Verbal fluency tests reliability in a Brazilian multicentric study, ELSA-Brasil

Valéria Maria de Azeredo Passos1, Luana Giatti1, Sandhi Maria Barreto1, Roberta Carvalho Figueiredo1, Paulo Caramelli1, Isabela Benseñor2, Maria de Jesus Mendes da Fonseca3, Nagela Valadão Cade4, Alessandra Carvalho Goulart5, Maria Angélica Nunes5, Márcia Guimarães de Mello Alves6, Ana Angélica Martins da Trindade7

ABSTRACT
Verbal Fluency Tests (VFT) are commonly used to assess executive functions and language in clinical and epidemiological studies. These tests were included to access cognitive function and predictors to its decline in ELSA-Brasil (Longitudinal Study of Adults’ Health), a cohort that investigates incidence and predictors of chronic diseases among 15,000 civil servants from six public educational institutions. **Objective:** To investigate the reliability of VFT scoring by supervisors of ELSA research centers, who independently judged 120 category (animals) and 120 phonemic (F letter) tests. **Method:** Their scores were compared to a reference standard score obtained by independent judgment of two experts. Intraclass correlation coefficient ratings reliability and Bland-Altman plot examined patterns of ratings disagreement. **Results:** Scores were very similar among ELSA centers and a high level of agreement was observed between each center and the reference standard. **Conclusion:** The high consistency of VFT scores confirms reliability and validity of the test and assures quality for its use in multicenter studies.

Key words: verbal fluency test, reliability, multicenter study, cohort.

Reprodutibilidade dos escores em testes de fluência verbal em estudo multicêntrico brasileiro

RESUMO
Introdução: Testes de fluência verbal (TFV) são muito utilizados para verificação de linguagem e função executiva em estudos clinico-epidemiológicos. Estes testes foram incluídos no ELSA-Brasil (Estudo Longitudinal de Saúde do Adulto), coorte que investiga incidência e fatores prognósticos de doenças crônicas em 15.000 servidores públicos de seis instituições de ensino. **Objetivo:** Investigar concordância dos escores dos TFV aferidos por supervisores dos centros de pesquisa, que interpretaram independentemente 120 TFV para categorias (animais) e 120 TFV fonêmica (letra F). **Método:** Os escores foram comparados a padrão-referência, definido por julgamento independente de dois especialistas. O índice de correlação intraclasse mediou concordância entre os escores e o gráfico de Bland-Altman espelhou padrões de discordância. **Resultados:** Os escores foram muito similares, com alto grau de concordância entre cada centro e o padrão-referência. **Conclusão:** A alta consistência dos escores revela importância de mecanismos de garantia de qualidade e normatização em estudos multicêntricos e permite ao ELSA avançar na investigação cognitiva.

Palavras-Chave: testes de fluência verbal, concordância, estudo multicêntrico, coorte.

1Faculdade de Medicina da Universidade Federal de Minas Gerais, Belo Horizonte MG, Brazil; 2Faculdade de Medicina da Universidade de São Paulo, São Paulo SP, Brazil; 3Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz, Rio de Janeiro RJ, Brazil; 4Universidade Federal do Espírito Santo, Vitória ES, Brasil; 5Faculdade de Medicina, Programa de Pós Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul, Porto Alegre RS, Brazil; 6Instituto de Saúde da Comunidade da Universidade Federal Fluminense, Niterói RJ, Brazil; 7Centro de Investigação ELSA-Bahia, Salvador BA, Brazil.
The pronounced and fast population ageing of low and middle-income countries is a reality. As a consequence, cognitive decline and dementia are now a leading cause of disability worldwide.

Cognitive tests are used as a proxy of intrinsic and complex neurological mechanisms. In this way, studies on cognitive assessment face a great range of challenges, from determining the most sensitive and specific diagnostic tests to verifying the applicability and comparability of results among different population and countries, with different languages, cultures and levels of education. Normative data by age, gender and educational level have been provided for some of these cognitive tests in Brazil, generally obtained by cross-sectional studies.

A major advantage of longitudinal over cross-sectional designs is that within-person cognitive changes can be directly measured. In Brazil, cognitive function and predictors to its decline is one of the interests of ELSA-Brasil (Escudo Longitudinal de Saúde do Adulto, Longitudinal Study of Adult’s Health), an ongoing cohort study designed to investigate the incidence and predictors of chronic non-communicable diseases, including cognitive decline and dementia, among 15,000 federal civil servants from six public educational and research institutions. Three previously validated cognitive tests to the Brazilian population were used in this study. The memory test from the Consortium to Establish a Registry for Alzheimer’s disease evaluates immediate and delayed recall of a list of 10 words and the score is based on the total number of recalled words in each trial. In the Trail Making Test B, executive function is evaluated by asking the participants to sequentially connect 26 circles containing numbers and letters; and the total score is given by time spent to correctly complete the test. Participants are also asked to perform semantic and phonemic Verbal Fluency Tests (VFT), where they have to say as many words as possible from a previous specified semantic category (animals) or beginning with a given letter (F) within one minute.

The performance in VFT needs to be evaluated by a trained investigator, in order to count the words excluding repetitions and intrusions. Besides the psychometric properties (reliability and validity) of these tests, multicenter studies also have to concern about the homogeneity of their interpretation. The present study intends to assess the interpreter reliability of VFT scoring by six different supervisors of the ELSA research centers.

**METHOD**

In order to obtain the quality assurance of ELSA cognitive tests, all efforts were made before data collection. A written manual with a detailed description of the procedures was used to train and certify the research staff. The feasibility and efficiency of the procedures were carried out through pretests and a pilot study.

During data collection, the participant responses to VFT were taped and written down. Periodic supervision and evaluation of the interviewers’ performance were done, in order to maintain quality control. A written manual was provided with all items that would be considered as errors: intrusions (inappropriate words for the given category or word) and perseverations (same word repeated) and general category errors (when a participant said a general and also a specific item, e.g. bird and nightingale). For this study, the six ELSA supervisors independently judged 120 category (animals) and 120 phonemic (F letter) tests, 40 from each center. The results of the scores were compared to a reference standard score obtained by two experts, a geriatrician (VMAP) and a neurologist (PC).

Data entry was carried out using the EPIDATA 3.1 software. The statistical analysis was carried out through the STATA version 10.0 software. The Intraclass Correlation Coefficient (ICC) was used to measure the reliability of ratings and Bland-Altman plot was included to examine the patterns of ratings disagreement.

The study was approved by the Institutional Review Board of the Federal University of Pelotas.

### Table

<table>
<thead>
<tr>
<th>ELSA Centers</th>
<th>Category Verbal Fluency Test (Animals)</th>
<th>Phonemic Verbal Fluency Test (F letter)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percentiles</td>
<td>ICC <em>(CI 95%)</em></td>
</tr>
<tr>
<td></td>
<td>25 50 75</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>15 18.5 22</td>
<td>0.982 (0.975-0.988)</td>
</tr>
<tr>
<td>2</td>
<td>14.5 18 22</td>
<td>0.979 (0.970-0.985)</td>
</tr>
<tr>
<td>3</td>
<td>14.5 18 22</td>
<td>0.987 (0.981-0.991)</td>
</tr>
<tr>
<td>4</td>
<td>15 18 22</td>
<td>0.986 (0.979-0.990)</td>
</tr>
<tr>
<td>5</td>
<td>15 18 22</td>
<td>0.980 (0.971-0.986)</td>
</tr>
<tr>
<td>6</td>
<td>14.5 18 22.5</td>
<td>0.986 (0.980-0.990)</td>
</tr>
</tbody>
</table>

ICC: intraclass correlation coefficient.
Board of all institutions participating in the study and all participants signed an informed consent.

RESULTS
Scores of VFT were very similar among ELSA centers and a high level of agreement was observed between each of the six centers and the reference standard (Table). Data were also analyzed using the Bland-Altman plot (Figure). The mean difference is small for both tests, and 95% of scores are located between the mean and two standard deviations. However, it can be observed a small underestimation of the category scores determined by supervisors.

DISCUSSION
VFT are commonly used to assess executive functions and language in clinical and epidemiological studies. Previous Brazilian studies have investigated normative, reliability and validity issues but, to the best of our knowledge, this is the first study to evaluate interrater agreement.

Most measurements in human sciences involve measurement error, but judgments made by humans are especially plagued by this problem. Since measurement error can affect analysis and interpretation of data, the validity of the inferences made from results of observational studies depends on the accuracy of its methods and procedures.

The high consistency of VFT scores among the ELSA research centers shows the importance of the implemented measures of quality assurance and control. Besides that, this result allows ELSA cognitive investigation to proceed upon its main objective, to identify social, psychological and biological predictors of cognitive decline in a Brazilian adult population. Its urges all multicentric studies to evaluate interrater agreement in order to guarantee the quality of data analyze.

REFERENCES