The Association of Sociodemographic Factors With Total and Item-Level Semantic Fluency Metrics

Magdalena Beran[1, 2], Emma L. Twait[1, 3, 4, 5], Annelot P. Smit[1], Marleen F. Posthuma[6], Demi van Dijk[6], Katherinne M. Rabanal[7], Dayanara Rosado[7], Roxanna J. Flores[7], Carolyn L. Qian[7], Shana S. Samuel[7], Gelan Ying[8], Richard Mayeux[7, 9, 10], Thomas T. van Sloten[2, 11], Miranda T. Schram[2, 12, 13], Jennifer J. Manly[7, 9, 10], Mirjam I. Geerlings[1, 3, 4, 5], and Jet M. J. Vonk[1, 14]

[1] Department of Epidemiology and Global Health, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University
[2] Department of Internal Medicine, School for Cardiovascular Diseases (Cardiovascular Research Institute Maastricht), Maastricht University
[3] Department of General Practice, Amsterdam University Medical Centers, University of Amsterdam
[4] Amsterdam Public Health, Aging and Later life, and Personalized Medicine, Amsterdam University Medical Centers, the Netherlands
[5] Amsterdam Neuroscience, Neurodegeneration, and Mood, Anxiety, Psychosis, Stress, and Sleep, Amsterdam University Medical Centers, the Netherlands
[6] Department of Linguistics, Faculty of Arts, University of Groningen
[7] Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Vagelos College of Physicians and Surgeons, Columbia University
[8] Department of Clinical and Health Psychology, University of Florida
[9] Getrude H. Sergievsky Center, Vagelos College of Physicians and Surgeons, Columbia University
[10] Department of Neurology, Vagelos College of Physicians and Surgeons, New York Presbyterian Hospital, Columbia University
[11] Department of Vascular Medicine, University Medical Center Utrecht
[12] School for Mental Health and Neuroscience, Maastricht University
[13] Heart and Vascular Centre, Maastricht University Medical Centre
[14] Department of Neurology, Memory and Aging Center, University of California San Francisco

Objective: We aimed to estimate the association of age, education, and sex/gender with semantic fluency performance as measured by the standard total number of words as well as novel item-level metrics and to descriptively compare associations across cohorts with different recruitment strategies and sample compositions. ***Method:*** Cross-sectional data from 2,391 individuals from three cohorts were used: Washington Heights/Inwood Columbia Aging Project, a community-based cohort; Second Manifestations of ARTerial disease-Magnetic Resonance, a clinic-based cohort; and African American Alzheimer's Disease Genetics Study, a volunteer-based cohort. Total number of correct words and six item-level semantic fluency metrics were included as main outcomes: average cluster size, number of cluster switches, lexical/Zipf frequency, age of acquisition, and lexical decision response time. General linear models were run separately in each cohort to model the association between sociodemographic variables and semantic fluency metrics. ***Results:*** Across cohorts, older age was associated with a lower total score and fewer cluster switches. Higher level of education was associated with naming more words, performing more cluster switches, and naming words with a longer lexical decision response time, lower frequency of occurrence, or later age of acquisition. Being female compared to male was associated with naming fewer words, smaller cluster sizes, naming words with a longer lexical decision response time, and lower age of acquisition. The effects varied in strength but were in a similar direction across cohorts. ***Conclusions:*** Item-level semantic fluency metrics—similar to the standard total score—are sensitive to the effects of age, education, and sex/gender. The results suggest geographical, cultural, and cross-linguistic generalizability of these sociodemographic effects on semantic fluency performance.

Key Points
Question: How are age, educational level, and sex/gender related to semantic fluency test performance in geographically, culturally, and linguistically diverse cohorts? *Findings:* Semantic fluency test performance varied in individuals with differing ages, educational level attainments, and sexes/genders in a similar manner across cohorts. *Importance:* Differences in age, educational level, and sex/gender need to be considered when rating semantic fluency test performance based on item-level metrics. *Next Steps:* Future studies are recommended to focus on creating normative data for semantic fluency test performance based on item-level metrics.

Keywords: animal fluency, cohort study, demographics, semantic fluency

Supplemental materials: https://doi.org/10.1037/neu0000955.supp

Semantic fluency is a widely used neuropsychological test to assess cognitive functioning and access to semantic memory (Henry et al., 2004). The test consists of naming as many items belonging to a prespecified category (e.g., animals) during a time interval (e.g., 1 or 2 min; Zemla et al., 2020). While traditionally the total number of unique words produced is used for semantic fluency test scoring, novel item-level metrics of this task have gained recent interest as more sensitive measures of cognitive performance (De Marco et al., 2021; Rofes et al., 2020; van den Berg et al., 2022; Vonk, Flores, et al., 2019). Importantly, performance on cognitive tests can vary considerably when used in ethnically and educationally diverse

populations (Arce Rentería et al., 2019; Avila et al., 2019; Manly et al., 1998; Vonk, Arce Rentería, et al., 2019). Advanced aging and low educational background might lead to lower test performance and could be misinterpreted as cognitive impairment if these factors are not appropriately taken into account (Mougias et al., 2019). However, sociodemographic effects on item-level metrics of semantic fluency are currently underexplored.

In semantic fluency, older age has been shown to be inversely associated with the total number of correct words in several studies (Brickman et al., 2005; Lanting et al., 2009; van Hooren et al., 2007; Zarino et al., 2014). Furthermore, a higher level of education has

repeatedly been shown to be associated with a higher total score on the animal fluency test (Brickman et al., 2005; van Hooren et al., 2007; Zarino et al., 2014). While the importance of taking age and level of education into account when rating semantic fluency test performance was highlighted in a recent review on cross-linguistic animal fluency test performance, the impact of sex on test performance was considered negligible (Ardila, 2020). This conclusion is only partially in line with previous studies that found conflicting results for sex or gender differences in semantic fluency test performance based on total scores or measures of word order (cluster size, cluster switches; Brucki & Rocha, 2004; Kosmidis et al., 2004; Sokołowski et al., 2020; Weiss et al., 2006). Up until now, the literature on this has only been summarized in young adults (Sokołowski et al., 2020).

However, the total score alone might not fully capture a participant's performance (Troyer et al., 1997). Based on the item-level data (i.e., which exact words are generated), several alternative measures for scoring and deciphering fluency data have been identified. These metrics can provide further insights into both healthy aging (Rofes et al., 2023; Troyer et al., 1997) and neurodegenerative disorders, such as (preclinical) Alzheimer's disease (Woods et al., 2016) and frontotemporal dementia (van den Berg et al., 2022). Examples of metrics that can be extracted at the item level of semantic fluency are lexical frequency, age of acquisition, lexical decision response time, clusters, and switches, among others.

Lexical frequency, also called word frequency, is a measure of how often a word occurs in daily language. Words with a lower frequency of occurrence take longer to process than words with a higher frequency (Brysbaert et al., 2018). Age of acquisition represents the age at which a certain word was learned (Ghyselinck et al., 2004). Importantly, earlier acquired words seem to be remembered faster than words acquired later (Catling et al., 2013). Lexical decision response time is a metric derived by usage of the lexical decision task. It indicates the time in milliseconds that it took participants to decide if a string of letters is a word or a nonword (Balota et al., 2007). The aforementioned metrics can be subsumed under the term psycholinguistic metrics. Additionally, metrics of word order have been identified. Troyer et al. (1997) proposed clustering and switching as two fundamental components underlying optimal verbal fluency performance. When analyzing the animal fluency test, clusters represent (sub)categories that can be used to group animals such as type of species (e.g., birds, fish), environment (e.g., zoo, jungle), or geographical location (e.g., Africa, the poles). The underlying hypothesis is that word retrieval involves searching for a subcategory in our semantic network and naming all items that belong to this subcategory until it is exhausted, followed by a switch to a different subcategory (Troyer et al., 1997).

The effects of sociodemographic factors on animal fluency performance are most often studied on total score or metrics of word order (clusters and switches), whereas literature on these effects on psycholinguistic metrics is scarce. Clustering and switching seem to be sensitive to aging as older individuals have been shown to produce larger sized clusters (Kosmidis et al., 2004; Lanting et al., 2009) and fewer (Kosmidis et al., 2004; Lanting et al., 2009; Troyer et al., 1997) switches than younger individuals. However, one study also found no association of age with cluster size (Troyer et al., 1997). With regard to sex/gender differences, most studies did not find clear evidence for an association for metrics of word order (Brucki & Rocha, 2004; Kosmidis et al., 2004; Sokołowski et al., 2020; Weiss et al., 2006). For the novel psycholinguistic metrics, no studies have been performed assessing the association of sex/gender

and these metrics yet. Evidence on the effect of level of education on clusters and switches points toward a beneficial effect of having obtained a higher educational level on the number of switches, but results for clustering measures remain inconclusive (Ardila, 2020; Brucki & Rocha, 2004; Kosmidis et al., 2004).

The sample composition of studies on this topic varies widely; sources of variation stem from different recruitment strategies as well as geographical and cultural differences in the population pool where the participants are sampled from. Other sources of variation could be differences in test instructions or differing categorization schemes and rules for defining clusters and cluster/category switches. In addition, differences might also stem from the operationalization of the age variable in prior studies. For example, the study by Troyer et al. (1997) that did not find an association of age with cluster size had a sample size of 95 individuals and dichotomized age with a large age gap between the groups ("young": 18–35 years; "old": 60–89 years), whereas the study by Kosmidis et al. (2004) that found an association of cluster size with age had a sample size of 300 individuals and used age as a continuous variable.

In order to achieve maximum sensitivity and specificity to cognitive impairment, scoring and interpretation of semantic fluency test performance should adapt to the population in which it is being used (Manly et al., 1998; Stern et al., 1992). Therefore, investigating associations between sociodemographic factors and semantic fluency metrics across different cohorts could aid in establishing the transportability and generalizability of study results. The aims of this study were threefold: (I) to identify the association of three sociodemographic factors (i.e., age, education, and sex/gender) with semantic fluency test performance based on the total number of correct words, (II) to identify the association of sociodemographic factors with alternative item-level metrics, and (III) to descriptively compare findings of the aforementioned analyses across three cohorts with different recruitment strategies and sample compositions. The investigation of the associations of age, education, and sex/gender with the total number of correct words (Aim I) is crucial as it provides the foundational understanding and a baseline which is necessary for interpreting Aim II and Aim III. This baseline serves as a reference point, allowing us to discern how variations in age, education, and sex/gender influence semantic fluency outcomes. The associations identified in Aim I provide the necessary context for understanding why certain item-level semantic fluency metrics may vary across different sociodemographic groups and how these variations manifest in the three cohorts.

Method

Study Population

Participants were derived from three different cohort studies. The Second Manifestations of ARTerial disease-Magnetic Resonance (SMART-MR) study is a clinic-based prospective cohort study from the University Medical Center Utrecht (Utrecht, the Netherlands) aimed at investigating brain changes on Magnetic Resonance Imaging in nondemented patients with symptomatic atherosclerotic disease (Geerlings et al., 2009). The Washington Heights–Inwood Columbia Aging Project (WHICAP) is a community-based longitudinal study from New York (United States) to investigate Alzheimer's disease and other types of dementia associated with Parkinson's disease and stroke (Stern et al., 1992). The African

American Alzheimer's Disease Genetics Study (AAG), a volunteer-based cohort, was established by multisite cooperation from Columbia University, North Carolina A&T State University, Vanderbilt University, and the University of Miami (United States; Hamilton et al., 2014). This study included AAG participants recruited at Columbia University only, as their semantic fluency performance was recorded and entered at the item level (Vonk, Flores, et al., 2019). Cross-sectional data were used from the first measurement time point with available item-level semantic fluency data. For the analyses, samples were restricted to nondemented participants aged 40 and above with valid item-level fluency and sociodemographic data. Participants with >20% of the total words being unidentifiable were excluded. This resulted in sample sizes of 711 (SMART-MR), 672 (WHICAP), and 1,056 (AAG) for the statistical analyses (see Figures 1–3).

The SMART-MR study was approved by the hospitals' ethics committee, and written informed consent was obtained from all participants. The WHICAP study and the AAG study were approved by the Institutional Review Board, and each participant provided written informed consent.

Semantic Fluency Metrics

The semantic fluency task was administered as part of a neuropsychological battery in each cohort. In all three cohorts, participant responses on the semantic fluency task were written down on paper and not audio recorded.

Total Word Count

Participants were given 1 (WHICAP, AAG) or 2 min (SMART-MR) to name as many animals as possible. Both

Figure 1
Flowchart SMART-MR Study Population

Note. SMART-MR = Second Manifestations of ARTerial disease-Magnetic Resonance.

supraordinates and subspecies (e.g., dog, poodle) were counted as acceptable responses. Perseverations and words in different languages were excluded from the total score. The same animal named in both the masculine and feminine form was counted as a perseveration. If participants recalled animals so fast that interviewers could not keep up with writing them down, answers were recorded as (+) and counted as valid responses.

In addition to the total word count, the following six item-level semantic fluency metrics were derived. Item-level metrics were prepared for the statistical analyses following the rules in Vonk et al. (2023) and Vonk, Flores, et al. (2019).

Cluster Size and Switches

Clusters refer to words belonging to the same semantic (or phonemic) subcategory. *Switches* occur when participants switch from one subcategory to another when the preceding subcategory is exhausted which involves cognitive search processes (Troyer et al., 1997). Animals were assigned to the following 35 subcategories based on an adapted scheme from Hills et al. (2012), Troyer et al. (1997), and Zemla et al. (2020): African, arachnid, Arctic, Asian, Australian, beasts of burden, birds, bovine, canine, deer, dinosaur, European, farm, feline, fish, forest, genus, insectivores, insects, jungle, marsupial, mythical creatures, North American, pets, primates, rabbits, reptile/amphibian, rodents, South American, unicellular organisms, used for fur, water, weasels, worms, and zoo. Animals could be assigned to more than one subcategory. Mean cluster size and the number of fluid cluster switches were derived using the Semantic Network and Fluency Utility tool (Zemla et al., 2020). To be considered a fluid cluster switch, the next word in a list does not share a category label with the previous word (Zemla et al., 2020).

Lexical Frequency and Zipf Frequency

Lexical frequency values were derived from SUBTLEX-NL, a database of Dutch word frequencies based on film subtitles (Keuleers et al., 2010), and the American equivalent SUBTLEX-US (Brysbaert & New, 2009). A lexical frequency value of 0.301 (corresponding to a frequency of 0.5 in 51 million words) was imputed for words without available frequencies in the database (Kuperman et al., 2012). As the lexical frequency metric is dependent on corpus size, a standardized measure (Zipf scale) was additionally investigated and compared to the SUBTLEX lexical frequency (Brysbaert et al., 2018; van Heuven et al., 2014). The interpretation of Zipf values is as follows: A value of 1–3 indicates a low-frequency word (less than 1 per million words), and a value of 4–7 is a high-frequency word ($\geq$10 per million words; Brysbaert et al., 2016). For reasons of comparability across corpora in different languages, the lowest possible Zipf value (=1; corresponding to a frequency of 1 per 100 million words) was entered for words without available frequencies in the database. Based on the calculation suggested by van Heuven et al. (2014), those words would have a Zipf value of 1.356 (corresponding to a frequency of 0 per million words). For SMART-MR, 3.6% (lexical frequency) and 5.7% (Zipf frequency) of all correct entries required imputation. For WHICAP, 2.4% (lexical frequency) and 2.5% (Zipf frequency) of all correct entries required imputation. For AAG, 0.1% (lexical frequency) and 0.1% (Zipf frequency) of all correct entries required imputation.
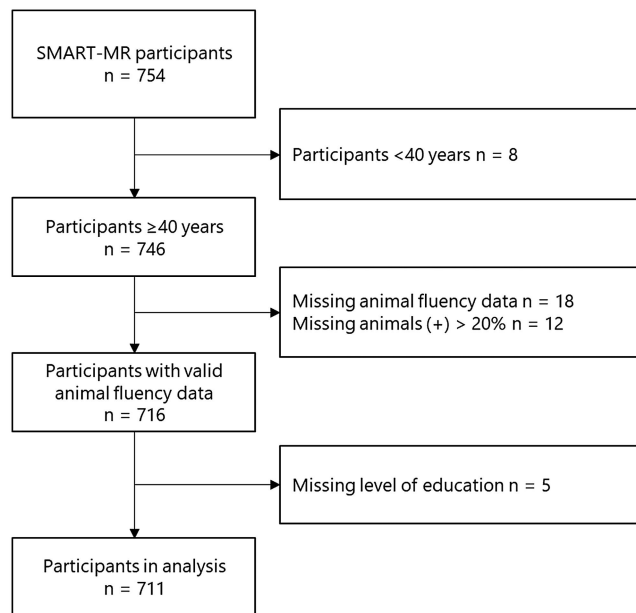
Figure 2

Flowchart WHICAP Study Population
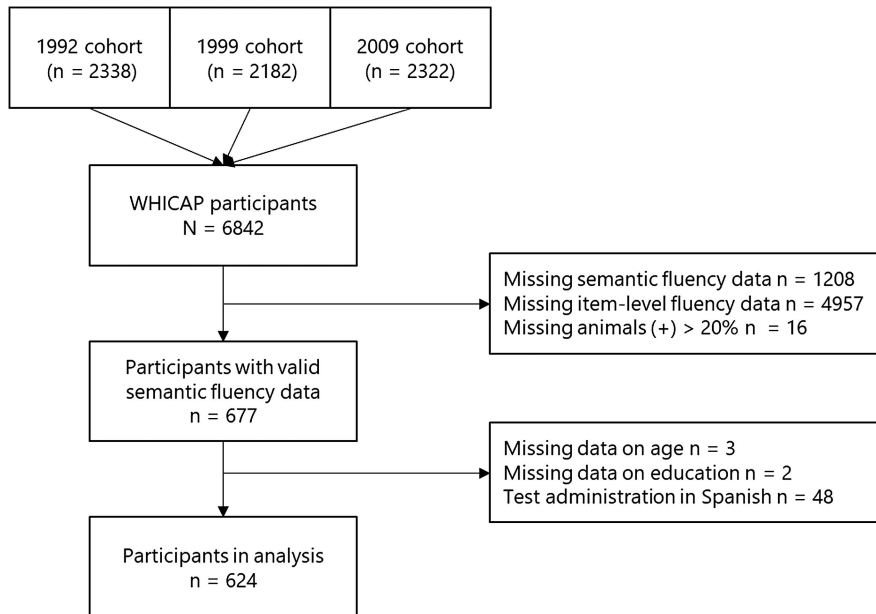
Note. WHICAP = Washington Heights–Inwood Columbia Aging Project.

Age of Acquisition

Age of acquisition was derived from the age of acquisition norms for 30,000 Dutch words (Brysbaert et al., 2014) and 50,000 English words (Brysbaert & Biemiller, 2017). An age of acquisition value of 12 was chosen for words without available values in the database due to the steep increase in vocabulary growth between age 5 and 12, followed by a levelling off (Beitchman et al., 2008). Of all correct entries, 4.5% (SMART-MR), 2.8% (WHICAP), and 0.2% (AAG) required imputation.

Lexical Decision Response Time

Lexical decision response times were obtained from the Dutch Lexicon Project 2 (Brysbaert et al., 2016) and the English Lexicon Project (Balota et al., 2007). The variable reflects the time in milliseconds to decide if a word is real or not (Keuleers et al., 2010). For words without response time values, the mean database values of 578.11 (Dutch words) and 784.1 (English words) were chosen for imputation. Of all correct entries, 9.1%

Figure 3

AAG Study Population
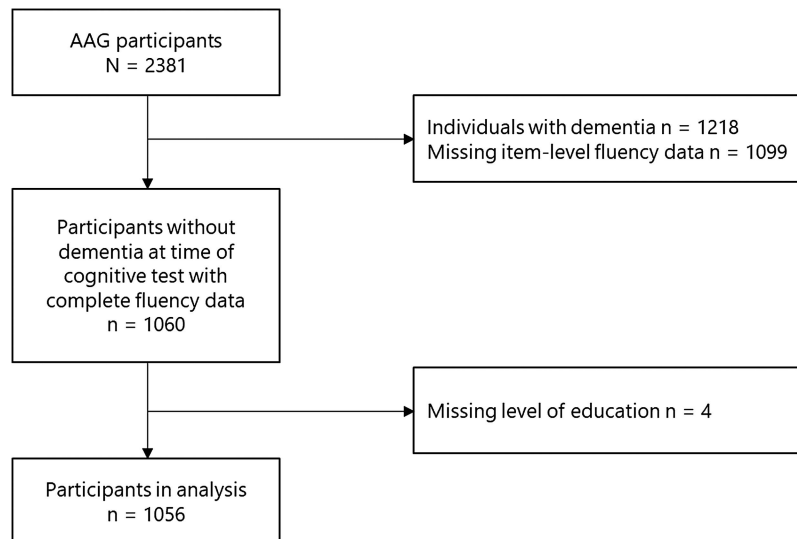
Note. AAG = African American Alzheimer's Disease Genetics Study.

(SMART-MR), 3.4% (WHICAP), and 1.1% (AAG) required imputation.

Lexical frequency and Zipf frequency values were reverse coded so that higher values indicated better test performance.

Main Determinants

The sociodemographic factors age, level of education, and sex/gender were considered as main determinants.

Age

Age was mean-centered in the analyses.

Education

In the Dutch cohort (SMART-MR), level of education has been assessed based on eight levels corresponding to the Dutch school system, ranging from no primary school to an academic degree: (a) no/only primary education, (b) lower vocational education, (c) secondary education, (d) (preparatory) secondary vocational education (=voorbereidend middelbaar beroepsonderwijs), (e) higher general secondary education (=hoger algemeen voortgezet onderwijs), (f) preparatory scientific education/gymnasium (=voorbereidend wetenschappelijk onderwijs), (g) higher vocational education, and (h) university education. In WHICAP and AAG, education ranged from 0 to 20 years.

The American education system includes elementary, middle, and high school phases before higher education, with a greater emphasis on a uniform curriculum for all students up to high school. The Dutch education system has a more structured secondary education phase with distinct tracks (voorbereidend middelbaar beroepsonderwijs, hoger algemeen voortgezet onderwijs, voorbereidend wetenschappelijk onderwijs), and the track a student continues on is based on the results of an aptitude test, their teacher's recommendation, and the opinion of the student and the student's parents. These tracks also differ in terms of their duration and curriculum. The voorbereidend middelbaar beroepsonderwijs track, for example, prepares students for a career in vocational or technical fields. voorbereidend wetenschappelijk onderwijs is the most academically challenging track, with a curriculum that focuses on theoretical subjects and prepares students for university education. It is therefore difficult to map Dutch levels of education to years of education. Consequently, we decided to categorize years of education in the American cohorts into low (<12 years) and high (≥12 years). The Dutch educational levels have been aggregated into low (Levels 1–6) and high (Levels 7 and 8).

Sex/Gender

Across all three cohorts, sex/gender was based on whether participants identified as male or female. As it was unknown whether an individual reported their biological sex or their gender identification, we refer to this variable as "sex/gender."

Clinical Characteristics

Stroke

In SMART-MR, stroke was defined as participants with a clinical history of brain ischemia at the first follow-up moment based on composite scoring made of self-reported previous ischemic stroke, previous history of carotid artery operation, or a physician diagnosis

at study inclusion of one among the following conditions: transient ischemic attack, brain infarct, ischemic stroke, cerebral ischemia, amaurosis fugax, or retinal infarct. Participants received a questionnaire every 6 months via post regarding hospitalization and outpatient clinic visits to establish the recurrence of new cardiovascular events including strokes until March 1, 2018. When a cardiovascular event was reported, the participant' documents were retrieved from the hospital archives and independently assessed by an endpoint committee in order to determine the nature of the event (Jaarsma-Coes et al., 2020). In WHICAP (Luchsinger et al., 2005) and AAG, history of a clinical stroke was ascertained by self-report from the participant or relatives, supplemented by a neurological examination or review of medical records.

Subjective Cognitive Complaints or Decline

In SMART-MR, subjective cognitive decline is based on a self-report questionnaire in which participants have rated their memory or concentration to be much worse compared to 5–10 years ago or compared to other people that are the same age. In AAG, subjective cognitive complaints were based on a consensus conference. Participants with cognitive complaints who did not meet the mild cognitive impairment (MCI) criteria were assigned the diagnosis "cognitive impairment not MCI" (Meier et al., 2012).

Mild Cognitive Impairment

In SMART-MR, MCI is based on the Petersen criteria (Petersen et al., 1999). It includes individuals with at least one cognitive domain z score below 1.5 SDs of the norm without self-reported memory problems and without impairment in (independent) activities of daily living. In WHICAP (Manly et al., 2008) and AAG (Meier et al., 2012), MCI was determined in a consensus conference based on the Petersen criteria (Petersen et al., 1999).

Global Cognitive Functioning

In SMART-MR and AAG, global cognitive functioning was assessed by use of the Mini-Mental State Examination. It consists of 11 items that generate scores between 0 and 30 and is used as a screening tool for cognitive impairment (Folstein et al., 1975).

Statistical Analysis

Data cleaning and descriptive analyses were performed in SPSS Version 26.0 (IBM Corp, 2019). Descriptive analyses and data visualization were performed in R Version 4.0.3 (R Core Team, 2020) using the cowplot, extrafont, flextable, forestplot, ggplot2, grid, gtsummary, haven, jtools, modelsummary, MplusAutomation, plyr, table1, and tidyverse packages (see Supplemental Material for further details). Mean differences in semantic fluency metrics across cohorts were compared using t tests for pairwise comparison. These comparisons are uncorrected for age, education, sex/gender, or other cohort characteristics. General linear models were run in Mplus Version 8 (Muthén & Muthén, 1998/2010). Analyses were conducted in each cohort separately for each determinant to model the crude association between sociodemographic variables and semantic fluency metrics using robust maximum likelihood estimation (Becker & Wu, 2007).

Since there is significant variability in the number of words generated by individuals during the semantic fluency task, a first set of sensitivity analyses was conducted by calculating item-level metrics as the mean value of each person's 10 lowest lexical/Zipf frequency, latest age of acquisition, and longest lexical decision response time. Participants were not eliminated from the analysis if they produced less than 10 valid responses. Additionally, a second set of sensitivity analyses was run with all sociodemographic determinants in one model to identify the effect sizes of each variable after adjusting for the others. In all analyses, the reverse-coded lexical/Zipf frequency values were used. Analyses were corrected for multiple comparison using the Benjamini–Hochberg procedure. A *p* value of <.05 was considered statistically significant.

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study, and we follow Journal Article Reporting Standards-Quant (Appelbaum et al., 2018). The code behind this analysis/simulation has been made publicly available at GitHub and can be accessed at "https://github.com/jmjvonk." This study's design and its analysis were not preregistered.

SMART-MR data are available on reasonable request (https://www.umcutrecht.nl/en/ucc-smart). Please send an email to Utrecht Cardiovascular Cohorts data request (uccdatarequest@umcutrecht.nl). After registration, the administrator will send an invite which grants access to the data request module. The data are not publicly available due to privacy or ethical restrictions.

For WHICAP, data are available on reasonable request to the WHICAP Publications Committee. Data requests should be submitted at https://cumc.co1.qualtrics.com/jfe/form/SV_6x5rRy14B6vpoqN.

Results

Data from 2,391 individuals were included in this study. Participants were, on average, the oldest in WHICAP (76 ± 7 years) compared to SMART-MR (62 ± 9 years) and AAG (70 ± 8 years; see Table 1). The majority of the study population was female in WHICAP (64%) and AAG (79%), whereas only 18% of SMART-MR participants were female. The statistical comparison

Table 1

Baseline Characteristics of the Study Population, Stratified by Cohort

Characteristic	SMART-MR (*n* = 711)	WHICAP (*n* = 624)	AAG (*n* = 1,056)
Sociodemographic characteristics			
Female sex/gender, *n* (%)	127 (17.9%)	399 (63.9%)	831 (78.7%)
Age, years	62.5 ± 9.1	75.5 ± 6.5	69.5 ± 7.6
Age, range in years	41–83	64–96	44–93
Race/ethnicity			
Non-Latinx Black	[a]	325 (52.1%)	1,056 (100%)
Non-Latinx White	[a]	248 (39.7%)	0 (0%)
Latinx	[a]	35 (5.6%)	0 (0%)
Other	[a]	16 (2.6%)	0 (0%)
Level of education, *n* (%)			
High school or lower	529 (74.4%)	247 (39.6%)	443 (42.0%)
College/university	182 (25.6%)	377 (60.4%)	613 (58.0%)
Employment status/occupation, *n* (%)			
Employed	243 (34.2%)	583 (95.3%)	n.a.
Unemployed	54 (7.6%)	9 (1.5%)	n.a.
Retired	297 (41.8%)	n.a.[b]	n.a.
Other	117 (16.5%)	20 (3.3%)	n.a.
Clinical characteristics			
Prior stroke, *n* (%)	185 (26.0%)	20 (3.2%)	33 (13.6%)[c]
Subjective cognitive complaints/decline, *n* (%)	61 (8.6%)	n.a.	45 (4.3%)
Mild cognitive impairment, *n* (%)	64 (9.0%)	109 (18.0%)	293 (27.8%)
MMSE total score (0–30)	28.5 ± 1.8	n.a.	28.5 ± 1.5[d]
Semantic fluency metrics			
Total score (correct)	29.9 ± 8.3[e]	15.6 ± 5.4[e]	15.5 ± 4.5[e]
Lexical frequency	2.2 ± 0.3	2.8 ± 0.3	2.9 ± 0.2
Lexical frequency, Ø for lowest 10 words	1.3 ± 0.5	2.5 ± 0.4	2.5 ± 0.3
Zipf frequency	3.6 ± 0.3	4.1 ± 0.3	4.2 ± 0.2
Zipf frequency, Ø for lowest 10 words	2.6 ± 0.6	3.7 ± 0.6	4.5 ± 0.2
Age of acquisition	6.6 ± 0.7	5.0 ± 0.7	4.8 ± 0.5
Age of acquisition, Ø for highest 10 words	8.9 ± 1.6	5.7 ± 1.2	5.5 ± 0.9
Lexical decision response time (milliseconds)	534.0 ± 11.5	638.2 ± 33.5	627.6 ± 23.0
Lexical decision response time, Ø for highest 10 words	578.9 ± 23.0	680.2 ± 59.3	665.6 ± 43.5
Mean cluster size	2.7 ± 1.2	1.9 ± 0.6	2.0 ± 0.7
Number of switches	10.9 ± 3.9	7.2 ± 2.9	8.2 ± 3.3

Note. Data are shown as mean ± standard deviations unless stated otherwise. SMART-MR = Second Manifestations of ARTerial disease-Magnetic Resonance; WHICAP = Washington Heights–Inwood Columbia Aging Project; AAG = African American Alzheimer's Disease Genetics Study; MMSE = Mini-Mental State Examination; n.a. = not applicable.
[a] Question asked: "Birth country of respondent." [b] Question asked: "Occupation during most of the career." [c] Data only available for *n* = 242. [d] Data only available for *n* = 133. [e] Test duration was 2 min in SMART-MR and 1 min in WHICAP and AAG.

of mean semantic fluency metric values across cohorts can be found in Supplemental Table S1. The correlation coefficients of age, education, sex/gender, and semantic fluency metrics can be found in Supplemental Table S2 and Figures S1–S3.

Age

Across cohorts, older higher age was associated with lower total scores and fewer cluster switches. Older age was also associated with shorter lexical decision response time and lower age of acquisition in WHICAP and AAG. A lower lexical frequency and Zipf frequency were associated with older age in WHICAP only. In the cohort with the oldest average age of participants (WHICAP), the strongest and most consistent associations of age with semantic fluency metrics were observed. The strongest association with age was found for the total score, followed by lexical frequency and lexical decision response time (Figure 4, Table 2). All associations survived multiple comparison correction.

Education

Across cohorts, a higher level of education was associated with higher total scores, more cluster switches, naming words with a lower lexical frequency, a higher age of acquisition, and higher lexical decision response times. The strength of the associations differed per cohort and semantic fluency metric. For example, SMART-MR and AAG showed similar associations of education with cluster size whereas SMART-MR and WHICAP showed similar associations for

lexical frequency. Overall, the strongest associations with education across metrics were observed in WHICAP (Figure 5, Table 2). All associations survived multiple comparison correction except for the association of level of education and Zipf frequency in SMART-MR.

Sex/Gender

For sex/gender, we found that being female was associated with lower total scores, smaller clusters, naming words with a lower age of acquisition, and a lower lexical decision response time across cohorts. In WHICAP and SMART-MR, being female was associated with naming words with a higher lexical/Zipf frequency. Being female was associated with the use of more switches, but this association was not statistically significant (Figure 6, Table 2). Across cohorts, the strongest association was found between being male and producing words with a higher average age of acquisition. All associations survived multiple comparison correction except for the association of sex/gender and cluster size in AAG.

Sensitivity Analyses

For the first sensitivity analysis, we calculated the mean of the psycholinguistic value of the 10 words with the lowest lexical/Zipf frequency, latest age of acquisition, or longest lexical decision response time instead of the mean across all words produced in the fluency task. Associations with age, education, and sex/gender remained the same for all item-level metrics in WHICAP and SMART-MR, and most metrics in AAG. The exception was that the

Figure 4

Association of Age (per 10 Years) With Semantic Fluency Metrics Across Cohorts
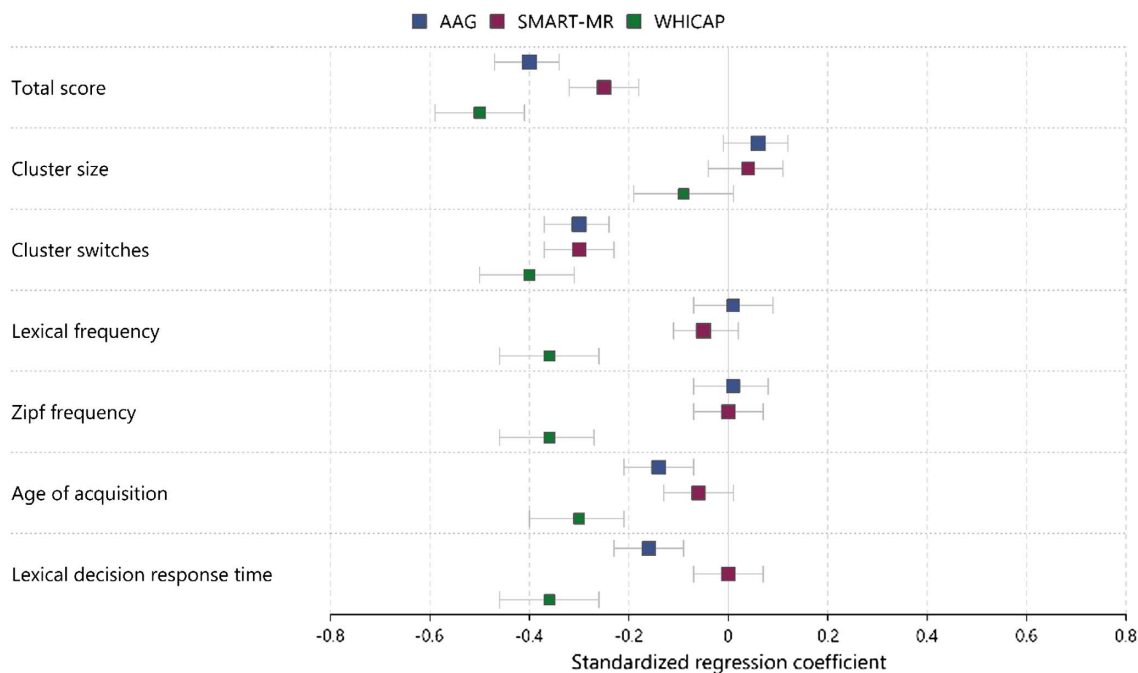
Note. AAG = African American Alzheimer's Disease Genetics Study; SMART-MR = Second Manifestations of ARTerial disease-Magnetic Resonance; WHICAP = Washington Heights–Inwood Columbia Aging Project. See the online article for the color version of this figure.

Table 2

Association of Sociodemographic Factors With Total Score and Item-Level Metrics of Semantic Fluency in SMART-MR, WHICAP, and AAG

	SMART-MR		WHICAP		AAG	
Variable	Estimate [95% CI]	*p*	Estimate [95% CI]	*p*	Estimate [95% CI]	*p*
Total score						
Age per 1 year increase	**−0.03 [−0.03, −0.02]**	**.000**	**−0.05 [−0.06, −0.04]**	**.000**	**−0.04 [−0.05, −0.03]**	**.000**
Level of education	**0.50 [0.36, 0.64]**	**.000**	**0.49 [0.36, 0.62]**	**.000**	**0.32 [0.22, 0.42]**	**.000**
Sex/gender	−0.15 [−0.29, 0.00]	.097	−0.14 [−0.28, −0.00]	.097	−0.15 [−0.28, −0.02]	.060
Cluster size						
Age per 1 year increase	0.00 [0.00, 0.01]	.431	−0.01 [−0.02, 0.00]	.132	0.01 [−0.00, 0.01]	.139
Level of education	0.02 [−0.10, 0.13]	.817	**0.18 [0.05, 0.31]**	**.021**	0.01 [−0.09, 0.11]	.899
Sex/gender	**−0.25 [−0.36, −0.14]**	**.000**	**−0.18 [−0.33, −0.04]**	**.035**	**−0.18 [−0.32, −0.04]**	**.038**
Cluster switches						
Age per 1 year increase	**−0.03 [−0.04, −0.02]**	**.000**	**−0.04 [−0.05, −0.03]**	**.000**	**−0.03 [−0.04, −0.02]**	**.000**
Level of education	**0.28 [0.13, 0.43]**	**.002**	**0.33 [0.19, 0.46]**	**.000**	**0.24 [0.14, 0.34]**	**.000**
Sex/gender	0.14 [−0.01, 0.30]	.127	0.01 [−0.12, 0.15]	.865	0.01 [−0.12, 0.13]	.938
Lexical frequency						
Age per 1 year increase	−0.01 [−0.01, 0.00]	.271	**−0.04 [−0.05, −0.03]**	**.000**	−0.00 [−0.01, 0.01]	.783
Level of education	**0.36 [0.25, 0.47]**	**.000**	**0.54 [0.41, 0.67]**	**.000**	**0.15 [0.05, 0.24]**	**.011**
Sex/gender	**−0.27 [−0.44, −0.09]**	**.012**	**−0.37 [−0.51, −0.23]**	**.000**	−0.09 [−0.23, 0.04]	.257
Zipf frequency						
Age per 1 year increase	0.00 [−0.01, 0.01]	.968	**−0.04 [−0.05, −0.03]**	**.000**	−0.00 [−0.01, 0.01]	.910
Level of education	**0.17 [0.03, 0.32]**	**.048**	**0.43 [0.29, 0.56]**	**.000**	−0.01 [−0.12, 0.09]	.825
Sex/gender	**−0.29 [−0.43, −0.14]**	**.001**	**−0.29 [−0.43, −0.15]**	**.001**	−0.11 [−0.23, 0.02]	.153
Age of acquisition						
Age per 1 year increase	−0.01 [−0.01, 0.00]	.137	**−0.03 [−0.04, −0.02]**	**.000**	**−0.01 [−0.02, −0.01]**	**.001**
Level of education	**0.34 [0.21, 0.48]**	**.000**	**0.54 [0.42, 0.67]**	**.000**	**0.25 [0.15, 0.35]**	**.000**
Sex/gender	**−0.48 [−0.61, −0.34]**	**.000**	**−0.42 [−0.56, −0.28]**	**.000**	**−0.23 [−0.36, −0.10]**	**.004**
Lexical decision response time						
Age per 1 year increase	0.00 [−0.01, 0.01]	.976	**−0.04 [−0.05, −0.03]**	**.000**	**−0.02 [−0.02, −0.01]**	**.000**
Level of education	**0.37 [0.25, 0.50]**	**.000**	**0.49 [0.36, 0.62]**	**.000**	**0.26 [0.16, 0.36]**	**.000**
Sex/gender	**−0.33 [−0.48, −0.18]**	**.000**	**−0.27 [−0.40, −0.13]**	**.001**	**−0.24 [−0.37, −0.12]**	**.001**

Note. Bolded numbers are significant at *p* < .05. Level of education: high school and lower versus college/university. Corrected *p* values for the results in SMART-MR are as follows (in order of occurrence in the table): .000, .000, .146, 0.503, .903, 0.000, .000, .004, .178, .335, .000, .021, .976, .078, .002, .180, .000, .000, .976, .000, .000. Corrected *p* values for the results in WHICAP are as follows (in order of occurrence in the table): .000, .000, .107, .139, .026, .041, .000, .000, .865, .000, .000, .000, .000, .000, .001, .000, .000, .000, .000, .000, .001. Corrected *p* values for the results in AAG are as follows (in order of occurrence in the table): .000, .000, .097, .209, .938, .067, .000, .000, .938, .938, .021, .337, .938, .938, .214, .002, .000, .008, .000, .000, .002. SMART-MR = Second Manifestations of ARTerial disease-Magnetic Resonance; WHICAP = Washington Heights–Inwood Columbia Aging Project; AAG = African American Alzheimer's Disease Genetics Study; Estimate = *Y*-standardized regression coefficient β; CI = confidence interval.

associations of all sociodemographic variables with lexical/Zipf frequency in AAG strengthened and became statistically significant. The association of age with age of acquisition also reached statistical significance in SMART-MR in this sensitivity analysis (see Supplemental Figures S4–S6).

For the second sensitivity analysis, we adjusted for the other two sociodemographic variables before testing the association of age, education, and sex/gender with semantic fluency metrics. The associations of sociodemographic variables with total score and item-level metrics remained the same after covariate adjustment. Solely in SMART-MR, the association of higher educational level with Zipf frequency was no longer statistically significant, whereas the association of being female and performing more cluster switches reached statistical significance (see Supplemental Figures S7–S9).

Discussion

In this study, we aimed to identify and compare associations of sociodemographic factors with total and item-level semantic fluency metrics across cohorts with different recruitment strategies and sample compositions. Differences across cohorts were compared descriptively. We found that the relationship of sociodemographic factors with semantic fluency metrics typically had the same direction of effect across cohorts, but the strength of association differed. Moreover, for each sociodemographic factor, the direction of effect was similar across the different semantic fluency metrics except for cluster size (age, education) or cluster switches (sex/gender).

Our results showed that older age was associated with a lower total score on the semantic fluency task, corresponding to previous research in which younger individuals produced more words on the semantic fluency task than older individuals (Kempler et al., 1998; Troyer, 2000; van Hooren et al., 2007). Older individuals also tended to produce fewer cluster switches than younger individuals in our analyses, similar to previous findings (Troyer et al., 1997). For cluster size, there was no evidence of an association with age. Previous literature on this topic was inconclusive (Kosmidis et al., 2004; Lanting et al., 2009; Troyer et al., 1997). This result might be explained by the preservation of vocabulary knowledge, part of crystallized intelligence, even as we age (Gordon et al., 2018). As such, the decrease in total number of words with older age may be caused by a decreased ability to execute cluster switches. With regard to psycholinguistic metrics, older individuals generally named words with a lower lexical decision response time and a lower age of acquisition, which are generally thought of as words

Figure 5
Association of Level of Education With Semantic Fluency Metrics Across Cohorts
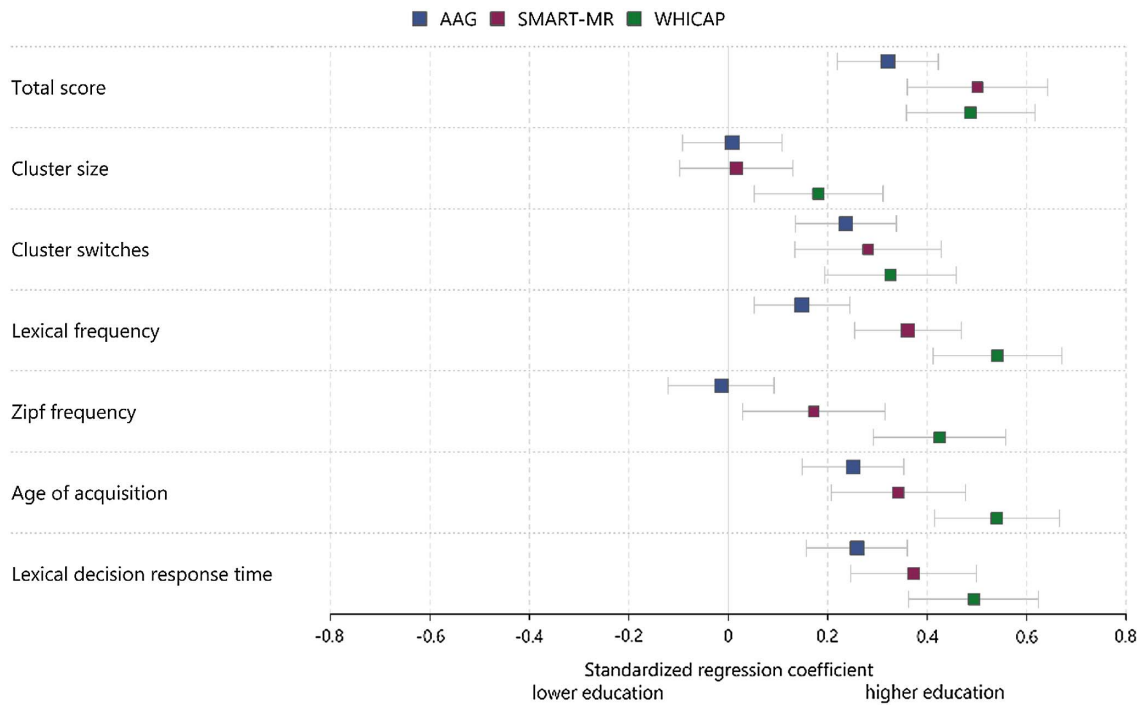
Note. AAG = African American Alzheimer's Disease Genetics Study; SMART-MR = Second Manifestations of ARTerial disease-Magnetic Resonance; WHICAP = Washington Heights–Inwood Columbia Aging Project. See the online article for the color version of this figure.

that are retrieved more easily (Barry et al., 1997; Ellis & Morrison, 1998). Of note, the participants included in this study ranged from 41 to 96 years. Age ranges from prior studies differed with some studies also focusing on middle-aged and older adults (Kempler et al., 1998; van Hooren et al., 2007). Other studies used a lifetime approach with extended age ranges to also include younger adults (Brickman et al., 2005; Kosmidis et al., 2004; Zarino et al., 2014), comparing younger to older adults and leaving out middle-aged adults (Troyer et al., 1997) or solely included younger adults (Lanting et al., 2009). As such, our results might be difficult to compare to study populations with different age ranges and cannot be translated and generalized to younger adults.

Across cohorts, a higher level of education was associated with better performance on the semantic fluency test as measured by both total score and item-level metrics. Interestingly, we observed a trend across cohorts with AAG showing the weakest and WHICAP showing the strongest association with semantic fluency metrics. This pattern might be explained by a higher average of educational attainment in AAG and SMART-MR than WHICAP. We had to quantify educational attainment as categorical and could not investigate education as a continuous variable due to different educational systems in the Netherlands (SMART-MR) versus the United States (AAG, WHICAP). We did not observe an association of cluster size with level of education in AAG and SMART. Previous studies found conflicting results, including evidence (Zhao et al., 2013) and no evidence (Kosmidis et al., 2004) for an

association between cluster size and education. For example, Zhao et al. (2013) observed a larger cluster size in individuals with a higher level of education. We did find an association of cluster size with education in WHICAP, which is the cohort with the widest distribution of educational attainment among participants.

We found that compared to women, men obtained higher scores on the semantic fluency test as measured by total score (except for SMART-MR) and item-level metrics, except for cluster switches and lexical/Zipf frequency in AAG, across all cohorts. In the past, conflicting results have been reported for sex/gender differences in semantic fluency task performance (Gawda & Szepietowska, 2013; Jebahi et al., 2022; Sokołowski et al., 2020; Weiss et al., 2006). A recent meta-analysis reported that sex/gender differences in semantic fluency task performance seem to differ based on the category used (Hirnstein et al., 2022). In line with our results, Hirnstein et al. (2022) reported a male advantage in the category "animals" with male participants naming more animals than female participants. Our results remain similar when adjusting the analysis of one sociodemographic factor for the other two sociodemographic factors. Due to its descriptive nature, our study did not investigate other determinants aside from age and education that might further explain these sex/gender differences.

The semantic fluency test was designed to be scored based on the total number of words generated. However, development and use of item-level metrics of semantic fluency in aging and dementia research have flourished in recent years, as it has been shown that

Figure 6
Association of Sex/Gender With Semantic Fluency Metrics Across Cohorts
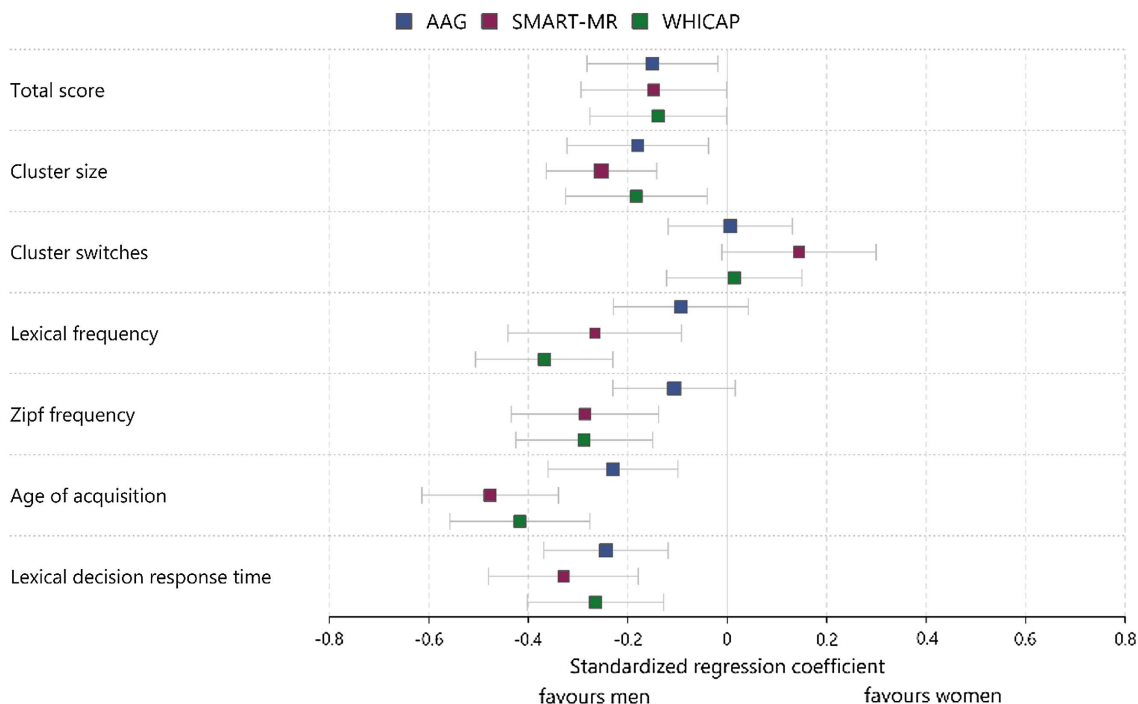
Note. AAG = African American Alzheimer's Disease Genetics Study; SMART-MR = Second Manifestations of ARTerial disease-Magnetic Resonance; WHICAP = Washington Heights–Inwood Columbia Aging Project. See the online article for the color version of this figure.

these metrics could play a valuable role in dementia care for early diagnosis (e.g., Vonk et al., 2023) and differential diagnosis (e.g., van den Berg et al., 2022). Our recent work showed that among several different item-level metrics, lexical frequency and age of acquisition were particularly robust in their association with future memory decline in older adults without dementia, even beyond existing cognitive tests and adjusted for the total score, and that these relationships were not moderated by education or race (Vonk et al., 2023). The current article further investigated the associations of these novel metrics with sociodemographic factors to establish the generalizability of these metrics, given the international interest to apply item-level metrics of semantic fluency (e.g., Rofes et al., 2020; Saranpää et al., 2022; Taler et al., 2020; van den Berg et al., 2022). Future studies may explore the item-level metric intercorrelations and differential informativeness in understanding and delineating the association between sociodemographic factors and semantic fluency performance.

Strengths of this study encompass the inclusion of three different large cohorts with different recruitment strategies, different sample compositions, and assessment in different languages, increasing the generalizability of our study results. The consistency in results between cohorts with different languages suggests that animal fluency tasks are comparable cross-linguistically, as previously suggested (Ardila, 2020). Another strength includes that we derived item-level metrics and conducted data analyses in the same manner across cohorts. Thereby, heterogeneity of study results due to different analytic approaches was minimized. A limitation of the

present study is that the data sets could not be pooled and analyzed in a unified or meta-analytic way due to harmonization challenges and differences in recruitment strategies across cohorts. For example, education was assessed and coded differently in the different cohorts; the U.S.-based cohorts captured years of education, whereas the Dutch-based cohort assessed the highest completed school degree or credential, hindering comparisons across cohorts. As a result, we had to dichotomize education in each cohort, which might have led to a loss of information. Further, the allotted time to generate responses differed across cohorts (2 min in SMART-MR vs. 1 min in WHICAP and AAG). The unadjusted means of the semantic fluency metrics differed between cohorts. Of note, differences were also found between WHICAP and AAG, both of which used a 1-min time span. Due to the many differences found across the three cohorts, it is not possible to determine the influence of this time difference on semantic fluency metrics. Future studies could experimentally investigate the effect of task time on semantic fluency metrics.

In conclusion, we showed that item-level semantic fluency metrics—similar to the standard total score—are sensitive to the effects of sociodemographic factors. The effects of sociodemographic factors on semantic fluency metrics varied in strength across cohorts but were in a similar direction. In addition to the similar direction of effects for different recruitment strategies across the cohorts, our results are also consistent across English and Dutch. As such, despite the geographic, cultural, and linguistic differences across the three cohorts, our results suggest generalizability of the effects of age, education, and sex/gender on semantic fluency test performance.

References

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Arce Rentería, M., Vonk, J. M. J., Felix, G., Avila, J. F., Zahodne, L. B., Dalchand, E., Frazer, K. M., Martinez, M. N., Shouel, H. L., & Manly, J. J. (2019). Illiteracy, dementia risk, and cognitive trajectories among older adults with low education. *Neurology*, *93*(24), e2247–e2256. https://doi.org/10.1212/WNL.0000000000008587

Ardila, A. (2020). A cross-linguistic comparison of category verbal fluency test (ANIMALS): A systematic review. *Archives of Clinical Neuropsychology*, *35*(2), 213–225. https://doi.org/10.1093/arclin/acz060

Avila, J. F., Vonk, J. M. J., Verney, S. P., Witkiewitz, K., Arce Rentería, M., Schupf, N., Mayeux, R., & Manly, J. J. (2019). Sex/gender differences in cognitive trajectories vary as a function of race/ethnicity. *Alzheimer's & Dementia*, *15*(12), 1516–1523. https://doi.org/10.1016/j.jalz.2019.04.006

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Barry, C., Morrison, C. M., & Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency, and name agreement. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *50*(3), 560–585. https://doi.org/10.1080/783663595

Becker, B. J., & Wu, M. J. (2007). The synthesis of regression slopes in meta-analysis. *Statistical Science*, *22*(3), 414–429. https://doi.org/10.1214/07-STS243

Beitchman, J. H., Jiang, H., Koyama, E., Johnson, C. J., Escobar, M., Atkinson, L., Brownlie, E. B., & Vida, R. (2008). Models and determinants of vocabulary growth from kindergarten to adulthood. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *49*(6), 626–634. https://doi.org/10.1111/j.1469-7610.2008.01878.x

Brickman, A. M., Paul, R. H., Cohen, R. A., Williams, L. M., MacGregor, K. L., Jefferson, A. L., Tate, D. F., Gunstad, J., & Gordon, E. (2005). Category and letter verbal fluency across the adult lifespan: Relationship to EEG theta power. *Archives of Clinical Neuropsychology*, *20*(5), 561–573. https://doi.org/10.1016/j.acn.2004.12.006

Brucki, S. M. D., & Rocha, M. S. G. (2004). Category fluency test: Effects of age, gender and education on total scores, clustering and switching in Brazilian Portuguese-speaking subjects. *Brazilian Journal of Medical and Biological Research*, *37*(12), 1771–1777. https://doi.org/10.1590/S0100-879X2004001200002

Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods*, *49*(4), 1520–1523. https://doi.org/10.3758/s13428-016-0811-4

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*, *150*, 80–84. https://doi.org/10.1016/j.actpsy.2014.04.010

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 441–458. https://doi.org/10.1037/xhp0000159

Catling, J., South, F., & Dent, K. (2013). The effect of age of acquisition on older individuals with and without cognitive impairments. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *66*(10), 1963–1973. https://doi.org/10.1080/17470218.2013.771689

De Marco, M., Blackburn, D. J., & Venneri, A. (2021). Serial recall order and semantic features of category fluency words to study semantic memory in normal ageing. *Frontiers in Aging Neuroscience*, *13*, Article 678588. https://doi.org/10.3389/fnagi.2021.678588

Ellis, A. W., & Morrison, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(2), 515–523. https://doi.org/10.1037/0278-7393.24.2.515

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

Gawda, B., & Szepietowska, E. M. (2013). Semantic and affective verbal fluency: Sex differences. *Psychological Reports*, *113*(1), 1258–1268. https://doi.org/10.2466/28.21.PR0.113x17z3

Geerlings, M. I., Appelman, A. P. A., Vincken, K. L., Mali, W. P. T. M., van der Graaf, Y., & the SMART Study Group. (2009). Association of white matter lesions and lacunar infarcts with executive functioning: The SMART-MR study. *American Journal of Epidemiology*, *170*(9), 1147–1155. https://doi.org/10.1093/aje/kwp256

Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica*, *115*(1), 43–67. https://doi.org/10.1016/j.actpsy.2003.11.002

Gordon, J. K., Young, M., & Garcia, C. (2018). Why do older adults have difficulty with semantic fluency? *Aging, Neuropsychology and Cognition*, *25*(6), 803–828. https://doi.org/10.1080/13825585.2017.1374328

Hamilton, J. L., Brickman, A. M., Lang, R., Byrd, G. S., Haines, J. L., Pericak-Vance, M. A., & Manly, J. J. (2014). Relationship between depressive symptoms and cognition in older, non-demented African Americans. *Journal of the International Neuropsychological Society*, *20*(7), 756–763. https://doi.org/10.1017/S1355617714000423

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

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

Hirnstein, M., Stuebs, J., Moè, A., & Hausmann, M. (2022). Sex/gender differences in verbal fluency and verbal-episodic memory: A meta-analysis. *Perspectives on Psychological Science*, *18*(1), 67–90. https://doi.org/10.1177/17456916221082116

IBM Corp. (2019). *IBM SPSS statistics for windows* (26.0).

Jaarsma-Coes, M. G., Ghaznawi, R., Hendrikse, J., Slump, C., Witkamp, T. D., van der Graaf, Y., Geerlings, M. I., de Bresser, J., & the Second Manifestations of ARTerial Disease Study Group. (2020). MRI phenotypes of the brain are related to future stroke and mortality in patients with manifest arterial disease: The SMART-MR study. *Journal of Cerebral Blood Flow and Metabolism*, *40*(2), 354–364. https://doi.org/10.1177/0271678X18818918

Jebahi, F., Abou Jaoude, R., & Ellis, C. (2022). Semantic verbal fluency task: The effects of age, educational level, and sex in Lebanese-speaking adults. *Applied Neuropsychology: Adult*, *29*(5), 936–940. https://doi.org/10.1080/23279095.2020.1821031

Kempler, D., Teng, E. L., Dick, M., Taussig, I. M., & Davis, D. S. (1998). The effects of age, education, and ethnicity on verbal fluency. *Journal of*

the *International Neuropsychological Society*, *4*(6), 531–538. https://doi.org/10.1017/S1355617798466013

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

Kosmidis, M. H., Vlahou, C. H., Panagiotaki, P., & Kiosseoglou, G. (2004). The verbal fluency task in the Greek population: Normative data, and clustering and switching strategies. *Journal of the International Neuropsychological Society*, *10*(2), 164–172. https://doi.org/10.1017/S1355617704102014

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

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

Luchsinger, J. A., Reitz, C., Honig, L. S., Tang, M. X., Shea, S., & Mayeux, R. (2005). Aggregation of vascular risk factors and risk of incident Alzheimer disease. *Neurology*, *65*(4), 545–551. https://doi.org/10.1212/01.wnl.0000172914.08967.dc

Manly, J. J., Jacobs, D. M., Sano, M., Bell, K., Merchant, C. A., Small, S. A., & Stern, Y. (1998). Cognitive test performance among nondemented elderly African Americans and whites. *Neurology*, *50*(5), 1238–1245. https://doi.org/10.1212/WNL.50.5.1238

Manly, J. J., Tang, M. X., Schupf, N., Stern, Y., Vonsattel, J. P. G., & Mayeux, R. (2008). Frequency and course of mild cognitive impairment in a multiethnic community. *Annals of Neurology*, *63*(4), 494–506. https://doi.org/10.1002/ana.21326

Meier, I. B., Manly, J. J., Provenzano, F. A., Louie, K. S., Wasserman, B. T., Griffith, E. Y., Hector, J. T., Allocco, E., & Brickman, A. M. (2012). White matter predictors of cognitive functioning in older adults. *Journal of the International Neuropsychological Society*, *18*(3), 414–427. https://doi.org/10.1017/S1355617712000227

Mougias, A., Christidi, F., Synetou, M., Kotrotsou, I., Valkimadi, P., & Politis, A. (2019). Differential effect of demographics, processing speed, and depression on cognitive function in 755 non-demented community-dwelling elderly individuals. *Cognitive and Behavioral Neurology*, *32*(4), 236–246. https://doi.org/10.1097/WNN.0000000000000211

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). (Original work published 1998)

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology*, *56*(3), 303–308. https://doi.org/10.1001/archneur.56.3.303

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rofes, A., Beran, M., Jonkers, R., Geerlings, M. I., & Vonk, J. M. J. (2023). What drives task performance in animal fluency in individuals without dementia? The SMART-MR study. *Journal of Speech, Language, and Hearing Research*, *66*(9), 3473–3485. https://doi.org/10.1044/2023_JSLHR-22-00445

Rofes, A., de Aguiar, V., Jonkers, R., Oh, S. J., DeDe, G., & Sung, J. E. (2020). What drives task performance during animal fluency in people with Alzheimer's disease? *Frontiers in Psychology*, *11*, Article 1485. https://doi.org/10.3389/fpsyg.2020.01485

Saranpää, A. M., Kivisaari, S. L., Salmelin, R., & Krumm, S. (2022). Moving in semantic space in prodromal and very early Alzheimer's disease: An item-level characterization of the semantic fluency task. *Frontiers in Psychology*, *13*, Article 777656. https://doi.org/10.3389/fpsyg.2022.777656

Sokołowski, A., Tyburski, E., Sołtys, A., & Karabanowicz, E. (2020). Sex differences in verbal fluency among young adults. *Advances in Cognitive Psychology*, *16*(2), 92–102. https://doi.org/10.5709/acp-0288-1

Stern, Y., Andrews, H., Pittman, J., Sano, M., Tatemichi, T., Lantigua, R., & Mayeux, R. (1992). Diagnosis of dementia in a heterogeneous population. Development of a neuropsychological paradigm-based diagnosis of dementia and quantified correction for the effects of education. *Archives of Neurology*, *49*(5), 453–460. https://doi.org/10.1001/archneur.1992.00530290035009

Taler, V., Johns, B. T., & Jones, M. N. (2020). A large-scale semantic analysis of verbal fluency across the aging spectrum: Data from the canadian longitudinal study on aging. *The Journals of Gerontology: Series B*, *75*(9), e221–e230. https://doi.org/10.1093/geronb/gbz003

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

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

van den Berg, E., Dijkzeul, J. C. M., Poos, J. M., Eikelboom, W. S., van Hemmen, J., Franzen, S., de Jong, F. J., Dopper, E. G. P., Vonk, J. M. J., Papma, J. M., Satoer, D., Jiskoot, L. C., & Seelaar, H. (2022). Differential linguistic features of verbal fluency in behavioral variant frontotemporal dementia and primary progressive aphasia. *Applied Neuropsychology: Adult*. Advance online publication. https://doi.org/10.1080/23279095.2022.2060748

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

van Hooren, S. A. H., Valentijn, A. M., Bosma, H., Ponds, R. W. H. M., van Boxtel, M. P. J., & Jolles, J. (2007). Cognitive functioning in healthy older adults aged 64–81: A cohort study into the effects of age, sex, and education. *Aging, Neuropsychology and Cognition*, *14*(1), 40–54. https://doi.org/10.1080/138255890969483

Vonk, J. M. J., Arce Rentería, M., Avila, J. F., Schupf, N., Noble, J. M., Mayeux, R., Brickman, A. M., & Manly, J. (2019). Secular trends in cognitive trajectories of diverse older adults. *Alzheimer's & Dementia*, *15*(12), 1576–1587. https://doi.org/10.1016/j.jalz.2019.06.4944

Vonk, J. M. J., Flores, R. J., Rosado, D., Qian, C., Cabo, R., Habegger, J., Louie, K., Allocco, E., Brickman, A. M., & Manly, J. J. (2019). Semantic network function captured by word frequency in nondemented APOE ε4 carriers. *Neuropsychology*, *33*(2), 256–262. https://doi.org/10.1037/neu0000508

Vonk, J. M. J., Geerlings, M. I., Avila-Rieger, J. F., Qian, C. L., Schupf, N., Mayeux, R., Brickman, A. M., & Manly, J. J. (2023). Semantic item-level metrics relate to future memory decline beyond existing cognitive tests in older adults without dementia. *Psychology and Aging*, *38*(5), 443–454. https://doi.org/10.1037/pag0000747

Weiss, E. M., Ragland, J. D., Brensinger, C. M., Bilker, W. B., Deisenhammer, E. A., & Delazer, M. (2006). Sex differences in clustering and switching in verbal fluency tasks. *Journal of the International Neuropsychological Society*, *12*(4), 502–509. https://doi.org/10.1017/S1355617706060656

Woods, D. L., Wyma, J. M., Herron, T. J., & Yund, E. W. (2016). Computerized analysis of verbal fluency: Normative data and the effects of repeated testing, simulated malingering, and traumatic brain injury. *PLOS ONE*, *11*(12), Article e0166439. https://doi.org/10.1371/journal.pone.0166439

Zarino, B., Crespi, M., Launi, M., & Casarotti, A. (2014). A new standardization of semantic verbal fluency test. *Neurological Sciences*, *35*(9), 1405–1411. https://doi.org/10.1007/s10072-014-1729-1

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

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