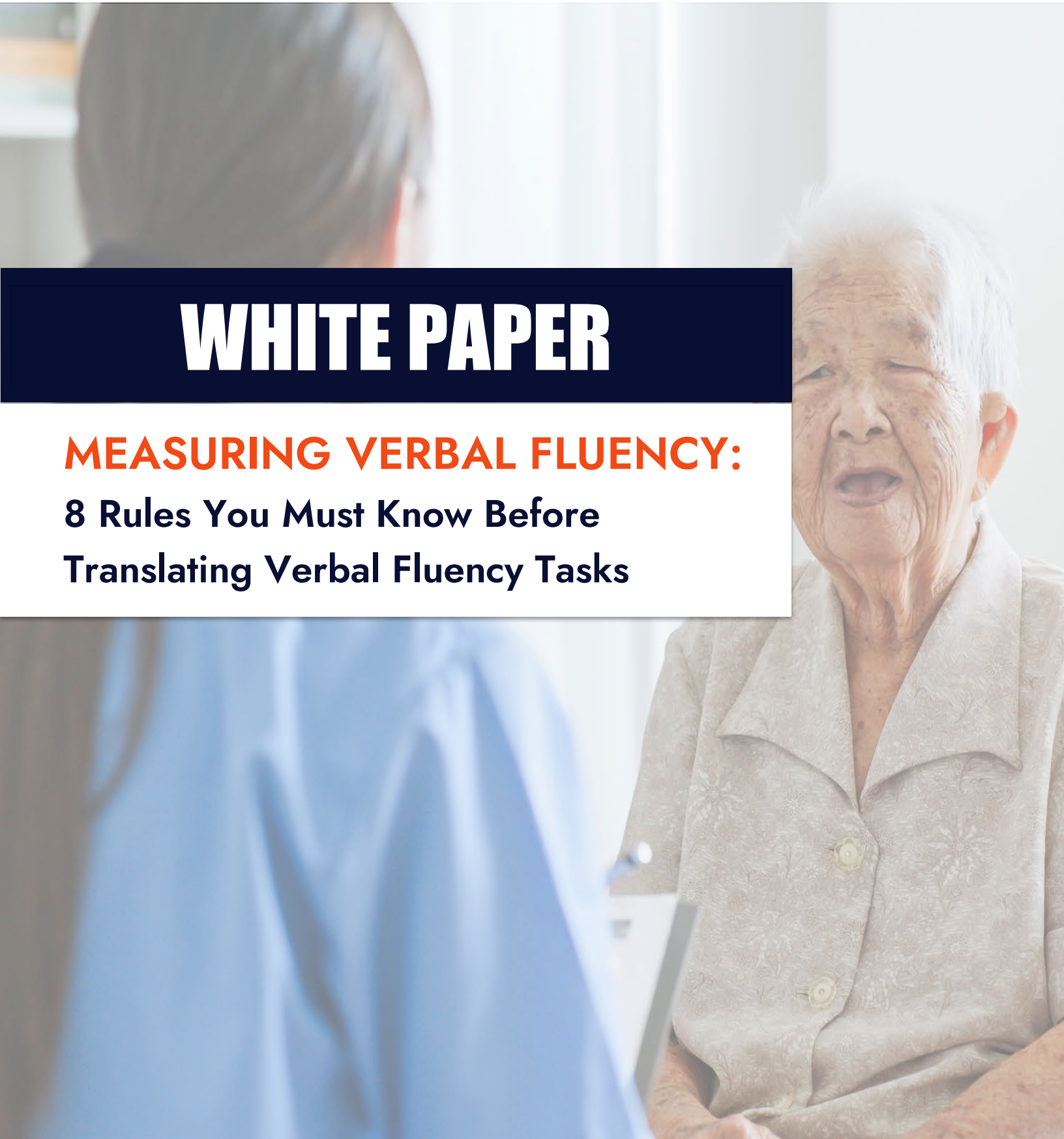
SANTIUM



WHITE PAPER

MEASURING VERBAL FLUENCY:

8 Rules You Must Know Before Translating Verbal Fluency Tasks

WHITE PAPER: TRANSLATING CLINICAL OUTCOME ASSESSMENTS

May 2026

MEASURING VERBAL FLUENCY: 8 RULES YOU MUST KNOW BEFORE TRANSLATING VERBAL FLUENCY TASKS

SANTIUM

Table of Contents

INTRODUCTION

Long before verbal fluency tasks became formal neuropsychological instruments, they were social category-word games. The most popular variant in the English-speaking world is *Scattegories*. In other countries, and in addition to *Scattegories*, the same concept games are sold as *Alto el lápiz* ("Pencils down"), *Tutti Frutti*, *Le Jeu du Bac* ("The Baccalaureate Game"), *Βρες το γράψτο* ("Find it, write it"), and so on. For centuries, the underlying verbal fluency game mechanic has been culturally widespread because the game exploits fundamental properties of lexical organization and semantic retrieval in a socially entertaining way.

Over time, clinicians realized that the game exposed something important about the brain: both letter (phonemic) fluency and category (semantic) fluency place entirely different structural demands on the brain. Some individuals could retrieve many words quickly and in an organized, strategic way. Others perseverated, stalled, repeated themselves, drifted outside the category or could not think of more than only a couple of words. Patients with frontal lobe damage, such as traumatic brain injury, often struggled with letter-based retrieval tasks that require sustained executive control, while patients with semantic memory disorders, like Alzheimer's disease, showed deterioration in category fluency.

The task was simple to administer, required no specialized equipment, and revealed measurable differences across neurological conditions. By the middle of the 20th century, verbal fluency tasks were used in formal neuropsychological assessment. Researchers standardized phonemic fluency procedures, such as the *Controlled Oral Word Association Test* (COWAT), commonly using the letters F, A, and S (for English speakers), while semantic fluency tasks, such as Animal Naming, became widely used in dementia and language production assessment.

Eventually, verbal fluency measures became embedded within larger clinical batteries, including the *Montreal Cognitive Assessment* (MoCA), the *Repeatable Battery for the Assessment of Neuropsychological Status* (RBANS), and numerous others that include only one type of the task instead of both.

Verbal fluency tasks appear to be deceptively simple to translate. Aside from straightforward instructions and routine scoring terminology, it is seemingly only about 3 alphabetic letters and a couple of low-register sentences. *This* is the translation trap! The ingenuity behind a super simple, but very (linguistically) fragile, measurement!

Verbal fluency tests are among the most structurally delicate instruments in cognitive assessment. The letters and semantic categories were selected based on a series of very specific design rules that are not universal to every language and culture. Hence, translating them without knowing the design rules will likely produce clinically unusable translations in many languages. And downstream, unusable data sets.

TYPES OF VERBAL FLUENCY TASKS

Verbal fluency tasks are generally divided into two major forms: phonemic fluency and semantic fluency. The distinction between phonemic and semantic fluency is clinically important because different neurological conditions impair them in different ways. Discrepancy scores, the performance gap between the two tasks, help clinicians differentiate and diagnose several major conditions.

PHONEMIC FLUENCY

Phonemic fluency tasks require participants to generate words beginning with a specific sound or letter. The classic English example is the F-A-S task, often referred to as the *Controlled Oral Word Association Test* (COWAT). Participants may be asked to produce as many words as possible beginning with *F* in sixty seconds, then repeat the process with A and S. The alternate letters are often used in subsequent testing sessions to reduce the participant's ability to memorize a group of words from session to session.

Say as many words as you can that begin with the letter F, in 60 seconds. Go!

Phonemic fluency places substantial demands on:

- Executive functioning – Higher-order cognitive processes involved in planning, rule-following, self-monitoring, inhibition, and strategic problem-solving.
- Strategic retrieval – The deliberate and organized search for information stored in memory to efficiently generate new words after the most immediately accessible responses are exhausted.
- Self-monitoring – The ability to track and evaluate one's own responses during a task (e.g., avoid repetitions, detect errors, and maintain adherence to the task rules).
- Inhibitory control – The ability to suppress automatic, repetitive, or inappropriate responses that violate the task rules.
- Working memory – The ability to temporarily hold and manipulate information during a task.
- Lexical search organization – The structured process of searching vocabulary stored in memory. Participants tend to organize retrieval by subcategories, sounds, meanings, or associations to quickly generate words.

The phonemic fluency task is highly unnatural because the brain does not file words alphabetically. It places substantial demand on the left dorsolateral prefrontal cortex and requires intense working memory management, systematic strategic searching, and sustained self-monitoring effort to avoid repetitions.

Relative to normed standards, weak performance in this task is commonly observed in participants with conditions such as ADHD, Parkinson's disease, frontotemporal dementia, vascular dementia, or left-hemisphere ischemic stroke.

SEMANTIC FLUENCY

Semantic fluency tasks rely on conceptual categories. Participants may be asked to name animals, fruits, musical instruments, occupations, tools, or supermarket items.

Name as many animals as you can in 60 seconds.
Go!

Semantic fluency relies more heavily on:

- Stored vocabulary – The ability to quickly and accurately retrieve known words from memory during speech or cognitive tasks.
- Semantic memory networks – Interconnected systems of stored meanings, concepts, and factual knowledge in long-term memory that help participants retrieve related words.
- Lexical-semantic organization – The way words and meanings are structured and connected within memory, which influences how efficiently participants can retrieve related concepts.
- Conceptual association – The mental linking of related ideas or concepts that helps participants navigate semantically related words.
- Category structure – The internal organization and boundaries of a conceptual group, such as animals or fruits. Some categories are broader, more familiar, or more interconnected than others.
- Stored world knowledge – General knowledge accumulated through experience and learning.

The brain naturally relies on automatic activation clustering, mainly through the left temporal lobe. Hence, semantic searching requires much less executive override from the frontal lobe because the words are already stored together in memory banks.

This area of the brain becomes significantly impaired with conditions such as semantic dementia, vascular dementia, dementia with Lewy Bodies, schizophrenia, Alzheimer's disease or Huntington's disease.

CORE RULES FOR DESIGNING VERBAL FLUENCY TASKS

RULE 1: LETTER SELECTION MUST BE EMPIRICALLY JUSTIFIED

Not all letters generate equivalent retrieval difficulty, especially across different languages.

The original letter selection process, developed by Dr. Arthur Benton for the *Controlled Oral Word Association Test* (COWAT), relies on distinct linguistic criteria:

THE GOLDILOCKS PRINCIPLE OF WORD FREQUENCY

The fundamental rule of letter selection is to normalize search-space density so that performance variance reflects pure executive function, rather than linguistic artifacts or automated semantic distribution.

Letters are selected based on the frequency with which they appear in the dictionary. In other words, the number of words in the dictionary that start with that specific letter.

Exclusion of Ultra-Rare Letters: In English, letters like X, Z, Q, or K have too few words. If a patient stalls out, a clinician cannot tell if it is due to a frontal lobe impairment or simply because the language lacks words starting with that letter.

Strategic Use of Hyper-Abundant Letters: In English, letters such as S or C produce extremely large quantities of words. To pass a phonemic test, the frontal lobe must enforce *sound-based* rules. For example, the letter C has two sounds in English: /s/ sound as in *city*, *ceiling*, or *cylinder*, and /k/ sound as in *cat*, *car* or *cold*. An impaired or under-stimulated frontal lobe will block out the /s/ sound words and miss half of the available lexical pool. Hence, high-density letters are used as "cognitive bait" because the prefrontal cortex must work twice as hard to suppress the brain's natural tendency to shift into semantic clustering and continue to follow the alphabetical task rule.

The "Just Right" Zone: Cues are selected from letters that fall into a moderate-frequency range (e.g., F, A, S or C, F, L), ensuring the task is challenging but entirely doable for a healthy brain.

Order of Progressively Increasing Difficulty

The classic F-A-S or C-F-L triads used with English speakers are intentionally arranged to scale up in cognitive demand.

The letters are ordered by associative difficulty. The first letter (e.g., F) provides a slightly easier baseline. The subsequent letters gradually require tighter executive constraints to force the frontal lobe to work harder in maintaining adherence to the task rules.

Rule Inhibition and Orthographic Triggers

Letters are also selected based on how easily they trigger rule violations, which helps clinicians measure a patient's inhibitory control, a core frontal lobe function. For example, the letter A is excellent for testing rule adherence because participants frequently attempt to list proper names (e.g., Alex, America), which are explicitly forbidden under standard test rules.

A damaged frontal lobe cannot inhibit these rapid, automatic associations, resulting in "intrusion errors" that provide critical diagnostic data.

Cross-Linguistic Equivalence

Considering the letter selection rules, this becomes significantly more complicated in translation. English triads, F-A-S and C-F-L, cannot be simply transferred for use in another language because letter and sound-based frequencies change dramatically from language to language.

The goal is to preserve an equal level of functional retrieval difficulty across languages spoken by patients whose cognitive functioning is being evaluated with this method. This is important for being able to conduct comparative analyses, especially when high-stakes data interpretability across borders is critical.

Researchers designing multilingual verbal fluency tasks, therefore, often evaluate:

- Word frequency distributions in the low register – The pattern of how commonly words appear in everyday, non-specialized language used by the general population.
- Corpus productivity – The frequency and diversity with which words or linguistic forms appear within a dictionary, or a large collection of real-world language data.
- Phonological accessibility in the target language – The ease with which speakers can access and produce words based on their sound structure within a particular language. Some sounds or letters naturally generate larger or smaller pools of accessible vocabulary.

- Morphological structure of candidate words starting with a specific letter – The way words are formed and modified through prefixes, suffixes, and grammatical changes. This is important because some languages can generate many related word forms from a single lexical root.
- Pilot normative and clinical performance – Preliminary testing conducted on healthy and clinical populations to evaluate how a verbal fluency task performs before full implementation. This helps identify unexpected difficulty differences, scoring issues, or cultural effects.
- Word retrieval patterns and volume based on the level of education – The tendency for word generation to vary according to educational background. Individuals with greater literacy exposure often produce larger word volumes and use more structured retrieval strategies.

Different languages, especially those coming from very different linguistic families than English, often require entirely different letter sets or testing mechanisms to maintain comparable cognitive load. Spanish distributes lexical frequency differently. Arabic introduces additional complexity because root morphology interacts with lexical generation in ways that differ substantially from Indo-European languages. Agglutinative languages such as Turkish or Finnish may artificially inflate or constrain production depending on how derivational forms of words are scored.

To maintain comparable cognitive workloads for participants taking the test in other languages, researchers use computerized lexical databases, such as CELEX or SUBTLEX, to calculate the exact frequency distributions within the target language.

Although some languages share the F-A-S triad with English, or a portion thereof, direct letter transfer across languages without prior literature review is fundamentally inappropriate. The default assumption must be that the task will need to be linguistically adapted until proven otherwise.

Using validated letter sets helps ensure that the ratio of automatic-to-controlled lexical search effort remains relatively equivalent and comparable across borders.

RULE 2: CATEGORIES MUST BE CULTURALLY RELEVANT AND FAMILIAR

Semantic categories must be familiar, semantically coherent, and broad enough to sustain retrieval during the testing interval. Animal Naming and Fruits remain one of the most widely used semantic fluency tasks because the categories are relatively universal and produce stable normative distributions across many populations.

Other categories may create cultural instability. A category such as “supermarket items” performs differently depending on:

- Urbanization
- Socioeconomic exposure
- Regional lifestyle
- Grocery shopping patterns and settings
- Cultural familiarity

Occupational categories can become particularly unreliable across generations and countries. Categories such as Tools and Kitchen Utensils rely on the participant's life experiences, hobbies, gender roles, or professional backgrounds and carry a sharp gender and occupational divide.

When category familiarity substantially differs across target populations, patients' score differences are more likely to reflect cultural differences, or the timing of exposure to local culture, rather than cognitive impairment. The consequence of this kind of misalignment is a possible misdiagnosis or prescribing an inappropriate treatment.

RULE 3: TIMING RULES MUST BE STANDARDIZED

Most verbal fluency tasks use a sixty-second response window. This convention is an intentional design feature. *Never* alter the duration of the task without testing the new time frame on a statistically significant number of participants. Five or ten participants are not reliable sample sizes for this purpose.

Research has shown that sixty seconds provides a useful balance between automatic retrieval and executive search demands. Shorter intervals may reduce sensitivity to change over multiple testing sessions, while longer intervals may introduce fatigue effects and alter retrieval dynamics and skew real scores.

Also, verbal fluency production is not consistent within that one-minute time frame. Participants usually generate words rapidly at the beginning of the task and progressively slow down toward the end of the minute. Early responses often reflect highly accessible lexical retrieval through words used with high frequency in daily life, while later responses require increasingly effortful search strategies.

Researchers sometimes analyze retrieval curves separately because different neurological conditions affect early and late retrieval in different ways.

RULE 4: SCORING PROCEDURES MUST BE EXPLICIT

Be specific about the scoring rules you want to follow. Check the test development documentation and ensure that your scoring rules are aligned with it. Do not change the scoring rules without a

statistically significant sample to confirm the validity and reliability of your desired scoring convention. Change tests, if necessary.

Common scoring decisions include:

- Whether proper nouns are allowed
- Whether repetitions count
- Whether inflected forms are accepted
- Whether derivational variants are accepted as distinct responses
- Whether corrections are permitted, and how quickly
- How intrusions are scored

For example, should "run," "runner," and "running" count as separate words?

Rules that apply to English may not necessarily apply to other languages with different lexical and grammatical structures. Different tests available on the market address these decisions based on how they were designed. It's important to become familiar with them because they will provide necessary insights to guide the drafting of adaptation instructions for researchers and translation teams.

Scoring inconsistencies caused by uninformed adaptation can produce substantial content validity and reliability problems, particularly in multicenter studies involving multiple raters.

RULE 5: CLUSTERING AND SWITCHING SHOULD BE CONSIDERED

In clinical practice, this should be phrased as "must be considered." However, when the target respondent group is highly specific in terms of, say, neurological profile, this rule may be simply specified in the protocol and applied to all participants. Whether this is the case is ultimately up to the discretion of researchers designing the study.

Clustering refers to producing semantically or phonologically related words in a sequence. *Switching* refers to jumping from cluster to cluster.

Traditional scoring procedures simply count the total number of correct responses, but modern verbal fluency analysis increasingly evaluates the organization of retrieval.

For example:

Cluster 1: dog, cat, hamster

Cluster 2: lion, tiger, leopard

The first cluster reflects domestic animals, while the second reflects large felines. Switching between clusters requires flexibility in executive functioning and strategic word search control.

Some neurological disorders impair switching more heavily than clustering, while others show the opposite pattern. This means that identical total scores can reflect very different retrieval strategies.

RULE 6: EFFECTS ACROSS EDUCATIONAL LEVELS MUST BE ADDRESSED

Education level strongly influences performance in verbal fluency.

Individuals with greater literacy exposure generally perform better on phonemic fluency tasks because these tasks rely more heavily on explicit lexical search strategies associated with both colloquial and formal language experience.

Semantic fluency is often somewhat more resistant to educational effects, although it remains influenced by vocabulary exposure and cultural knowledge.

For this reason, well-designed verbal fluency scores are interpreted relative to normative datasets stratified by:

- Age
- Education
- Language
- Cultural background
- Sometimes sex and regional variation, where gender differences and gender-specific roles are a result of social, socioeconomic and culturally patriarchal differences within a society

A score that appears impaired in one population may be entirely normal in another.

RULE 7: LINGUISTIC STRUCTURE CHANGES THE NATURE OF THE TASK

The linguistic structure of the target language directly affects the administration mechanism and scoring of the verbal fluency task. Translation alone is inadequate. The task must be adapted to local verbal fluency testing norms to be usable and functional, and clinically meaningful.

WORD FORMATION

Languages with extensive inflectional morphology require particularly careful scoring rules because participants may rapidly generate multiple grammatical forms from a single lexical root.

In highly synthetic languages, this can artificially increase production totals unless derivational boundaries are carefully defined.

English Example

An English-speaking participant using the letter F might produce:

- *farm*
- *farmer*
- *farming*

Most English-language protocols would consider these to be derivational variants rather than fully independent lexical items. However, in some cases, they could count as scorable items.

Turkish Example

In agglutinative languages such as Turkish, however, a single lexical root can generate a very large number of grammatically valid "surface" forms through suffixation.

Consider the root *ev* (house or home):

- ev (house)
- evler (houses)
- evlerimiz (our houses)

Without carefully and explicitly defined scoring rules, participants in certain countries may appear artificially more fluent simply because the language structure allows extensive morphological expansion.

Arabic Example

Arabic presents a different challenge. Arabic vocabulary is structured around consonantal root systems, and when the verbal fluency task is not adapted, phonemic retrieval interacts with morphology that works very differently from Indo-European languages. Participants may retrieve semantically related words through shared root structures in ways that can be more challenging for Arabic participants and not easily comparable to English scores.

Differences in Writing Systems

Orthography introduces additional complications. English and most Indo-European language phonemic fluency tasks are usually letter-based because orthography and phonology align relatively closely for initial sounds. In other languages, this relationship does not always exist.

Some researchers, therefore, prefer sound-based or syllabic rather than letter-based instructions depending on the linguistic system involved.

Languages with logographic writing systems, such as Chinese and Japanese, present a particularly important example because their writing systems are not alphabetic, but syllabic. Traditional letter-based phonemic fluency tasks are not directly transferable to non-alphabetic languages in the same way.

This illustrates a central principle of cognitive assessment adaptation, where the goal is to preserve the cognitive construct and its relevance to the target responders rather than the administration and scoring mechanism of the original verbal fluency task.

RULE 8: RATER INSTRUCTIONS MUST BE CONSISTENT AND EXPLICIT

Even small wording differences can alter participant behaviour in other languages. The clinical value of verbal fluency tasks depends on their structural consistency and psychometric sensitivity. Weak standardization damages that sensitivity, and translating verbal fluency tasks without understanding the design rules behind their construct may destroy the usefulness of the task in the target country.

The administration instructions must clearly define:

- Whether proper nouns are allowed
- Whether repetitions count
- Whether inflected forms are acceptable

- When timing begins
- Whether self-corrections are permitted
- Administrators must avoid cueing participants during administration
- When a discontinue rule can be applied

IMPLICATIONS FOR LINGUISTIC VALIDATION AND COGNITIVE ASSESSMENT

Verbal fluency tasks illustrate a broader problem in multilingual cognitive assessment. Many cognitive instruments appear deceptively simple to translate, yet their constructs are heavily dependent on the target language structure and its speakers' language-specific lexical retrieval processes.

Literal translation alone, especially an eloquently rephrased translation, cannot preserve measurement equivalence in certain languages.

A clinimetrically observant adaptation requires understanding of:

- Linguistic structure
- Cognitive task demands
- Cultural semantics
- Executive functioning requirements
- Measurement theory
- Administration methods and expected participant behaviour
- Verbal fluency test design rules

INSTRUCTING LANGUAGE SERVICE PROVIDERS

As multilingual research continues to expand, especially with an increased focus on the patient voice, the scientific importance of informed translations through linguistic and cultural adaptations also continues to increase.

The average translation service provider is not trained in measurement theory, especially in test-specific design methods. Those who have been translating clinical trial materials for several years certainly bring with them experience through trial and error, but the level of adaptation required for verbal fluency tasks is, in most cases and through no fault of their own, well beyond their awareness and expertise.

Translators are trained in generating translations that remain faithful to the meaning and flow naturally in the target language. They are not qualified to adapt clinical outcomes assessments to function equally in foreign languages and cultures. And they should not be expected to figure it out without explicit guidance from the sponsor. In principle, a literature review of this type is not an appropriate assignment for a translator, unless the translator is explicitly trained in linguistics in the target language and is made aware before the project that this will be a part of their scope of work.

It is ultimately the responsibility of the sponsor to conduct literature reviews, educate and inform their translation teams on the precise scope of work that is required to successfully adapt and validate a verbal fluency task in the languages they wish to use in studies.

In turn, language service providers have the ethical and moral responsibility to either decline the project when they do not have the competency to do this work, or if they accept it, to translate according to the instructions provided by the sponsor. A simply linguistically equivalent translation of a verbal fluency task without research into how verbal fluency is tested in the target language, regardless of language, cannot be an acceptable deliverable for use with real patients.

ABOUT SANTIUM

Santium supports the translation and cultural adaptation of neurocognitive assessments through scientifically and psychometrically informed translation, linguistic validation, evidence-based cognitive interviewing, and measurement-sensitive adaptation.

By combining expertise in clinical outcomes measures, clinimetrics, and multilingual research operations, Santium helps sponsors, researchers, and assessment developers preserve measurement integrity across more than 140 languages.

REFERENCES

AERA, APA, & NCME. (2014). Standards for educational and psychological testing. American Educational Research Association.

Ardila, A. (2005). Cultural values underlying psychometric cognitive testing. Neuropsychology Review, 15(4), 185–195.

Ardila, A., Ostrosky-Solís, F., & Bernal, B. (2006). Cognitive testing toward the future: The example of semantic verbal fluency. International Journal of Psychology, 41(5), 324–332.

Benton, A. L., Hamsher, K., & Sivan, A. B. (1994). Multilingual Aphasia Examination (3rd ed.). AJA Associates.

Borkowski, J. G., Benton, A. L., & Spreen, O. (1967). Word fluency and brain damage. Neuropsychologia, 5(2), 135–140.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). Delis-Kaplan Executive Function System (D-KEFS). Psychological Corporation.

Henry, J. D., & Crawford, J. R. (2004). A meta-analytic review of verbal fluency deficits in depression. Journal of Clinical and Experimental Neuropsychology, 27(1), 78–101.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). Neuropsychological Assessment (5th ed.). Oxford University Press.

Levelt, W. J. M. (1989). Speaking: From intention to articulation. MIT Press.

Loonstra, A. S., Tarlow, A. R., & Sellers, A. H. (2001). COWAT metanorms across age, education, and gender. Applied Neuropsychology, 8(3), 161–166.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). Handbook of normative data for neuropsychological assessment (2nd ed.). Oxford University Press.

Mungas, D., Reed, B. R., Crane, P. K., Haan, M. N., & González, H. (2004). Spanish and English neuropsychological assessment scales: Relationship to demographics, language, cognition, and independent function. Neuropsychology, 18(3), 466–475.

Paap, K. R., Johnson, H. A., & Sawi, O. (2014). Are bilingual advantages dependent upon specific tasks or specific bilingual experiences? Journal of Cognitive Psychology, 26(6), 615–639.

Radanovic, M., Diniz, B. S., Mirandez, R. M., Novaretti, T. M. S., Flacks, M. K., Yassuda, M. S., & Forlenza, O. V. (2009). Verbal fluency in the detection of mild cognitive impairment and Alzheimer's disease among Brazilian Portuguese speakers. International Psychogeriatrics, 21(6), 1081–1087.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary (3rd ed.). Oxford University Press.

Troyer, A. K. (2000). Normative data for clustering and switching on verbal fluency tasks. Journal of Clinical and Experimental Neuropsychology, 22(3), 370–378.

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. Neuropsychology, 11(1), 138–146.

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.

Translating Clinical Outcome Assessments:
Measuring Verbal Fluency – 8 Rules You Must Know Before Translating Verbal Fluency Tasks
Author: Monika Vance
May 2026

Santium Media Corporation
13-3120 Rutherford Road
Suite 339
Toronto, ON L4K 0B2
Canada

Worldwide Inquiries:
Phone: +1.833.726.8486
Email: info@santium.com

santium.com