

Clustering and switching patterns in semantic verbal fluency among Chinese older adults: associations with cognitive function and dementia screening

Juqiang Chen

Shanghai Jiao Tong University

Yangqing Wu

Shanghai Jiao Tong University

Xuhao Zhao

Zhejiang University

Xin Xu

Zhejiang University

Mengyue Wu

mengyuewu@sjtu.edu.cn

Shanghai Jiao Tong University

Article

Keywords: verbal fluency, clustering, switching, Chinese elders

Posted Date: February 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8879783/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Verbal fluency tasks can aid in the early detection of dementia, although the underlying cognitive mechanisms, particularly in Mandarin-speaking populations, remain underexplored. Prior studies often emphasize number of words generated, overlooking more complex components such as clustering and switching and their associations with cognitive functions like memory and orientation. To address this gap, we employed Tongyici Cilin, a comprehensive Mandarin synonym knowledge base, to quantify clustering and switching in large-scale verbal fluency data. Our results show that measures such as number of words generated, and switching frequency are generally associated with memory and orientation, especially among middle-aged and older Mandarin speakers with lower education levels. People with dementia exhibited significant reductions in number of words generated, switching frequency and cluster size, reflecting impaired lexical organization and reduced cognitive flexibility. Similar but milder deficits were observed in the MCI group, suggesting that disruptions in word generation and switching emerge in early cognitive decline and worsen with disease progression. These findings underscore the value of analysing clustering and switching in the verbal fluency task in capturing distinct cognitive processes affected by dementia, thus enhancing early detection and monitoring.

1. Introduction

Dementia represents a rapidly growing global public health challenge, with prevalence rates steadily rising among individuals over the age of 65 (Jia et al., 2020). While Alzheimer's disease (AD) and Mild Cognitive Impairment (MCI) are characterized by progressive neurodegeneration, subtle linguistic deficits, particularly in lexical access and semantic organization, often manifest years before a formal clinical diagnosis (Amieva et al., 2008). Consequently, language-based assessments have emerged as critical tools for early detection. Among these, the verbal fluency test, which requires individuals to generate words within a specific category (e.g., animals) under time constraints, is widely regarded as a sensitive probe of the lexical-semantic domain (Amieva et al., 2008; Salehi et al., 2017).

Despite its expediency and simplicity, the verbal fluency test involves complex cognitive processes. Individuals must access and retrieve words stored in their mental lexicon. To maximize output, participants need to switch between subcategories (e.g., moving from "pet animals" to "wild animals" in a semantic task) while inhibiting previously mentioned words to avoid repetition. The task also requires monitoring output to ensure it meets task criteria and tracking time, thereby involving executive functions, inhibition, attention, and working memory (Aita et al., 2019; Laisney et al., 2009; Ojeda et al., 2010). This task has been successfully used to examine different pathologies, including AD (Bertola et al., 2014; Fagundo et al., 2008; Monsch et al., 1992; see Henry et al., 2004 for a review), Parkinson's disease (Farzanfar et al., 2018; Torralva et al., 2015), stroke (Schmidt et al., 2019) and traumatic brain injury (TBI, see Henry & Crawford, 2004 for a review).

The most basic evaluation of verbal fluency performance is the total count of correct words generated. Different error types, such as repetitions or intrusions (e.g., words from another semantic category), are also analyzed, particularly in populations with language or cognitive impairments. Two additional features of task performance, i.e., clustering and switching, were employed to reveal more in-depth cognitive processes underlying the task (Troyer, 2000; Troyer et al., 1997, 1998). Clustering refers to the production of words within subcategories (e.g., pet animals in the animal fluency task) and switching refers to the ability to shift from one subcategory to a new subcategory (e.g., from pet animals to farm animals). It is proposed that clustering should reflect the process of accessing and using mental lexicon whereas switching should reflect a searching process. These two components are further hypothesised to be related to different brain regions and functions: clustering is related to temporal lobe functions such as verbal memory and word storage, while switching depends on frontal lobe processes such as strategic search, cognitive flexibility, and shifting (Troyer et al., 1998).

In a seminal study (Troyer et al., 1997), younger participants in a semantic fluency task generated more words and switched more frequently than older participants, though no group difference was observed in cluster size. Furthermore, Fagundo et al., (2008) conducted both cross-sectional and longitudinal studies to examine verbal fluency in three groups of participants, i.e., participants with memory complaints, AD patients, and healthy controls. The cross-sectional study revealed that AD patients produced significantly fewer correct words, smaller cluster sizes, and exhibited less switching than both the group with memory complaints and healthy controls. The group with memory complaints also showed deficits in word production and clustering but not in switching compared to healthy controls. In the longitudinal follow-up, individuals from the memory complaints group who progressed to AD exhibited a further decline in correct word count and cluster size, but not in switching performance, compared to those who did not develop dementia. The variability in clustering and switching behaviors with age and cognitive status remains incompletely understood, highlighting the need for further investigation..

Moreover, few studies have compared verbal fluency data among non-English-speaking individuals with dementia, MCI, and healthy controls. A study of Cantonese speakers (Mok et al., 2004) found significant differences in fluency scores between control and AD groups, as well as between mild and moderate AD groups.. Verbal fluency scores correlated significantly with the Cantonese version of the Mini-Mental State Examination scores for both AD and control groups. However, they did not calculate cluster size and switching times for the verbal fluency data. In another study with Mandarin Chinese participants (Zhao et al., 2013), they compared verbal fluency performance in terms of correct words, cluster size, and switching times among participants with AD, vascular dementia (VaD), MCI, vascular cognitive impairment without dementia (VCIND) and cognitively normal senior controls (NC). The general pattern was that control participants performed better than participants with MCI/VCIND, who in turn performed better than participants with AD/VaD in the total verbal fluency score and all sub-scores. Subcategory and switching times were able to discriminate between AD and VaD.

In the present study, we explored the use of clustering and switching measures derived from semantic verbal fluency data within a screening project for Mandarin-speaking individuals with MCI and AD.. The

study was conducted in two phases. In Phase 1, we analyzed data from 9670 community-dwelling older adults to establish population norms and elucidate the distinct effects of age and education on clustering and switching strategies. In Phase 2, we validated the diagnostic utility of these metrics in a clinically diagnosed subset (N = 315), including No Cognitive Impairment (NCI), Mild Cognitive Impairment (MCI), and Dementia groups based on their performance in Montreal Cognitive Assessment (MoCA), a comprehensive neuropsychological evaluation (Nasreddine et al., 2005), followed by the Clinical Dementia Rating (CDR) scale (Morris, 1993) from licensed professionals. By integrating these fine-grained semantic features into a class-weighted LASSO regularized machine learning model, we found that adding clustering and switching metrics improves the sensitivity of cognitive screening beyond traditional word counts, yielding a high negative predictive value that establishes the tool’s utility for the scalable exclusion of cognitively healthy individuals..This study innovates by utilizing a comprehensive Chinese semantic knowledge base (Tongyici Cilin) for objective metric calculation and evaluating the added value of these metrics within a machine learning framework for scalable community screening.

2. Results

A total of 11,583 participants were enrolled in a large-scale project aimed at developing a cognitive screener for dementia detection in China. This two-phase study first established norms and correlations in a large community sample (Phase 1: N = 9,670), then validated the discriminative and predictive utility of verbal fluency metrics in a clinically diagnosed cohort (Phase 2: N = 315). The demographic characteristics of the community sample are summarized in Table 1.

Table 1
Demographic characteristics of the community-dwelling sample.

Characteristic		Value (N = 9,670)
Age (years)	Mean (SD)	65.0 (8.8)
	Range	18–80
Gender, n (%)	Male	3,510 (36.3%)
	Female	6,160 (63.7%)
Education (years)	Mean (SD)	9.3 (3.8)
	Range	0–16

Phase 1 focused on the interplay between verbal fluency (word count, mean cluster size, switching frequency), cognitive scores (memory, orientation), and demographics (age, education). Phase 2 compared these fluency metrics across diagnostic groups (NCI, MCI, Dementia) and evaluated their performance in a predictive model for cognitive status.

2.1 Correlational Analyses in the Community Sample

To elucidate the interaction between verbal fluency and cognitive function, we first conducted correlation analyses among age, years of education, memory, the orientation score, the number of generated words, mean cluster size and switch frequency. As detailed in Table 2, age was negatively correlated with years of education and two fluency measures (word count and switching frequency), suggesting that older adults generally produce fewer words and switch less frequently than the young, consistent with previous studies (Lanting et al., 2009; Zhao et al., 2013). Age was also negatively correlated with memory task scores and orientation task scores. Conversely, education was positively correlated with all three animal fluency task measures, specifically number of words, mean cluster size, switches and two cognitive scores, i.e., memory and orientation, while the number of generated words was positively correlated with both switch frequency and mean cluster size. The global structure of these associations is visualized in Fig. 1, where the width and color intensity of the connecting lines reflect the magnitude of the Pearson correlation coefficients. Notably, word count and switching frequency were positively correlated with memory and orientation scores, whereas mean cluster size was not significantly correlated with either cognitive measure.

Table 2

Pearson correlation coefficients among demographic variables, semantic verbal fluency measures, and cognitive scores in the community sample (Phase 1, N = 9,670). Age: chronological age; Education: years of formal schooling; No. Words: total correct animal names; Switch frequency: number of transitions between semantic subcategories; Cluster size: mean number of words per semantic cluster; Memory: delayed recall task score (0–5); Orientation: temporal/spatial orientation task score (0–6). *p < .05.

	Age	Education	No. Words	Switch frequency	Cluster size	Memory
Age						
Education	-0.24*					
No. Words	-0.25*	0.35*				
Switch frequency	-0.29*	0.35*	0.87*			
Cluster size	0.0035	0.045*	0.15*	-0.14*		
Memory	-0.23*	0.29*	0.34*	0.36*	0.0079	
Orientation	-0.037*	0.13*	0.14*	0.12*	0.018	0.1*

To further explore how age and education modulate these dynamics, we stratified the dataset into four age groups, each with three education level: low (< 6 years), intermediate (6–12 years), and high (> 12 years), as shown in Fig. 2.

A positive correlation was observed between the number of words generated and both memory and orientation scores across all four age groups, but only among individuals with low and intermediate

education levels. However, this pattern was not consistent in the high-education group. For these participants, the correlations with orientation were non-significant in the 50–59, 70–79, and 80-plus age cohorts. Similarly, the correlation with memory was non-significant for the high-education group aged 80 and older.

Across the 50–59, 60–69, and 70–79 age cohorts, switch times demonstrated a positive correlation with memory scores at all levels of education; this correlation also extended to the intermediate education group within the oldest cohort (80+). Conversely, correlations with orientation were more circumscribed, reaching significance solely among the low and intermediate education participants in the 60–69 and 70–79 cohorts, and the intermediate education participants in the 80-plus cohort.

Cluster sizes were negatively correlated with memory scores for the intermediate education group in the 50–59 age cohort. However, no significant correlation was found between cluster sizes and orientation scores for any education level across all age groups.

2.2 Group Differences in Verbal Fluency Performance in the Clinical Cohort

We next analyzed data from the clinically diagnosed subset (Phase 2, $N = 315$; NCI = 203, MCI = 94, Dementia = 18) to examine differences in verbal fluency across cognitive status groups. Significant main effects of cognitive status were observed for all three verbal fluency scores (see Fig. 3, number of generated words: $F(2, 306) = 19.14, p < 0.001$; mean cluster size: $F(2, 198) = 3.86, p = 0.02$; switches: $F(2, 309) = 15.55, p < 0.001$).

Pairwise multiple comparisons with Tukey adjustments were conducted to examine group differences (Table 3). For the number of words generated, the dementia group produced significantly fewer words than both the MCI and NCI groups. The MCI group also generated significantly fewer words than the NCI group. Regarding switching behavior, the dementia group showed significantly fewer switches than the NCI group, and the MCI group had fewer switches than the NCI group. For mean cluster size, a significant difference was observed between the dementia and NCI groups. However, differences in cluster size between MCI and NCI, as well as between dementia and MCI, were not statistically significant.

Table 3

Results of post-hoc pairwise comparisons (Tukey-adjusted) for verbal fluency measures across diagnostic groups. Estimates represent the mean difference between groups, with standard error (SE), degrees of freedom (df), t-statistic, and p-value.

Models	Groups	Estimates	SE	df	t	p
Words generated	Dementia - MCI	-2.66	0.99	308.72	-2.68	0.0211
	Dementia - NCI	-5.08	0.99	309.63	-5.14	< 0.001
	MCI - NCI	-2.42	0.52	306.09	-4.63	< 0.001
Switch times	Dementia - MCI	0.84	0.84	308.23	-1.44	0.3216
	Dementia - NCI	0.83	0.83	309.24	-3.98	< 0.001
	MCI - NCI	0.44	0.44	310.69	-4.80	< 0.001
Clusters	Dementia - MCI	0.09	0.09	308.02	-2.17	0.0774
	Dementia - NCI	0.09	0.09	255.89	-2.75	0.0177
	MCI - NCI	0.04	0.04	164.71	-1.13	0.4940

Post-hoc pairwise comparisons were conducted to identify specific group differences. The estimated differences and their 95% confidence intervals for each verbal fluency measure are visualized in Fig. 4.

These results collectively indicate that cognitive decline, as observed in the dementia and MCI groups, is associated with a reduction in verbal fluency performance, specifically in terms of the number of words generated, the ability to switch between semantic categories, and the mean cluster size. The findings highlight the sensitivity of these verbal fluency measures in distinguishing between different levels of cognitive impairment.

2.3 Predictive Modeling of Diagnostic Status

To assess the predictive utility of verbal fluency measures for diagnostic classification, particularly in the presence of severe class imbalance (NCI > > Dementia), we employed a class-weighted LASSO regularized multinomial logistic regression. Weights were inversely proportional to class frequencies to penalize the misclassification of the minority class (Dementia), thereby prioritizing clinical sensitivity over raw accuracy. We compared a Baseline model, which includes only age, education, and number of words generated, against a Full model adding mean cluster size and switching frequency to the baseline predictors.

The Full model, incorporating both demographic and fine-grained verbal fluency metrics, achieved a robust multiclass Area Under the Curve (AUC) of 0.81, indicating strong overall separability between diagnostic groups despite the imbalanced data distribution.

Critically, the application of class weighting successfully addressed the sensitivity limitations observed in unweighted approaches. As detailed in Table 4, the Full model achieved a sensitivity of 66.7% for detecting Dementia. While this strategic adjustment resulted in a moderate overall accuracy (57.9%) and a lower Positive Predictive Value (PPV = 12.5%) due to the low prevalence of dementia in the sample, it yielded a remarkable Negative Predictive Value (NPV) of 97.6%. This high NPV indicates that the model is exceptionally reliable as a "rule-out" screening tool: individuals classified as "NCI" by the model are highly likely to be truly cognitively healthy.

Table 4

Classification performance metrics of the class-weighted LASSO regularized multinomial logistic regression model on the held-out test set. Metrics are shown for each diagnostic class: No Cognitive Impairment (NCI), Mild Cognitive Impairment (MCI), and Dementia. PPV: Positive Predictive Value; NPV: Negative Predictive Value.

Class	Sensitivity (Recall)	Specificity	Precision (PPV)	NPV	F1-Score	Balanced Accuracy
NCI	0.632	0.895	0.923	0.548	0.750	0.763
MCI	0.438	0.805	0.467	0.786	0.452	0.621
Dementia	0.667	0.741	0.125	0.976	0.211	0.704

For identifying Mild Cognitive Impairment (MCI), the model exhibited moderate performance (sensitivity = 43.8%, specificity = 80.5%), suggesting that while clustering and switching metrics provide valuable signals, the semantic deficits in the prodromal MCI stage are more subtle and heterogeneous than in established dementia.

The mechanics of feature selection and model performance are visualized in Fig. 5. The cross-validation process (a) identified the optimal penalty parameter λ that minimizes weighted deviance. The coefficient path (b) explicitly reveals the contribution of the proposed metrics. Notably, "Mean Cluster Size", the green line, exhibits a steep negative trajectory with a large magnitude, indicating that smaller cluster sizes are a potent predictor of cognitive decline, independent of "Word Count" (orange line). "Switching" (purple line) also displays a stable negative coefficient, confirming that reduced cognitive flexibility contributes to the decision boundary. Finally, the ROC comparison for the specific task of distinguishing Dementia from NCI (c) demonstrates that the Full Model (orange line, AUC = 0.94) maintains a competitive edge over the Baseline (grey line, AUC = 0.93), particularly in the high-sensitivity regions critical for screening purposes.

3. Overall Discussion

3.1 Correlations of verbal fluency and cognitive performance as a function of age and education

The findings of Phase 1 reveal that the correlations between verbal fluency and cognitive function measures vary across different age and education groups.

Age was negatively correlated with the number of generated words and switch times, consistent with previous findings suggesting that older people generally produce fewer words and make fewer switches (Troyer et al., 1997). Education was positively correlated with all three verbal fluency measures (number of words, mean cluster size, switches) and both cognitive scores (memory and orientation). More importantly, the number of generated words and switches were positively correlated with memory and orientation scores, whereas the mean cluster size was not significantly correlated with either of the two cognitive measures.

To further explore the roles of age and education, we regrouped the participants and found that the correlation between the three verbal fluency measures and the two cognitive performance measures show slightly different patterns across different age and education level groups. Specially, there are notable positive correlations between the number of words generated and both memory and orientation, particularly among those with low-to-intermediate education levels. This suggests that as individuals age, the ability to generate words in verbal fluency tasks becomes more closely linked to their cognitive functions, particularly in areas such as memory and orientation. The fact that these correlations are stronger in individuals with lower education levels might indicate that education serves as a cognitive reserve, buffering against age-related cognitive decline. Thus, for those with less education, the verbal fluency task may better reflect their cognitive status.

Switching frequency showed a similar positive correlation with memory across different education levels suggesting that the ability to switch between clusters in a verbal fluency task is a cognitive skill that may be a robust indicator of cognitive function in aging populations across different levels of education. The additional correlation with orientation in age groups (60–69; 70–79) with the intermediate-to-low education further supports the idea that switching frequency, which reflects cognitive flexibility, is particularly important for assessing cognitive function in these individuals.

4.2 Verbal fluency performance and cognitive decline

The analysis in Phase 2 provides a deeper exploration of verbal fluency performance as a function of cognitive decline by comparing three groups: No Cognitive Impairment (NCI), Mild Cognitive Impairment (MCI), and Dementia.

First, the dementia group produced significantly fewer words compared to both the MCI and NCI groups, indicating the impact of cognitive decline on verbal fluency. This result aligns with existing literature (Epker et al., 1999; Fagundo et al., 2008; Gomez & White, 2006; Mok et al., 2004; Zhao et al., 2013), suggesting a progressive deterioration in lexical production abilities as cognitive impairment worsens. The fact that the MCI group generated significantly fewer words than the NCI group indicates that even early stages of cognitive decline can adversely affect verbal fluency. The gradual reduction in the number of words generated across these groups suggests that verbal fluency tasks can effectively differentiate

between varying levels of cognitive impairment and may serve as a useful tool for early detection and monitoring of cognitive decline.

The analysis of clustering and switching behaviours in the verbal fluency task offers important insights into the cognitive processes affected by different stages of cognitive decline. Clustering, which refers to the production of words within subcategories, and switching, the ability to shift between these subcategories, are both critical components of verbal fluency that reveal how individuals access and organize lexical information. Our findings revealed differences of varying degrees in both clustering and switching among participants with NCI, MCI, and dementia, consistent with previous studies (Epker et al., 1999; Fagundo et al., 2008; Gomez & White, 2006; Zhao et al., 2013). Specifically, switching, which relies on cognitive flexibility and strategic search, showed significant declines across the groups. The dementia group exhibited the lowest switching frequency, indicating a reduced ability to efficiently shift between subcategories. The MCI group displayed an intermediate level of performance, lower than the NCI group but comparable with the dementia group.

In addition, individuals with dementia showed a marked reduction in mean cluster size compared to the NCI groups. This suggests that as cognitive decline progresses, the ability to generate semantically related words within specific subcategories deteriorates, reflecting a breakdown in the retrieval of lexical information stored in memory. Although statistical tests showed mixed results for MCI group differences, the integration of clustering and switching into the predictive modeling framework provided critical clinical insights. By employing a class-weighted approach, our Full Model demonstrated a substantial improvement in sensitivity for detecting dementia compared to traditional unweighted methods. While this adjustment resulted in a lower positive predictive value, this trade-off is clinically justifiable for a first-line community screening tool. The model achieved a high Negative Predictive Value, suggesting it functions effectively as a "rule-out" instrument. In practice, this means the tool can safely exclude the vast majority of cognitively healthy older adults, allowing limited medical resources to be focused on the smaller subset of flagged individuals for comprehensive diagnosis.

The moderate sensitivity observed for MCI further highlights the nuanced nature of prodromal cognitive decline. Unlike the overt semantic breakdown seen in dementia, MCI deficits may be more heterogeneous. This suggests that while SVF metrics are powerful markers for established dementia, detecting the subtle signals of MCI may require combining these automated linguistic metrics with other cognitive modalities to improve screening precision.

It should be noted that the validity of measures, such as switching times and cluster size depends on the appropriateness of defining word categories. Previous studies rely heavily on schemes that were often developed for English-speaking populations. These schemes may not fully capture the semantic relationships in other languages, potentially leading to inaccuracies when applied cross-culturally. Our use of semantic knowledge base like Cilin can provide relatively effective classification of words in our automated calculation of verbal fluency measures. Future studies could further explore this approach in other languages and also examine other computational ways of representing lexical knowledge (Alacam

et al., 2022; Kim et al., 2019) to ensure that the observed differences in verbal fluency are truly reflective of cognitive decline rather than artifacts of linguistic mismatches.

4. Methods

4.1 Participants and study design

A total number of 11,583 participants joined a large-scale project which aims to develop cognitive screener for dementia detection in China. In Phase 1, we selected participants with an age range between 18 and 80 and removed those with missing data. The dataset comprised a total of 9670 participant (3510 males and 6160 females) with a mean age of 65 years (SD = 8.8). The mean level of education was 9.3 years (SD = 3.8) ranging from 0 to 16 years. This study received approval from the Medical Ethics Committee at Zhejiang University School of Public Health (ZGL202101-1), and written informed consent was obtained from all participants or their legally authorized representatives. To further explore the intricate dynamics between verbal fluency measures and the two cognitive measures as modulated by age and years of education, we regrouped the dataset into four age groups (see Table 5), and each with three education levels, low (< 6 yrs), intermediate (6–12 yrs), and high (> 12 yrs).

Table 5
Demographic characteristics of the community-dwelling older adult sample, stratified by age group and education level (Phase 1). Data are presented as mean (M) and standard deviation (SD) for age and years of education within each stratum.

Groups	Education	N	Age	Education		
			M	SD	M	SD
50–59	low	336	55.1	2.8	4.9	2.3
	mid	879	55.5	2.8	9	0
	high	942	55.2	2.9	13.1	1.8
60–69	low	822	65.3	2.7	4.4	2.7
	mid	1789	65.1	2.7	9	0
	high	1304	63.8	2.7	13.1	1.8
70–79	low	1194	73.9	2.7	4.4	2.7
	mid	1109	73.2	2.7	9	0
	high	752	74.3	2.7	13.6	2
> 80	low	83	80	0	3.8	2.9
	mid	58	80	0	9	0
	high	61	80	0	13.7	2

To further investigate whether there were significant differences of verbal fluency measures among participants with and without cognitive impairment, we used a subset of data ($n = 315$) from Phase 1 with diagnosis from cognitive evaluations (Phase 2). These participants have undergone comprehensive cognitive evaluations, including the CDR (Clinical Dementia Rating) scale (Morris, 1993), and a comprehensive neuropsychological evaluation (Silva & Lee, 2023) from licensed professionals. These participants were then classified into three groups: No Cognitive Impairment (NCI), Mild Cognitive Impairment (MCI), or Dementia. Table 6 summarizes the demographic characteristics of the clinical cohort. Notably, the Dementia group was older ($M = 79.3$) and had fewer years of education ($M = 3.7$) compared to the NCI and MCI groups. These demographic disparities underscore the necessity of controlling for age and education as random effects in our subsequent Linear Mixed-Effects Models to isolate the specific impact of cognitive status on verbal fluency performance.

Table 6
Demographic and clinical characteristics of the participants in the diagnostic validation phase (Phase 2). The cohort (N = 315) was classified into No Cognitive Impairment (NCI), Mild Cognitive Impairment (MCI), and Dementia groups based on comprehensive clinical assessment.

Groups	N	Age		Education	
		M	SD	M	SD
Dementia	18	79.3	9.8	3.7	4.1
MCI	94	74.2	8.6	5	4.3
NCI	203	69.3	7.4	8.5	3.6

4.2 Digital Assessments of Semantic Verbal Fluency Task

The data we used had been obtained via a Digital Cognitive Screener (or DCS, see Zhao et al., 2022 for details) which included three tasks. The first task was a delayed memory test in which participants have to recall 5 words (i.e., face, silk, chrysanthemum, hotel, red) with a total score of 5. The second was an orientation task in which participants need to answer six questions of current year, month, week, date and their current city, with a total score of 6. The third is a semantic verbal fluency task in which sixty seconds were allotted for participants to generate animal names.

The DCS platform employs voice-recognition technology and is designed for utilization by non-professional facilitators, including caregivers or social workers (Zhao et al., 2022). The platform autonomously evaluates participants' performance by first converting the test audio into text through automatic speech recognition (ASR). The system also includes a name entity recognition model to identify crucial details, such as year, day, and address. The extracted information is cross-referenced with the accurate standard answer. For every question, the system autonomously assesses the response and assigns a corresponding score. A previous study has confirmed the validity of this system in terms of the consistency between automated scoring and the conventional manual approach (Zhao et al., 2022).

4.3 Calculating clusters and switches

Many tools (e.g., Zemla et al., 2020) have been developed to facilitate the analysis of verbal fluency data. Scheme files are required to compute cluster sizes and switches in these tools. There exist several English animal category schemes from previous studies (Hills et al., 2012; Troyer et al., 1997). However, these schemes rely on subjective judgements and may not be valid across different languages and cultures.

We employed Tongyici Cilin (or Cilin in short), an existing comprehensive semantic knowledge base as an objective and comprehensive way of defining semantic categories. Cilin is a dictionary created by HIT

IR-Lab and organized in a five-tier hierarchical tree structure (see Table 7). The first tier indicates major classes and the second tier is the intermediate level, e.g., Bh for all plants and Bi for all animals. The third and fourth tiers further narrow down the category. For example, horse is marked with Bi06A whereas donkey is marked with Bi06B. The fifth tier further distinguishes subcategories, e.g., foals (= Bi06A02) and mares (= Bi06A03). The words grouped at the terminal nodes have the same or similar meaning, e.g., “马”, “驴”, “骡” (= Bi06B01, donkey in Chinese). This dictionary was originally created for and has been widely used in Chinese natural language processing (Wang et al., 2018).

Table 7
Hierarchical structure of Tongyici Cilin (Cilin) with an example code. The five-tier structure (Major class, Intermediate class, Minor class, Group, Atomic group) was used to objectively define semantic clusters for animal names in the verbal fluency task.

Level	1	2	3	4	5
Category	Major class	Intermediate class	Minor class	Group	Atomic group
Code	B	i	0 6	A	0 1

Based on Cilin, clusters were defined as groups of successively generated words belonging to the same semantic subcategory with the same first three levels. For example, horse and donkey were treated as belonging to the same category. Cluster size was counted from the first word in each cluster, and the mean cluster size was calculated for each participant. Switches were calculated as the number of transitions between clusters, including single words.

4.4 Statistical analyses

The analysis followed a structured, multi-stage approach, beginning with variable definition and descriptive summaries, followed by relational, comparative, and predictive modeling, and concluding with robustness checks.

For both phases, the predictors were the number of animal names produced (measuring overall verbal productivity), the switching frequency (i.e., a discrete count representing the number of times a participant transitioned between different semantic subcategories, measuring cognitive flexibility) and the mean cluster size (i.e., the average number of words produced within a single semantic subcategory). The dependent variable for the analyses in Phase 2 was participant's cognitive status, i.e., No Cognitive Impairment (NCI), Mild Cognitive Impairment (MCI), or Dementia.

In Phase 1, to explore the interrelationships between variables, Pearson correlation analyses were conducted and faceted heatmaps were used to examine how these correlations varied across different age and education subgroups. A network graph was generated to visualize the the direction and strength of these correlations.

In Phase 2, we built three separate LMER models (via the *lme4* package in R, Bates et al., 2010) with number of generated words, mean cluster size and switching frequency as the dependent variable

respectively. All models included the overall cognitive function level (NCI, MCI, dementia) of the subjects as a fixed factor and age and years of education were considered as random factors with random intercepts to maximize the models' generalizability across the data, following the suggestion by Barr et al., (2013). We utilized the Kenward-Roger degrees of freedom approximation (Halekoh & Hojsgaard, 2014) and the *Anova* function from the *car* package in *R*, specifying the test as "F". Main effects for multilevel factors are reported instead of comparisons with the baseline level using t-tests. Therefore, the reported main effects were averaged across all levels of the other effects and can be directly used to test whether these verbal fluency measures can detect group differences. Post-hoc pairwise comparisons with Tukey's adjustment were used to identify specific group differences.

To assess the predictive power of verbal fluency measures for cognitive impairment, we used a supervised machine learning approach: LASSO-regularized multinomial logistic regression. The model was trained and tuned via 10-fold cross-validation on a training subset, with final performance evaluated on a held-out test set. Performance metrics included overall accuracy, Cohen's Kappa, class-specific sensitivity and specificity, and multiclass Area Under the Curve..

5. Conclusion

The present study demonstrates that verbal fluency measures—particularly the number of words generated and switching frequency—correlate with cognitive functions such as memory and orientation among Mandarin-speaking middle-aged and older adults, especially those with lower education levels. Significant differences in these measures were observed across NCI, MCI, and Dementia groups. Individuals with dementia exhibited marked reductions in word production, switching frequency, and mean cluster size, reflecting impaired lexical organization and cognitive flexibility. Similar, though less severe, deficits were present in the MCI group, indicating that disruptions in word generation and switching emerge early in cognitive decline and progress with dementia severity. Machine learning analysis confirmed that adding clustering and switching metrics to a baseline model (age, education, word count) enhances dementia detection sensitivity. The model's high negative predictive value supports its potential as a cost-effective, automated triage tool for community screening. By reliably ruling out cognitively healthy individuals, such a tool could streamline the diagnostic pathway, ensuring that those with potential impairments receive timely and comprehensive clinical assessment.

Declarations

6. Data availability

The data that support the findings of this study are available from authors, Xin Xu and Mengyue Wu, upon reasonable request.

7. Code availability

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

8. Competing interests

All authors declare no financial or non-financial competing interests.

9. Funding Declaration

This study received no funding.

Author Contribution

J.C. and M.W. conceptualized the work. X.X. and X.Z. were responsible for data collection. J.C. and Y.W. developed the code and conducted the computational analysis. Y.W. performed the machine learning modelling. J.C. wrote the first draft of the manuscript, and all authors participated in the revision of the manuscript. All authors read and approved the final version.

References

1. Aita, S. L., Beach, J. D., Taylor, S. E., Borgogna, N. C., Harrell, M. N., & Hill, B. D. (2019). Executive, language, or both? An examination of the construct validity of verbal fluency measures. *Applied Neuropsychology: Adult*, 26(5), 441–451. <https://doi.org/10.1080/23279095.2018.1439830>
2. Alacam, Ö., Schüz, S., Wegrzyn, M., Kißler, J., & Zarrieß, S. (2022). Exploring Semantic Spaces for Detecting Clustering and Switching in Verbal Fluency. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Eds.), *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 178–191). International Committee on Computational Linguistics. <https://aclanthology.org/2022.coling-1.16>
3. Amieva, H., Le Goff, M., Millet, X., Orgogozo, J. M., Pérès, K., Barberger-Gateau, P., Jacqmin-Gadda, H., Dartigues, J. F., Amieva, H., Le Goff, M., Millet, X., Orgogozo, J. M., Pérès, K., Barberger-Gateau, P., Jacqmin-Gadda, H., & Dartigues, J. F. (2008). Prodromal Alzheimer's disease: Successive emergence of the clinical symptoms. *Annals of Neurology*, 64(5), 492–498. <https://doi.org/10.1002/ana.21509>
4. Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
5. Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. Springer.
6. Bertola, L., Mota, N. B., Copelli, M., Rivero, T., Diniz, B. S., Romano-Silva, M. A., Ribeiro, S., & Malloy-Diniz, L. F. (2014). Graph analysis of verbal fluency test discriminate between patients with

- Alzheimer's disease, mild cognitive impairment and normal elderly controls. *Frontiers in Aging Neuroscience*, 6. <https://doi.org/10.3389/fnagi.2014.00185>
7. Epker, M. O., Lacritz, L. H., & Munro Cullum, C. (1999). Comparative Analysis of Qualitative Verbal Fluency Performance in Normal Elderly and Demented Populations. *Journal of Clinical and Experimental Neuropsychology*, 21(4), 425–434. <https://doi.org/10.1076/jcen.21.4.425.890>
 8. Fagundo, A. B., López, S., Romero, M., Guarch, J., Marcos, T., & Salamero, M. (2008). Clustering and switching in semantic fluency: Predictors of the development of Alzheimer's disease. *International Journal of Geriatric Psychiatry*, 23(10), 1007–1013. <https://doi.org/10.1002/gps.2025>
 9. Farzanfar, D., Statucka, M., & Cohn, M. (2018). Automated Indices of Clustering and Switching of Semantic Verbal Fluency in Parkinson's Disease. *Journal of the International Neuropsychological Society*, 24(10), 1047–1056. <https://doi.org/10.1017/S1355617718000759>
 10. Gomez, R., & White, D. (2006). Using verbal fluency to detect very mild dementia of the Alzheimer type. *Archives of Clinical Neuropsychology*, 21(8), 771–775. <https://doi.org/10.1016/j.acn.2006.06.012>
 11. Halekoh, U., & Hojsgaard, S. (2014). A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models—the R package pbkrtest. *Journal of Statistical Software*, 59(9), 1–30.
 12. Henry, J. D., & Crawford, J. R. (2004). A Meta-Analytic Review of Verbal Fluency Performance in Patients With Traumatic Brain Injury. *Neuropsychology*, 18(4), 621–628. <https://doi.org/10.1037/0894-4105.18.4.621>
 13. Henry, J. D., Crawford, J. R., & Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis. *Neuropsychologia*, 42(9), 1212–1222. <https://doi.org/10.1016/j.neuropsychologia.2004.02.001>
 14. Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, 119(2), 431–440. <https://doi.org/10.1037/a0027373>
 15. Jia, L., Du, Y., Chu, L., Zhang, Z., Li, F., Lyu, D., Li, Y., Li, Y., Zhu, M., Jiao, H., Song, Y., Shi, Y., Zhang, H., Gong, M., Wei, C., Tang, Y., Fang, B., Guo, D., Wang, F., ... Qiu, Q. (2020). Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China: A cross-sectional study. *The Lancet Public Health*, 5(12), e661–e671. [https://doi.org/10.1016/S2468-2667\(20\)30185-7](https://doi.org/10.1016/S2468-2667(20)30185-7)
 16. Kim, N., Kim, J.-H., Wolters, M. K., MacPherson, S. E., & Park, J. C. (2019). Automatic Scoring of Semantic Fluency. *Frontiers in Psychology*, 10. <https://doi.org/10.3389/fpsyg.2019.01020>
 17. Kogan, M., & Jeong, H. S. (2018). Chapter 11—Alzheimer Disease. In D. Rakel (Ed.), *Integrative Medicine (Fourth Edition)* (Fourth Edition, pp. 95–107.e4). Elsevier. <https://doi.org/10.1016/B978-0-323-35868-2.00011-6>
 18. Laisney, M., Matuszewski, V., Mézenge, F., Belliard, S., Sayette, V., Eustache, F., & Desgranges, B. (2009). The underlying mechanisms of verbal fluency deficit in frontotemporal dementia and

- semantic dementia. *Journal of Neurology*, 256(7), 1083–1094. <https://doi.org/10.1007/s00415-009-5073-y>
19. LANTING, S., HAUGRUD, N., & CROSSLEY, M. (2009). The effect of age and sex on clustering and switching during speeded verbal fluency tasks. *Journal of the International Neuropsychological Society*, 15(2), 196–204. <https://doi.org/10.1017/S1355617709090237>
 20. Mok, E. H. L., Lam, L. C. W., & Chiu, H. F. K. (2004). Category Verbal Fluency Test Performance in Chinese Elderly with Alzheimer's Disease. *Dementia and Geriatric Cognitive Disorders*, 18(2), 120–124. <https://doi.org/10.1159/000079190>
 21. Monsch, A. U., Bondi, M. W., Butters, N., Salmon, D. P., Katzman, R., & Thal, L. J. (1992). Comparisons of Verbal Fluency Tasks in the Detection of Dementia of the Alzheimer Type. *Archives of Neurology*, 49(12), 1253–1258. <https://doi.org/10.1001/archneur.1992.00530360051017>
 22. Morris, J. C. (1993). The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology*, 43(11), 2412–2414. <https://www.neurology.org/doi/abs/10.1212/WNL.43.11.2412-a>
 23. Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4), 695–699. <https://doi.org/10.1111/j.1532-5415.2005.53221.x>
 24. Salehi, M., Reisi, M., & Ghasisin, L. (2017). Lexical Retrieval or Semantic Knowledge Which One Causes Naming Errors in Patients with Mild and Moderate Alzheimer's Disease. *Dementia and Geriatric Cognitive Disorders Extra*, 7(3), 419–429. <https://doi.org/10.1159/000484137>
 25. Schmidt, C. S. M., Nitschke, K., Bormann, T., Römer, P., Kümmerer, D., Martin, M., Umarova, R. M., Leonhart, R., Egger, K., Dressing, A., Musso, M., Willmes, K., Weiller, C., & Kaller, C. P. (2019). Dissociating frontal and temporal correlates of phonological and semantic fluency in a large sample of left hemisphere stroke patients. *NeuroImage: Clinical*, 23, 101840. <https://doi.org/10.1016/j.nicl.2019.101840>
 26. Silva, M. A., & Lee, J. M. (2023). Neurocognitive testing. In C. A. Kushida (Ed.), *Encyclopedia of Sleep and Circadian Rhythms (Second Edition)* (Second Edition, pp. 552–557). Academic Press. <https://doi.org/10.1016/B978-0-12-822963-7.00047-5>
 27. Torralva, T., Laffaye, T., Báez, S., Gleichgerricht, E., Bruno, D., Chade, A., Ibañez, A., Manes, F., Gershanik, O., & Roca, M. (2015). Verbal Fluency as a Rapid Screening Test for Cognitive Impairment in Early Parkinson's Disease. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 27(3), 244–247. <https://doi.org/10.1176/appi.neuropsych.14060139>
 28. Troyer, A. K. (2000). Normative Data for Clustering and Switching on Verbal Fluency Tasks. *Journal of Clinical and Experimental Neuropsychology*, 22(3), 370–378. [https://doi.org/10.1076/1380-3395\(200006\)22:3;1-V;FT370](https://doi.org/10.1076/1380-3395(200006)22:3;1-V;FT370)
 29. Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology*, 11(1), 138–146. <https://doi.org/10.1037/0894-4105.11.1.138>

30. Troyer, A. K., Moscovitch, M., Winocur, G., Alexander, M. P., & Stuss, D. (1998). Clustering and switching on verbal fluency: The effects of focal frontal- and temporal-lobe lesions. *Neuropsychologia*, 36(6), 499–504. [https://doi.org/10.1016/S0028-3932\(97\)00152-8](https://doi.org/10.1016/S0028-3932(97)00152-8)
31. Wang, X., Wang, X., & Liu, Z. (2018). Cilin-Based Semantic Similarity Calculation of Chinese Words. In F. Qiao, S. Patnaik, & J. Wang (Eds.), *Advances in Intelligent Systems and Computing* (vol 690; pp. 240–244). Springer International Publishing. https://doi.org/10.1007/978-3-319-65978-7_37
32. Zemla, J. C., Cao, K., Mueller, K. D., & Austerweil, J. L. (2020). SNAFU: The Semantic Network and Fluency Utility. *Behavior Research Methods*, 52(4), 1681–1699. <https://doi.org/10.3758/s13428-019-01343-w>
33. 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(1), 75–82. <https://doi.org/10.1007/s12264-013-1301-7>
34. Zhao, X., Hu, R., Wen, H., Xu, G., Pang, T., He, X., Zhang, Y., Zhang, J., Chen, C., Wu, X., & Xu, X. (2022). A voice recognition-based digital cognitive screener for dementia detection in the community: Development and validation study. *Frontiers in Psychiatry*, 13. <https://doi.org/10.3389/fpsy.2022.899729>

Figures

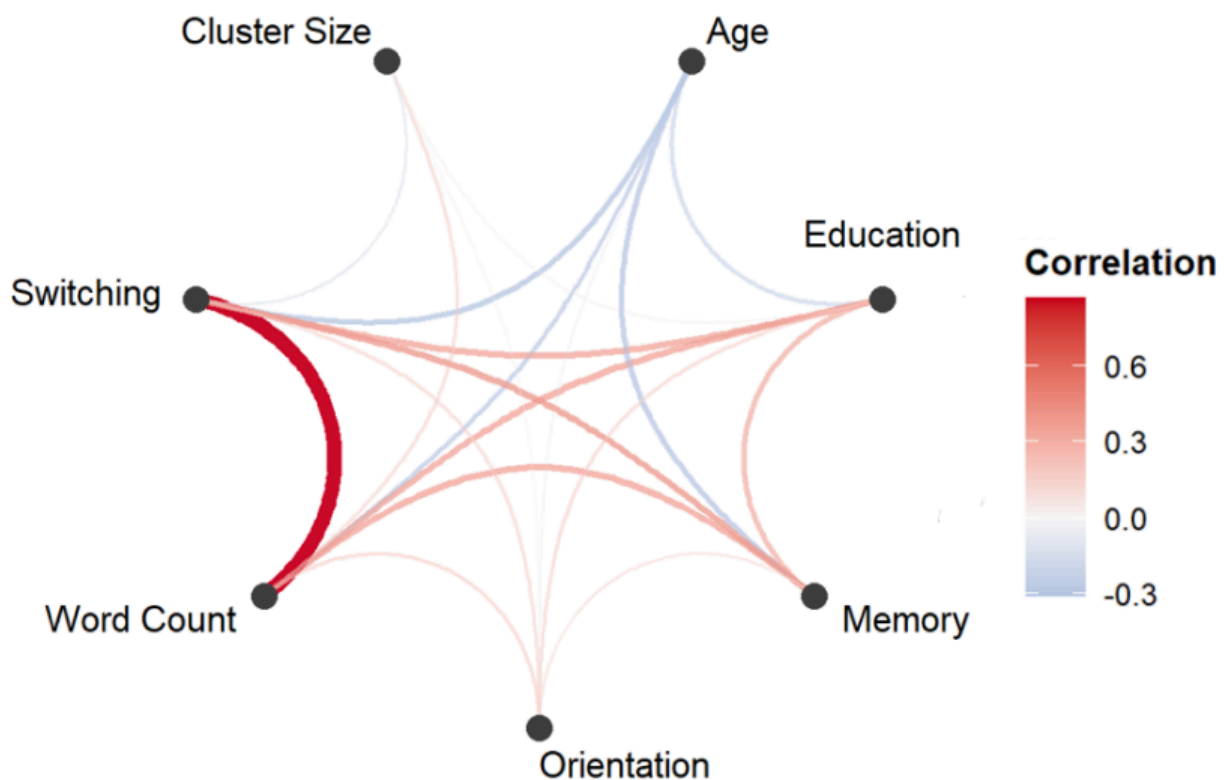


Figure 1

Network visualization of correlations among verbal fluency metrics, cognitive performance, and demographic factors among community-dwelling older adults (Phase 1, N=9,670). Line width and color intensity represent the magnitude of Pearson correlation coefficients (see Table 2). The strongest associations were observed between word count, switching frequency, and memory/orientation scores.

Correlations by Age Group and Education Level

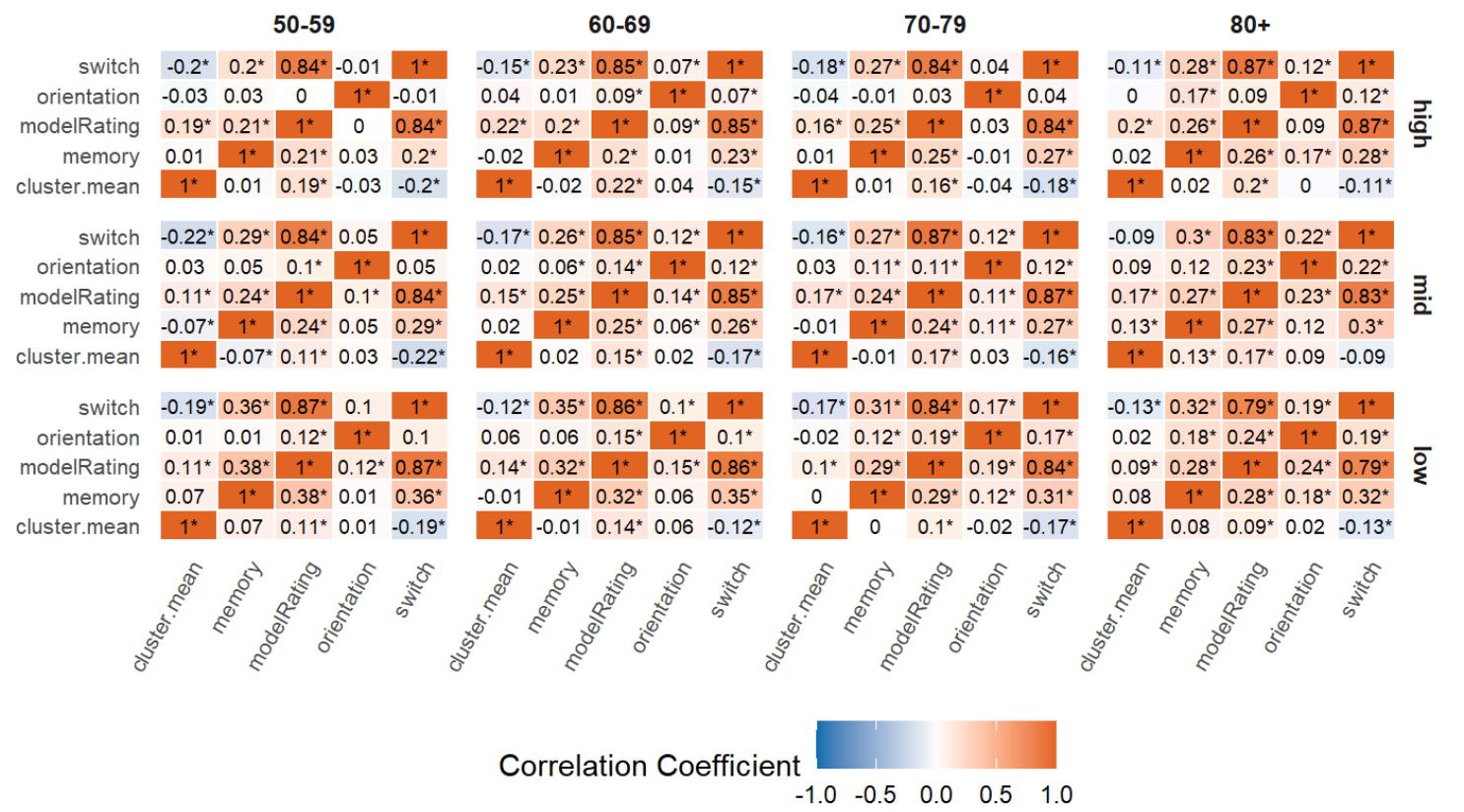
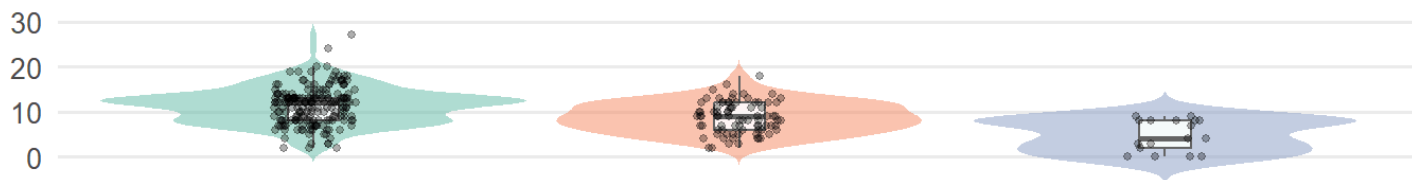


Figure 2

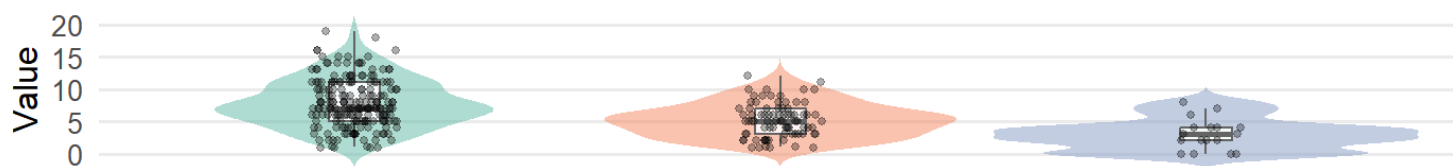
Heatmaps of correlations between verbal fluency performance and cognitive measures, stratified by age group and education level. Positive correlations were most consistent for word count and switching, particularly among individuals with low-to-intermediate education.

Verbal Fluency Performance across Diagnostic Groups

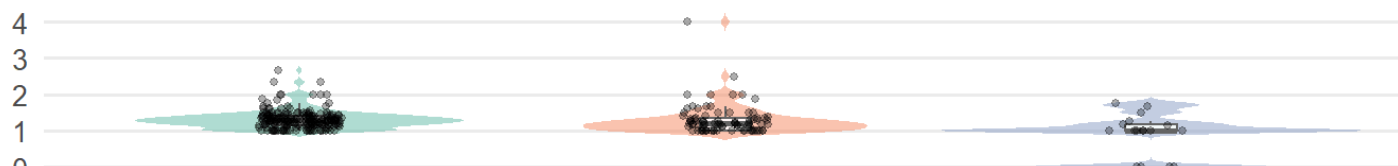
Number of words



Switch times



Cluster sizes



NCI

MCI

Dementia

Diagnostic Group

Figure 3

Group differences in semantic verbal fluency performance across diagnostic categories (Phase 2, N=315). Violin plots with overlaid boxplots show the distribution of (A) total number of words generated, (B) switching frequency, and (C) mean cluster size for the No Cognitive Impairment (NCI), Mild Cognitive Impairment (MCI), and Dementia groups. Significant declines are evident across groups, particularly for word count and switching.

LMER Model Estimates of Group Differences

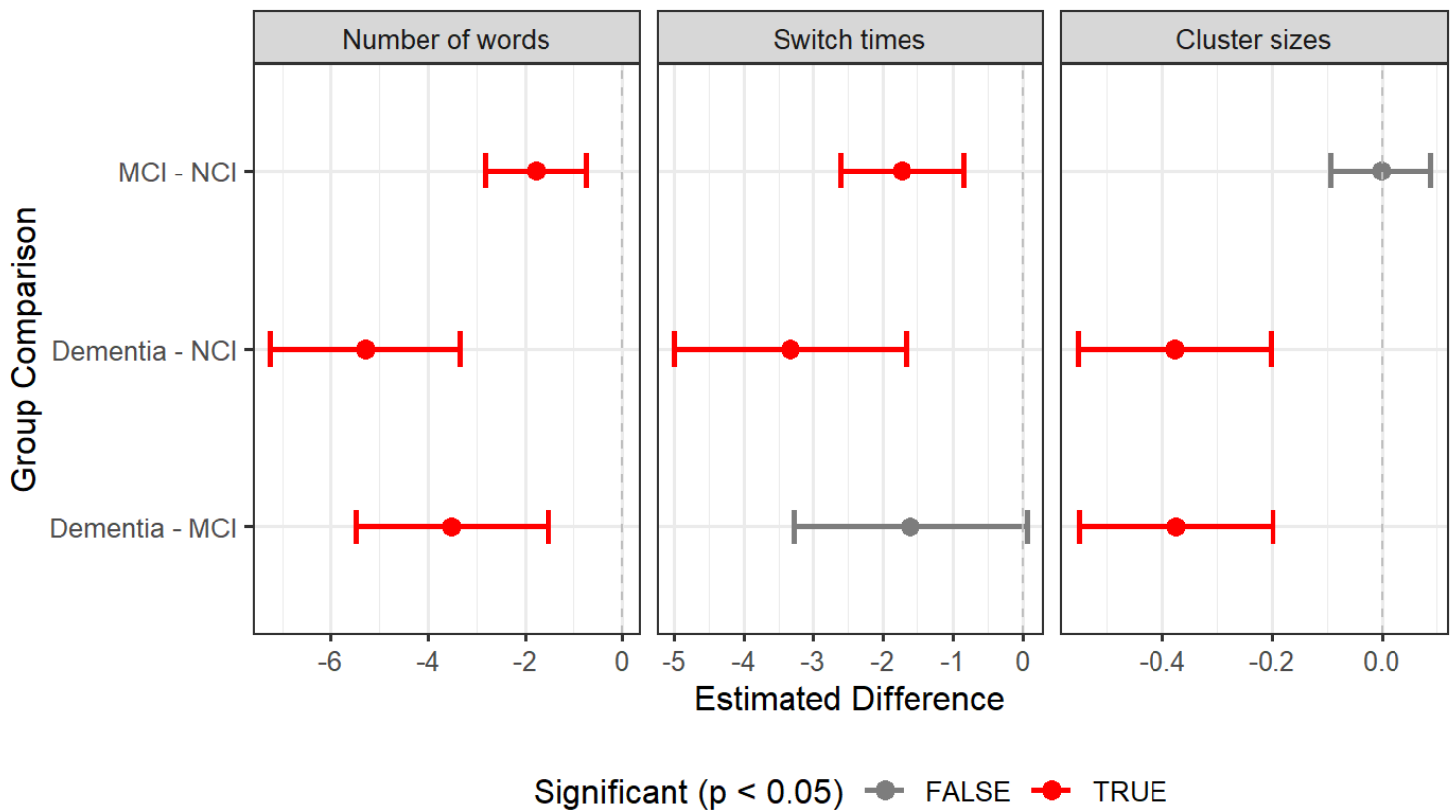


Figure 4

Pairwise comparisons of verbal fluency measures between diagnostic groups. Estimated differences (with 95% confidence intervals) from linear mixed-effects models are shown for (A) number of words, (B) switching frequency, and (C) mean cluster size.

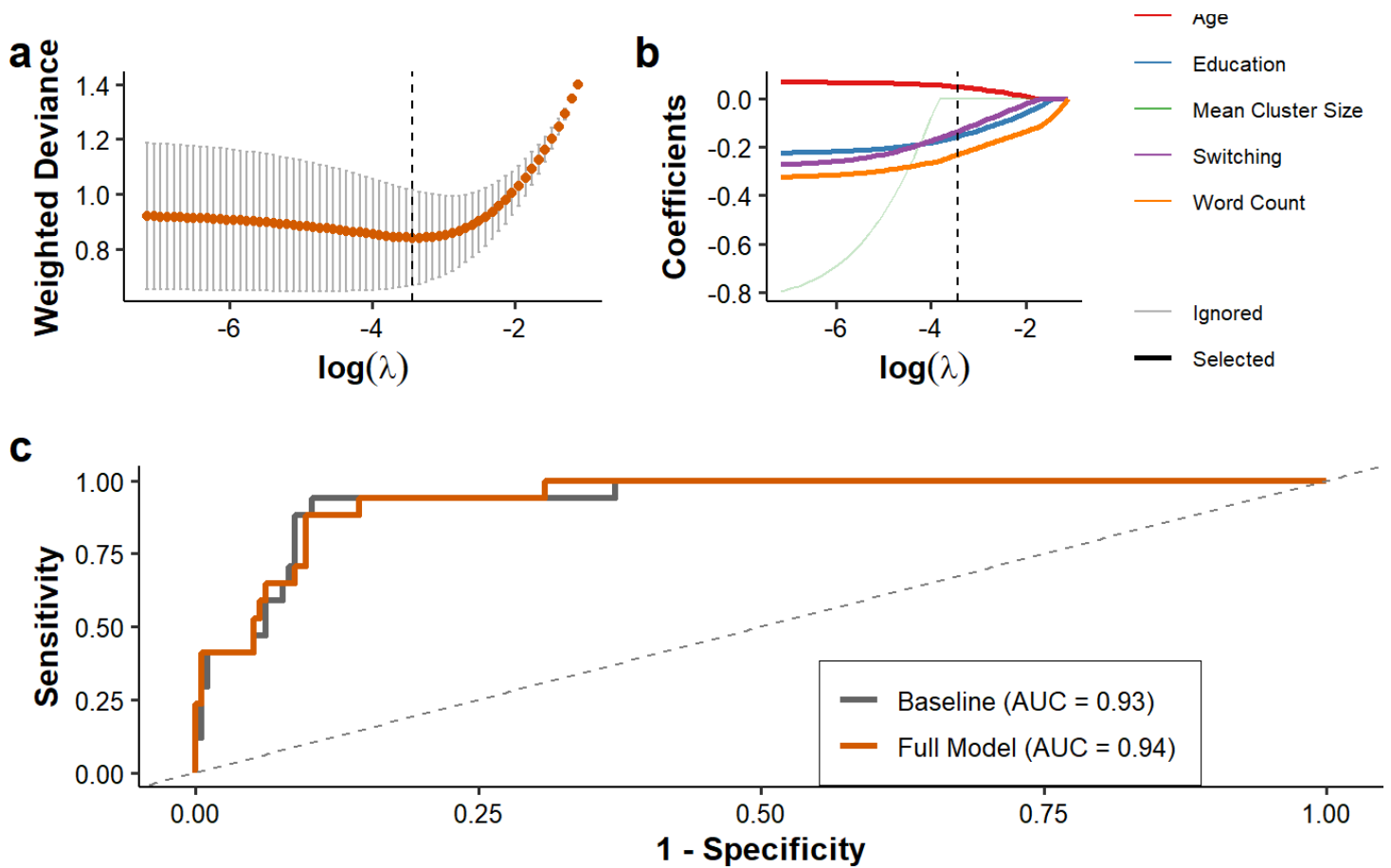


Figure 5

Feature selection and performance of the class-weighted LASSO model for diagnostic prediction. (A) Cross-validation curve for selecting the optimal regularization parameter (λ). (B) Coefficient paths showing how model coefficients shrink as λ increases; note the strong negative coefficients for Mean Cluster Size and Switching. (C) Receiver operating characteristic (ROC) curves comparing the Full model (with clustering/switching) and the Baseline model for distinguishing Dementia from the NCI group.