# VALIDITY CONVERGENT AND RELIABILITY TEST-RETEST OF THE REY AUDITORY VERBAL LEARNING TEST

Sabrina de Sousa Magalhães, Leandro Fernandes Malloy-Diniz, Amer Cavalheiro Hamdan


## Abstract

*Objective:* This study provides evidence for the validity and reliability of the Rey Auditory Verbal Learning Test (RAVLT).

*Method:* The reliability was measured by internal consistency and the test-retest method with a mean interval of 35 days. To determine the convergent and divergent validity, it was compared the performance on the RAVLT with the Benton Visual Retention Test (BVRT) and the Trail Making Test (TMT), respectively. The test was been taken by 34 healthy participants of both genders, ages 17 to 40 and 11.2±0.7 years of education.

*Results:* All test-retest correlation coefficients achieved significance, ranging between 0.36 and 0.68. The A2 measure obtained the weakest correlation ($r$ =0.28), and the sum of A1-A5 ($r$ =0.68) was the strongest. The rest of the measures obtained moderate correlations. The value of the Cronbach's Alpha coefficient was 0.80. The two RAVLT measures, the sum of A1-A5 and A7, did not significantly correlate with the TMT measures. In contrast, these measures did exhibit significant but modest correlations with the measures of BVRT (ranging from 0.37 to 0.44).

*Conclusions:* The results had adequate divergent and convergent validity and good reliability in terms of internal consistency. The evidence collected in this study indicates that RAVLT is a valid and reliable psychometric instrument in neuropsychological assessment.


**Key words:** neuropsychological tests, psychometrics, reliability, validity of tests

**Declaration of interest**: none

Sabrina de Sousa Magalhães – Department of Psychology, Federal University of Paraná, Paraná, Brazil
Leandro Fernandes Malloy-Diniz – Laboratory of Neuropsychological Investigations – National Institute of Science and Technology of Molecular Medicine, Federal University of Minas Gerais, Minas Gerais, Brazil and Department of Psychology, Federal University of Minas Gerais, Minas Gerais, Brazil
Amer Cavalheiro Hamdan – Department of Psychology, Federal University of Paraná, Paraná, Brazil

**Corresponding author**:
Amer Cavalheiro Hamdan, UFPR – Department of Psychology
Praça Santos Andrade, 50, Mestrado em Psicologia CEP: 80020-240, Curitiba, Paraná, Brazil
E-mail address: achamdan@ufpr.br

## Introduction

In the context of neuropsychological assessment, there are two fundamental psychometric properties to legitimate the efficiency of an instrument: validity and reliability. Validity refers to what the test measures, specifically the degree to which it actually measures what it allegedly measures, according to empirical verifications. Reliability is not about what is being measured but whether the measure given by the instrument is consistent. In other words, if the test shows constancy in the scores for the same individual, the test corresponds to an estimative of measurement error (Mitrushina et al. 2005). The factors of reliability are listed as follows: consistency (the test provides equal or similar scores for the same individual); the ability to replicate (within a certain margin of error, the test can be refused and repeated); and confidence (the test presents the same results to the same individual) (Hogan 2007).

The Rey Auditory Verbal Learning Test (RAVLT) is an instrument used to evaluate episodic declarative memory, immediate memory span, verbal learning, susceptibility to proactive and retroactive interferences, retention of information, recall and memory recognition (Lezak et al. 2004, Strauss et al. 2006). In Brazil, Malloy-Diniz et al. (2007) developed an RAVLT version composed of a list of high frequency, one or two-syllable Portuguese substantives, which served as stimuli to assess the performance of Brazilian adults and senior citizens. The demographic variables that commonly influence the performance on the RAVLT are gender, educational level and age (Magalhães and Hamdan 2010). Normative data are available (Malloy-Diniz et al. 2007, Magalhães and Hamdan 2010), however evidence regarding its psychometric properties is still preliminary.

There is evidence that the RAVLT has construct validity when compared with other measures of verbal learning and memory, such as the California Verbal

Submitted September 2011, Accepted July 2012

Learning Test (Stallings et al. 1995). Several studies conducted in Brazil also presented evidence of RAVTL construct validity. For example, Fichman et al. (2010) showed that some of the RAVLT indices present significant and positive correlations with a test of visual learning. De Paula et al. (2012) demonstrated that the RAVLT has a bifactorial structure, which is related to the processes of verbal learning and retrieval. Nonetheless, despite the existence of a few studies on RAVLT construct validity in the elderly Brazilian population, we found no studies that evaluated the reliability properties or construct validity of this test in other age groups.

The main objective of this study was to assess the psychometric properties of the reliability and validity of the RAVLT. The additional objectives are listed as follows: (a) to evaluate the confidence of the RAVLT scores by comparing the performance in two distinct experimental situations (test-retest); (b) to analyze the internal consistency reliability of the instrument; (c) to appraise the criterion validity of the RAVLT by comparing it with other neuropsychological tests to confirm that the test evaluates memory and not another cognitive function. This study aims to contribute evidence for the applicability of RAVLT by determining whether the Brazilian adaptation is a valid psychometric instrument in neuropsychological assessment.

## Method

### *1. Participants*

The participants were 34 male and female undergraduates between 17 to 40 years in age and were recruited by requests for volunteer participation during visits to undergraduate psychology classes. All participants signed two copies of the Volunteer Participation Consent and Confidentiality Terms. The research was conducted according to the Helsinki Declaration of ethical principles.

The research used the following exclusion criteria to select only healthy individuals: a) a history or presence of psychiatric disturbances, b) diabetes, c) heart problems or any related pathologies, d) past or current use of psychoactive drugs, especially those with already related side effects that disturb mnemonic functions, and e) abuse of illicit substances according to the Brazilian law. These criteria were investigated through the "Questionnaire of demographic data, health and cultural conditions" and the Mini Mental State Exam (MMSE). The inclusion criterion for participation in the MMSE sample was a score of at least 24 points.

**Table 1** presents the demographic data from the sample. The study included 34 participants of bother genders (35% men and 65% women) with an age range between 17 and 40 (mean 20.7±4.5). The education variable was computed using the number of years the subject studied according to the Brazilian school system. All participants had reached at least the high school level, and the average education time varied from 11 to 15 years of formal study, including repetitions of the same grade. The MMSE score varied from 25 to 30 points, with the mean score of 29.1±1.1 points.

### *2. Instruments*

The purpose and procedures of the instruments are listed as follows:

1) The "Questionnaire on demographic data, health and cultural conditions" was used to gather general personal, cultural and health information about the individual, thus gathering data for the sample characterization.

2) The Mini Mental State Exam (MMSE) screened for cognitive impairment. The MMSE is a useful screening instrument to estimate the level of performance of individuals aged 18 to 85+ (Strauss et al. 2006) because it is highly specific (Bertolucci et al. 1994, Hancock and Larner 2011, Milian et al. 2011) and allows for reliable detection of cognitive impairment among healthy individuals within a modest margin of error, although it does not provide a differential diagnosis. A cut-off of 24 points was used for our Brazilian sample.

**Table 1.** *Demographic characterization of the sample (n = 34)*

| | **Mean (SD)** | **CI 95%** | |
|---|---|---|---|
| | | **Minimum** | **Maximum** |
| **Age (years)** | 20.7 (4.5) | 19.1 | 22.3 |
| **Gender (Male/Female)** | 12 / 22 | - | - |
| **%** | 35.3 / 64.7 | - | - |
| **Education level (years)** | 11.2 (0.7) | 10.9 | 11.4 |
| **MMSE score** | 29.1 (1.1) | 28.7 | 29.5 |

Note: SD = standard deviation; CI = confidence interval; MMSE = Mini Mental State Exam.

3) A Portuguese language version of the Rey Auditory Verbal Learning Test (RAVLT) was used (Malloy Diniz et al. 2007). The test consists of five consecutive oral presentations of 15 concrete nouns (List A), followed by a free-recall after each presentation (trials A1, A2, A3, A4 and A5). An interference list (List B) with 15 different concrete nouns is presented, followed by an immediate recall of this new list (trial B1). Next, the recall of List A post-interference takes place (trial A6). After a 20-minute interval, a delayed recall of List A is required (trial A7). The last trial consists of an oral presentation of a 50-noun list, which is composed by List A and B and 20 phonologically or semantically similar words to the previous lists, and then recognition of the words in List A. The score for each trial corresponds to the number of correctly recalled words (Strauss et al. 2006). The rates of proactive interference (B1/A1) and retroactive interference (A6/A5) are also calculated; the rates represent the susceptibility to previous and later activities, respectively, involving the presentation of the content about to be recalled. Proactive interference can be observed when the subject's recall of List B is affected by previously learning List A. Retroactive interference occurs when the retention or recall of the post-distractor list is affected by learning the distractor list, which is assumed to confuse one's memory of the learned list. If the interference ratio is equal to 1, no interference effect is observed; however, if it is less than 1, interference is demonstrated. Facilitation instead of interference can be verified if the interference ratio is greater than 1 (Geffen and Geffen 2000).

4) The Trail Making Test (TMT) evaluates executive function and is composed by two activities. In part A (TMTA), the participant is asked to connect, in ascending order, 25 numbered circles randomly arranged on the page. In part B (TMTB), the participant is required to alternate circles of numbers and letters (numbers in crescent order, letters alphabetically). The score of the test reflects the time (in seconds) taken to complete the cycle (Hamdan and Hamdan 2009, Strauss et al. 2006).

5) The Benton Visual Retention Test (BVRT) evaluates immediate and delayed visual memory. Using form C, the test was composed of 10 cards, each featuring a complex figure. On the first two cards, the figure is composed of simple geometric forms; the eight other cards features two large geometric forms with a small geometric figure located in the cards' peripheries (Strauss et al. 2006). In the immediate recall, the participant must draw a replica, as exact as possible, of the figure on the card after a period of 10 seconds of studying the card with no opportunity to consult it afterward. The procedure is repeated for all 10 cards. After a time interval usually of 10 minutes, the participant is requested to draw all the cards he or she still remembers, in any order, with no opportunity to consult the cards. Each drawing is scored with 0, 1 or 2 points. The maximum score is given to an identical reproduction; the score of 1 to those with at least two correct components; and a score of zero to drawings that do not have sufficient components from the original card (and do not attempt to earn the intermediate score) or contain more mistakes than what is acceptable. Afterward, these scores were added, and the immediate (Be1) and delayed (Be2) recall scores were obtained (maximum of 20 points).

## *3. Design*

All participants completed all instruments. Each test was administered to the same group of participants twice to evaluate the score stability over time (test-retest reliability). The present study contained a mean interval of 35 days±8.9 between the two test sessions. The interval length was 31 days at minimum and 38 days at maximum.

The equivalence of the test scores across two administrations and between all tests was first evaluated prior to the assessment of scale reliability. Dependent *t* tests were performed for the RAVLT if the means calculated from the same form from both testing sessions were significantly different. Internal consistency reliability was estimated by intraclass correlation coefficients (ICC) and coefficient a. The *k*-sample significance test for independent coefficients proposed by Hakstian and Whalen (1976) was adopted to statistically evaluate the effects of the number of items on coefficient a. The probability of making a Type I error on each test was constrained at .01.

To access criterion validity, the two most reliable RAVLT measures were chosen *a posteriori*, which were the Sum A1-A5 and A7. Convergent validity was estimated comparing those RAVLT items with the scores on the Benton Visual Retention Test (BVRT), which is also an instrument for episodic memory assessment. Although the BVRT was developed for visual memory assessment, some authors have argued that this test assesses both visual and verbal memory because its items (geometrical forms) can be verbalized (Strauss et al. 2006). For divergent validity, the same RAVLT items were compared with two measures of the TMT, specifically, the TMTA, which demands mainly attentional resources and the TMTB, which includes an executive function component.

The first data collection began after the presentation, reading, and discussion of possible doubts about the Volunteer Participation Consent and Confidentiality Terms, which was followed by the "Questionnaire on demographic data, health and cultural conditions" and the MMSE. After these screening instruments, the RAVLT was administered. During the twenty-minute interval necessary to collect data for the A7 trial and Recognition, two other neuropsychological tests, the TMT and immediate recall of the BVRT, were given. Both instruments are assumed to have no influence on the RAVLT scores because they are nonverbal tasks and do not demand semantic processing. The main concern here was to provide distractor tasks without confounding the stimulus and the processing demand of each task, in particular.

Immediately after the Recognition trial of the RAVLT, the late recall of the BVRT was administered. In the retest session, the demographic questionnaire and MMSE were not applied, and the session began with the participant's permission and the administration of the RAVLT. All experimental sessions lasted approximately 50 minutes, and all activities were completed in a single session.

## Results

**Table 2** presents the performance on the RAVLT (test and retest session), Intraclass Correlation Coefficient (ICC) and standard error of the mean (SEM). The correlations varied from 0.36 to 0.68. The weakest correlation was found in the A2 measure, and the strongest was found in the Sum A1-A5. The indexes A1, A5, B1 and A7 obtained medium correlation, varying between 0.30 and 0.49. The A3, A4, Sum A1-A5, A6 and Recognition reached a high correlation rate. The proactive and retroactive interference did not present any significant correlation between the two sessions of testing. The susceptibility to interference was small, except for those related to the expressive proactive interference in the retest.

The paired t-test was performed to compare the differences between the mean scores in both situations (test-retest) and verify the presence of any disparity between them. The results are listed as follows: A1: t(33)=-9.13, p<0.001, d=3.18; A2: t(33)=-6.16, p<=0.001, d=2.15; A3: t(33)=-6.54, p=<0.001, d=2.28; A4: t(33)=-3.01, p<0.001, d=1.05; A5: t(33)=-4.13, p<0.001, d=1.43; Sum A1 to A5: t(33)=-10.48, p<0.001, d=3.65; B1/A1: t(33)=6.83, p<0.001, d=2.38; A6/A1: t(33)=-0.70, p=0.4875, d=0.24; B1: t(33)=0.58, p=0.5641, d=0.20; A6: t(33)= -4.61, p<0.001, d= 1.60 and A7: t(33)= -3.14, p<0.001, d= 1.09. Measures B1 and Recognition were the only ones that did not differ in both situations (p>0,01); all the others demonstrate significant differences in their results from both experimental situations. The *d* measures reached small effect sizes (Cohen 1988).

The F test for ICC are listed as follows: A1: F(1, 33)=83.39, p<0.001; A2: F(1, 33)=37.99, p<0.001; A3: F(1, 33)= 42.73, p<0.001; A4: F(1, 33)= 9.08, p=0.0049; A5: F(1, 33)=17.03, p<0.001; Sum A1 to A5: F(1, 33)= 109.76, p<0.001; B1/A1: 46.71, p<0.001; A6/A1: F(1, 33)= 0.49, p=0.488; B1: F(1, 33)= 0.34, p=0.564; A6: F(1, 33)= 21.21, p<0.001 and A7: F(1, 33)= 9.85, p=0.0035.

The RAVLT measures analyzed with the Cronbach's Alpha Coefficient Reliability, deleting each item in turn, are indicated in **table 3**. The alpha reliability of the RAVLT was 0.84, and the standardized alpha was 0.88. The test condition coefficients varied from 0.78 to 0.82, and the retests varied between 0.81 and 0.86. **Table 4** shows the Pearson's correlation between the

**Table 2.** *Performance on the RAVLT (test and retest session), Intraclass Correlation Coefficient (ICC) and SEM*

| | Test<br>Mean (SD) | Retest<br>Mean (SD) | *ICC* | SEM |
|---|---|---|---|---|
| A1 | 7,2 (1,5) | 10,2 (2,1) | 0.71 | 1.38 |
| A2 | 10,2 (2,1) | 12,5 (1,7) | 0.52 | 1.56 |
| A3 | 11,9 (1,5) | 13,6 (1,4) | 0.55 | 1 |
| A4 | 12,9 (1,6) | 13,5 (1,3) | 0.19 | 0.89 |
| A5 | 12,8 (1,5) | 13,8 (1,2) | 0.32 | 1.06 |
| Sum | 54,9 (6,2) | 63,6 (5,9) | 0.76 | 3.42 |
| B1/A1 | 0,98 (0,3) | 0,66 (0,2) | 0.57 | 0.19 |
| A6/A5 | 0,95 (0,1) | 0,97 (0,1) | -0.02 | 0.11 |
| B1 | 6,8 (1,6) | 6,6 (1,7) | -0.02 | 1.25 |
| A6 | 12,1 (1,8) | 13,3 (1,5) | 0.37 | 1.13 |
| A7 | 12,2 (1,8) | 13,1 (1,6) | 0.21 | 1.20 |
| Rec | 14,6 (0,7) | 14,6 (0,9) | -0,01 | 0.52 |

Note: SD = standard deviation; ICC = Intraclass Correlation Coefficient; SEM = standard error of the mean; Sum = sum of scores across the five acquisition trials (A1–A5); B1/A1 = proactive interference; A6/A5 = retroactive interference; Rec = recognition trial.

**Table 3.** *Cronbach's Alpha Coefficient for the RAVLT measures in the two experimental conditions*

| | Test | | | Retest | | |
|---|---|---|---|---|---|---|
| | **Alpha** | **Sd.Alpha** | **r(item, total)** | **Alpha** | **Sd.Alpha** | **r(item, total)** |
| A1 | 0.79 | 0.83 | 0.67 | 0.82 | 0.88 | 0.61 |
| A2 | 0.79 | 0.83 | 0.60 | 0.81 | 0.86 | 0.75 |
| A3 | 0.78 | 0.81 | 0.78 | 0.81 | 0.85 | 0.85 |
| A4 | 0.79 | 0.82 | 0.68 | 0.82 | 0.86 | 0.72 |
| A5 | 0.79 | 0.83 | 0.66 | 0.82 | 0.87 | 0.66 |
| Sum | 0.82 | 0.80 | 0.94 | 0.86 | 0.85 | 0.96 |
| B1/A1 | 0.82 | 0.88 | -0.23 | 0.83 | 0.87 | 0.47 |
| A6/A5 | 0.82 | 0.84 | 0.26 | 0.84 | 0.90 | 0.02 |
| B1 | 0.81 | 0.83 | 0.32 | 0.85 | 0.89 | 0.15 |
| A6 | 0.78 | 0.81 | 0.73 | 0.82 | 0.86 | 0.68 |
| A7 | 0.79 | 0.82 | 0.62 | 0.81 | 0.86 | 0.76 |
| Rec | 0.81 | 0.84 | 0.35 | 0.83 | 0.87 | 0.61 |

Note: Sd.Alpha = Standard Alpha; Sum = sum of scores across the five acquisition trials (A1 – A5);; B1/A1 = proactive interference; A6/A5 = retroactive interference; A7 = RAVLT delayed recall measure; Rec = recognition trial.

**Table 4.** *Validity coefficients between the RAVLT, Trail Making Test and Benton Visual Retention Test*

| | Sum | A7 | B1/A1 | A6/A5 | TTA | TTB | Be1 |
|---|---|---|---|---|---|---|---|
| Sum | - | - | | | - | - | - |
| A7 | 0,57** | - | | | - | - | - |
| B1/A1 | -0,33 | -0,18 | - | | - | | |
| A6/A5 | 0,12 | 0,45** | 0,24 | - | | - | |
| TTA | -0,03 | 0,01 | 0,22 | 0,02 | - | - | - |
| TTB | -0,19 | -0,15 | 0,22 | 0,01 | 0,68** | - | - |
| Be1 | 0,44** | 0,27 | -0,12 | -0,04 | -0,31 | -0,58** | - |
| Be2 | 0,37* | 0,39* | -0,23 | 0,02 | 0,01 | -0,13 | 0,31 |

Note: Sum = sum of scores across the five acquisition trials (A1 – A5); A7 = RAVLT delayed recall measure; B1/A1 = proactive interference; A6/A5 = retroactive interference; TTA = part A score from TMT; TTB = part B score from TMT; Be1 = immediate recall score from BVRT; Be2 = delayed recall score from BVRT; * = p<0,05; ** = p<0,01.

RAVLT, the TMT and the BVRT; these indices represent the measurement of divergent and convergent validity, respectively. Once the BVRT provided the immediate and delayed memory rates, they were compared with equivalent measures on the RAVLT, specifically Sum A1-A5 (the most reliable between the immediate recall indices) and A7. The measures of susceptibility to proactive and retroactive interferences usually correlate with measures of executive function (Mitrushina et al. 2005) and attention; therefore, they were also included in the analysis.

The validity coefficient between Sum A1-A5 and A7 of the RAVLT and the two TMT measures demonstrated very weak and non-significant correlations, which is evidence that they do not evaluate the same function. None of the interference measures of the RAVLT reached significant correlation with the TMT measures. In contrast, regarding the BVRT measures, the RAVLT mostly exhibited validity coefficients of medium magnitude. The immediate memory measures of the tests (Sum A1-A5 and Be1) achieved a 0.44 correlation; the delayed memory measures of both tests (A7 and Be2) reached a correlation of 0.39. The Sum A1-A5 also demonstrated a positive correlation of 0.37 with the Be2 measure but did not verify a significant correlation between the delayed memory recall of RAVLT (A7) and the immediate memory BVRT measure (Be1).

## Discussion

This study provides evidence for the validity and reliability of the RAVLT. Some factors and conditions are described (Hogan 2007) as non-reliable sources that affect the measurements of the test and bring it below a reliable level. Among the numerous sources that result in non-systematic variations in test scores, three are highlighted: (a) different criteria of test correction adopted by the person that grades the test; (b) changes in test application procedure; (c) the personal condition of the examinees, i.e., the participant's temporary self-state (characteristics not related to the ones under evaluation, such as health and mood conditions, willingness and level of engagement), which might influence the participant's performance. This study aims to contribute evidence for the applicability of the RAVLT by determining whether this Brazilian adaptation is a valid psychometric instrument in neuropsychological assessment.

The RAVLT includes clear and specific instructions, which minimize possible variations between different researchers during the application and correction process. Thus, this study aimed to neutralize this non-reliable source. Procedures were standardized, and all examiners were trained in the test application policies related to physical environment preparation, instructions and test administration. The objective was to keep all these factors constant to prevent them from interfering with both test situations. It has been reported that the examiner has an important role in participant performance, and his/her preparation for the administration of the neuropsychological instruments should not be neglected (Van den Burg and Kingma 1999). Finally, the third source of non-systematic error was verified through test-retest methodology.

The weakest reliability correlation was observed in the A2 measure ($r$ = 0.36), and the strongest was found in the Sum A1-A5 ($r$ = 0.68). Researchers who have studied the psychometric properties of the RAVLT in different language versions all find the Sum A1-A5 as the most reliable measure of the instrument, although with different coefficients: $r$ = 0.77 (Geffen and Geffen 2000), $r$ = 0.79 (Knight et al. 2007) e $r$ = 0.70 (Van den Burg and Kingma 1999). This finding was expected because the reliability is proportional to the test extension. Nevertheless, these data are not unanimous. In another study, although the Sum A1-A5 demonstrated a significant correlation of $r$ = 0.55, between the test and retest session, the higher correlation rate was found in the A1 trial ($r$ = 0.69) (Rezvanfard et al. 2011).

The second highest correlation rate obtained in the present study was for the A4 measure ($r$ = 0.64), followed by the Recognition ($r$ = 0.59) and the A3 measure ($r$ = 0.53). Evidence indicates that subsequent trials of the RAVLT are usually more reliable than the first ones (Van den Burg & Kingma, 1999). The exception found in this study was the A5 measure, which was expected to reach a high correlation rate but obtained an average score of ($r$ = 0.41). In contrast, another study obtained a 0.61 rate with this same trial (Van den Burg and Kingma 1999).

The proactive and retroactive interference rates were quantified by the ratio between the two analyzed scores instead of subtraction. This method (Geffen and Geffen 2000) takes into account the different performance levels of the participants. The results showed no correlation between these rates in both experimental situations. Even when alternate forms of the test were used, a correlation between those interference measures was not reported (Rezvanfard et al. 2011). Thereby, the data suggest that the measures of higher reliability and stability levels are Sum A1-A5, A3, A4, A6 and Recognition because they obtained high correlation rates. Other measures pointed to moderate reliability.

The RAVLT exhibited two measures, B1 and Recognition, which showed no significant differences between both test situations. It is possible to infer that the learning process did not influence these two measures; therefore, their original results were maintained. After List B is shown one time in the first session, we can assume that it retains its novelty. In the case of Recognition, this measure may be influenced by the ceiling effect because the mean of the scores is lower than one standard deviation of the maximum score (Uttl 2005). The Recognition task may not serve as a good measure of mnemonic skills of healthy participants. This task is proposed to assist in identifying persons with suspected retrieval problems that might benefit from mnemonic hints.

The results indicate that in all other RAVLT trials, the effect of the practice can be credited to two different main factors: (a) retention of specific test material (in this case, words that compose the lists); (b) a metamemoric factor, which considers how the exposure to a similar task may improve performance strategies (Mitrushina et al. 2005). If one considers implementing a different interval between test-retest sessions, this data could be distinct. Knight et al. (2007), when performing an analysis of the mean scores of the test-retest sessions (with a 1 year interval between applications), found significant differences only for the measures A5, Sum A1-A5 and A7. In addition, if the samples are also different, another data configuration might be expected.

The Cronbach's Alpha Coefficient represents the equivalence degree regarding to realized measurement (Hogan 2007), it is a homogeneity index of the items. The high value to this coefficient ($r = 0.80$) indicates a good internal consistency of the items that compose the RAVLT. Therefore, all the trials are reliable measures for learning processes and verbal memory. This measure has a value close to the found for the same Portuguese version of the test ($r = 0.85$) (Malloy-Diniz et al. 2007).

The RAVLT correlates moderately with other learning measures (Strauss et al. 2006, Kessels et al. 2006, Helmstaedter et al. 2009), verbal memory (Strauss et al. 2006, Helmstaedter et al. 2009) and spacial memory (Strauss et al. 2006, Kessels et al. 2006), therefore, have moderate convergent validity. Medium magnitude evidence has been ascertained to the convergent validity of the RAVLT with the BVRT, what means that it, indeed, evaluates episodic memory. High and moderate magnitude correlations (minimum $r = 0.47$ e maximum $r = 0.54$) were obtained by Duff et al. (2005) when comparing the measures A6 and A7 of the RAVLT with immediate and delayed visual memory instruments. Probably, the correlations between RAVLT and BVRT were not of a higher rate due to the different structures of the tests, as they measure distinct aspects of episodic memory. Whereas RAVLT assess mainly verbal memory, BVRT assess visual memory and is related to other cognitive skills as psychomotor abilities (Strauss et al. 2006). Nonetheless, is important to note that the mnemonic variables are together, regardless of being verbal or nonverbal measures (Strauss et al. 2006), which justifies the choice of BVRT in the present study.

The RAVLT demonstrated divergent validity compared with the TMT, which supports the evidence that both measure different cognitive functions. However, other studies have presented contradictory evidence regarding this correlation. Duff et al. (2005) found a high and significant correlation of 0.51 between A7 and the performance on part B of the TMT. This finding suggests a mutual relationship between executive functions and verbal memory, which was not verified in the present study. Although the proactive and retroactive interference measures of the RAVLT were expected to correlate with the TMT because they are well-known executive measures, they did not exhibit significant correlation. Additionally, for the German version of the RAVLT, Helmstaedter et al. (2009) did not find any significant correlation between A7 and the tests that measure attention and fluency, both aspects of executive function. It is important to highlight the methodology differences regarding the sample composition, chosen instrument and data analysis of the mentioned studies. The range in which the measures of the RAVLT correlate with executive functions still requires additional empirical evidence.

## Conclusions

Learning and memory difficulties are the most common complaints in patients with neurological impairment and significantly impact the daily life activities and social functionality of these patients (Messinis et al. 2007). The RAVLT is widely known to offer objective methods to help identify these deficits. Furthermore, repeated evaluations are often frequent and necessary in the clinical context to closely follow the progress of degenerative conditions, the effect of prescribed drugs or recovery after intervention (Uchiyama et al. 1995).

It is valid to note that all measurements obtained by a neuropsychological test are subject to the influence of non-reliable sources. Each investigation method seeks to deal with one or more of the non-systematic variations of the test scores. Therefore, one can assert that once each reliability measures encompasses a part of each possible source of variation, a single reliability measure does not exist for a test (Hogan 2007).

Among the resultant limits of the test-retest methodology choice for reliability measurement are: (a) the difficulty of applying the test due to the fact that participants need to cooperate to reestablish the same application context in a second round; and, more importantly, (b) decisions concerning the effects that the first application might have on the second one. The learning effect is a very important point to consider because of its proclivity to inflate the reliability coefficient and thus overestimate the measured attribute.

There are vast descriptions in the literature about the RAVLT's sensitivity to the effect of learning in a second application of the same version of the test. After successive applications, a small, but significant enhancement in the number of recalled words (1 or two words mean per trial) is expected (Mitrushina et al. 2005). The effect of the practice is diminished when the participants are not exposed to the same list, which is why the use of alternative lists is suggested to minimize this effect (Uchiyama et al. 1995, Strauss et al. 2006, Van den Burg and Kingma 1999, Beglinger et al. 2005, Mitrushina et al. 2005, Knight et al. 2007). Those studies that have applied the test-retest methodology utilizing alternate forms of the RAVLT found a lesser effect of practice (Strauss et al. 2006, Van den Burg and Kingma 1999, Mitrushina et al. 2005, Rezvanfard et al. 2011). However, these data are not unanimous; Uchiyama et al. (1995) demonstrated a significant effect of practice on two distinct versions of the RAVLT, using a 1-year interval between trials. Those researchers therefore suggested the need for longitudinal normative data for interpretation of the retest data (Uchiyama et al. 1995).

Strauss et al. (2006) observed that with intervals greater than 1 year, the RAVLT demonstrates moderate test-retest reliability. In this study, the one-month interval between applications was demonstrated to be acceptable. However, this interval time is inferior when observed in clinical practice, where the monitoring and successive evaluations are performed with longer time intervals, which varying from 3 months to 1 year, for example.

In the Portuguese language and Brazilian culture, a second version that meets the RAVLT construction criteria is unfortunately not yet available. This is a problem for which Brazilian researchers will have to find viable solutions. An alternative would be to use the first list published by Malloy-Diniz et al. (2000).

However, this list is a direct translation of the English version and demonstrates several limitations, such as not taking into account the word frequency in the Portuguese language and the number of syllables that the newest version has presented (Malloy-Diniz et al. 2007).

The RAVLT evaluates the participant's ability to encode, consolidate, store and recall verbal information. The most reliable measures in terms of stability and low measurement error (greater reliability) were, in regressive order, the Sum A1-A5, A4, Recognition, A6 and A3. The total score performance represented by Sum A1-A5 is substantially more reliable than the items considered individually.

Compared with other RAVLT psychometrical property studies (Strauss et al. 2006, Van den Burg and Kingma 1999, Geffen and Geffen 2000, Duff et al. 2005, Kessels et al. 2006, Knight et al. 2007, Helmstaedter et al. 2009), the current research achieves modest, although relatively high, correlation values. The small sample size might have been a contributing factor to the lower reliability coefficient expression.

To establish validity, an integrated collection of evidence is required for the appropriate interpretation of the analyzed instrument's score. The results indicate that the verbal memory measured by the RAVLT positively correlates with the visual memory evaluated by BVRT but does not correspond to the executive functions assessed by the TMT. Although a vast amount of evidence suggests that memory and executive functions correlate with each other (Tremont et al. 2000), this study did not aim to verify the magnitude of their interaction. As such, the chosen methodology effectively supported the proposed objectives and verified that the RAVLT measures mnemonic attributes rather than ones related to executive functions.

Reliability and validity are distinct but complementary constructs and are important in the field of psychometrics. A test can be valid without being reliable; however, a test cannot be reliable if it is not valid (Hogan 2007). Therefore, the relationship between these two properties is a fundamental consideration because the validity of the test depends considerably on its degree of reliability.

We aimed to determine whether the RAVLT is a safe source of information about the mnemonic construct. The collected results suggest that the RAVLT shows good reliability and validity, and therefore this psychometric instrument can be considered to provide valid measurement of episodic verbal memory in neuropsychological evaluation. This study proposed to contribute psychometrical evidence of the validity and reliability of the RAVLT, but because of the relatively small sample, the outcomes must be interpreted with caution. New investigations are expected to supplement these results and strengthen the data that has thus far collected.

## References


Beglinger LJ, Gaydos B, Tangphao-Daniels O, Duff K, Kareken DA, Crawford J, Fastenau PS, Siemers ER (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology* 20, 517-29.

Bertolucci PH, Brucki SM, Campacci SR, Juliano Y (1994). The Mini-Mental State Examination in a general population: impact of education status. *Arquivos de Neuropsiquiatria* 52, 1, 1-7.

Cohen J (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, New Jersey.

De Paula JJ, Melo LPC, Nicolato R, Moraes EN, Bicalho MA, Hamdan AC, Malloy-Diniz LF (2012). Reliability and construct validity of the Rey-Auditory Verbal Learning Test in Brazilian elders. *Revista de Psiquiatria Clínica* 39, 1, 19-23.

Duff K, Schoenberg MR, Scott JG, Adams, RL (2005). The relationship between executive functioning and verbal and visual learning and memory. *Archives of Clinical Neuropsychology* 20, 111-22.

Fichman HC, Dias LBT, Fernandes CS, Lourenço R, Caramelli P, Nitrini R (2010). Normative data and construct validity of the Rey Auditory Verbal Learning Test in a Brazilian Elderly population. *Psychology & Neuroscience* 3, 1, 79-84.

Geffen G, Geffen L (2000). *Auditory Verbal Learning Test (AVLT)*: computerised scoring program and population norms. Acer Press, Australia.

Hakstian AR, Whelan TE (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika* 41, 219-231.

Hamdan, AC. Hamdan EMLR (2009). Effects of age and education level on the Trail Making Test in A healthy Brazilian sample. *Psychology & Neuroscience* 2, 2, 199-203.

Hancock P, Larner AJ (2011). Test Your Memory test: diagnostic utility in a memory clinic population. *International Journal of Geriatric Psychiatry* 26, 976-80.

Helmstaedter C, Wietzke J, Lutz MT (2009). Unique and shared validity of the ''Wechsler logical memory test'', the ''California verbal learning test'', and the ''verbal learning and memory test'' in patients with epilepsy. *Epilepsy Research* 87, 203-12.

Hogan TP (2007). *Psychological testing*: a practical introduction (2nd ed.). Hoboken, NJ: John Wiley and Sons.

Kessels RPC, Nys GMS, Brands AMA, van den Berg E, van Zandvoort MJE (2006). The modified Location Learning Test: Norms for the assessment of spatial memory function in neuropsychological patients. *Archives of Clinical Neuropsychology* 21, 841-46.

Knight RG, McMahon J, Skeaff, CM, Green, TJ (2007). Reliable Change Index scores for persons over the age of 65 tested on alternate forms of the Rey AVLT. *Archives of Clinical Neuropsychology* 22, 513-18.

Lezak MD, Howieson DB, Loring DW (2004). *Neuropsychological assessment*, 4th ed. Oxford University Press, New York.

Magalhães SS, Hamdan AC (2010). The Rey Auditory Verbal Learning Test: normative data for the Brazilian population and analysis of the influence of demographic variables. *Psychology & Neuroscience* 3, 1, 85-91.

Malloy-Diniz LF, Lasmar VAP, Gazinelli LSR, Fuentes D, Salgado, JV (2007). The Rey Auditory-verbal Learning Test: applicability for the brazilian elderly population. *Revista Brasileira de Psiquiatria* 29, 4, 324-29.

Malloy-Diniz LFM, Cruz MF, Torres VM, Cosenza RM (2000). O teste de Aprendizagem Auditivo-Verbal de Rey: normas para uma população Brasileira. *Revista Brasileira de Neurologia* 36, 79-83.

Messinis L, Tsakona I, Malefaki S, Papathanasopoulos P (2007). Normative data and discriminant validity of Rey's Verbal Learning Test for the Greek adult population. *Archives of Clinical Neuropsychology* 22, 739-52.

Milian M, Leiherr AM, Straten G, Müller S, Leyhe T, Eschweiler GW (2011). The Mini-Cog versus the Mini-Mental State Examination and the Clock Drawing Test in daily clinical practice: screening value in a German Memory Clinic. *International Psychogeriatrics* 15, 1-9.

Mitrushina MN, Bone KB, D'Elia LF (2005). *Handbook of normative data for neuropsychological assessment*, 2nd ed. Oxford University Press, New York.

Rezvanfard M, Ekhtiari, H, Noroozian M, Rezvanifar A, Nilipour R, Javan GK (2011). The Rey Auditory Verbal Learning Test: alternate forms equivalency and reliability for the Iranian adult population (Persian version). *Archives of Iranian Medicine* 14, 2, 104-9.

Stallings GA, Boake C, Sherer M (1995). Comparison of the California Verbal Learning Test and the Rey Auditory Verbal Learning Test in head-injured patients. *Journal of Clinical and Experimental Neuropsychology* 17, 5, 706-712.

Strauss E, Sherman EMS, Spreen O (2006). *A compendium of neuropsychological tests*: *administration, norms, and commentary*, 3rd ed. Oxford University Press, New York.

Tremont G, Halpert S, Javorsky DJ, Stern RA (2000). Differential impact of executive dysfunction on verbal list learning and story recall. *Clinical Neuropsychology* 14, 3, 295-302.

Uchiyama CL, D'Elia LF, Dellinger AM, Becker JT, Selnes OA, Wesch JE, Chen BB, Satz P, can Gorp W, Miller EN (1995). Alternate forms of the Auditory-Verbal Learning Test: issues of test comparability, longitudinal reliability, and moderating demographic variables. *Archives of Clinical Neuropsychology* 10, 2, 133-45.

Uttl B (2005). Measurement of Individual Differences: lessons from memory assessment in Research and Clinical Practice. *Psychological Science* 16, 6, 460-67.

Van den Burg W, Kingma A (1999). Performance of 225 Dutch School Children on Rey's Auditory Verbal Learning Test (AVLT): Parallel Test-Retest Reliabilities with an Interval of 3 Months and Normative Data. *Archives of Clinical Neuropsychology* 14, 6, 545-59.