گارگاه‌های آموزشی مرکز اطلاعات علمی

مقاله نویسی علوم انسانی

اصول تنظیم قراردادها

آموزش مهارت های کاربردی در تدوین و چاپ مقاله
The Rey Auditory Verbal Learning Test: Alternate Forms Equivalency and Reliability for the Iranian Adult Population (Persian Version)

Mehrnaz Rezvanfard MD1, Hamed Ekhtiari MD1,2, Maryam Noroozian MD2, Alireza Rezvanifar MSc1, Reza Nilipour PhD3, Gelavizh Karimi Javan MSc3

Abstract
Background: Increasing demand for memory assessment in clinical settings in Iran, as well as the absence of a comprehensive and standardized task based upon the Persian culture and language, requires an appropriate culture- and language-specific version of the commonly used neuropsychological measure of verbal learning and memory, the Rey Auditory Verbal Learning Test (RAVLT).

Methods: The Persian adapted version of the original RAVLT and two other alternate word lists were generated based upon criteria previously set for developing new word lists. A total of 90 subjects (three groups of 30 persons), aged 29.75±7.10 years, volunteered to participate in our study and were tested using the original word list. The practice effect was assessed by retesting the first and second groups using the same word list after 30 and 60 days, respectively. The test-retest reliability was evaluated by retesting the third group of participants twice using two new alternate word lists with an interval of 30 days.

Results: The re-administration of the same list after one or even two months led to significant practice effects. However, the use of alternate forms after a one-month delay yielded no significant difference across the forms. The first and second trials, as well as the total, immediate, and delayed recall scores showed the best reliability in retesting by the alternate list.

Conclusion: The difference between the generated forms was minor, and it seems that the Persian version of the RAVLT is a reliable instrument for repeated neuropsychological testing as long as alternate forms are used and scores are carefully chosen.

Keywords: alternative lists, Iranian population, Persian version, practice effect, Rey auditory verbal learning test reliability

Introduction

Memory complaints in outpatient settings, particularly in the elderly populations, appear to be the most frequent reason for neuropsychological referral1 and should be borne in mind in tandem with the fact that deficits in memory processes are begotten by a variety of common neurological and psychiatric disorders.1-6 According to the latest national population estimates of 2007 (Statistical Center of Iran), Iran is one of the youngest populations in the world (77.41% of the population is less than 40 years old). The anticipated increase in the elderly population and concomitant rise in demand for memory assessment prompted us to address the current lack of a comprehensive and standardized task based upon the Persian culture and language by developing an appropriate culture- and language-specific version of commonly used neuropsychological measures of verbal learning and memory, e.g., the Rey auditory verbal learning test (RAVLT). The RAVLT is generally utilized for memory evaluation in scientific research and in clinical practice1 in that it is easily administrable and confers the encoding, consolidating, storing, and retrieving of verbal information via a five-trial presentation of a 15-word list (list A), a single presentation of an interference list (list B), two post-interference recall trials (one immediate and one delayed), and recognition of the target words presented with distracters. Measures essential for an understanding of the kind and severity of a patient’s memory deficits can, consequently, be straightforwardly obtained.1 The original French words7 and their order have been translated into English without modification. Other language versions such as Flemish,8 Spanish,9 Hebrew,10 German,11 Chinese,12 and Greek13 are also provided.

The following linguistic criteria were established for the generation of word lists and additional forms14,15:

- The chance of occurrence and frequency of the selected words based upon a large scale contemporary database in common usage with Thorndike-Lorge Tables as...
guidelines. Only the most frequent words were used because of word usage and the likelihood of recall effect. 

- The imagery value using Pavio’s Table. Words with high rates of imagery value were selected.
- Word length measured by the number of letters and syllables. Only one- or two-syllable nouns were used to generate the lists.
- Control of any obvious semantic ambiguity and/or phonetic similarities or associations between the words on the same list.

The assessment of change is often critical in the neuropsychological evaluation, whether in research or in clinical settings. The repeated use of instruments confounds this assessment inasmuch as practice effects (gains in performance by prior experiences with the test) have been demonstrated for many measures. Practice effects are particularly expected with memory testing on account of the fact that the learning gained during the preliminary evaluations tends to be transferred to the following ones. A practice effect may be related to the fact that patients explicitly remember test items or the testing format (test sophisticated effects) previously presented. It is possible to minimize the practice effect by using alternate forms of a test instead of retesting by the same form. Several alternative forms have been developed for the RAVLT such as those created by: Lezak; Shapiro & Harrison; Crawford et al.; Majdan et al.; and Geffen et al. They have been proven useful, particularly in longitudinal or sequential research designs where baseline performance, experimental manipulation, and recovery time are expected to feature within clinical populations.

Measurement error is another effect of significance in so far as it can baffle a patient’s actual score in a test-retest design even when alternative forms are employed. Measurement error is often construed as the concept of reliability, which incorporates test-retest reliability (TRR). A score with a good TRR is free of measurement error unrelated to chance variance, and the second session of testing is likely to result in subjects having scored the highest remaining amongst the best. It is also worthy of note that highly significant reliability coefficients were reported while studying different alternative lists of the RAVLT.

The participants were assessed using the Persian adapted list of the RAVLT and two other alternate forms, (Lezak and Shapero & Harisson), based on the original English word lists reviewed by Hawkins et al. (2004). The original English words, presented in Table 1, were initially translated into Persian without any change. Of the 90 words from the six original English word lists, 80 which were in accordance with desired linguistic criteria were retained from both lists A and B. Ten new words adapted to the Persian language were chosen for the development of the Persian lists. Finally, the word items on these newly formed Persian lists were compared with the original lists for consistency in terms of word length (as measured by number of phoneme counts in Persian), and all were one- or two-syllable concrete nouns. There were no obvious semantic or phonetic associations or similarities between the words on the same list and they were chosen from amongst frequently occurring words in the Persian language. The probability of the occurrence of the word in common usage in the Persian language was ascertained using the World Wide Web (WWW) as a database (from six billion contemporary Persian words on four million websites), Google as the search engine, and a Google-based applet, which provided word counts on the Web pages. This method was also employed by Cilibrasi (2007) to automatically compute the similarity distance between words and phrases, and its performance was in agreement with that of the Word Net database. Recognition lists were constructed using target words and adding 20 new semantically associated or phonetically similar words as distracters. The above criteria helped establish form equivalence between the three new lists.

Materials and Methods

Participants

Ninety volunteers ages 18 to 50 (29.75±7.10 years) participated in the study after providing written informed consent. The study population was comprised of 44 men and 46 women, with educational levels ranging from 2 to 20 years (11.56±4.91). The exclusion criteria were history of brain surgery or psychiatric symptoms and any type of medical condition or using a medication that might affect cognitive functioning.

Word selection procedures

The participants were assessed using the Persian adapted list of the RAVLT and two other alternate forms, (Lezak and Shapero & Harisson), based on the original English word lists reviewed by Hawkins et al. (2004). The original English words, presented in Table 1, were initially translated into Persian without any change. Of the 90 words from the six original English word lists, 80 which were in accordance with desired linguistic criteria were retained from both lists A and B. Ten new words adapted to the Persian language were chosen for the development of the Persian lists. Finally, the word items on these newly formed Persian lists were compared with the original lists for consistency in terms of word length (as measured by number of phoneme counts in Persian), and all were one- or two-syllable concrete nouns. There were no obvious semantic or phonetic associations or similarities between the words on the same list and they were chosen from amongst frequently occurring words in the Persian language. The probability of the occurrence of the word in common usage in the Persian language was ascertained using the World Wide Web (WWW) as a database (from six billion contemporary Persian words on four million websites), Google as the search engine, and a Google-based applet, which provided word counts on the Web pages. This method was also employed by Cilibrasi (2007) to automatically compute the similarity distance between words and phrases, and its performance was in agreement with that of the Word Net database. Recognition lists were constructed using target words and adding 20 new semantically associated or phonetically similar words as distracters. The above criteria helped establish form equivalence between the three new lists.

Procedure

All tests were administered by qualified examiners (psychology experts) trained in AVLT administration at quiet locations in accordance with Lezak’s instructions. The participants were screened at the beginning of the testing session for exclusion criteria such as health problems via a standardized interview before being categorized into three groups of 30 persons and matched based upon demographic variables that might affect performance on the auditory learning test such as age, gender, and education level.

The first group was tested based upon the Persian adapted version of Rey’s (word list 1) in two test sessions with intersession intervals of 30 days (±3); the second group received the Rey’s 60 days (±5) after the initial testing; and
the participants in the third group were evaluated in three
test sessions via the Persian adapted version of the Rey’s
(word list 1), Lezak’s (word list 2), and Shapiro & Haris-
son’s (word list 3), respectively, with intersession intervals
of 30 days (±3). All tests and retests were performed in a
fixed sequence (non-counterbalanced design), and the delay
intervals were chosen on the basis of previous literature.25

Description of the auditory verbal learning test and related
measures
The test was administered in accordance with the proce-
dure described by Lezak (2004).1 In brief, subjects are asked
to recall as many words on an initial oral 15-word list (list A)
as possible, in any order. This procedure was repeated five
times consecutively (trials 1 – 5 or acquisition trials), with
the score for each trial being the number of words correctly
recollected.
Participants were subsequently presented with a second list
(list B); which was a measure of proactive interference, or
the degree to which old learning can meddle with new learn-
ing. In the next step, participants were asked to remember as
many words from list A as possible without reading it again;
or immediate recall (ImmRec) and is believed to stimulate a
retroactive interference situation in which new learning tan-
gles with the recollection of old information. Delayed free
recall (DelRec) was thereafter administered to the partici-
pants in the same manner as ImmRec (e.g., no access to list
A) but following a 20 minute time interval. Subjects were
then presented with a yes/no recognition trial of 50 words
comprised of 15 target words from list A, 15 words from list
B, and 20 words similar to word lists A and B in phonologi-
cal or semantic terms.

The immediate memory span is usually measured based
on the number of words recollected on the first trial (Trial
1). The learning curve is the variation in performance over
the course of the five free-recall trials, with its slope pro-
viding a measure of the verbal learning test.26 This slope is
generally characterized as the difference between recall on
the fifth trial and that on the first trial (Trial 5-Trial 1). Serial
positions are obtained by adding up the number of words re-
called across the five trials in blocks of three; in other words
they can make a total amount of fifteen. Serial Position 1
(SP1) to Serial Position 5 (SP5) are utilized by studying the
serial positions of the words recalled across the trials. Nor-
mal subjects are likely to recall the initial words (Primacy)
but also the latest ones (Recency).1

The following is a list of other scores that are calculated:
the total learning (TL=sum of words recalled in the five

---

Table 1. Item characteristics of the original AVLT and new adapted Persian word lists.

<table>
<thead>
<tr>
<th>Original list 1 (Rey)</th>
<th>Adapted Persian version</th>
<th>Original list 2 (Lezak)</th>
<th>Adapted Persian version</th>
<th>Original list 3 (Shapiro &amp; Harison)</th>
<th>Adapted Persian list 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Drum Desk</td>
<td>بیز هنیه</td>
<td>Book Bowl</td>
<td>کنسه کتاب</td>
<td>Street Baby</td>
<td></td>
</tr>
<tr>
<td>Curtain Ranger</td>
<td>سربراز پرده</td>
<td>Flower Down</td>
<td>سحر گل</td>
<td>افیاپوسی علف</td>
<td></td>
</tr>
<tr>
<td>Bell Bird</td>
<td>پرده زیب</td>
<td>Train Judge</td>
<td>قاضی قطر</td>
<td>قصر درب</td>
<td></td>
</tr>
<tr>
<td>Coffee Shoe</td>
<td>کافی کهی</td>
<td>Rag Grant</td>
<td>هدیه فرش</td>
<td>لب دست</td>
<td></td>
</tr>
<tr>
<td>School Stove</td>
<td>بخاری مردسم</td>
<td>Meadow Insect</td>
<td>حضرت جیبک</td>
<td>مهربان ساره</td>
<td></td>
</tr>
<tr>
<td>Parent Mountain</td>
<td>ماه چتر</td>
<td>Harp Plane</td>
<td>ماهیچه آواز</td>
<td>پیراهن همسر</td>
<td></td>
</tr>
<tr>
<td>Moon Glasses</td>
<td>میکا ماه</td>
<td>Salt County</td>
<td>کشور نمک</td>
<td>پیامبر یکه</td>
<td></td>
</tr>
<tr>
<td>Garden Towel</td>
<td>دود باغ</td>
<td>Finger Pool</td>
<td>استخر انگشت</td>
<td>شهر سکه</td>
<td></td>
</tr>
<tr>
<td>Hat Cloud</td>
<td>ایر گل</td>
<td>Apple Seed</td>
<td>دانه سیب</td>
<td>شهر پسر</td>
<td></td>
</tr>
<tr>
<td>Farmer Boat</td>
<td>قابی دمتان</td>
<td>Chimney Sheep</td>
<td>گوشت پنجره</td>
<td>سکه شکرک</td>
<td></td>
</tr>
<tr>
<td>Turkey Gum</td>
<td>دندان کره‌پیک</td>
<td>Log Coat</td>
<td>دوست لوله</td>
<td>ارتش انتق</td>
<td></td>
</tr>
<tr>
<td>Color Pencil</td>
<td>میشت رنگ</td>
<td>Key Bottle</td>
<td>اسپری پیک</td>
<td>دوست لوله</td>
<td></td>
</tr>
<tr>
<td>House Church</td>
<td>مسجد خانه</td>
<td>Rattle Peach</td>
<td>ساختمان شتر</td>
<td>طوفان پوست</td>
<td></td>
</tr>
<tr>
<td>River Fish</td>
<td>نیش رود</td>
<td>Gold Chair</td>
<td>روستای آتش</td>
<td>روستای آتش</td>
<td></td>
</tr>
<tr>
<td>Nose Lamb</td>
<td>بینه بیمه</td>
<td>Button Meal</td>
<td>دستان سائیت</td>
<td>شلوار پر سانتی</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lake Building</td>
<td></td>
</tr>
</tbody>
</table>

| Word length (mean)   | 5.1 5.0 5.0 5.1 4.9 4.8 4.9 5.7 4.5 4.9 5.4 5.3 |

A: < 50 – 100 in 10^-6 words; AA: >100 in 10^-6 words
free-recall trials), the learning over trials (LOT=TL-5* Trial 1), the proactive interference score (PIS=List B-Trial 1), the retroactive interference score (RIS=ImmRec-Trial 5), and the forgetting score (FR=DelRec-ImmRec). The recognition task generates the customary hits (Hit=number of words recognized from list A), misses (Miss=number of words not recognized from the list), correct rejection (CR=number of words not included on list A which were correctly identified as not being on the list), and false alarms (FA=numbers of words specified as being on list A which were not present).

Statistical analysis

With respect to the normal distribution of the scores, the effects of having administered the same list at different intervals were analyzed with the paired t-test. The practice effect and alternate forms equivalency were evaluated using the analysis of variance (ANOVA) for repeated measures. Taking place in pairs, comparisons were made via the Post Hoc Tukey test. The reliability of the tests was estimated by using Pearson’s correlation coefficients across the three test sessions, and with interclass coefficients (ICC) using the absolute agreement definition. No values were excluded from the analysis.

Results

The new and original lists did not statistically differ with regards to any word item matching variables (Table 1). Table 2 demonstrates that administration of the same list after an interval of one or even two months yielded a significant practice effect on the word recall trials and some of the derived scores. The use of alternate forms after a one month delay, however, bore no significant difference across the forms. Figures 1a, 1b, and 2a, 2b show learning and serial positions curves utilizing the same list after one- or two-month intervals.

<table>
<thead>
<tr>
<th>Interval: 1 month (n=30)</th>
<th>Interval: 2 months (n=30)</th>
<th>Interval: 1 month (n=30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>List 1</td>
<td>List 1</td>
<td>List 1</td>
</tr>
<tr>
<td><strong>7.03(2.43)</strong></td>
<td>9.40(2.55)**</td>
<td><strong>7.03(1.68)</strong></td>
</tr>
<tr>
<td><strong>9.83(2.32)</strong></td>
<td>11.73(2.20)**</td>
<td><strong>9.33(2.15)</strong></td>
</tr>
<tr>
<td><strong>11.23(2.30)</strong></td>
<td>12.77(1.81)**</td>
<td><strong>10.90(2.20)</strong></td>
</tr>
<tr>
<td><strong>12.23(2.24)</strong></td>
<td>13.00(1.38)**</td>
<td><strong>11.80(1.89)</strong></td>
</tr>
<tr>
<td><strong>12.57(1.98)</strong></td>
<td>13.23(1.55)**</td>
<td><strong>12.03(2.24)</strong></td>
</tr>
<tr>
<td><strong>52.57(9.74)</strong></td>
<td>58.07(13.03)**</td>
<td><strong>51.07(8.00)</strong></td>
</tr>
</tbody>
</table>

**Table 2.** Descriptive statistics for the RAVLT scores and reliability coefficients in different test-retest administrations.
month intersession intervals. Figures 3a and 3b illustrate the equivalency of AVLT learning curves employing various alternate lists after a one-month intersession interval.

According to Table 2, the first and second trials, as well as the total, immediate, and delayed recall scores are the most reliable, whereas trial 3 and list B recall have modest reliability and the acquisition trials (trials 4 and 5) and recognition scores display poor ICCs.

Discussion

Our results showed that utilization of the same list resulted in the production of a significant practice effect in re-evaluation even after a two-month delay; repeated administration of alternate lists after a one-month intersession interval removed this undesirable effect. Insignificant difference between the mean scores of the trials across the three Persian adapted AVLT forms denotes equivalency and similarity in terms of difficulty level between these word lists.

In accordance with the results of the present study, previous direct comparisons of the two administration modes have indicated negligible improvement on alternate adaptations of verbal memory tasks, while a considerable practice effect is observed with the use of the same version.25,27,28 Lezak also asserted that some patients were likely to exhibit an outstandingly good recall of word lists even after more than a year.1 Shapiro and Harrison (1990)15 recommended that subjects not be retested twice in a row within days of their initial evaluations even with alternative lists on account of the type of practice effect attributable to the meta-memor- ic influence, which may be the consequence of an earlier experience with the testing format. Our results are also in line with those reported by Crawford, et al.,25 who found no practice effects when retesting normal subjects four weeks after the initial examination with an alternate word list. Furthermore, similar to previous studies,1 our subjects showed normal primacy and recency patterns in serial position curves (Figures 1a, 2a, and 3a).

In the present study, Trial 1, Trial 2, Total, ImmRec, and DelRec scores showed the best reliability correlation (range: 0.56 to 0.70) whereas other acquisition trials were less reliable and recognition scores displayed poor ICCs. Test reliability across the forms was similar to that by Lemay, et al.,25 who reported the best reliability for Total, ImmRec, and DelRec and modest reliability for acquisition trials and poor ICCs for recognition and derived scores. Some studies have yielded better correlation coefficients in retesting by alternate lists,22,29 which could be partly attributed to different methods of design and analysis used by the researchers. We did not exclude any out of range scores, thus this difference is not unexpected. Duration intervals should also be regarded as one of the most important factors, which could influence the reliability coefficients; bearing in mind that exceeded correlations as reported by Ryan et al. (1986)22 may be attributed to a relatively brief test-retest interval (140 min). Dikmen (1999)30 suggested that poor reliability in memory tests might not necessarily be related to a measurement error but rather to the variable nature of memory. Low reliability
for verbal memory scores has also been observed for most tests used in clinical practice\textsuperscript{30–32}. To sum up, there is enough data to suggest that the difference between forms is sufficiently minor to be able to ignore it in clinical application. It seems that the Persian version of the RAVLT is a reliable instrument for repeated neuropsychological testing as long as alternate forms are used and scores chosen carefully. The results of the present study are useful for both clinical trials and clinical purposes such as pre- and post-assessments. Be that as it may, prior to general clinical application, normative data is required on larger samples of both clinical and normal subjects and for different age groups and education levels. Multiple studies with counter-balance design are needed to determine reliability, equivalency, and relative difficulty levels more definitively among different clinical populations.

Acknowledgment

This work was supported by a grant from Tehran University of Medical Sciences to Maryam Noroozian. We thank the staff of INCAS neurocognitive laboratory for their help in participant recruitment and management. We also thank Gholamreza Esmaeeli Djavid for his consultations for statistical analysis. We declare no conflicts of interest.

References
