”) and different endings for the same word (e.g., “love,” “lover,” “loving”). Performance is usually expected to be higher on category fluency tasks than on letter fluency tasks (Luo et al., 2010; Shao et al., 2014; Gonzalez-Recober et al., 2023), perhaps

<sup>☆</sup> This article is part of a Special issue entitled: ‘Communication in psychosis’ published in Schizophrenia Research: Cognition.

\* Corresponding author.

E-mail address: [csunghye@ldc.upenn.edu](mailto:csunghye@ldc.upenn.edu) (S. Cho).

<https://doi.org/10.1016/j.scog.2025.100407>

Received 1 August 2025; Received in revised form 12 November 2025; Accepted 20 November 2025

Available online 26 November 2025

2215-0013/© 2025 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

because category fluency tasks resemble familiar processes in daily life, like listing items from a grocery list.

Although these tasks are simple on the surface, both category and letter fluency tasks necessitate a series of complex cognitive processes, including searching the mental lexicon for correct responses (e.g., “father” for an F-letter fluency), suppressing relevant but incorrect responses (e.g., “mother” is semantically similar to “father,” but does not start with “f”), and keeping track of previously mentioned items to avoid repetition. Therefore, task “performance” (i.e., total number of correct responses or scores) on these fluency tasks has been considered a useful measure of various dimensions of cognition. Previous studies have also found that participants tend to produce semantically or phonetically related responses in “clusters.” For example, in an animal category task, participants may list farm animals (e.g., chicken, cow, pig, goat) and switch to birds (e.g., owl, blue jay, crow, pigeon). Alternatively, in an F-letter task, participants may cluster by one set of sounds (e.g., fry, frigid, fresh) and then morph into another set (e.g., fish, fin, finish). Prior research has shown that the number of cluster switches and the size of clusters are closely related to performance on verbal fluency tasks. Furthermore, recent studies have found that, beyond total scores and the number/size of clusters, other metrics from the tasks (e.g., response time between two consecutive responses, the first latency time, word frequency, etc.) can provide valuable insight into task performance (Shao et al., 2014).

Schizophrenia Spectrum Disorder (SSD) is a neuropsychiatric condition characterized by conceptual disorganization, language disturbance, and difficulties in abstract reasoning. Prior research has consistently shown that individuals with SSD perform more poorly on verbal fluency tasks than demographically matched healthy volunteers (HVs) (Goldberg et al., 1998; Elvevåg et al., 2002; Van Beilen et al., 2004; Lundin et al., 2022; Lundin et al., 2023) (see Tan et al (Tan et al., 2020), for summaries and meta-analysis). In particular, individuals with SSD tend to struggle more with category fluency tasks than with letter fluency tasks, showing that individuals with SSD showed a greater performance difference from HV in category fluency than letter fluency (Lundin et al., 2022; Bokar and Goldberg, 2003; Meshulam-Gately et al., 2009; Nour et al., 2023). This may be due to the impaired ability to organize speech in SSD, as listing semantically related words requires coherent and organized thinking and reasoning. Prior research has linked reduced semantic fluency to core symptoms of SSD, including thought disorder (Goldberg et al., 1998), disorganization (Lundin et al., 2022), and impaired associative learning (Nour et al., 2023).

Lower performance in verbal fluency has also been observed in other psychiatric conditions, such as unipolar depression. For example, a meta-analysis review of 42 studies has shown that verbal fluency performance is affected in individuals with depression due to their executive dysfunction (Henry and Crawford, 2005), which is one of the most prominent clinical characteristics associated with depression. Prior research has also reported that severe depressive symptoms were associated with lower verbal fluency performance in this population (Yin et al., 2024). While some prior research has shown that both category and letter fluency were similar in their sensitivity for individuals with depression (Henry and Crawford, 2005), others have shown that the performance of phonemic fluency is less affected in individuals with depression (Hammar et al., 2011).

Previous studies have shown mixed results in the number/size of clusters. Some studies have reported that individuals with SSD produced fewer switches with similar cluster sizes (Elvevåg et al., 2002; Lundin et al., 2022; Moore et al., 2006; Piras et al., 2019), while others have found a similar number of switches with smaller cluster sizes than HVs (Van Beilen et al., 2004). A few recent studies have also found that higher semantic similarity in verbal fluency responses was associated with lower speech disorganization and higher social functioning (Lundin et al., 2022; Elvevåg et al., 2007; Pauselli et al., 2018), and that variance in semantic modulation in verbal fluency tasks correlated with abstract conceptual reasoning and the negative dimension of formal thought

disorder (Nour et al., 2023).

Few studies examining verbal fluency in schizophrenia have included more than two groups: SSD and HVs. Previous studies including another psychiatric comparison group (e.g., depression) have been limited by small sample sizes, and they only compared the total scores without in-depth analyses. One study with 26 participants with schizophrenia, 26 healthy control participants, and 16 participants with unipolar depression found that both psychiatric groups performed worse than HVs on semantic fluency tasks (Lafont et al., 1998); participants with schizophrenia also showed smaller clusters of words than HVs. However, this study did not compare the schizophrenia group to the depression group. Another study with 30 participants in each group also found decreased performance as well as hemodynamic activation in the prefrontal cortex in schizophrenia and depression, with no difference between the two groups (Xiang et al., 2021). Comparing the verbal fluency performance of individuals with SSD to those with other mental health conditions may yield a better understanding of SSD by clarifying which characteristics are unique to SSD and which are shared across disorders. To our knowledge, no previous study has compared SSD to another psychiatric group using objective, computational measurements of semantic and phonemic distance.

In this study, we compared verbal fluency performance among individuals with SSD ( $n = 64$ ), individuals with unipolar depression ( $n = 70$ ), and HVs ( $n = 37$ ). Examining both groups allows us to identify which linguistic and acoustic markers are transdiagnostic (shared across conditions) versus diagnosis-specific. For example, while both groups may exhibit reduced verbal fluency performance, these may stem from different mechanisms: difficulties in thought organization in schizophrenia versus psychomotor retardation in depression. By directly comparing these groups, we can better disentangle diagnosis-specific versus shared pathways in verbal fluency performance, thereby enhancing the clinical interpretability and potential diagnostic utility of our findings.

We hypothesized that both individuals with depression and SSD would show lower performance than HVs due to shared characteristics such as decreased motivation, anhedonia, and cognitive difficulties such as concentration and decision making. However, we expected individuals with SSD to perform worse than those with depression, as SSD is usually associated with larger effect sizes for cognitive impairment (Ma et al., 2021; Barch, 2009), and may be further impacted by the presence of more pronounced conceptual disorganization. We next explored whether impaired performance in depression would be associated with similar changes in clustering and pauses during the tasks. We also hypothesized that the quantitative measures derived from automated processing of the verbal fluency tasks would be related to clinical assessments of speech disturbance.

## 2. Methods

### 2.1. Participants & assessment

Participants were recruited at the Zucker Hillside Hospital in Glen Oaks, NY. The participants recruited for this study presented with a range of mental health conditions, as part of three larger studies (“Remora” (Mehta et al., 2025), “LPOP” (Tang et al., 2025), “ACES” (Tang et al., 2024)). Diagnoses were obtained by a combination of clinical records and standardized interviews, and are based on the DSM-5. All participants performed a letter-guided fluency (“F”) and a category-guided (“Animal”) fluency tasks for one minute, respectively. Participants were given a standardized set of instructions for each task, and had the opportunity to ask clarifying questions before beginning. If participants reported difficulty thinking of words within the one minute timeframe, they were encouraged by the assessor to keep trying until the one minute cutoff. Among the 171 participants, there were 64 individuals with a primary diagnosis of schizophrenia spectrum disorders (SSD; including schizophrenia ( $n = 33$ ), schizoaffective disorder ( $n =$

16), schizophreniform disorder ( $n = 3$ ), unspecified psychotic disorder ( $n = 11$ ), 70 with unipolar depressive disorders (Dep; including major depressive disorder and persistent depressive disorder), and 37 healthy volunteers (HV). Comorbidities with other disorders were permitted; 14 participants with SSD as their primary diagnosis also had a diagnosis of a unipolar depressive disorder. Individuals with developmental (e.g., autism), neurodegenerative (e.g., dementias), or other medical conditions that significantly affect speech were excluded from the study. All participants provided informed consent, and the research received ethical approval from the Institutional Review Board at the Feinstein Institutes for Medical Research.

Table 1 presents participants' demographic and clinical characteristics by group. In these samples, we previously examined latent factors in the clinical ratings for speech and language disturbances in some of these participants (Tang et al., 2023a), novel approaches to linguistic analysis (Nikzad et al., 2022; Palominos et al., 2025), as well as markers of social cognition (Tang et al., 2023b), but here we present new and unique analyses focused on fluency task findings. In addition to the fluency tasks, participants also underwent clinical assessments and provided speech samples for picture description, paragraph reading, and open-ended narrative tasks, which were not analyzed in the current manuscript. Trained clinical assessors rated observable speech and language disturbances for the encounter on the Scale for the Assessment of Thought Language and Communication (TLC). We represented the TLC ratings as three factor scores as previously described in Tang et al. (2023): two disorganized dimensions (incoherence and inefficiency), and one negative dimension for impaired expressivity. Participants also underwent several neuropsychological assessments, including Brief Symptom Inventory (BSI), Scale for the Negative Symptoms (SANS), Brief Psychiatric Rating Scale (BPRS), Quality of Life Scale (QLS), Patient Health Questionnaire-9 (PHQ-9), Columbia-Suicide Severity Rating Scale (CSSRS), and/or Generalized Anxiety Disorder (GAD), depending on the study they participated. The complete clinical characteristics and medication information of the participants are summarized in Table S1 in the Supplementary Materials.

## 2.2. Data processing

All audio files were manually transcribed by trained annotators. For these analyses, we isolated participant speech during the task itself and omitted any extraneous discussions with the assessor (e.g., asking clarifying questions). We automatically aligned each word from the transcripts to audio signals using a forced aligner (Yuan and Liberman,

**Table 1**

Demographic and clinical characteristics of participants. *P*-values in the last column show ANOVA results for group differences in a given demographic or clinical characteristic. SSD: Schizophrenia spectrum disorder, Dep: Depression, HV: Healthy volunteer, TLC: Thought, Language, and Communication ratings, WRAT: Wide Range Achievement Test.

|                    | Dep ( $n = 70$ ) | HV ( $n = 37$ ) | SSD ( $n = 64$ ) | <i>p</i> -value |
|--------------------|------------------|-----------------|------------------|-----------------|
| Age (years)        | 22.9 (5.1)       | 30.1 (4.8)      | 26.5 (6.2)       | < 0.001         |
| Sex (%)            |                  |                 |                  | < 0.001         |
| - Female           | 50 (71.4 %)      | 20 (54.1 %)     | 22 (34.4 %)      |                 |
| - Intersex         | 0 (0.0 %)        | 0 (0.0 %)       | 1 (1.6 %)        |                 |
| - Male             | 20 (28.6 %)      | 17 (45.9 %)     | 41 (64.1 %)      |                 |
| Race (%)           |                  |                 |                  | 0.044           |
| - Asian            | 14 (20.0 %)      | 5 (13.5 %)      | 7 (10.9 %)       |                 |
| - Black            | 10 (14.3 %)      | 8 (21.6 %)      | 28 (43.8 %)      |                 |
| - Multiple         | 4 (5.7 %)        | 4 (10.8 %)      | 5 (7.8 %)        |                 |
| - Native American  | 1 (1.4 %)        | 0 (0.0 %)       | 0 (0.0 %)        |                 |
| - Other            | 5 (7.1 %)        | 1 (2.7 %)       | 4 (6.2 %)        |                 |
| - Pacific Islander | 0 (0.0 %)        | 1 (2.7 %)       | 2 (3.1 %)        |                 |
| - Unknown          | 1 (1.4 %)        | 0 (0.0 %)       | 1 (1.6 %)        |                 |
| - White            | 35 (50.0 %)      | 18 (48.6 %)     | 17 (26.6 %)      |                 |
| Education (years)  | 14.8 (1.8)       | 16.8 (1.6)      | 13.9 (2.2)       | < 0.001         |
| TLC (total)        | 4.1 (4.7)        | 2.6 (3.9)       | 14.8 (15.5)      | < 0.001         |
| WRAT (total)       | 49.1 (5.5)       | 52.0 (3.9)      | 46.6 (5.7)       | < 0.001         |

2008), which provided timestamps for each of the words produced. The automated processing pipelines for letter and category fluency tasks that we developed (Gonzalez-Recober et al., 2023; Cho et al., 2021) first counted the number of correct responses without repetitions, proper nouns, and numbers. Semantic and phonemic distances between two consecutive, correct responses were then measured, respectively. We employed pretrained word embeddings from fastText (Bojanowski et al., 2017) and calculated the cosine distances between two correct responses for semantic distance, following previous studies (Gonzalez-Recober et al., 2023; Lundin et al., 2022; Elvevåg et al., 2007; Cho et al., 2021; Lundin et al., 2020). A Levenstein distance (Levenshtein, 1966) was calculated for phonemic distance between two consecutive, correct responses using the words' dictionary pronunciation from the CMU pronouncing dictionary (Carnegie Mellon Speech Group, 2014), following a similar approach in a previous study (Lundin et al., 2022). We calculated the mean and standard deviation (SD) values of semantic and phonemic distances from all participants. We also identified semantic and phonemic clusters for each of the tasks, where cluster boundaries were defined when semantic/phonemic distances between two consecutive words were larger than 1 SD from the mean of the HV group. The number of words within two adjacent cluster boundaries were counted to establish the size of each semantic or phonemic cluster. We also measured the pause duration between two correct responses as well as the first response time (the end of the interviewers' speech to the start of the first correct response from the participants), using the timestamps from the forced aligner outputs. More details about the analytical pipelines can be found in our previous studies (Gonzalez-Recober et al., 2023; Cho et al., 2021).

## 2.3. Statistical considerations

To investigate group differences in task performance and the linguistic measures during category or letter fluency tasks, we first checked if the measures met requirements for parametric tests using Levene's tests and residual visualization. When a given variable did not meet requirements for parametric tests, we log-transformed the variable before building models. For task scores and each linguistic variable, we built ANCOVA models where each linguistic variable or task score was included as a dependent variable and the diagnostic groups were included as an independent variable. Education level was covaried to account for group differences in all models. We also experimented with additional covariates (sex and age) in the model, but those covariates did not change the results. Therefore, we reported the results of simpler models with covarying education levels only. When Group showed significant effect in the ANCOVA models, we compared estimated marginal means by Group (or Group by Task for task scores) for a posthoc test using the emmeans package in R, controlling for education levels. *P*-values adjusted for multiple comparisons ( $n = 3$ ) with false discovery rate were reported in the results.

To examine if changes in the cluster sizes during the fluency tasks differ by group, we built linear mixed-effects models, which included the cluster size as a dependent variable and the interaction of group and the order of clusters as fixed effects. Individuals were added as a random intercept, and education levels were covaried in all models. We also examined if semantic and phonetic distances changed by the function of pause duration between two consecutive words, using linear mixed-effects models. In these models, we included phonemic or semantic distance between two words as a dependent variable and pause duration between the words as an independent variable. Education levels were covaried, and individuals were included as a random effect.

Lastly, we investigated the relationships between clinical ratings of speech impairment and linguistic measures of verbal fluency in patients with schizophrenia or depression using linear regression models. Three-factor scores were calculated from TLC for inefficiency, incoherence, and impaired expressivity, based on a previous study (Tang et al., 2023a). These factor scores in the SSD and depression groups were

associated with verbal fluency scores and other linguistic variables, including numbers and sizes of semantic/phonemic clusters, mean semantic/phonemic distances, mean pause duration, and first latency duration, in linear regression models. Education levels were covaried in all models, and results were reported in the results section, only when a fitted model was significant. We also note that we did not include medication information in our analyses because (1) the available data were incomplete, and (2) medication use was highly dependent on diagnosis, which could obscure group differences in verbal fluency tasks.

### 3. Results

#### 3.1. Fluency task performance

The groups significantly differed in total scores for both letter and category fluency tasks (Category:  $F(2,167) = 20.79, p < 0.001$ ; Letter:  $F(2,167) = 4.14, p = 0.018$ ; Fig. 1). Individuals with SSD produced the fewest correct responses in the category fluency task; group differences were significant compared to the depression group ( $p < 0.001$ ) and HVs ( $p < 0.001$ ). Individuals with depression exhibited an intermediate effect and produced significantly fewer correct responses than HVs ( $p = 0.025$ ). For the letter fluency task, patients with SSD also produced fewer correct responses than HVs ( $p = 0.013$ ), but they did not differ from the depression group ( $p = 0.251$ ). Individuals with depression also did not differ from HVs ( $p = 0.17$ ).

#### 3.2. Semantic and phonemic clusters

The groups significantly differed in the numbers of semantic and phonemic clusters in the category tasks (semantic:  $F(2,167) = 6.21, p = 0.003$ ; phonemic:  $F(2,167) = 8.75, p < 0.001$ ; Fig. 2A-B). The SSD group produced fewer semantic and phonemic clusters compared to the HV group (semantic:  $p = 0.002$ ; phonemic:  $p < 0.001$ ) and the depression group (semantic:  $p = 0.04$ ; phonemic:  $p = 0.006$ ) during the category fluency task. The difference between the depression and HV groups was significant in the number of semantic clusters ( $p = 0.04$ ) and marginally significant in the number of phonemic clusters ( $p = 0.052$ ) during the category fluency task. The SSD group also produced fewer semantic clusters during the letter fluency tasks than HVs ( $p = 0.013$ ; Fig. 2C), yet all the other comparisons in the number of clusters during the letter fluency task were not significant (all  $ps \geq 0.1$ ; Fig. 2C-D).

The average size of the phonemic clusters in the category fluency task significantly differed by group ( $F(2,167) = 4.22, p = 0.016$ ), where the mean size of semantic clusters produced by individuals with SSD was

smaller than those of HVs ( $p = 0.026$ ) and the depression group ( $p = 0.026$ ); Fig. 2F), but the depression group did not significantly differ from HV's. Also, the average size of semantic clusters in the letter fluency significantly varied by group ( $F(2,167) = 3.3, p = 0.039$ ). The depression group produced larger semantic clusters than HVs ( $p = 0.037$ ; Fig. 2G). All the other pairwise group comparisons in the mean and SD of the cluster sizes were not significant in both category and letter fluency tasks ( $p \geq 0.1$  for all combinations). Furthermore, the change in the size of clusters during the category and letter fluency tasks did not significantly vary by group ( $p \geq 0.1$  for all combinations; figure not shown).

#### 3.3. Pause duration

The latency duration from the end of the interviewers' prompt to the beginning of the first correct response significantly varied by group in the category fluency task ( $F(2,189) = 7.1, p = 0.001$ ). In the letter fluency task, similar trends were observed, but the overall group effect was trend-level ( $F(2,189) = 2.7, p = 0.07$ ; Fig. 3A-B). The SSD group was significantly slower in listing the first correct response compared to the HV ( $p = 0.005$ ) and depression groups ( $p = 0.002$ ) in the category fluency task. The depression and HV groups did not differ from each other ( $p = 0.64$ ). In addition, participants' mean pause duration between two correct responses significantly differed by group in both category ( $F(2,189) = 16.26, p < 0.001$ ) and letter fluency tasks ( $F(2,189) = 9.5, p < 0.001$ ; Fig. 3C-D). Both SSD and depression groups produced significantly longer pauses compared to HVs in both category ( $p < 0.001$  for both comparisons) and letter fluency tasks (Depression:  $p < 0.001$ ; SSD:  $p = 0.008$ ). Participants with SSD did not differ from those with depression in either the category ( $p = 0.67$ ) or letter fluency tasks ( $p = 0.13$ ).

#### 3.4. Semantic and phonemic distance by time

The mean and SD of semantic and phonemic distances between two consecutive correct responses did not differ by group in both category and letter fluency tasks ( $p > 0.1$  for all comparisons; figure not shown). However, the association between pause duration and semantic/phonemic distance between two correct responses significantly varied by group in the category fluency task. Both semantic and phonemic distances increased as a function of pause duration in the HV group (semantic:  $\beta = 0.01, p < 0.001$ ; phonemic:  $\beta = 0.78, p < 0.001$ ; Fig. 4A-B); this was significantly different from the pattern in the SSD (semantic:  $\beta = -0.01, p = 0.004$ ; phonemic:  $\beta = -0.62, p = 0.004$ ) and depression groups (semantic:  $\beta = -0.01, p = 0.005$ ; phonemic:  $\beta = -0.58, p = 0.001$ ) where there was no significant change in semantic or phonemic distance as a function of pause duration. However, none of the associations were significant in the letter fluency task ( $p > 0.1$  for all groups; Fig. 4C-D).

#### 3.5. Relationship with clinical ratings for speech and language disturbances

SSD patients with impaired expressivity produced fewer correct responses ( $\beta = -0.04, p = 0.018$ ) and a longer first latency ( $\beta = 0.07, p = 0.002$ ) during the category fluency task, which did not significantly differ from the trend of the depression group (score:  $p = 0.32$ , first latency:  $p = 0.2$ ; Fig. 5A).

Individuals with increased speech incoherence in the SSD group produced fewer correct responses (i.e., lower scores) in the category task ( $\beta = 0.29, p = 0.026$ ), and this trend was marginally different from that of the depression group ( $\beta = -0.26, p = 0.186$ ; Fig. 5A). SSD patients with increased speech incoherence had a longer latency for their first correct response during the category fluency task ( $\beta = -0.05, p = 0.025$ ), which marginally differed from the depression group ( $\beta = 0.05, p = 0.112$ ). Lastly, SSD patients with increased inefficiency tend to produce fewer

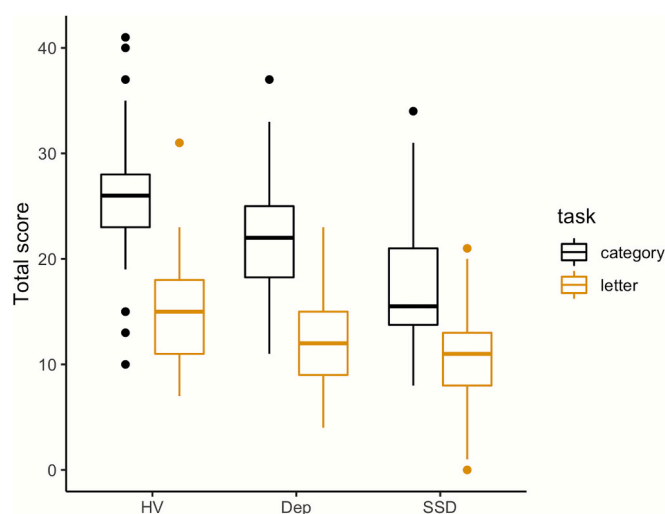


Fig. 1. Total scores of category and letter fluency tasks by group. SSD: Schizophrenia spectrum disorder, Dep: Depression, HV: healthy volunteer.

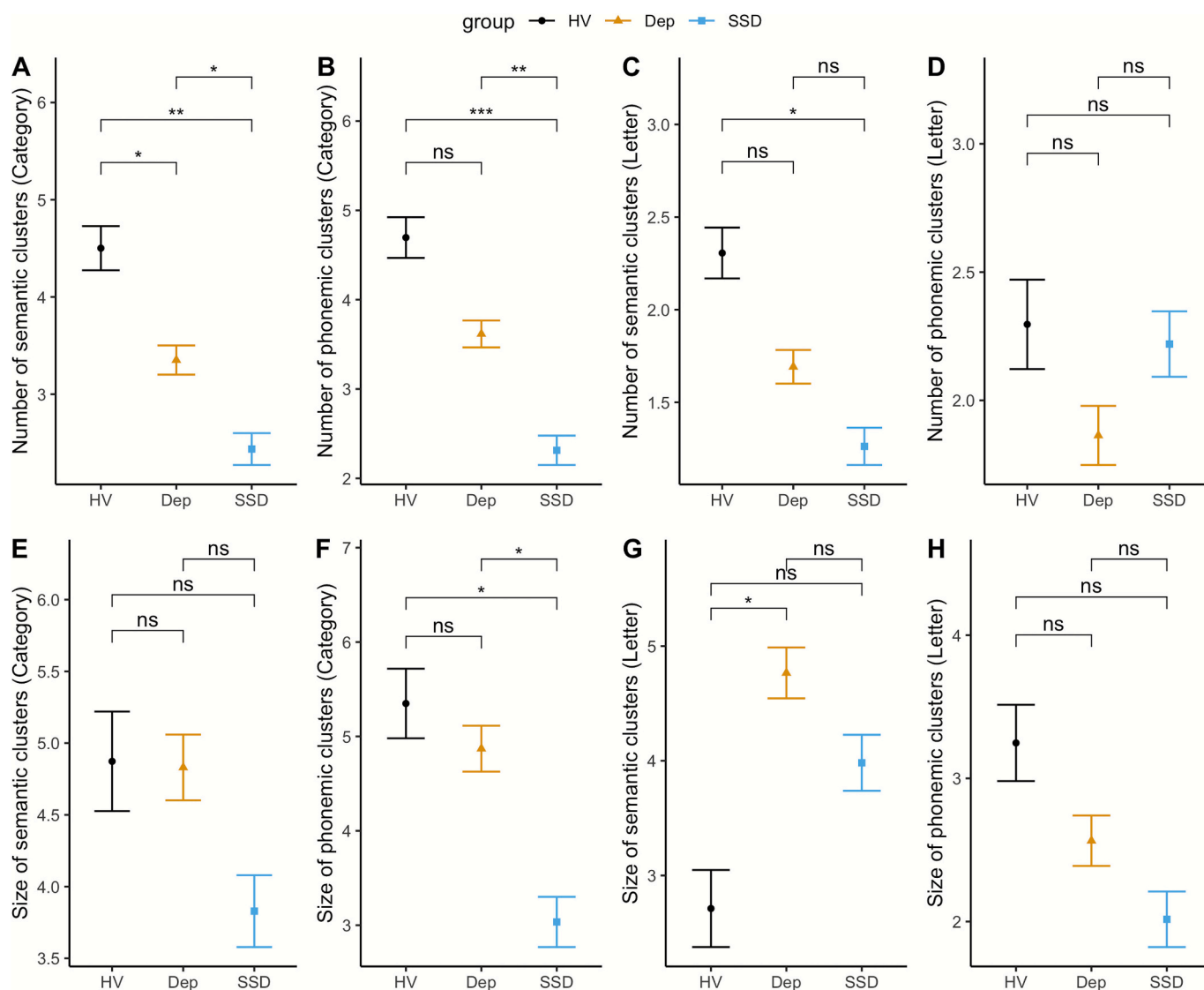


Fig. 2. Estimated marginal means and SD of the number (A-D) and size (E-H) of semantic and phonemic clusters by group and task. SSD: Schizophrenia spectrum disorder, Dep: Depression, HV: healthy volunteer, \*\*\*:  $p < 0.001$ , \*\*:  $p < 0.01$ , \*:  $p < 0.05$ , ns: not significant.

correct responses during the letter fluency task ( $\beta = -0.08$ ,  $p = 0.016$ ), which was significantly different from the depression group ( $\beta = 0.1$ ,  $p = 0.013$ ). All the other associations between TLC factor scores and language variables (e.g., semantic/phonemic distances between two words, number of clusters, and mean size of clusters) were not significant after covarying education levels.

#### 4. Discussion

In this study, we examined the performance of category- and letter-guided fluency tasks across three different groups (SSD vs. Depression vs. HV), using automated analysis pipelines. Our results demonstrated that the groups differed on many metrics, with the category fluency task generally showing more pronounced group distinctions than the letter fluency task. This is consistent with findings from several prior studies (Lundin et al., 2022; Bokar and Goldberg, 2003; Meshulam-Gately et al., 2009; Nour et al., 2023), and is suggestive that when time constraints are present, category fluency may be the more clinically informative task to administer in most circumstances when assessing psychiatric disorders. Specifically, the SSD group produced fewer correct responses than HVs on both tasks, but performance differences with the depression group were only significant in the category fluency task. We also

observed more significant pairwise group differences in the number and size of clusters, as well as pause duration, in the category fluency task compared to the letter fluency task. Further investigation of pause duration revealed a similar pattern. Lastly, the first latency duration and overall scores during the category fluency were significantly associated with impaired expressivity in the SSD group, while average pause duration and overall score during the category fluency were significantly associated with speech incoherence in the SSD group. We discuss the key findings below.

We hypothesized that individuals with SSD and depression would perform worse than HVs on both category and letter fluency tasks, with SSD individuals performing worse than those with depression. While results revealed a graded pattern of group differences on both tasks – SSD showing the lowest scores, followed by depression – significant differences between the SSD and depression groups were observed only in the category fluency task. Previous studies comparing these three groups have not demonstrated this intermediate effect in depression. However, given the consistent understanding that overall cognitive impairment follows the pattern SSD > Depression > HV, the absence of this pattern in prior studies may be attributable to their smaller sample sizes. Our correlation analyses further showed that the clinical dimension of impaired expressivity, which characterizes the decreased verbal

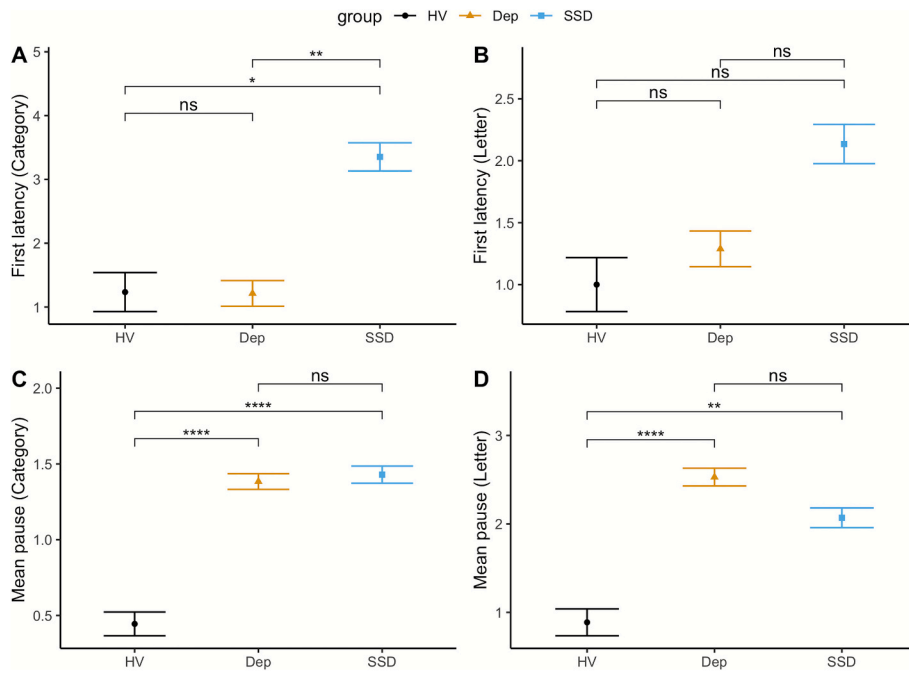


Fig. 3. Estimated marginal means and standard deviation of the first latency duration and average pause duration by group and task. SSD: Schizophrenia spectrum disorder, Dep: Depression, HV: healthy volunteer, \*\*\*\*:  $p < 0.0001$ , \*\*\*:  $p < 0.001$ , \*\*:  $p < 0.01$ , \*:  $p < 0.05$ , ns: not significant.

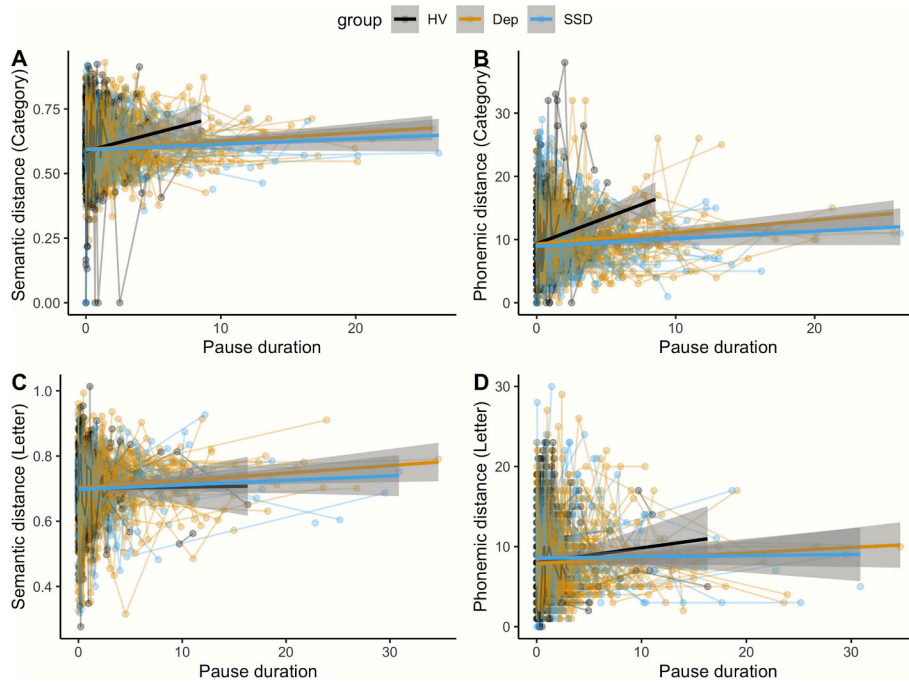


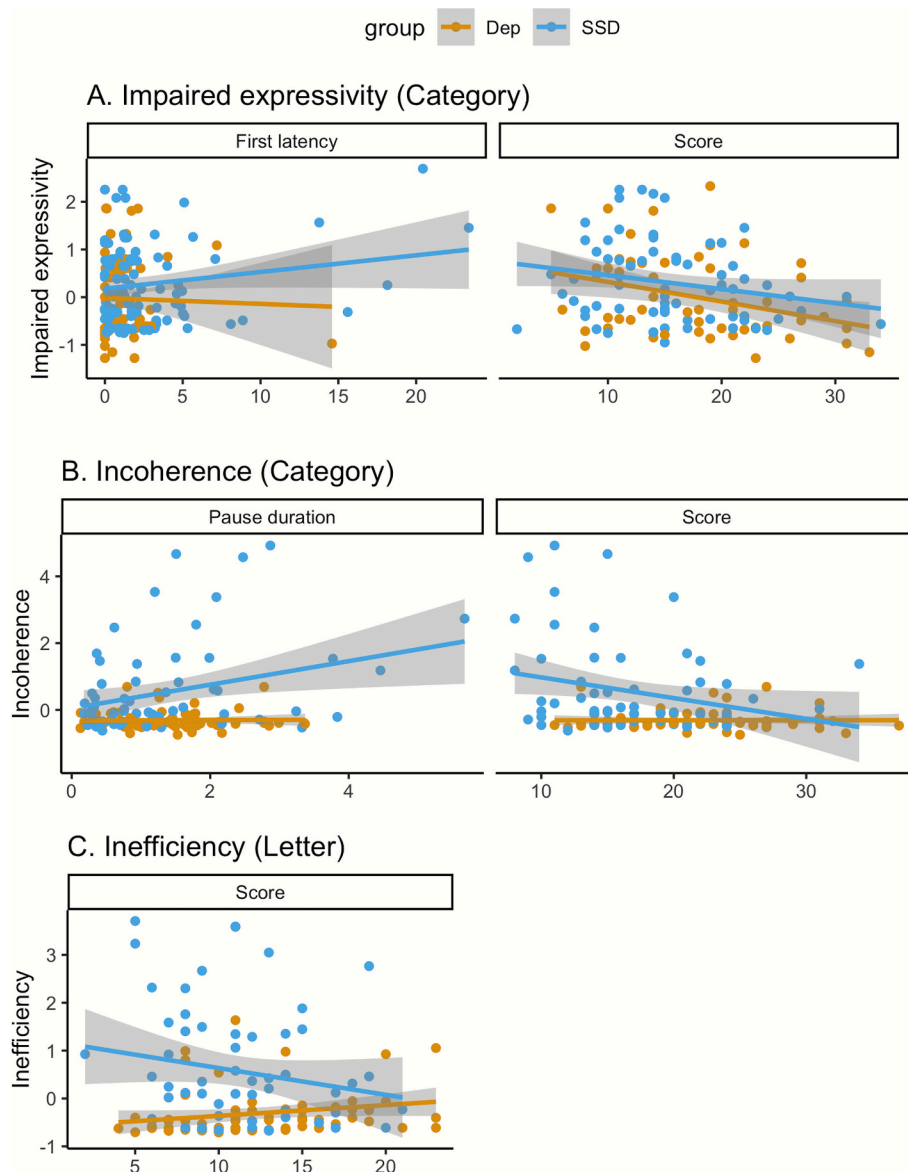
Fig. 4. Associations between pause duration and semantic/phonetic distances between two correct responses by group and task. SSD: Schizophrenia spectrum disorder, Dep: Depression, HV: healthy volunteer.

output and expressiveness common to both SSD and depression, was significantly associated with category fluency scores in both SSD and depression groups.

However, incoherence, a disorganized dimension of speech and language impairment, was significantly associated with category fluency performance only in the SSD group and was uncommon in the depression group. These findings align with our hypothesis that both SSD and depression groups showed worse performance than HVs likely due to shared negative symptoms, while the SSD group's additional

conceptual disorganization further impacted performance. Another possible explanation involves the restricted score range typically found in letter fluency tasks. In our study, all groups performed worse on the letter fluency task compared to the category fluency task. This limited variability in letter fluency scores may have reduced the sensitivity of the task to detect between-group differences. Future research is needed to disentangle the effects of task type on performance differences.

All pairwise comparisons reaching significance for the number of semantic clusters produced during the category fluency task, suggested a



**Fig. 5.** Associations between TLC factor scores and linguistic variables in the SSD and depression groups. SSD: Schizophrenia spectrum disorder, Dep: Depression.

graded effect in the ability to switch between clusters, with individuals with SSD being most severely affected. In addition, the number and size of phonemic clusters in the category fluency task further differentiated the SSD group from the others. These findings are consistent with prior studies reporting that individuals with SSD produce fewer clusters of similar size during category fluency tasks (Elvevåg et al., 2002; Lundin et al., 2022; Moore et al., 2006; Piras et al., 2019). Relatedly, our analysis of pause duration and semantic/phonemic distances revealed that, in individuals with SSD or depression, semantic and phonemic distances did not increase with longer pause durations, unlike in HVs, who showed a positive relationship between these variables. Together, these findings suggest that individuals with SSD or depression, despite taking more time between responses, often remained within the same cluster rather than switching to a new one – leading to overall fewer, but similarly sized, clusters.

Average pause duration between two correct responses differentiated HVs from the other two groups on both fluency tasks, consistent with previous research showing that longer switching times in SSD contribute to reduced performance (Nour et al., 2023). Moreover, our results revealed that average pause duration during the category fluency task was significantly associated with increased speech incoherence only

in SSD. If we make the common assumption that increased pause time is related to increased difficulty with producing the next word, this raises the possibility that there may be underlying impairments in the ability of people with SSD to navigate semantic space, leading to greater difficulty in producing associated words during the fluency task, and perhaps also manifesting as incoherence during less constrained speech. This interpretation would be consistent with the findings of Nour et al (Nour et al., 2023), who found that performance on the semantic verbal fluency task was associated with the degree to which participants were preferentially guided by semantic relationships vs. orthographic ones, and which was further associated with hippocampal markers of associative learning on magnetoencephalogram (MEG). In contrast, no such association between pause duration and incoherence was found in the depression group, highlighting a potential difference in the underlying mechanisms of increased pause duration in this group. Future studies should explore the causes of increased pause durations in depression.

First latency (i.e., the time from the end of the interviewer's prompt to the beginning of the first correct response) was significantly longer in the SSD group for category fluency compared to the other groups, with no difference between depression and HV's. This increased latency was significantly associated with impaired expressivity in the SSD group, but

not in the depression group, even though lower category fluency performance correlated with increased first latency overall. This suggests that factors other than expressivity or disorganization may contribute to decreased first latency in depression. Further research is needed to investigate these potential factors.

Of note, we did not replicate the finding from prior research (Lundin et al., 2022; Nour et al., 2023) of a group-wise difference in mean semantic distance between consecutive responses. We also did not find a significant difference in average phonemic distance for either task. This may be due to a shorter task time being used in this study (1 min), versus the longer duration in the prior studies (3 vs. 5 min). At any rate, the fact that a robust difference in mean semantic distance was not found here suggests that the primary effect of the disorders (SSD and depression) is not best described by this metric. Rather, the fact that robust results exist for the decreased number of semantic clusters suggests that the impairment is manifest more in the ability (or lack thereof) to switch between semantic clusters. Lundin and colleagues (Lundin et al., 2023) have previously characterized this phenomenon as an optimal foraging task, where an active cognitive process is required for decisions to switch to a new cluster.

Limitations of the study include the smaller sample size of healthy volunteers, incomplete information on medication use, and the absence of some of the clinical ratings for the depression group. The results of this study should be interpreted in light of these limitations. Despite these limitations, our findings and methods may contribute to early detection and risk assessment efforts. Integrating automated speech metrics into larger research frameworks could also enhance understanding of shared and distinct cognitive mechanisms across different conditions. As these approaches evolve, they may help refine data-driven and dimensional models for various mental health conditions.

## 5. Conclusion

This study demonstrated that individuals with SSD and depression show distinct patterns in verbal fluency tasks, with category fluency revealing more robust group differences than letter fluency. The SSD group consistently performed worse than both HVs and the depression group, particularly in measures related to clustering, pause duration, and latency. Key findings suggest that clinical dimensions of impaired expressivity and speech incoherence are related to poor fluency performance in SSD, and that longer pause durations may reflect difficulties in switching between semantic categories. In contrast, the depression group showed more subtle patterns, largely limited to category fluency, and without strong associations to speech incoherence. Overall, the results underscore the greater sensitivity of category fluency tasks and highlight the value of automated speech analysis for detecting distinct cognitive and linguistic patterns across psychiatric populations.

Our findings can inform neuropsychological assessment by providing objective, language-based markers that complement traditional cognitive testing, offering sensitive indicators of executive and semantic organization difficulties. Furthermore, these findings can guide cognitive remediation efforts, for instance by targeting semantic network organization and verbal retrieval strategies during therapy. Speech-based or language-focused digital tools could also be developed to monitor treatment progress or cognitive changes over time in both schizophrenia and depression populations.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.scog.2025.100407>.

## CRedit authorship contribution statement

**Sunghye Cho:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Yan Cong:** Writing – review & editing, Data curation. **Aarush Mehta:** Writing – review & editing. **Amir H. Nikzad:** Writing – review & editing. **Sarah A. Berretta:** Writing – review & editing, Data curation.

**Leily M. Behbehani:** Writing – review & editing, Data curation. **Mark Liberman:** Writing – review & editing. **Sunny X. Tang:** Writing – review & editing, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sunny X. Tang reports financial support was provided by National Institute of Mental Health. Sunny X. Tang reports financial support was provided by Brain and Behavior Research Foundation. Sunny X. Tang reports financial support was provided by Winterlight Labs Inc. Sunny X. Tang reports a relationship with North Shore Therapeutics that includes: board membership and equity or stocks. Sunny X. Tang reports a relationship with Psyryn that includes: board membership and equity or stocks. Sunny X. Tang reports a relationship with Catholic Charities Neighborhood Services that includes: consulting or advisory. Sunny X. Tang reports a relationship with LB Pharmaceuticals that includes: consulting or advisory. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We thank the participants for their time and contributions. Funding was received from NIMH/NIDCD K23 MH130750 (SXT) and the Brain and Behavior Research Foundation Young Investigator Grant (SXT). Data collection for the ACES and LPOP projects was funded by Winterlight Labs and provided a portion of the raw data used in this manuscript, but all analyses were conceived of and conducted independently.

## References

- Barch, D.M., 2009. Neuropsychological abnormalities in schizophrenia and major mood disorders: similarities and differences. *Curr. Psychiatry Rep.* 11 (4), 313–319. <https://doi.org/10.1007/s11920-009-0045-6>.
- Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics* 5, 135–146. [https://doi.org/10.1162/tacl\\_a\\_00051](https://doi.org/10.1162/tacl_a_00051).
- Bokat, C.E., Goldberg, T.E., 2003. Letter and category fluency in schizophrenic patients: a meta-analysis. *Schizophr. Res.* 64 (1), 73–78. [https://doi.org/10.1016/S0920-9964\(02\)00282-7](https://doi.org/10.1016/S0920-9964(02)00282-7).
- Carnegie Mellon Speech Group, 2014. The Carnegie Mellon University Pronouncing Dictionary. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
- Cho, S., Nevler, N., Parjane, N., et al., 2021. Automated analysis of digitized letter fluency data. *Front. Psychol.* 12 (July). <https://doi.org/10.3389/fpsyg.2021.654214>.
- Elvevåg, B., Fisher, J.E., Gurd, J.M., Goldberg, T.E., 2002. Semantic clustering in verbal fluency: schizophrenic patients versus control participants. *Psychol. Med.* 32 (5), 909–917. <https://doi.org/10.1017/S0033291702005597>.
- Elvevåg, B., Foltz, P.W., Weinberger, D.R., Goldberg, T.E., 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophr. Res.* 93 (1–3), 304–316. <https://doi.org/10.1016/j.schres.2007.03.001>.
- Goldberg, T.E., Aloia, M.S., Gourovitch, M.L., Missar, D., Pickar, D., Weinberger, D.R., 1998. Cognitive substrates of thought disorder, I: the semantic system. *Am. J. Psychiatry* 155 (12), 1671–1676. <https://doi.org/10.1176/ajp.155.12.1671>.
- Gonzalez-Recober, C., Nevler, N., Shellikeri, S., et al., 2023. Comparison of category and letter fluency tasks through automated analysis. *Front. Psychol.* 14 (October), 1212793. <https://doi.org/10.3389/fpsyg.2023.1212793>.
- Hammar, Å., Strand, M., Årdal, G., Schmid, M., Lund, A., Elliott, R., 2011. Testing the cognitive effort hypothesis of cognitive impairment in major depression. *Nord. J. Psychiatry* 65 (1), 74–80. <https://doi.org/10.3109/08039488.2010.494311>.
- Henry, J.D., Crawford, J.R., 2005. A meta-analytic review of verbal fluency deficits in depression. *J. Clin. Exp. Neuropsychol.* 27 (1), 78–101. <https://doi.org/10.1080/138033990513654>.
- Lafont, V., Medecin, I., Robert, P.H., et al., 1998. Initiation and supervisory processes in schizophrenia and depression. *Schizophr. Res.* 34 (1–2), 49–57. [https://doi.org/10.1016/S0920-9964\(98\)00084-X](https://doi.org/10.1016/S0920-9964(98)00084-X).
- Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* 10 (8), 845–848. [https://doi.org/10.1016/S0074-7742\(08\)60036-7](https://doi.org/10.1016/S0074-7742(08)60036-7).
- Lundin, N.B., Todd, P.M., Jones, M.N., Avery, J.E., O'donnell BF, Hetrick WP., 2020. Semantic search in psychosis: modeling local exploitation and global exploration. *Schizophr. Bull. Open* 1 (1), 1–11. <https://doi.org/10.1093/schizbullopen/sgaa011>.



- Lundin, N.B., Jones, M.N., Myers, E.J., Breier, A., Minor, K.S., 2022. Semantic and phonetic similarity of verbal fluency responses in early-stage psychosis. *Psychiatry Res.* 309 (January), 114404. <https://doi.org/10.1016/j.psychres.2022.114404>.
- Lundin, N.B., Brown, J.W., Johns, B.T., et al., 2023. Neural evidence of switch processes during semantic and phonetic foraging in human memory. *Proc. Natl. Acad. Sci.* 120 (42), 2017. <https://doi.org/10.1073/pnas.2312462120>.
- Luo, L., Luk, G., Bialystok, E., 2010. Effect of language proficiency and executive control on verbal fluency performance in bilinguals. *Cognition* 114 (1), 29–41. <https://doi.org/10.1016/j.cognition.2009.08.014>.
- Ma, M., Zhang, Y., Zhang, X., Yan, H., Zhang, D., Yue, W., 2021. Common and distinct alterations of cognitive function and brain structure in schizophrenia and major depressive disorder: a pilot study. *Front. Psych.* 12 (July), 1–10. <https://doi.org/10.3389/fpsy.2021.705998>.
- Mehta, A., Nikzad, A.H., Cong, Y., Cho, S., Pradhan, S., Sunny, X., 2025. Sentiment in speech is associated with symptom severity in psychosis. *Sentiment in speech is associated with symptom severity in psychosis. Cogn. Neuropsychiatry* 30 (3), 199–210.
- Meshulam-Gately, R.I., Giuliano, A.J., Goff, K.P., Faraone, S.V., Seidman, L.J., 2009. Neurocognition in first-episode schizophrenia: a meta-analytic review. *Neuropsychology* 23 (3), 315–336. <https://doi.org/10.1037/a0014708>.
- Moore, D.J., Savla, G.N., Woods, S.P., Jeste, D.V., Palmer, B.W., 2006. Verbal fluency impairments among middle-aged and older outpatients with schizophrenia are characterized by deficient switching. *Schizophr. Res.* 87 (1–3), 254–260. <https://doi.org/10.1016/j.schres.2006.06.005>.
- Nikzad, A.H., Cong, Y., Berretta, S., et al., 2022. Who does what to whom? graph representations of action-predication in speech relate to psychopathological dimensions of psychosis. *Schizophrenia* 8 (1), 1–10. <https://doi.org/10.1038/s41537-022-00263-7>.
- Nour, M.M., McNamee, D.C., Liu, Y., Dolan, R.J., 2023. Trajectories through semantic spaces in schizophrenia and the relationship to ripple bursts. *Proc. Natl. Acad. Sci.* 120 (42), 2017. <https://doi.org/10.1073/pnas.2305290120>.
- Palominos, C., Kirdun, M., Nikzad, A.H., et al., 2025. A single composite index of semantic behavior tracks symptoms of psychosis over time. *Schizophr. Res.* 279 (October 2024), 116–127. <https://doi.org/10.1016/j.schres.2025.03.038>.
- Pauselli, L., Halpern, B., Cleary, S.D., Ku, B., Covington, M.A., Compton, M.T., 2018. Computational linguistic analysis applied to a semantic fluency task to measure derailment and tangentiality in schizophrenia. *Psychiatry Res.* 263 (December 2017), 74–79. <https://doi.org/10.1016/j.psychres.2018.02.037>.
- Piras, F., Piras, F., Banaj, N., et al., 2019. Cerebellar GABAergic correlates of cognition-mediated verbal fluency in physiology and schizophrenia. *Acta Psychiatr. Scand.* 139 (6), 582–594. <https://doi.org/10.1111/acps.13027>.
- Shao, Z., Janse, E., Visser, K., Meyer, A.S., 2014. What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Front. Psychol.* 5 (JUL), 1–10. <https://doi.org/10.3389/fpsyg.2014.00772>.
- Tan, E.J., Neill, E., Tomlinson, K., Rossell, S.L., 2020. Semantic memory impairment across the schizophrenia continuum: a meta-analysis of category fluency performance. *Schizophr. Bull. Open* 1 (1), 1–14. <https://doi.org/10.1093/schizbullopen/sgaa054>.
- Tang, S.X., Hänsel, K., Cong, Y., et al., 2023a. Latent factors of language disturbance and relationships to quantitative speech features. *Schizophr. Bull.* 49, S93–S103. <https://doi.org/10.1093/schbul/sbac145>.
- Tang, S.X., Cong, Y., Nikzad, A.H., et al., 2023b. Clinical and computational speech measures are associated with social cognition in schizophrenia spectrum disorders. *Schizophr. Res.* 259 (February), 28–37. <https://doi.org/10.1016/j.schres.2022.06.012>.
- Tang, S., Massouh, N., Spilka, M., et al., 2024. Machine learning classification across multiple psychiatric disorders using objective speech and language features. In: *Neuropsychopharmacology*, Vol 49, pp. 65–235. <https://doi.org/10.1038/s41386-024-02011-0>.
- Tang, S.X., Spilka, M.J., John, M., et al., 2025. Automated speech and language markers of longitudinal changes in psychosis symptoms. *NPP—Digit. Psychiatry Neurosci.* 3 (1). <https://doi.org/10.1038/s44277-025-00034-z>.
- Van Beilen, M., Pijnenborg, M., Van Zomeren, E.H., Van Den Bosch, R.J., Withaar, F.K., Bouma, A., 2004. What is measured by verbal fluency tests in schizophrenia? *Schizophr. Res.* 69 (2–3), 267–276. <https://doi.org/10.1016/j.schres.2003.09.007>.
- Xiang, Y., Li, Y., Shu, C., Liu, Z., Wang, H., Wang, G., 2021. Prefrontal cortex activation during verbal fluency task and tower of london task in schizophrenia and major depressive disorder. *Front. Psych.* 12 (October), 1–8. <https://doi.org/10.3389/fpsy.2021.709875>.
- Yin, J., John, A., Cadar, D., 2024. Bidirectional associations of depressive symptoms and cognitive function over time. *JAMA Netw. Open* 7 (6), 1–13. <https://doi.org/10.1001/jamanetworkopen.2024.16305>.
- Yuan, J., Liberman, M., 2008. Speaker identification on the SCOTUS corpus. In: *Proceedings - European Conference on Noise Control*, pp. 5687–5690. <https://doi.org/10.1121/1.2935783>.