



Systematic administration and analysis of verbal fluency tasks: Preliminary evidence for reliable exploration of processes underlying task performance

Nana Lehtinen, Ida Luotonen & Anna Kautto

To cite this article: Nana Lehtinen, Ida Luotonen & Anna Kautto (2021): Systematic administration and analysis of verbal fluency tasks: Preliminary evidence for reliable exploration of processes underlying task performance, Applied Neuropsychology: Adult, DOI: [10.1080/23279095.2021.1973471](https://doi.org/10.1080/23279095.2021.1973471)

To link to this article: <https://doi.org/10.1080/23279095.2021.1973471>



© 2021 The Author(s). Published with license by Taylor & Francis Group, LLC



[View supplementary material](#)



Published online: 20 Sep 2021.



[Submit your article to this journal](#)



Article views: 380



[View related articles](#)



[View Crossmark data](#)

Systematic administration and analysis of verbal fluency tasks: Preliminary evidence for reliable exploration of processes underlying task performance

Nana Lehtinen , Ida Luotonen , and Anna Kautto 

Department of Psychology and Speech-Language Pathology, University of Turku, Turku, Finland

ABSTRACT

Verbal fluency (VF) tasks are typically scored by the number of acceptable words generated within an allotted time (i.e., total score). However, total scores do not provide insight into verbal and executive processes underlying VF task performance. Further analyses have been implemented to increase the analytical power of VF tasks, but systematic scoring guidelines are needed. We generated instructions for administration, scoring, and analyses of total scores, errors, temporal parameters, clustering, and switching with strong inter-rater reliability. To investigate the reliability of the proposed analysis, we modeled the performance of Finnish-speaking older adults ($N = 50$) in phonemic (/k/, /a/, and /p/) and semantic (animals) categories. Our results are in line with previous studies: We observed a higher performance on semantic than phonemic fluency ($p \leq 0.001$, $d = 0.91$) and significant effects for education ($p \leq 0.001$, $d = 1.11$) and gender ($p \leq 0.001$, $d = -1.11$), but not for age ($p = 0.10$, $d = 0.48$). Most errors were repetitions. Performance declined over the allotted time frame as measured in 15-s segments (all $ps < 0.001$ with medium to large effect sizes). Task congruent clustering and switching were productive strategies (all $ps < 0.001$ with large effect sizes), and participants generated task discrepant clusters in both phonemic ($p = 0.004$, $d = 0.69$) and semantic tasks ($p = 0.66$, $d = 0.18$). The results substantiate the proposed method, providing evidence that these guidelines are a reliable starting point for VF task performance analyses in various clinical populations investigating VF task performance in depth.

KEYWORDS

Administration; clustering; error analysis; scoring; switching; temporal analysis; verbal fluency

Introduction



Verbal fluency (VF) tasks are widely used for clinical assessment and research purposes in multiple fields, such as speech pathology, neuropsychology, linguistics, and medicine (Strauss et al., 2006). In a VF task, the participant is asked to produce as many words as possible following a specific category in a specified time frame, often 60 s.


The most common VF task types are phonemic verbal fluency (PVF) and semantic verbal fluency (SVF). In the PVF, the participant is asked to produce words beginning with a specific phoneme or letter. This task is also referred to as phonemic fluency, Controlled Oral Word Associations (COWA), or the FAS test. In the semantic verbal fluency task (SVF), also referred to as semantic fluency or category fluency, the participant is asked to generate words belonging to a specific semantic category, such as “animals” or “clothes” (Strauss et al., 2006). Regardless of language context, VF task performance is considered to assess verbal knowledge and executive control. All VF tasks engage language processing, require maintaining focus on the task and selecting words that meet given criteria while inhibiting

unsuitable candidates (Shao et al., 2014; Whiteside et al., 2016).

The Cattell–Horn–Carroll Theory of Cognitive Abilities (CHC) classifies word fluency (FW) as a major narrow ability in retrieval fluency (Gr) under the broad ability of long-term storage and retrieval (Glr) (Schneider & McGrew, 2018). The CHC does not differentiate semantic and phonemic fluency per se. However, there is evidence suggesting a distinction between semantic and phonological fluency within the framework (Jewsbury & Bowden, 2017), and multiple neurocognitive studies support this distinction. PVF is typically considered to engage strategic cognitive organization, initiation, inhibition, and maintenance of effort without the support of the hierarchical organization of semantic memory (Barry et al., 2008; Santos Nogueira et al., 2016; Strauss et al., 2006). SVF is considered to rely on a more automatic systematic semantic search based on semantic categorization, hierarchical mental lexicon, and memory organization resembling everyday use of language (e.g., generating a shopping list; Patra et al., 2020; Strauss et al., 2006).

VF task performance is most typically evaluated by the total score, calculated as the number of acceptable words

CONTACT Nana Lehtinen  nana.lehtinen@utu.fi  Department of Psychology and Speech-Language Pathology, University of Turku, Assistentinkatu 7, Turku 20500, Finland.

 Supplemental data for this article can be accessed at [publisher's website](#).

© 2021 The Author(s). Published with license by Taylor & Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

generated in the given time frame (Strauss et al., 2006; Thiele et al., 2016) with normative data typically describing higher total scores for SVF than PVF tasks (Cavaco et al., 2013; Santos Nogueira et al., 2016; Strauss et al., 2006). While VF total scores can differentiate between healthy subjects and clinical groups, they also reflect verbal and executive processes underlying task performance. Evaluating total scores as the only metric does not provide insight into these processes, limiting the analytical and explanatory power of VF tasks (Becker & Salles, 2016; Johns et al., 2018; Oberg & Ramírez, 2006; Thiele et al., 2016). To investigate processes underlying VF performance, a growing body of literature is implementing additional analyses on temporal parameters, errors, and clustering and switching strategies using both traditional (Thiele et al., 2016) and computational approaches (Johns et al., 2018; Kim et al., 2019).

Norms for total scores and additional measures described above have been published in various languages, including effects of age, education, and gender (Ardila, 2020; Cavaco et al., 2013; Goral, 2004; Oberg & Ramírez, 2006; Olabarrieta-Landa et al., 2017; Quaranta et al., 2016; Santos Nogueira et al., 2016; Vicente et al., 2021). However, many studies fail to give accurate descriptions of administration, scoring, and analysis, making it challenging to compare and contrast studies and their outcomes (Olabarrieta-Landa et al., 2017; Thiele et al., 2016). A systematic approach to VF task analysis would increase the reliability and validity of studies implementing these simple to administer tasks across populations and languages (Becker & Salles, 2016; Thiele et al., 2016). In the following, we highlight aspects of VF task analyses implemented in previous literature and outline a suggestion for a systematic approach to administer, score and analyze phonemic and semantic verbal fluency tasks to increase the analytical and explanatory power of semantic and phonemic VF tasks.

Selecting categories for PVF and SVF tasks

In PVF, it is typical to include three trials. In English, FAS is the most common letter combination; other popular combinations are CFL and PWR. These combinations are selected from “easy letters” (Borkowski et al., 1967) and can, with some reservation, be used interchangeably (Ross, 2003; Strauss et al., 2006). Despite being intended initially for English speakers, a combination of FAS is used in studies conducted in other languages as well (for an overview, see Olabarrieta-Landa et al., 2017). However, it has been shown that letters with high frequency in the target language yield a higher number of words in VF tasks, and selecting language-specific categories for PVF is strongly recommended (Mardani et al., 2019; Oberg & Ramírez, 2006; Tombaugh et al., 1999). A short version of PVF consists of only one trial, often “B” for English (Harrison et al., 2000). As internal reliability between letters is high, it can be justified to reduce the number of trials. However, three phonemic trials are often preferred as it provides a more reliable measure of overall fluency ability (Oberg & Ramírez, 2006; Strauss et al., 2006; Tombaugh et al., 1999).

In SVF, a widely used category across languages is “animals” (Olabarrieta-Landa et al., 2017; Strauss et al., 2006). Other categories, such as “clothing,” “vehicles,” or “items in a supermarket,” are also used. Multiple categories are applied to parallel the number of categories in PVF, but combining data from different semantic categories is more complex than combining multiple phonemic categories. Demographic influences and various cultural settings can influence semantic memory organization, and semantic category size and content can vary between populations (Abwender et al., 2001; Olabarrieta-Landa et al., 2017; Roberts & Dorze, 1997; Rosselli et al., 2002; Strauss et al., 2006; Troyer, 2000). As a semantic category, “animals” are culturally and linguistically relatively neutral, but as all categories, requires specific guidelines for scoring (e.g., how to score variations of the same animal; Olabarrieta-Landa et al., 2017; Pekkala et al., 2009; Roberts & Dorze, 1997). Category “animals” is also included in many neuropsychological test batteries, such as the Western Aphasia Battery (WAB; Kertesz, 1982) and the Consortium to Establish a Registry for Alzheimer’s Disease (CERAD; Morris et al., 1989), making it a clinically appropriate choice for a semantic category.

Analyzing VF tasks

Performance in a verbal fluency task is typically analyzed by the total score (i.e., calculating the total number of acceptable words generated during the allotted time) (Strauss et al., 2006; Thiele et al., 2016). Higher total scores are generally associated with higher education, especially in PVF tasks (Oberg & Ramírez, 2006; Santos Nogueira et al., 2016; Tallberg et al., 2008; Tombaugh et al., 1999; Troyer, 2000), and total scores tend to decline with age, especially in SVF tasks (Ardila, 2020; Goral, 2004; Lanting et al., 2009; Santos Nogueira et al., 2016; Strauss et al., 2006; Tallberg et al., 2008; Tombaugh et al., 1999; Troyer, 2000). For gender, studies have shown no effect or a minor female advantage with emphasis on PVF (Scheuringer et al., 2017).

Temporal parameters of VF task performance are typically analyzed by the number of words generated during shorter time segments (10, 15, or 20-s segments) within the total time. Typically, participants produce most words in the early stages of the task using a semi-automatic rapid retrieval process. As time progresses, lexical retrieval becomes more effortful, with fewer and more infrequent words being generated toward the later segments of the task, demonstrating the function of time (Crowe, 1998; Fernaeus & Almkvist, 1998; Venegas & Mansur, 2011). Education has a positive effect in the early time segments in both PVF and SVF with no age effect found (Venegas & Mansur, 2011). In clinical populations, variation in the function of time has been shown to differentiate underlying mechanisms of VF performance in aphasia (Bose et al., 2017) and to have predictive power in Alzheimer’s Disease diagnosis (Fernaeus et al., 2008; Venegas & Mansur, 2011). It has also been suggested that as most words are generated in early segments of the task, a shorter total time (30 s) could have enough power

to differentiate healthy and patient populations (Fernaes et al., 2008; Kim et al., 2011).

Error types in VF typically include repetitions, categorical errors, and non-items (Thiele et al., 2016). While errors are relatively scarce in normative data (Crowe, 1998; Gollan et al., 2011), the number and type of errors in varied populations carry value in regards to their research objectives, such as perseverations in Alzheimer's studies (Pekkala et al., 2008) and language intrusions in bilingualism studies (Gollan et al., 2011).

Clustering and switching are strategies needed for optimal fluency performance (Strauss et al., 2006; Thiele et al., 2016; Troyer, 2000; Troyer et al., 1997). Analysis of clustering and switching has been applied for multiple research objectives, such as differential diagnostics in neuropsychological populations (Johns et al., 2018; Thiele et al., 2016; Troyer, 2000; Troyer et al., 1997) and cross-linguistic fluency strategies in bilingualism studies (Roberts & Dorze, 1997; Rosselli et al., 2002). While many studies apply predetermined subcategories for clustering following Troyer et al. (1997), other methods for clustering and calculation switches are also frequently applied. Different methods of determining clusters and calculating switches can result in varied outcomes (Abwender et al., 2001; Thiele et al., 2016).

Clustering refers to the ability to produce words within subcategories (e.g., "words with two same initial phonemes" or "pets"), and it relies on phonemic analysis in PVF and semantic categorization and semantic memory in SVF (Strauss et al., 2006; Thiele et al., 2016; Troyer, 2000; Troyer et al., 1997). Task congruent clustering is a relatively automatic process with participants naturally using phonemic clustering in PVF and semantic clustering in SVF (Troyer et al., 1997). In addition, task discrepant clustering (semantic clustering in PVF, phonemic clustering in SVF) is a prevalent strategy representing automatic semantic activation or the use of intentional and effortful cognitive strategies in PVF (Abwender et al., 2001; Sung et al., 2013). Considering different strategies for clustering, multiple qualities can be attached to one word, and surrounding words often define the intended subcategory (Becker & Salles, 2016). Switching is the ability to move to a new subcategory when the previous subcategory is exhausted. Compared to clustering, switching is considered being a more effortful process. It is considered to involve higher cognitive functions, such as cognitive flexibility and strategic search processes (Patra et al., 2020; Strauss et al., 2006; Thiele et al., 2016; Troyer, 2000; Troyer et al., 1997). Clustering and switching are closely related as the method of calculating clusters determine the number of switches.

In general, higher education predicts a larger cluster size and more switches in both task types, potentially due to a more robust semantic network or larger vocabulary size (Pereira et al., 2018; Troyer et al., 1997). Typically older adults generate larger clusters, possibly reflecting a more extensive vocabulary (Troyer et al., 1997). Older adults also switch less than younger adults signaling an age-related decline in higher executive functions, especially in SVF (Lanting et al., 2009; Troyer, 2000; Troyer et al., 1997). Men

tend to generate larger cluster sizes, with women switching more, especially in SVF (Lanting et al., 2009; Weiss et al., 2006).

VF tasks are analyzed across languages and cultures. Thus, language and culture-specific details should be considered when analyzing data and interpreting the results from different populations (Ardila, 2020; Becker & Salles, 2016; Kim et al., 2019; Oberg & Ramirez, 2006). Language-specific scoring guidelines can be essential in languages with productive compounding or extensive use of inflectional and derivational morphemes (Tallberg et al., 2008). In semantic clustering, predefined subcategories can be too narrow or broad to reveal culturally and linguistically unique lexical retrieval strategies (Becker & Salles, 2016; Roberts & Dorze, 1997). Unique retrieval strategies include, for example, dialectical variation and the influence of various cultural settings, as shown in studies investigating bilingual performance in VF tasks (Gollan et al., 2002; Rosselli et al., 2002). Detecting these subtle strategies requires thorough familiarization with the data from the population in question (Olabarrieta-Landa et al., 2017). Computational approaches typically train semantic models to analyze semantic variables using extensive written language corpora of the target language (Johns et al., 2018; Taler et al., 2020; Tröger et al., 2019) and it should be noted that subtle language variations can be underrepresented or omitted in written texts (Kim et al., 2019).

To summarize, VF tasks are a valuable and widely used tool for research and clinical purposes. An extensive amount of studies in multiple research areas have been conducted with varied categories utilizing both manual and computational approaches. Computational modeling has enabled great strides in VF task analysis via automatization, broadening our understanding of large-scale trends of human behavior (e.g., Kim et al., 2019; Taler et al., 2020). However, while detailed manual VF task analysis is time-consuming and can include inconsistencies (Kim et al., 2019), the need for manual analysis remains in smaller-scale studies where computational resources are not available and for specific populations and clinical purposes. Thus, to expedite the process of manual analysis and to increase the reliability and validity of verbal fluency task analysis across studies, comprehensive and precise instructions for administration, scoring, and analyses are warranted.

Aim of this study

In this study, we outlined detailed instructions for administering, scoring, and analyzing semantic and phonemic VF tasks to provide reliable tools for in-depth analysis in various clinical and research settings ([Supplementary Appendix A](#), Instruction Manual for Administration and Scoring Verbal Fluency Tasks). To investigate the reliability of the proposed method, we demonstrated the analysis in a sample of middle-aged and older Finnish-speaking adults.

To investigate the reliability of the proposed method, we described overall task performance and aimed to show whether task type (PVF vs. SVF), age, education, or gender

predicts (1) the total score and (2) the number of words generated in four 15-s segments. We also investigated (3) the frequency of errors and error types in PVF and SVF. For clustering and switching strategy use, we first investigated (4) whether task congruent cluster size and the number of switches or their interactions predict the total score. Second, we investigated (5) if the frequency of task discrepant clusters predicts total score.

We expected our results to be in line with previous literature. Based on previous research, we expected to see a higher overall score for SVF than PVF and participants with higher education to generate higher total scores than participants with lower education, especially in PVF. As the age range in our data was rather small, we expected minimal, if any, negative effect for age in total scores or performance in shorter time segments. For temporal parameters, we expected to find a systematic decline in the number of words generated in four 15-s segments. We expected participants with higher education to generate proportionally more words in the first 15-s segment of the task than participants with lower education. For clustering and switching, we expected the use of both strategies to contribute to a higher score. We expected a positive effect of education for task congruent cluster size and the number of task discrepant clusters as well as for the number of switches. We also expected to see more task discrepant clusters generated in the PVF task than in the SVF task with potential predictive power for higher total scores, especially in PVF. As for total scores, we expected age to have a minimal, if any, positive effect on cluster size and a negative effect on switching. We expected very little or no effect for gender overall.

Materials and methods

Participants

A sample of 50 middle-aged and older Finnish speakers with a higher proportion of women than men participated in this study. Sample mirrors the age and gender distribution of neurological deficits in populations (Roy-O'Reilly & McCullough, 2018) and thus serves the purpose of this study as a starting point for future studies consisting of larger clinical and control groups. Based on the calculations described below, the sample size was considered sufficient to demonstrate the proposed analysis methods and to show the effects of the size to have clinical relevance. Background information, health and language history were obtained in an interview setting by a graduate student in speech-language pathology via a comprehensive questionnaire. Participants were community-dwelling monolingual individuals who self-reported no history of language-related deficits or diagnoses (dyslexia, stroke, other neurological disorder) or hearing impairment. Participants who were not able to reliably report meeting the criteria described above were excluded.

Data for this study were collected as a part of an ongoing project on Finnish language attrition consisting of monolingual native Finnish speakers and participants with immigrant backgrounds. Data for monolingual performance in

Table 1. Demographic characteristics of participants and group internal comparisons for participant groups.

| Demographic variable | Total N = 50 | Education <12 years n = 27 | Education >12 years n = 23 | p |
|----------------------|-----------------|----------------------------------|----------------------------------|------------------|
| Age | | | | |
| Mean (SD) | 62.58 (7.59) | 62.23 (7.84) | 62.96 (7.43) | .75 ^a |
| Range | 49–79 | 49–79 | 52–79 | |
| Gender | | | | |
| Female n (%) | 35 (70) | 20 (74) | 15 (65) | .50 ^b |
| Male n (%) | 15 (30) | 7 (26) | 8 (35) | |

Note. Education < 12 years = no academic degree; Education > 12 years = academic degree.

^aGroups were compared using Two sample t-test.

^bGroups were compared using Chi-square goodness of fit test.

four verbal fluency tasks investigated in this study were extracted from a data pool of language tasks, including five verbal fluency tasks (one concrete semantic task (animals), one abstract semantic task (emotions), three phonemic categories (/k/, /a/, /p/). The research was conducted in accordance with the principles stated in the Declaration of Helsinki and the University of Turku Ethics committee approved all experimental procedures. All participants provided a written voluntary informed consent to participate in the study. They were informed of their right to withdraw at any time and did not receive compensation for participation. Demographic characteristics of participants and group internal comparisons are presented in Table 1.

Verbal fluency tasks

Data were collected as a part of a larger test battery as described above. Here, we report three phonemic verbal fluency tasks (/k/, /a/, /p/) and the concrete semantic verbal fluency task (animals). The order of VF tasks within the test battery was fixed (semantic, phonemic).

Categories

Phonemic verbal fluency tasks were localized into the Finnish language by using the most frequent word-initial consonants of Finnish /k/ (15,242 words) and /p/ (10,640 words) and the most frequent initial vowel /a/ (4,361 words) (Kielitoimiston Sanakirja, 2021; Leskinen, 1989), allowing comparison to earlier studies and providing a reference point for future studies.

For the semantic verbal fluency tasks, the category selected was “animals.” This category is commonly used, culturally and linguistically relatively neutral, and included in many neuropsychological test batteries.

Administration and scoring

For a detailed instructions manual for administration and scoring, see [Supplementary Appendix A](#). All tasks were completed in a quiet environment in a single session. Participants were encouraged to take short breaks in between tasks when needed. Responses were recorded for later verification and scoring. A research assistant transcribed audio tracks, and the authors verified transcripts.

For each trial, participants were asked to produce as many individual words as possible in 60 seconds.

Our goal was to have as simple and straightforward task instructions as possible. We approached this by including the phrase “individual words” in our instructions to discourage participants from using inflections (e.g., *kirja* [book], *kirjani* [my book]) but not to inhibit the use of derivational words that carry independent semantic meaning (e.g., *kirja* [book], *kirjasto* [library]) or compound words (e.g., *kirjakauppa* [book store]). This was a language-specific choice as Finnish has rich derivational and inflectional morphology, and it uses compounding productively to form new words (Helasvuo, 2008; Tyysteri, 2015). Using the wording “individual words” was also considered to guide participants not to produce multiple numerals without adding complexity to the instruction. In addition, we chose to use the word “letter” instead of “phoneme” in PVF, even to assess phonemic, not spelling, fluency. Finnish has strong orthographic transparency, and the words letter and sound are strongly interchangeable (Suomi et al., 2006). The more common word “letter” was selected to simplify instructions as much as possible. The only restriction was the use proper of names.

All verbal fluency tasks were scored for (1) total score, (2) number of acceptable words generated in 15-s segments, (3) number of errors and error types, (4) mean cluster size for task congruent clusters (semantic per semantic and phonemic per phonemic), and the number of switches calculated from task congruent clusters as well as for (5) number of task discrepant clusters (semantic clusters in PVF and vice versa).

We chose to investigate the distribution of words in 15-s segments to demonstrate the function of time comprehensively and to allow a simple combination for 30-s segment analysis if needed. We described the frequency of errors and error types to screen for error frequency and variety in our sample. Lastly, for clustering and switching, we based our analysis on task congruent clusters following Troyer et al. (1997) but chose to use naturally occurring clusters instead of fixed subcategories and to apply the rule for the smallest possible cluster in the analysis. In addition, we tracked the use of task discrepant clusters as suggested by Sung et al. (2013) and Abwender et al. (2001). In regards to detailed rules and instructions on coding semantic and phonemic clusters, see [Supplementary Appendix A](#). Briefly, naturally occurring clusters were determined by calculating the number of words generated in individual subcategories for each participant. Under the semantic condition, naturally occurring clusters are typically taxonomic subcategories of animals (e.g., birds or big cats). However, they can also be formed by environmental semantic connections (e.g., farm animals), geographical semantics (e.g., African animals), or visual semantics (e.g., snake, eel). Under the phonemic condition, clusters are typically formed by the same two initial phonemes or rhyming words, but they can also be formed by structurally similar words that differ only by one sound or vowel sounds.

Table 2. Intraclass correlation coefficients and their 95% confidence intervals for mean cluster sizes and the number of switches.

| Variable | ICC <i>n</i> = 140 | 95% CI | |
|-----------------------------|-----------------------|--------|------|
| | | LL | UL |
| Phonemic cluster size | .97 | .96 | .98 |
| Number of phonemic switches | >.99 | .99 | >.99 |
| Semantic cluster size | .79 | .73 | .84 |
| Number of semantic switches | .96 | .95 | .98 |

Note. ICC: intraclass correlation coefficient; CI: confidence interval; LL: lower limit; UL: upper limit.

Naturally occurring clusters include all culturally and linguistically unique strategies participants might use (Becker & Salles, 2016; Roberts & Dorze, 1997) and account for individual semantic networks and their possible influence on semantic categorization (Morais et al., 2013). Determining naturally occurring clusters eliminates the need for predefined categories mirroring computational approaches that extract information from natural language (e.g., Kim et al., 2019). Following the rule for the smallest possible cluster size allows for specific tracking of switching. We are aware that analyzing task congruent and task discrepant clusters separately does not account for the effect task discrepant clusters might have on switching. However, to accurately track the use of both cluster types, they are scored and analyzed separately, and only the number of switches based on task congruent clusters is analyzed.

Inter-rater reliability for cluster size and number of switches

To verify the reliability of the analysis for cluster size, two raters coded the data following instructions in [Supplementary Appendix A](#). Based on a larger dataset (3 tasks) and high inter-rater reliability for phonemic clustering in literature (Becker & Salles, 2016; Ross, 2003) 60% of the PVF data ($n = 90$; $n = 30$ for each phoneme) were used for inter-rater reliability analysis. Due to a smaller dataset (1 task) and a semantic component of the analysis, 100% of the SVF data ($N = 50$) were used for inter-rater reliability analysis. We calculated the Intraclass Correlation Coefficient (ICC) for all, task congruent and task discrepant, phonemic, and semantic cluster sizes in both verbal fluency tasks using R (R Core Team, 2019) and “psych” package (Revelle, 2020). We selected a two-way random-effects model with a single measurement and absolute agreement to show the magnitude of agreement achieved between two raters for these measures as we aim to generalize the reliability of the results to raters who want to use the same analysis in their clinical or research work (Koo & Li, 2016).

For the phonemic cluster size, the number of phonemic switches, and the number of semantic switches, the ICC analysis showed an excellent degree of reliability between the two raters as ICCs are above 0.90 (Table 2) (Koo & Li, 2016). For the semantic cluster size, the reliability between the raters was good; ICC value between 0.75 and 0.90 (Table 2) (Koo & Li, 2016). A blind review by a third rater showed that lower ICC in the semantic clustering analysis resulted primarily from differences between the raters in weak cluster size where there were multiple semantically

acceptable ways to form clusters (e.g., see [Supplementary Appendix A](#)). As subjective semantic categorization will always be present in semantic clustering analysis, this result was determined to demonstrate acceptable semantic variation between raters.

Data analysis

R software (R Core Team, 2019) with packages `dplyr` (Wickham et al., 2019), `tidyr` (Wickham, 2020), `lme4` (Bates et al., 2015), and `lmerTest` (Kuznetsova et al., 2017) were used in data cleanup and analyses. Packages `sjPlot` (Lüdtke, 2018), `jtools` (Long, 2020), `ggeffects` (Lüdtke, 2018), and `ggplot2` (Wickham, 2016) were used in tables and figures and packages `effect size` (Ben-Shachar et al., 2020) and `EMAtools` (Kleiman, 2017) for calculating effect size estimates. Model assumptions were checked using “`check_model`”-function from package `performance` (Lüdtke et al., 2020). Our data and analysis scripts are available at <https://osf.io/kh8f3/>.

We modeled PVF and SVF performance together. This allowed us to investigate the main effects of task type (the differences in performance between the tasks) but also the effects of demographic variables in PVF and SVF separately as well as in the combined data of both task types. To answer our research question on task type and participant age, education, and gender predicting the total score, we employed a linear mixed-effects model with the total score as a response variable, and task type, gender (female/male), age, education (high/low) and all two-level interactions (interactions of gender, age or education, and task type) as predictors. For modeling purposes, the age variable was scaled and centered to the sample mean so that the estimates would reflect the performance in mean age rather than at 0 years. This was considered to be more informative in this context. To supplement the analysis, we provide descriptive data on total scores for in-age bonds of 49–59, 60–69, and 70–79. The two-level education variable was centered between high and low values so that the model estimates would better reflect performance in the whole population, regardless of the education level. Participant IDs within each level of task type were used as random factors to account for individual variation in performance.

To address our research question on task type predicting the words produced during each 15-s segment of the task, we modeled the number of words produced as a function of each 15-s segment. Participant intercept and individual slope for task type were applied as random factors. To select the most parsimonious model fit to our data, models with participant background information (gender, age, education) as additional predictors were compared to the model with 15-s segments only as a predictor using analysis of variance and BIC values. Based on this model selection procedure, the model without participant background information turned out to be the most parsimonious fit for the data.

The frequency and type of errors are described as raw scores. Mean values are reported to enable comparison between the tasks due to the small number of errors in the

data. The number of participants who generated errors is described for all tasks separately.

To address our research question on clustering and switching, we modeled the total score as a function of task congruent cluster size, the number of switches, and their interactions. Separate models were used for PVF and SVF tasks. In the phonemic model, participant intercept and slope for the task (/k/, /a/, or /p/) were used as random factors. Since the semantic dataset only had one task per participant, using a mixed-effects model was unnecessary, and we employed a simple linear regression instead. To control the effects of participant background variables, we also considered participant education, age, and gender and all their interactions as potential predictors in the models. We performed a similar model selection than in the 15-s segment model, which suggested the model with no background variables as additional predictors to be the most parsimonious fit for the phonemic task performance and the model with education as an additional predictor to be the most parsimonious fit for the semantic task performance.

To investigate the use of task discrepant clusters, we modeled the total score in the phonemic tasks as a function of semantic clusters and the total score in the semantic task as a function of the number of phonemic clusters. In the phonemic model, we also included trial order as a predictor in the model to investigate whether the position of the phonemic trial had an effect on the number of semantic clusters as the order of trials was not randomized. All participants used task discrepant clustering in both task types, there were 39 phonemic trials from 30 participants with no task discrepant clusters. Trials with no task discrepant clusters were excluded from the phonemic model. Based on a similar model selection procedure as described above, we chose the models with no background variables as the most parsimonious ones.

We performed power calculations to ensure that the sample size provided adequate power for our outcome measures. However, power calculation for statistical methods we used is not straightforward: because of the model selection procedure, we did not know the exact number of predictors before model fitting. Also, there is no exact method to calculate power for linear mixed models (containing random effects part). We did not have prior estimates about the random effects (i.e., within-subjects variation) available from previous studies to perform appropriate power simulations. Thus, we based our sample size estimations on power calculations for linear regression with similar sample sizes, effect sizes, and the number of predictors. These power calculations suggested a statistical power >0.8 for all models except those examining clusters and switches or task discrepant clusters in the semantic task. In these models, the statistical power was 0.59 and 0.77, respectively. The statistical power was lower in these models because we only had data from one semantic task for each participant as compared to 111–800 data points in other models. In addition, as determining exact degrees of freedom for the test statistics estimated by linear mixed models is difficult, it is also problematic to determine unambiguous p -values (see Baayen et al., 2008). Hence, the statistical significance at the 0.05

Table 3. Descriptive statistics of total scores and temporal parameters in semantic and phonemic verbal fluency tasks.

| Variable | Phonemic /k/ | Phonemic /a/ | Phonemic /p/ | Semantic animals |
|-------------------------------|--------------|--------------|--------------|------------------|
| Total score 0–60 s | | | | |
| Mean | 19.38 | 14.68 | 17.44 | 25.96 |
| SD | 6.12 | 5.65 | 5.50 | 5.90 |
| Range | 5–35 | 5–31 | 4–30 | 12–43 |
| Time segment 0–15 s | | | | |
| Mean | 7.06 | 5.72 | 6.26 | 10.76 |
| SD | 2.37 | 2.25 | 2.20 | 2.14 |
| Range | 2–12 | 1–11 | 2–12 | 6–15 |
| Percentage of total score (%) | 36 | 39 | 36 | 41 |
| Time segment 16–30 s | | | | |
| Mean | 4.48 | 3.54 | 4.24 | 6.06 |
| SD | 2.04 | 1.94 | 1.69 | 2.13 |
| Range | 0–9 | 0–8 | 1–8 | 0–10 |
| Percentage of total score (%) | 23 | 24 | 24 | 23 |
| Time segment 31–45 s | | | | |
| Mean | 4.12 | 2.98 | 3.48 | 4.94 |
| SD | 1.90 | 1.64 | 1.66 | 2.36 |
| Range | 0–9 | 0–7 | 0–7 | 0–13 |
| percentage of total words (%) | 21 | 20 | 20 | 19 |
| Time segment 46–60 s | | | | |
| Mean | 3.72 | 2.44 | 3.46 | 4.20 |
| SD | 1.51 | 1.57 | 1.85 | 2.36 |
| Range | 0–7 | 0–7 | 0–8 | 0–10 |
| Percentage of total score (%) | 19 | 17 | 20 | 16 |

Note. Total score = total number of acceptable words generated in a 60-s trial; Time segment = number of acceptable words generated in a 15-s time segment within a 60-s trial.

level in this article is indicated by $|t| > 1.96$. However, since many readers might be more familiar with p -values than t -values, rough estimates for p -values are also provided in model summaries (Supplementary Appendix B).

Results

For the sake of brevity, all full model summaries with effect size estimates and supplementary information on age bands are presented in Supplementary Appendix B.

Total score

Modeling the total score as a function of task type, participant gender, education, and age revealed main effects of task type, 8.47, $t = 8.96$, 95% CI [6.62, 10.33], gender, -5.06 , $t = -3.78$, 95% CI $[-7.69, -2.44]$ and education, 4.55, $t = 3.75$, 95% CI [2.18, 6.93]. Higher numbers of correct words were associated with semantic task type, higher education, or being female. Age was not significantly associated with task performance in our sample, -1.00 , $t = -1.63$, 95% CI $[-2.20, 0.20]$. Interactions between the task type and education, -0.21 , $t = -0.14$, 95% CI $[-3.31, 2.88]$, task type and age, 0.53, $t = 0.67$, 95% CI $[-1.03, 2.10]$ or task type and gender, 1.03, $t = 0.59$, 95% CI $[-2.38, 4.44]$ were not significant. Descriptive statistics of raw scores in all four fluency tasks are presented in Table 3.

15-s segments

Modeling words produced during each 15-s segment revealed that frequency in producing acceptable words

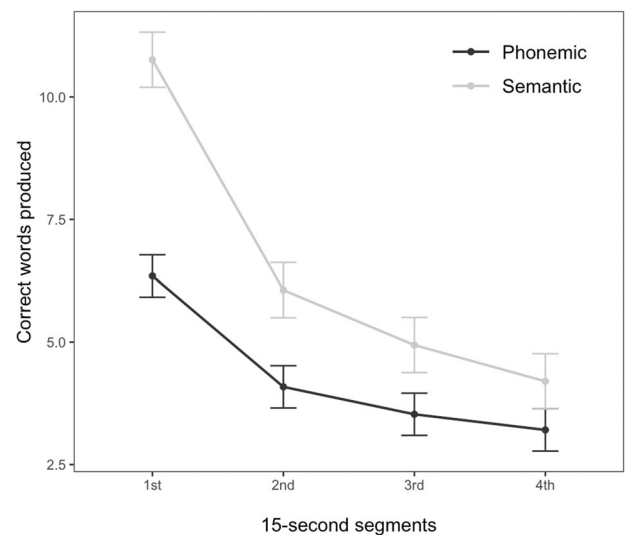


Figure 1. Predicted values of words produced in the four 15-s segments in phonemic and semantic verbal fluency tasks. Error bars represent 95% confidence intervals.

decreased during the task (Figure 1). This decrease was more substantial in semantic, as opposed to phonemic task type. Descriptive statistics of raw scores in all four fluency tasks are presented in Table 3.

Errors

In PVF, roughly half of the participants generated errors in all three trials (/k/ 44% [$n = 22$]; /a/ 58% [$n = 29$]; /p/ 50% [$n = 25$]) with most errors being repetitions and categorical errors. In the SVF task, 28% of participants ($n = 14$) generated errors of which most were repetitions. The total number of errors and distribution of error types are presented in Table 4.

Task congruent clustering and switching

The number of switches, 6.22, $t = 15.39$, 95% CI [5.43, 7.02] (phonemic); 7.59, $t = 10.40$, 95% CI [6.11, 9.06] (semantic), and the mean cluster size, 1.72, $t = 4.14$, 95% CI [0.91, 2.54] (phonemic); 4.65, $t = 6.86$, 95% CI [3.29, 6.02] (semantic); were associated with the total score in both tasks. The semantic task model also included participant education as a predictor. The main effect of education, 2.98, $t = 2.87$, 95% CI [0.88, 5.08], was consistent with the observation in the total score model reported earlier in this chapter. Two- and three-level interactions of education, number of switches, and cluster size were not significant. Descriptive statistics of raw scores in all four fluency tasks are presented in Table 5, and all interactions are presented in Figure 2.

Task discrepant clustering

Participants generated task discrepant clusters in both task types. Numerically, most task discrepant clusters were generated under phoneme /k/ ($M = 2.98$, $SD = 2.01$), followed by /p/ ($M = 1.7$, $SD = 1.17$). Least task discrepant clusters in

Table 4. Number of participants who generate errors, total number of errors, distribution of error types and in semantic and phonemic verbal fluency tasks.

| Errors | Phonemic /k/ | | Phonemic /a/ | | Phonemic /p/ | | Semantic animals | |
|----------------------|--------------|----|--------------|----|--------------|----|------------------|----|
| | | % | | % | | % | | % |
| <i>n</i> with errors | 22 | 44 | 29 | 58 | 25 | 50 | 14 | 28 |
| Total no. of errors | 39 | | 44 | | 35 | | 20 | |
| Mean | 1.77 | | 1.52 | | 1.40 | | 1.43 | |
| SD | 1.02 | | 0.95 | | 0.50 | | 0.65 | |
| Range | 1–4 | | 1–4 | | 1–2 | | 1–3 | |
| Error type | | | | | | | | |
| Repetition | 26 | 67 | 25 | 57 | 27 | 77 | 19 | 95 |
| Categorical | 12 | 31 | 16 | 36 | 7 | 20 | 0 | 0 |
| Nonword | 1 | 3 | 3 | 7 | 1 | 3 | 1 | 5 |

Note. *N* = 50.

Table 5. Descriptive statistics of task congruent cluster size, number of task discrepant clusters and switches in semantic and phonemic verbal fluency tasks.

| Variable | Phonemic /k/ | Phonemic /a/ | Phonemic /p/ | Semantic animals |
|---|--------------|--------------|--------------|------------------|
| Task congruent cluster size ^a | | | | |
| Mean | 2.62 | 2.24 | 2.56 | 2.7 |
| SD | 0.83 | 0.34 | 0.78 | 0.37 |
| Range | 2.00–5.50 | 2.00–3.25 | 2.00–5.60 | 2.00–3.75 |
| <i>n</i> with clusters ^b | 46 | 42 | 47 | 50 |
| % ^c | 92 | 84 | 94 | 100 |
| Number of task discrepant clusters ^d | | | | |
| Mean | 2.98 | 0.80 | 1.70 | 0.90 |
| SD | 2.01 | 1.18 | 1.17 | 1.15 |
| Range | 0–7 | 0–5 | 0–4 | 0–5 |
| <i>n</i> with clusters ^b | 46 | 22 | 43 | 26 |
| % ^c | 92 | 44 | 86 | 52 |
| Switches ^e | | | | |
| Mean | 13.88 | 11.52 | 11.70 | 11.60 |
| SD | 5.95 | 3.90 | 3.97 | 3.07 |
| Range | 2–26 | 1–20 | 3–22 | 5–20 |

Note. *N* = 50 in all conditions.

^aSVF semantic clusters, PVF phonemic clusters.

^bNumber of participants who generated clusters.

^cPercentage of participants who generated clusters.

^dSVF phonemic clusters, PVF semantic clusters.

^eNumber of switches calculated from task congruent clusters.

PVF were generated under vowel /a/ ($M = 0.8$, $SD = 1.18$) and this was in line with task discrepant cluster frequency in the semantic category ($M = 0.9$, $SD = 1.15$). In addition to the group mean values, it is worth noting that in the PVF category /k/ 92% ($n = 46$) of participants, in /p/ 86% ($n = 43$) of participants, and in /a/ 44% ($n = 22$) of participants generated semantic clusters. In SVF, 52% of the participants ($n = 26$) generated phonemic clusters. In PVF, the number of semantic clusters was a significant predictor for the total score, 1.42 , $t = 6.60$, 95% CI [1.00, 1.84]. The interaction of the number of semantic clusters and trial order, 0.02 , $t = 0.07$, 95% CI [−0.52, 0.56], was not a significant predictor for the use of task discrepant clustering nor was the main effect of trial order a significant predictor for the total score, -0.30 , $t = -0.43$, 95% CI [−1.68, 1.08]. In the semantic task, the number of phonemic clusters did not predict the total score of 0.52 , $t = 0.44$, 95% CI [−1.89, 2.93]. Descriptive statistics of raw scores in all four fluency tasks are presented in Table 5.

Discussion

This study describes a comprehensive analysis of phonemic and semantic verbal fluency tasks for clinical and research

purposes. In addition to total scores, we demonstrate an analysis for temporal parameters, errors, and clustering and switching with strong inter-rater reliability in a sample group of 50 older healthy participants with the aim of providing a starting point for future studies. As discussed below, our results align with earlier literature, supporting the proposed method as a reliable starting point to analyze linguistic and cognitive processes underlying VF performance in varied clinical groups.

In our dataset of monolingual middle-aged and older adults, participants generated a higher total score in the semantic than in the phonemic tasks, in line with multiple normative datasets (for an overview, see Strauss et al., 2006). In the PVF, the trial total scores reflected the category size of word-initial phonemes in the Finnish language, as expected (Gollan et al., 2002). We evaluated education on a 2-tier scale and found a positive association to performance on both fluency types in line with Pereira et al. (2018). As hypothesized, we found no association between age and task performance in either task type. The lack of age effect is likely due to our sampling process resulting in the relatively small age range in our data (Ardila, 2020). While we found no significant age effect, the trends in our data shown in

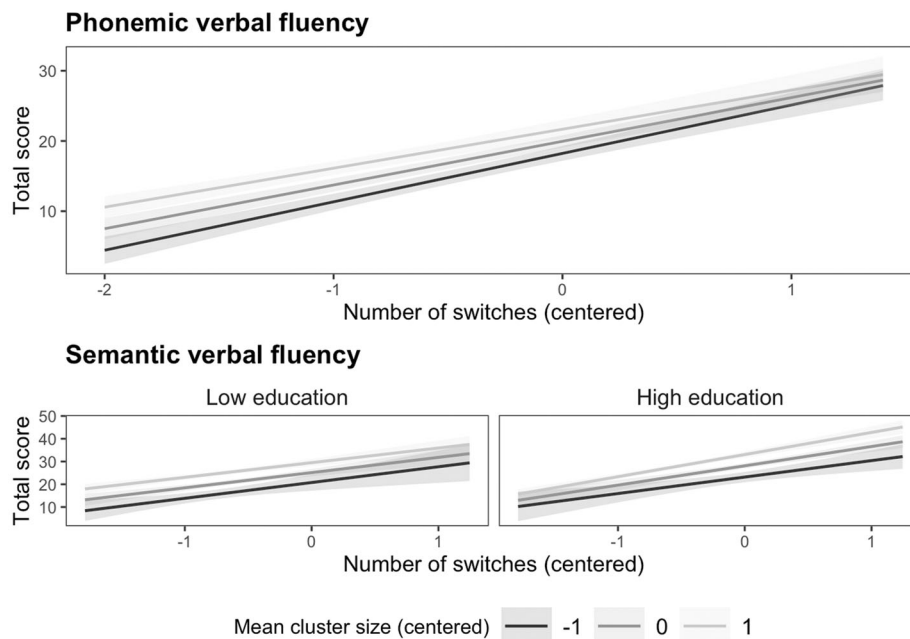


Figure 2. Estimates for the total score of phonemic and semantic tasks as a function of all predictor variables included in the models. For the sake of clarity, the continuous variable of centered mean cluster size is estimated on three levels. Error bars represent 95% confidence intervals. Note that due to the model selection procedure used, semantic but not phonemic model had education level as a predictor variable.

Supplementary Appendix B can be of value for clinical observations.

As expected, modeling task performance in 15-s segments revealed a performance decrease during both the task types. Following Crowe (1998), we infer that word retrieval becomes more effortful as time passes, reducing the number of words produced. Our results replicated results from Venegas and Mansur (2011), with performance in the first quartile being strongly associated with the overall score. In addition, the results corroborated with Kim et al. (2011), who suggest that 30-s total time for VF tasks can be a relevant approach to differentiate stroke patients with and without aphasia. Thus, it is tempting to suggest that it might be clinically efficient to screen patients using a very short VF task (i.e., 15 s) under specific circumstances. However, due to the limited sample size and focus on older adults across these studies, future studies with larger datasets and varied populations are needed to verify the relevance of this finding for clinical use.

Participants generated errors following error profiles described by Crowe (1998) and Gollan et al. (2011), with the most common error type being repetitions. Contrary to findings in Crowe (1998), we found more errors in PVF than in SVF. We did not randomize the order of trials but presented the semantic trial first, followed by phonemic trials. Thus, some errors in PVF could have been due to active semantic retrieval strategies in the first phonemic trial in PVF. However, no errors stemming from semantic strategies were detected. We conclude that generating errors with an emphasis on repetitions is a part of a verbal retrieval process in a healthy older monolingual population across tasks (Crowe, 1998; Gollan et al., 2011). A more significant number of errors, or a different distribution of error types than found here, can be an atypical finding, and as Thiele et al.

(2016) point out, can yield insight into various processes underlying task performance in clinical populations.

We found both task congruent clustering and switching to be productive strategies in phonemic and semantic VF tasks following Troyer et al. (1997), but no interactions between task congruent mean cluster size and switching were detected. Thus, the efficacy of switching as a strategy was not dependent on cluster size, nor was the efficacy of cluster size dependent on the number of switches in either task type. Here we point out that a deficit in the use of one of the strategies (clustering or switching), can lead to more extensive use of the other strategy.

Task discrepant clustering was common in phonemic trials. Depending on the trial, 44–98% of the participants generated task discrepant clusters. Our results align with Abwender et al. (2001), supporting the importance of including task discrepant clusters in VF task analysis. Notably, the number of semantic clusters predicted the total score in PVF, but the number of phonemic clusters in SVF did not. We think there are three possible explanations. First, as we had data from only one SVF trial (compared to three PVF trials), it is possible that we did not have sufficient statistical power to detect small effects. However, we are confident that effects large enough to have practical significance would have been observed in our sample. Second, our results support the notion that automatic semantic activation plays an essential role in both semantic and phonemic tasks, resulting in a more pronounced use of semantic clusters in phonemic tasks (Sung et al., 2013). Third, the use of semantic clustering in a phonemic task can be an additional, intentional strategy resulting from participants reaching out to their hierarchical semantic memory in an effortful phonemic task (Abwender et al., 2001). In the future, exploring the temporal distribution of task discrepant

clusters in PVF could provide evidence if semantic clustering is used throughout the task suggesting automatic semantic activation or if semantic strategies are used in the latter time segments suggesting a more intentional strategy use.

Here, we must consider if conducting the SVF trial before the PVF trials primed participants for semantic retrieval in PVF. If so, we would expect to see most semantic clusters in the first phonemic trial /k/. Statistical modeling of cluster frequency showed that the number of task discrepant clusters in phonemic trials does not differ between phonemic trials /k/, /a/, and /p/. Also, the proportion of participants who generated semantic clusters in phonemic trials lines up with phoneme category size rather than the order of trials presented. This suggests that the overall number of words available in a category can influence how semantic associations are activated in a phonemic VF task, reflected in our data as the number of participants who generated task discrepant clusters. Even with no difference between phoneme trials in the number of task discrepant clusters generated, we can not exclude the possibility of task order impacting task performance. Thus, we recommend randomizing tasks when implementing multiple VF tasks for research to minimize the possible effect task order might have on the performance.

In this study, we tracked the use of both cluster types, task congruent and task discrepant clusters, separately to simplify the coding for clinical use. Thus, our approach did not include switches stemming from task discrepant clustering. Our results show that both cluster types are evident and significant factors in VF task performance. We acknowledge the challenges including switching from both cluster types in the analysis but suggest that in the future, a combination of the two clustering analyses would be essential in determining the most reliable analyzing method for switching, especially in research settings.

Some limitations of the current study include screening patients' cognitive health via a self-reporting questionnaire and interview without standardized methods and a small sample size with a narrow age range. Based on rough estimates of statistical power, the sample size was sufficient in all models except clustering models in the semantic task. The statistical power was lower in these models due to only one semantic trial vs. three trials in the phonemic task. However, these preliminary findings on the semantic task can be valuable as guidelines to future studies, and even with limited sample size, we are confident that any effects of the size to have clinical relevance would be observed with this sample size and analyzing method. In future studies, standardized methods for cognitive screening should be included for reliable sampling. Exploring the suggested analysis in larger data samples of healthy participants with a broader age range as well as across languages and in varied clinical groups is needed to solidify our findings.

In the following, we further discuss the proposed method. The following aspects relate but are not limited to the Finnish language. Considerations for other languages should always be made in relation to the language and culture in

question (e.g., for Spanish, see Olabarrieta-Landa et al., 2017).

In the administration for PVF tasks, we used the word “letter” instead of “sound.” This was a deliberate, language-specific choice as the Finnish language has a strong letter-phoneme correspondence, and the use of the word “letter” is more common than the word “sound” in everyday language. However, there were some instances in the data where participants generated words that had the correct word-initial sound but that are spelled differently (e.g., /panaani/ with unvoiced /p/ vs. correct spelling /banaani/ with voiced /b/; [banana]), a common occurrence in spoken Finnish with a dialectical variation. As we did not evaluate spelling fluency these words were considered deviations from the task condition only when participants indicated that they had produced the word in error (e.g., “No, that begins with a different letter”). To eliminate potential confusion for the participants and simplify the analysis, we recommend using the word “sound” for PVF instructions in all languages following Olabarrieta-Landa et al. (2017) when assessing phonemic fluency. Using the word “sound” will also eliminate potential errors due to lack of spelling knowledge and reduce potential errors and lower total scores stemming from lower education. In our task instructions, we used the phrase “individual words” to discourage participants from producing multiple numerals in phonemic trials. This proved to be an effective strategy as very few participants included a string of numerals in PVF. Thus, we argue that the instruction is precise enough for VF task purposes without a restriction for sequential numerals.

Our clustering analysis consisted of naturally occurring clusters for each participant in both VF task types. In PVF, this meant specific rules to include all phonemic clustering strategies participants utilized in their output. In line with earlier literature, the inter-rater reliability for phonemic clustering was excellent between two raters (Becker & Salles, 2016; Ross, 2003; Troyer et al., 1997). For the PVF task, we did not allow lexical categories (verbs, adjectives, particles) as a basis for clustering due to interlanguage challenges in scoring demonstrated by Rosselli et al. (2002). However, during analysis, we did observe some participants using lexical categories as a productive word search strategy (e.g., *kiirehtiä* [to hurry], *keittää* [to boil], *kutittaa* [to tickle]). In future studies, it could be worthwhile to investigate if lexical categories could be included as an individual productive strategy for PVF utilizing language-specific guidelines for scoring.

In SVF, naturally occurring strategies allowed us to analyze subcategories in an area of expertise (e.g., birds of prey, aquatic birds), clustering based on geographical semantics (giraffe, monkey), and visual semantics (snake, eel). It also facilitated the use of context to determine the intended category for clustering analysis (e.g., forest animals, animals typical to Lapland; Becker & Salles, 2016). We chose this approach to include all language and culture-specific as well as individual clustering strategies resulting in a precise analysis of clustering and switching. Our inter-rater analysis indicated good reliability between raters and was deemed

sufficiently objective, but the margin of error between raters was more prominent in the semantic than in phonemic clustering. Computational approaches aim to create objective semantic variables to reduce subjectivity (Kim et al., 2019; Tröger et al., 2019), but manual analysis of semantic clustering strategies includes a subjective semantic component (Tröger et al., 2019). This variability due to subjective semantic interpretation is important to bear in mind while applying these rules in clinical and research settings. To minimize this variability in clinical and research settings where analysis is done manually, we have included a sample protocol and instructions on training raters in **Supplementary Appendix A**, as suggested by Ross (2003).

In conclusion, this study provides a starting point for a comprehensive analysis of VF performance. Future studies applying the suggested method in varied clinical groups will further test and solidify the method. Currently, we are working on implementing these analyses to investigate lexical processes underlying performance in VF tasks in bilingualism, aphasia, and Alzheimer's Disease. A consistent approach in administration, scoring, and analyzing VF data across studies is needed to enable systematic insight into cognitive processes underlying or possibly hindering optimal performance in different populations. We are hopeful that our research will be beneficial in determining specific and straightforward scoring rules for phonemic and semantic VF tasks.

Acknowledgments

We would like to thank Pirjo Korpilahti and Professor Marja-Liisa Helasvuo for their scientific advice, valuable comments, and support. We would also like to thank research assistants at the Department of Psychology and Speech-Language Pathology, University of Turku, for their assistance in data collection and analysis. Finally, we are grateful to the participants of this study for the generous donation of their time and effort in taking part in this research.

Disclosure statement

The authors do not have conflicts of interest regarding this research study.

Funding

This work was supported in part by the Alfred Kordelin Foundation under a grant awarded to the first author and University of Turku Graduate School wages awarded to the second and third authors, as well as an anonymous endowed fund to the University of Turku, Department of Speech-Language Pathology.

ORCID

Nana Lehtinen  <http://orcid.org/0000-0002-3606-1302>
 Ida Luotonen  <http://orcid.org/0000-0003-3310-3180>
 Anna Kautto  <http://orcid.org/0000-0003-4885-9167>

Data availability statement

The data supporting the findings of this study are openly available in Verbal Fluency at <https://osf.io/kh8f3/>.

References

- Abwender, D. A., Swan, J. G., Bowerman, J. T., & Connolly, S. W. (2001). Qualitative analysis of verbal fluency output: Review and comparison of several scoring methods. *Assessment*, 8(3), 323–338. <https://doi.org/10.1177/107319110100800308>
- 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>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Barry, D., Bates, M. E., & Labouvie, E. (2008). FAS and CFL forms of verbal fluency differ in difficulty: A meta-analytic study. *Applied Neuropsychology*, 15(2), 97–106. <https://doi.org/10.1080/09084280802083863>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Becker, N., & Salles, J. F. d. (2016). Methodological criteria for scoring clustering and switching in verbal fluency tasks. *Psico-USF*, 21(3), 445–457. <https://doi.org/10.1590/1413-82712016210301>
- Ben-Shachar, M. S., Makowski, D., & Lüdtke, D. (2020). *Compute and interpret indices of effect size*. CRAN. <https://github.com/easystats/effectsize>
- Borkowski, J. G., Benton, A. L., & Spreen, O. (1967). Word fluency and brain damage. *Neuropsychologia*, 5(2), 135–140. [https://doi.org/10.1016/0028-3932\(67\)90015-2](https://doi.org/10.1016/0028-3932(67)90015-2)
- Bose, A., Wood, R., & Kiran, S. (2017). Semantic fluency in aphasia: Clustering and switching in the course of 1 minute: Semantic fluency in aphasia. *International Journal of Language & Communication Disorders*, 52(3), 334–345. <https://doi.org/10.1111/1460-6984.12276>
- Cavaco, S., Gonçalves, A., Pinto, C., Almeida, E., Gomes, F., Moreira, I., Fernandes, J., & Teixeira-Pinto, A. (2013). Semantic fluency and phonemic fluency: Regression-based norms for the Portuguese population. *Archives of Clinical Neuropsychology*, 28(3), 262–271. <https://doi.org/10.1093/arclin/act001>
- Crowe, S. F. (1998). Decrease in performance on the verbal fluency test as a function of time: Evaluation in a young healthy sample. *Journal of Clinical and Experimental Neuropsychology*, 20(3), 391–401. <https://doi.org/10.1076/jcen.20.3.391.810>
- Fernaues, S.-E., & Almkvist, O. (1998). Word production: Dissociation of two retrieval modes of semantic memory across time. *Journal of Clinical and Experimental Neuropsychology*, 20(2), 137–143. <https://doi.org/10.1076/jcen.20.2.137.1170>
- Fernaues, S.-E., Östberg, P., Hellström, Å., & Wahlund, L.-O. (2008). Cut the coda: Early fluency intervals predict diagnoses. *Cortex*, 44(2), 161–169. <https://doi.org/10.1016/j.cortex.2006.04.002>
- Gollan, T. H., Montoya, R. I., & Werner, G. A. (2002). Semantic and letter fluency in Spanish-English bilinguals. *Neuropsychology*, 16(4), 562–576. <https://doi.org/10.1037/0894-4105.16.4.562>
- Gollan, T. H., Sandoval, T., & Salmon, D. P. (2011). Cross-language intrusion errors in aging bilinguals reveal the link between executive control and language selection. *Psychological Science*, 22(9), 1155–1164. <https://doi.org/10.1177/0956797611417002>
- Goral, M. (2004). First-language decline in healthy aging: Implications for attrition in bilingualism. *Journal of Neurolinguistics*, 17(1), 31–52. [https://doi.org/10.1016/S0911-6044\(03\)00052-6](https://doi.org/10.1016/S0911-6044(03)00052-6)
- Harrison, J. E., Buxton, P., Husain, M., & Wise, R. (2000). Short test of semantic and phonological fluency: Normal performance, validity and test-retest reliability. *The British Journal of Clinical Psychology*, 39(2), 181–191. <https://doi.org/10.1348/014466500163202>

- Helasvuo, M.-L. (2008). Aspects of the structure of Finnish. In *Research in Logopedics, Speech and Language Therapy in Finland. Communication disorders across languages* (pp. 9–18). Multilingual Matters.
- Jewsbury, P. A., & Bowden, S. C. (2017). Construct validity of fluency and implications for the factorial structure of memory. *Journal of Psychoeducational Assessment*, 35(5), 460–481. <https://doi.org/10.1177/0734282916648041>
- Johns, B. T., Taler, V., Pisoni, D. B., Farlow, M. R., Hake, A. M., Kareken, D. A., Unverzagt, F. W., & Jones, M. N. (2018). Cognitive modeling as an interface between brain and behavior: Measuring the semantic decline in mild cognitive impairment. *Canadian Journal of Experimental Psychology*, 72(2), 117–126. <https://doi.org/10.1037/cep0000132>
- Kertesz, A. (1982). *The western aphasia battery*. Psychological Corporation.
- Kielitoimiston sanakirja. (2021). *Kielitoimiston sanakirja*. <https://www.kielitoimistonanakirja.fi>
- Kim, H., Kim, J., Kim, D. Y., & Heo, J. (2011). Differentiating between aphasic and nonaphasic stroke patients using semantic verbal fluency measures with administration time of 30 seconds. *European Neurology*, 65(2), 113–117. <https://doi.org/10.1159/000324036>
- Kim, N., Kim, J.-H., Wolters, M. K., MacPherson, S. E., & Park, J. C. (2019). Automatic scoring of semantic fluency. *Frontiers in Psychology*, 10, 1020. <https://doi.org/10.3389/fpsyg.2019.01020>
- Kleiman, E. (2017). *EMAtools: Data management tools for real-time monitoring/ecological momentary assessment data*. CRAN. <https://CRAN.R-project.org/package=EMAtools>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- 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>
- Leskinen, H. (1989). Tietoja sananalkuisten grafeemien ja grafeemikombinaatioiden yleisyyssuhteista. *Virittäjä*, 93(3), 401. <https://journal.fi/virittaja/article/view/38309>
- Long, J. A. (2020). *jtools: Analysis and presentation of social scientific data*. CRAN. <https://cran.r-project.org/package=jtools>
- Lüdecke, D. (2018). *sjPlot—data visualization for statistics in social science*. Zenodo. <https://doi.org/10.5281/ZENODO.1308157>
- Lüdecke, D., Makowski, D., & Waggoner, P. (2020). *Performance: Assessment of regression models performance*. CRAN. <https://CRAN.R-project.org/package=performance>
- Mardani, N., Jalilevand, N., Ebrahimipour, M., & Kamali, M. (2019). Clustering and switching strategies in verbal fluency tasks: Comparison between amyotrophic lateral sclerosis (als) and healthy controls. *Journal of Rehabilitation Sciences & Research*, 6(1), 21–26. <https://doi.org/10.30476/jrsr.2019.44718>
- Morris, J. C., Heyman, A., Mohs, R. C., Hughes, J. P., van Belle, G., Fillenbaum, G., Mellits, E. D., & Clark, C. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology*, 39(9), 1159–1159. <https://doi.org/10.1212/wnl.39.9.1159>
- Morais, A. S., Olsson, H., & Schooler, L. J. (2013). Mapping the structure of semantic memory. *Cognitive Science*, 37(1), 125–145. <https://doi.org/10.1111/cogs.12013>
- Oberg, G., & Ramírez, M. (2006). Cross-linguistic meta-analysis of phonological fluency: Normal performance across cultures. *International Journal of Psychology*, 41(5), 342–347. <https://doi.org/10.1080/00207590500345872>
- Olabarrieta-Landa, L., Torre, E. L., López-Mugartza, J. C., Bialystok, E., & Arango-Lasprilla, J. C. (2017). Verbal fluency tests: Developing a new model of administration and scoring for Spanish language. *NeuroRehabilitation*, 41(2), 539–565. <https://doi.org/10.3233/NRE-162102>
- Patra, A., Bose, A., & Marinis, T. (2020). Performance difference in verbal fluency in bilingual and monolingual speakers. *Bilingualism: Language and Cognition*, 23(1), 204–218. <https://doi.org/10.1017/S1366728918001098>
- Pekkala, S., Albert, M. L., Spiro, A., III, & Erkinjuntti, T. (2008). Perseveration in Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, 25(2), 109–114. <https://doi.org/10.1159/000112476>
- Pekkala, S., Goral, M., Hyun, J., Obler, L. K., Erkinjuntti, T., & Albert, M. L. (2009). Semantic verbal fluency in two contrasting languages. *Clinical Linguistics & Phonetics*, 23(6), 431–445. <https://doi.org/10.1080/02699200902839800>
- Pereira, A. H., Gonçalves, A. B., Holz, M., Gonçalves, H. A., Kochhann, R., Joaette, Y., Zimmermann, N., & Fonseca, R. P. (2018). Influence of age and education on the processing of clustering and switching in verbal fluency tasks. *Dementia & Neuropsychologia*, 12(4), 360–367. <https://doi.org/10.1590/1980-57642018dn12-040004>
- Quaranta, D., Caprara, A., Piccininni, C., Vita, M. G., Gainotti, G., & Marra, C. (2016). Standardization, clinical validation, and typicality norms of a new test assessing semantic verbal fluency. *Archives of Clinical Neuropsychology*, 31(5), 434–445. <https://doi.org/10.1093/arclin/acw034>
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Revelle, W. (2020). *Psych: Procedures for psychological, psychometric, and personality research*. CRAN. <https://CRAN.R-project.org/package=psych>
- Roberts, P. M., & Dorze, G. L. (1997). Semantic organization, strategy use, and productivity in bilingual semantic verbal fluency. *Brain and Language*, 59(3), 412–449. <https://doi.org/10.1006/brln.1997.1753>
- Ross, T. P. (2003). The reliability of cluster and switch scores for the Controlled Oral Word Association Test. *Archives of Clinical Neuropsychology*, 18(2), 153–164. <https://doi.org/10.1093/arclin/18.2.153>
- Rosselli, M., Ardila, A., Salvatierra, J., Marquez, M., Luis, M., & Weekes, V. A. (2002). A cross-linguistic comparison of verbal fluency tests. *The International Journal of Neuroscience*, 112(6), 759–776. <https://doi.org/10.1080/00207450290025752>
- Roy-O'Reilly, M., & McCullough, L. D. (2018). Age and sex are critical factors in ischemic stroke pathology. *Endocrinology*, 159(8), 3120–3131. <https://doi.org/10.1210/en.2018-00465>
- Santos Nogueira, D., Azevedo Reis, E., & Vieira, A. (2016). Verbal Fluency Tasks: Effects of age, gender, and education. *Folia Phoniatrica et Logopaedica*, 68(3), 124–133. <https://doi.org/10.1159/000450640>
- Scheuringer, A., Wittig, R., & Pletzer, B. (2017). Sex differences in verbal fluency: The role of strategies and instructions. *Cognitive Processing*, 18(4), 407–417. <https://doi.org/10.1007/s10339-017-0801-1>
- Schneider, W. J., & McGrew, K. S. (2018). The Cattell–Horn–Carroll theory of cognitive abilities. In *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–163). The Guilford Press.
- Shao, Z., Janse, E., Visser, K., & Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, 5, 772. <https://doi.org/10.3389/fpsyg.2014.00772>
- Strauss, E., Sherman, E. M. S., Spreen, O., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). Oxford University Press.
- Sung, K., Gordon, B., Yang, S., & Schretlen, D. J. (2013). Evidence of semantic clustering in letter-cued word retrieval. *Journal of Clinical and Experimental Neuropsychology*, 35(10), 1015–1023. <https://doi.org/10.1080/13803395.2013.845141>
- Suomi, K., Toivanen, J., & Ylitalo, R. (2006). *Fonetiikan ja suomen äänneopin perusteet*. Gaudeamus.
- 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, *Psychological Sciences and Social Sciences*, 75(9), e221–e230. <https://doi.org/10.1093/geronb/gbz003>
- Tallberg, I. M., Ivachova, E., Jones Tinghag, K., & Östberg, P. (2008). Swedish norms for word fluency tests: FAS, animals and verbs. *Scandinavian Journal of Psychology*, 49(5), 479–485. <https://doi.org/10.1111/j.1467-9450.2008.00653.x>
- Thiele, K., Quinting, J. M., & Stenneken, P. (2016). New ways to analyze word generation performance in brain injury: A systematic review and meta-analysis of additional performance measures. *Journal of Clinical and Experimental Neuropsychology*, 38(7), 764–781. <https://doi.org/10.1080/13803395.2016.1163327>
- Tombaugh, T. N., Kozak, J., & Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of Clinical Neuropsychology*, 14(2), 167–177.
- Tröger, J., Linz, N., König, A., Robert, P., Alexandersson, J., Peter, J., & Kray, J. (2019). Exploitation vs. exploration-computational temporal and semantic analysis explains semantic verbal fluency impairment in Alzheimer's disease. *Neuropsychologia*, 131, 53–61. <https://doi.org/10.1016/j.neuropsychologia.2019.05.007>
- 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)
- 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>
- Tyysteri, L. (2015). *Aamiaiskahvilasta öökkätarjontaan. Suomen kirjoitetun yleiskielen morfosyntaktisten yhdyssanarakenteiden produktiivisuus* [Unpublished doctoral dissertation]. University of Turku.
- Venegas, M. J., & Mansur, L. L. (2011). Verbal fluency: Effect of time on item generation. *Dementia & Neuropsychologia*, 5(2), 104–107. <https://doi.org/10.1590/S1980-57642011DN05020008>
- Vicente, S. G., Benito-Sánchez, I., Barbosa, F., Gaspar, N., Dores, A. R., Rivera, D., & Arango-Lasprilla, J. C. (2021). Normative data for verbal fluency and object naming tests in a sample of European Portuguese adult population. *Applied Neuropsychology: Adult*. Advance online publication. <https://doi.org/10.1080/23279095.2020.1868472>
- Weiss, E. M., Ragland, J. D., Bressinger, 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>
- Whiteside, D. M., Kealey, T., Semla, M., Luu, H., Rice, L., Basso, M. R., & Roper, B. (2016). Verbal fluency: Language or executive function measure? *Applied Neuropsychology: Adult*, 23(1), 29–34. <https://doi.org/10.1080/23279095.2015.1004574>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.
- Wickham, H. (2020). *tidyr: Tidy messy data*. CRAN. <https://CRAN.R-project.org/package=tidyr>
- Wickham, H., François, R., Henry, L., & Müller, K. (2019). *dplyr: A grammar of data manipulation*. CRAN. <https://CRAN.R-project.org/package=dplyr>