Journal of Clinical and Experimental Neuropsychology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ncen20

Development, psychometric properties, and pilot norms of the first Arabic indigenous memory test: The Verbal Memory Arabic Test (VMAT)

Pia Zeinoun, Natali Farran, Samia J. Khoury & Hala Darwish

To cite this article: Pia Zeinoun, Natali Farran, Samia J. Khoury & Hala Darwish (2020) Development, psychometric properties, and pilot norms of the first Arabic indigenous memory test: The Verbal Memory Arabic Test (VMAT), Journal of Clinical and Experimental Neuropsychology, 42:5, 505-515, DOI: 10.1080/13803395.2020.1773408

To link to this article: https://doi.org/10.1080/13803395.2020.1773408

Published online: 22 Jun 2020.

JOURNAL OF CLINICAL AND EXPERIMENTAL NEUROPSYCHOLOGY
2020, VOL. 42, NO. 5, 505–515
https://doi.org/10.1080/13803395.2020.1773408

Development, psychometric properties, and pilot norms of the first Arabic indigenous memory test: The Verbal Memory Arabic Test (VMAT)

Pia Zeinoun[a,b], Natali Farran[c], Samia J. Khoury[d,e] and Hala Darwish[c,d]

[a]Department of Psychology, American University of Beirut, Beirut, Lebanon; [b]Department of Psychiatry, American University of Beirut Medical Center, Beirut, Lebanon; [c]Hariri School of Nursing, American University of Beirut, Beirut, Lebanon; [d]Nehme and Therese Tohme Multiple Sclerosis Center, American University of Beirut Medical Center, Beirut, Lebanon; [e]Abu-Haidar Neuroscience Institute, American University of Beirut Medical Center, Beirut, Lebanon

ABSTRACT

Background: Tests of verbal memory (list learning) are particularly useful for clinicians and researchers globally, yet there are no psychometrically robust tests that are built indigenously for Arabic-speaking populations, which comprise more than 370 million.
Objective: To develop a verbal memory Arabic test using a systematic procedure of item selection and then provide evidence of validity and reliability in an Arabic-speaking sample in Lebanon.
Method: In study 1, we conducted a word prototypicality study (n = 77), and identified 932 words across 7 semantic categories. Following quantitative analyzes and qualitative judgments by an expert panel, we selected a sufficient number of words and categories, and constructed and piloted the items, instructions, and protocol for the Verbal Memory Arabic Test (VMAT). In study 2, we administered the VMAT on a community sample (n = 199; screened for depression and cognitive impairment) and patients with Multiple Sclerosis (n = 16).
Results: Scores decreased with age as expected, they discriminated well between healthy and clinical populations (matched on age, sex, and years of education), and showed acceptable consistency within items and across time. *Conclusions*: The VMAT is the first Arabic test developed indigenously. It can be used in clinical and research settings with Arabic-speaking populations to assess verbal learning.

ARTICLE HISTORY
Received 4 November 2019
Accepted 18 May 2020

ARTICLE HISTORY

KEYWORDS
Verbal Memory Arabic Test; Arabic norms; assessment; learning and memory; normative data; test development

Introduction

Testing verbal memory is essential for neuropsychological evaluations and diagnoses and is used in clinical and research settings (Lezak, 2004). Although English-speaking and European countries have a longstanding neuropsychological assessment tradition, these practices are new in Arab-speaking nations. Arabic is the fifth most spoken language in the world; 23 countries have Arabic as their official language (Lewis et al., 2013). The population of the Arab region is north of 370 million in addition to more than 24 million immigrants from the region living outside their country of birth in 2015 (International Organization for Migration, 2016; Lewis et al., 2013). The need to serve these populations has led to an increased reliance on appropriate cognitive/neuropsychological tests.

However, most of the tests used on Arab-speaking populations lack cultural, linguistic, and psychometric robustness. Zeinoun et al. (2020) reviewed the translation/adaptation process and psychometric methods used to develop 124 clinical tests in Arabic and concluded that more than 90% of the studies used dated translation/adaptation methods, not supported by current guidelines such as those of the ITC and the Standards for Educational and Psychological Testing (Zeinoun et al., 2020). Also, a recent study that reviewed neuropsychological measures in Arabic found that the second most used test in research is the Wechsler Memory Scale (the first most used was the Trail Making Test, which examines visual attention and task switching) (Fasfous et al., 2017). All the studies reviewed however, except one, used the memory test without referring to any translation or adaptation process to Arabic and with none reporting norms (Fasfous et al., 2017). Similarly, studies using verbal memory tests like the California Verbal Learning Test (CVLT) for adults or children do not report any data on how they were adapted to Arabic. These studies' results are alarming, as they imply that studies investigating memory and other clinical outcomes among Arab-speakers may have reached conclusions based on measurements not supported by evidence.

Given the sheer size of Arabic-speakers in Arab and non-Arab countries, and the lack of robust Arabic

CONTACT Hala Darwish hd30@aub.edu.lb Hariri School of Nursing, American University of Beirut, Beirut 1107 2020, Lebanon

memory tests, it is imperative to develop a valid and reliable Arabic memory test. This study's aim was to create a rigorous verbal memory measure for adults whose native language is Arabic, for use in clinical and research settings with adequate evidence of validity and reliability.

Why construct a new test?

There are three ways tests can be developed for a new language and culture (Hambleton et al., 2004). One way is *adoption*, whereby an existing test is transferred into the new language, such as by translation. A more common method is *adaptation*, which alters an existing test to fit the target language and culture. Finally, *assembly* pertains to the construction of an instrument from the ground up. The method choice depends on several factors, such as the type of test, how, where, and on whom it will be used, the extent to which one is interested in having equivalence for multi-site comparisons, and the type of bias that the test may be expected to have – construct, method or item bias (Van de Vijver & Poortinga, 2005).

When deciding whether to use adaptation or assembly, we considered all these factors and focused on item bias – a possible source of systematic error when specific test items function differently in a new culture/language. If we were to adapt an existing verbal memory test, item bias could lead to differences in scores between the original and new test version, which are not reflective of the test taker's true (latent) abilities. For example, an item such as "raccoon" taken from the popular CVLT could be translated into Arabic, but raccoons are unfamiliar to most people living in the Arab region. Then, if we were to culturally adapt this word, by replacing raccoon with an alternative Arabic word for an existing animal such as rabbit, we would need to provide evidence that the item *rabbit* carries the same conceptual equivalence, frequency, range of semantic associations, and familiarity for Arabs as a raccoon for individuals living in North America. Otherwise, items would have differential difficulty and function across samples (Sireci et al., 2005).

An additional obstacle to translating/adapting words from an existing test is Arabic itself – a diglossic language, with various spoken and written varieties. The spoken language and dialect vary across, and within countries, to the extent it can be unintelligible across nations. Therefore, a simple translation/adaptation to "Arabic" would not address this issue Daouk-Öyry et al. (2016). Instead, the former authors suggest that tests items should be developed as "blueprints" and then slightly modified and subjected to pilot studies in different Arab countries, to ensure their appropriateness to the local vernaculars.

Given all these factors, and in line with our aim to develop a culturally and linguistically appropriate test, we developed an indigenous test from the ground up using the method of assembly.

The following studies were approved by the Institutional Review Board (IRB) of the American University of Beirut (AUB), all research assistants were trained on the ethics of research, and all participants consented to participate in the research.

Study 1: Test development

This qualitative study identified the most appropriate semantic categories, and words to be used in List A (15 words), List B (15 words), and Recognition list (45 words).

Method

Participants

We recruited a convenience sample of adults (N = 77) from all Lebanese governorates – Beirut, South, Beqaa, Mount Lebanon, and North. Participant age ranged from 18 to 80 years old (M = 31; SD = 15).

Procedure

Trained research assistants approached potential participants in public areas, such as markets and sidewalks. Consenting participants were asked to generate as many Arabic words as possible, within 90 seconds, for 7 semantic categories (animals, vegetables, stationery, clothes, musical instruments, transportation methods, and furniture). The research assistants noted generated words verbatim. After about 70 participants, we had gathered 932 words and reached saturation with no new words recorded in any category.

Data analysis

Data were analyzed using Microsoft Excel. Unique terms consisted of 207 stationery items, 161 furniture, 160 clothing items, 147 animals, 109 transportation vehicles, 68 musical instruments, and 80 vegetables. Selecting the most appropriate words followed an elimination process – we first set exclusion criteria to remove words, then chose words from the remaining pool.

Exclusion criteria

First, we excluded prototypical words from each semantic category (e.g., tomato for vegetables). If a verbal memory test includes prototypical words of a semantic category, then the test-taker might respond correctly, not due to true recall but because it is the first thing that comes to mind. Therefore, we excluded the first 3 to 5 most frequently generated words in each category and kept words with medium frequency. Rare words with a frequency equal or less than 3 were also excluded.

Next, we removed English or French words (total 42%) and kept Arabic words. Through this exercise, we eliminated the entire "transportation" and "musical instruments" categories, as 50% and 60% of their terms were non-Arabic (English/French) respectively. This is not unusual since trilingual mixing is frequent in Lebanon.

Then, focusing on the remaining five categories, we removed words judged by the authors to be too colloquial and particular to the Lebanese dialect, or too formal and literary. On one hand, colloquial words vary across Arab regions and countries than words in Modern Standard Arabic (MSA), leading to possible differential understanding among test-takers. On the other hand, people with low education levels may not understand words in MSA, used in formal text. For example, the colloquial Lebanese word for curtain is/برداية/[Berdei], but this term is not common in other Arab countries. Conversely, the more formal term is/ستارة/[/Sitara], which is not a common term in Lebanon and the Levant region. Both variations were excluded. Along the same reasoning line, we excluded terms with more than one colloquial variation and synonym (e.g.,طقية, ابوعة, برنيطة [*taqye, abou'a, bernayta/hat*], and many possible pronunciations.

Third, we removed terms specific to a particular region (e.g., Kangaroo), culture (e.g., شروال [*sirwal/sirwal pants*]) or religion (e.g., حجاب [*hijab/veil*]. Finally, we removed 2-word terms (e.g., بصل اخضر [*basal akhdar/onion sprouts*]), and terms that could arguably not belong to the assigned category (e.g., بطاطا [*batata/potato*]). After the second and third steps, the categories "clothing" and "furniture" resulted in less than 10 words, and were excluded entirety.

Selection of categories and list A and list B words

Focusing on the remaining words in the categories of *stationery, animals, vegetables*, the authors selected words to be included in List A and B, by consensus. The choice had to be simultaneously for List A and B because they had to be different from each other so they are not purposively confusing, but also similar in difficulty. We chose words with similar frequency, length, and imagery vividness (judged qualitatively among the authors). We also avoided words that could be associated with each other (e.g., zucchini and eggplant because they are cooked together in a traditional dish), rhymed with each other, and started with the same letter. Within each category, we attempted to diversify terms by including different animals and vegetable types and colors. The frequency in which words of lists A and B appeared are indicated in Table 1.

Selection of terms for recognition list

In concordance with other established measures (ex. CVLT-II), the VMAT includes a recognition trial comprised of target words and distractors. In total, the recognition list includes 45 words (15 target words from List A, 10 distractor words from the interference list B, and 20 distractor words that sound similar to List A). The 20 new words are homonyms, commonly associated with the target words, and are from superordinate categories, related and unrelated to target words. However, a distractor word could not serve in more than one type (i.e., could not be both a homonym and from a shared category).

Additionally, these words followed the same exclusion criteria listed earlier. We included 12 unrelated words, 4 words that are commonly associated, 2 related words, and 2 homonyms. This variability in distractor words discriminates between different types of recognition errors, hence help in the interpretation of memory impairments in patients. For example, patients with Alzheimer's dementia are expected to give more false positive responses for semantically related words (of a shared category), than patients with vascular dementia (Davis et al., 2002).

Table 1. Frequency of words per semantic category.

Category	List A	List B
Animals		
Word 1	33	4
Word 2	10	14
Word 3	40	35
Word 4	33	24
Word 5	17	45
Stationary		
Word 6	8	22
Word 7	5	27
Word 8	9	57
Word 9	46	8
Word 10	34	6
Vegetables		
Word 11	50	6
Word 12	42	29
Word 13	44	20
Word 14	38	28
Word 15	18	12

Results

Compilation of instrument

The items of List A, B, and the Recognition List were randomized using the Research Randomizer software (Urbaniak & Plous, 2013). The layout of the beta test protocol was designed on Adobe Illustrator. The authors also wrote a brief administration manual and test instructions in Arabic using clear and accessible language.

Pilot testing

The purpose of field-testing an instrument is to elicit feedback from participants (and test administrators) and mend possible issues with protocol, instructions, and items (Kristjansson et al., 2003). Author NF, a graduate student in psychology trained by the first author, administered the VMAT on 12 Arabic-speaking participants (4 males, 8 females), who were nurses and physicians, aged between 21 and 48 years old (M = 30.75; SD = 7.39). Following the administration, participants engaged in cognitive interviewing with NF and gave feedback on comprehensibility and clarity of items. Following participant feedback, we revised 2 terms from List A, 2 from list B, and 1 from the recognition list[1].

Final instrument

The VMAT consists of 5 learning trials of list A, followed by the interference list (list B), immediate and delayed recall trials of list A, with and without cues (delayed recall trial is administered after 15 minutes), and a recognition trial. All lists were verified post-randomization, to have the correct number of words within each semantic category (lists A and B), and the correct number of distracter and target words (Recognition list). This test format is similar to known verbal memory tests and takes about 35 minutes to complete. The test protocol and stimuli can be made available with permission and by request from the authors.

Participants first learned the words through 5 trials. In each trial, the examiner read list A in an even pace (1 second between two words) and asked the subject to recall them (the proceeding trial started if the respondent cannot recall any more words or 15 seconds have passed after his/her last word recall). The interference list was then administered after the 5 learning trials. In this trial also, the list was read once to the participant, and he/she was asked to recall it. Immediate recall was assessed following the interference list whereby participants were asked to recall list A freely, without hints. Then, the cued trials prompted the participant to recall all words in the vegetables, animals, and stationery category. If the respondent could not recall any more words or 3 seconds have passed after his/her last word recall, the administrator proceeded to the next category. After a 15-minute delay, long-term memory was assessed through free and cued recalls of list A (a similar procedure to the short delay-free and cued recall trials). Finally, the recognition trial was administered, the administrator read the list of words, and the participant indicated whether the word belongs to the first list, or not.

Discussion

The VMAT is the first indigenously developed memory test in Arabic (Zeinoun et al., 2020). By using qualitative data, we have strong evidence of validity for *test content* and *response processes* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). We are confident the items, instructions, scoring guidelines, and overall content, were understood and used as intended, by both participants and examiners. The comments we received during the pilot phase were informative and aided us in making minor changes to the protocol. Additionally, the rigorous linguistic and cultural analysis of the items, the systematic selection, and the data gathered in-field testing, lends strong evidence that test items are culturally and linguistically appropriate to be used in Arabic-speaking samples in Lebanon. It should be noted however, that the sample was not screened for cognitive impairment, and information on health status (ex. Presence of psychiatric illnesses) was not collected. Study 2 will examine additional psychometric properties in a convenience sample of adults known to be healthy or to have verbal learning and memory impairment.

Study 2

The purpose of this study was to obtain initial evidence of reliability and validity. The VMAT was administered on healthy participants (n = 199), of which a subsample was re-tested (n = 56). Additionally, we investigated the instrument's ability to detect differences between groups known to differ on the latent trait (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). For this, the VMAT was administered on a group of patients diagnosed with Multiple Sclerosis (MS) known to often have memory impairment (Chiaravalloti & DeLuca, 2008).

Methods

Procedures

This study was part of a larger project that aimed to validate a cognitive battery for MS patients in Arabic. Before testing, all participants provided written informed consent.

Training for data collectors

We recruited and trained seven data collectors who were either in their senior year or holders of an undergraduate degree in psychology. Training consisted of a three-hour practical workshop that included a live demonstration of flawed and flawless administrations. Then, the data collectors were observed in three mock administrations, until no errors were detected. Authors PZ and NF supervised the training and ongoing data collection.

Sampling and inclusion criteria

We recruited participants from the community via flyers, social media, and snowball techniques, and attempted to purposefully target potential participants from different areas of Lebanon and age groups. There was no incentive for participating. Although this is a pilot norming study, and we did not aim for a representative sample, we conducted a-posteriori tests to gauge the representativeness of our convenience sample against the Lebanese population. We compared our sample distribution against the distribution of the population in the six Lebanese governorates (Central Administration of Statistics in Lebanon, 2010), using a chi-square goodness-of-fit test (Agresti, 2002). Our sample was similar in distribution to the population of 4 out of 6 governorates but underrepresented individuals from Bekaa, and overrepresented individuals from Beirut (χ^2 (df = 5) = 127.098, p <.001).

To ensure that the community population was relatively free of impairment, we excluded those with a self-reported history of neurological disorders, traumatic brain injury, psychiatric disorders (including alcohol and drug abuse/dependence), and those taking medications that affect cognition, in the past three months (such as antipsychotics, anticonvulsants, or antidepressants). Since our sample was embedded in a larger study that assessed for depression and cognitive impairment, we also excluded those who screened positive for depression on the Hopkins Anxiety and Depression Checklist – Arabic (HSCL-25) (Fares et al., 2019; Mahfoud et al., 2013), and cognitive impairment on the Arabic Montreal Cognitive Assessment (MoCA) (cut-off score of 26 for individuals below the age of 60, and 24 with those≥60 years) (Carson et al., 2018; Nasreddine et al., 2005; Rahman & El Gaafary, 2009).

We also selected patients with Multiple Sclerosis (MS) from the American University of Beirut Medical Center, Nehme, and Therese Tohme Multiple Sclerosis center, as part of an ongoing study. This subsample was used to determine the discrimination ability of the VMAT scores between matched groups that differ on illness status (i.e., community versus control). To calculate the required sample size of MS patients matched on age, sex, and years of education, we used the G-POWER software (Faul et al., 2008) and set criteria for a 2-tailed independent sample t-test, with an alpha of 0.05, power of 0.80, The desired effect size was 0.87 based on previous studies that compared verbal memory between MS patients and healthy controls (Stegen et al., 2010). The calculation yielded a sample size of 17 MS patients and 33 healthy participants. However, such matched data was only available using 16 patients who were used for later analysis.

Participants

Our community sample (n = 199) consisted of two-thirds women (62%), with an average age of 37.9 (SD = 16.63). Table 2 shows the characteristics of the sample. A subsample of 56 community participants was retested after an average of 22.13 days (SD = 3.71). Those who were retested were more likely to be younger than those who were not (Mdn = 25.645, U = 2885, p = 0.002), male, (χ^2(df = 1) = 14.197, p < 0.001), and living in Beirut (χ^2 (df = 5) = 13.766, p = 0.017). This is not unusual, since most of the participants who agreed to return to the study site for a retest, were living in the medical center's vicinity (Beirut). The subsamples did not differ on any other variable (education, employment, native language, or Arabic language proficiency).

The MS sample (n = 16) consisted of 13 females and 3 males, with a mean age of 34.69 (SD = 11.89), and 13.94 years of education (SD = 3.07). The matched healthy sample (n = 32), consisting of 6 males and 26 females, with a mean age was 34.59 (SD = 12.07), and 14.25 (SD = 2.48), years of education, on average.

Statistical analysis

Two research assistants entered the data, and 30% of it was reentered for quality assurance. All analyses were performed on SPSS-25, significance was set at p < 0.05, and two-tails were used. When scores violated the assumption of normality (Shapiro-Wilk p < 0.005), non-parametric alternatives were used.

We derived primary scores/indices (*total number of words recalled in the first trial, and over the five trials,*

Table 2. Sample characteristics.

Demographics	N = 199	%	M	SD
Age	-	-	37.90	16.633
Age Range				
16–35.9	102	51.30%		
36–65.9	82	41.20%		
≥ 66	15	7.50%		
Sex				
Male	76	38.20%		
Female	123	61.8%		
Reported Native language				
Arabic	111	55.78%		
English or other	2	1%		
Reported language fluency (scale 1 to 5)				
Arabic	-	-	4.96	0.20
English	-	-	4.125	1.329
French	-	-	2.636	2.128
Highest Attained Education				
Elementary to Intermediate	7	3.50%		
Some High School or Vocational Education	18	9%		
Completed High School	51	25.60%		
University	123	61.80%		
Employment				
Employed	115	57.80%		
Unemployed	84	42.20%		
Marital Status				
Married	76	38.191%		
Divorced or Separated	10	5.026%		
Widowed	5	2.513%		
Single	94	47.236%		
In a relationship	14	7.035%		
Annual Income (in USD)				
< 6,000 $	41	20.603%		
6,000 $ – < 10,000 $	11	5.528%		
10,000 $ – < 15,000 $	14	7.035%		
15,000 $ – < 25,000$	25	12.563%		
25,000 $ – < 50,000 $	23	11.558%		
≥ 50,000 $	15	7.538%		
Unsure	26	13.065%		
No Answer	44	22.111%		
Governorate				
Mount Lebanon	47	23.60%		
Beirut	42	21.10%		
North Lebanon	14	7%		
Nabatiyeh	4	2%		
South	3	1.50%		
Bekaa	1	0.50%		
Physical Activity Levels				
None	55	27.60%		
1–5 times per month	12	6%		
1–2 times per week	46	23.10%		
3–4 times per week	46	23.10%		
5 or more times per week	40	20.10%		
Alcohol Intake During the Month prior to Testing				
None	98	49.20%		
1–3 times per month	21	10.60%		
1 time per week	29	14.60%		
2–3 times per week	40	20.10%		
4–5 times per week	9	4.50%		
6–7 times per week	2	1%		

number of words recalled in the short and long, free and cued recall trials, and recognition discriminability) by calculating the arithmetic mean of scores across participants. The index of recognition discriminability[2] was defined as the ability to endorse the 15 target items *and* reject all 30 distractors. This was calculated based on Corwin's (1994) signal detection theory, which uses hit and false-positive rates, rather than absolute numbers. The d′ raw score on the VMAT can range from +4.00 (15 hits, 0 false positives) to −4.00 (0 hits, 30 false positives) (Corwin, 1994).

Secondary scores/indices included learning slope, clustering methods, intrusions, and repetitions, serial positions, and recognition patterns computed as follows: The learning slope- that reflects an improvement in the number of words recalled from trials 1 to 5-, was calculated using the two-slope method (Delis et al., 2000). Semantic and serial clustering were derived by contrasting observed organization against an expected organization.[3] The extent to which the participant recalled words from the first, end or middle of List A was computed as the percent of total correct responses for the first or last 4 words of list A, and 7 middle words, related to the total of correct responses for the first 5 trials.

We used Spearman's correlation to examine test-retest reliability and Cronbach's alpha and split-half procedures to measure internal consistency (computed onto whether or not a word was said). To compare scores between community and MS samples, we used one-way ANOVAs and Cohen's *d* effect sizes. Finally, we conducted Multiple Regression Analyses (MRA); using the enter method, to examine which independent variables of age (in years), education (less or more than 12 years of education), and sex, predicted VMAT scores, and to derive regression-based norms (Oosterhuis et al., 2016). Several dependent variables[4] did not meet assumptions of MRA and could not be corrected, and were therefore excluded from further analysis. Variables that met all assumptions and used as MRA outcomes were: Trial 1, Total score on trials 1 to 5, Short Delay Free Recall, and Long Delay Free Recall. Reliable Change Indices (RCIs) were also computed, with 90% and 95% Confidence Intervals (CI).

Results

VMAT performance scores

A meaningful test of learning and memory should include various scores and indices (composite scores) to account for the different neurocognitive processes of learning and memory. Thus, we calculated primary scores that give an overview of encoding and retrieval, and secondary processes scores for a granular understanding of memory processes such as methods of clustering information (e.g., semantic versus serial clustering), learning slope, and errors in recall and discrimination (e.g., intrusions, repetitions, recognition discriminability). These are reported in Table 3.

Reliability

Temporal consistency

The test scores of the retested sample (*n* = 56)were relatively constant across time. For the primary measures, test-retest reliability coefficients (*Rho*) were adequate and ranged between 0.46 and 0.65 (Table 4). Selected process measures showed slightly lower scores. For the measures which had statistically significant results, Spearman's Rho ranged between 0.317 (repetitions of trial 1) and 0.467 (interference trial; list B). The data from Table 4 can be used to calculate the *Reliable Change Index*, which aids clinicians to understand whether a change in score across time is within expected limits.

Internal consistency

The test items showed adequate internal consistency. Cronbach's alpha for trials 1 to 5, short delay (free and cued trials), and long delay (free and cued trials) was 0.83, 0.74, and 0.78 respectively. For split-half reliability, the Spearman-Brown coefficient, with equal variances assumed, was 0.49 for combined scores of T1 and T3, 0.93 for combined T2 and T4, and 0.71 for T3 and T5.

Validities

Known group differences

One way of demonstrating the discrimination ability of a test score is to evaluate the extent to which the score could adequately differentiate between matched participants that vary only on the latent trait measured (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). MS patients scored lower than community participants on most VMAT segments. In particular, MS patients scored significantly lower than healthy participants on trial 1 ($F(1, 46) = 5.930$, $p = 0.019$), trial 4 ($F(1, 46) = 5.791$, $p = 0.020$), trial 5 ($F(1, 46) = 5.677$, $p = 0.021$), total score on trials 1 to 5 ($F(1, 46) = 5.235$, $p = 0.027$), short delay free recall ($F(1, 46) = 5.368$, $p = 0.025$), and long delay free recall ($F(1, 46) = 9.450$, $p = 0.004$). There were no significant differences on trials 2 and 3. Effect sizes were large for Trial 1 ($d = 0.7$), and Short and Long Delay Free Recall ($d = 0.6$).

Relationships with known predictors

By examining which demographic variables weigh in on the prediction of VMAT scores, we can provide evidence that the scores are related to expected demographics. We ran four multiple regressions (one for each outcome). We found, as expected, age was consistently the main significant predictor in all MRA, and was associated with a decrease in scores. It predicted 5.5% of the variance on trial 1 ($F(3, 194) = 3.794$, $p = 0.011$), 11.9% on short delay free recall ($F(3, 194) = 8.707$, $p < 0.001$), and 13.8% on long delay free recall trial ($F(3, 195) = 10.418$, $p < 0.001$). For the total scores on trials 1 to 5, age *and* sex explained 14.4% of the variance ($F(3, 194) = 10.887$, $p < 0.001$), so being younger ($B = -0.152$, $p < 0.001$), and female ($B = 2.511$, $p = 0.036$), led to improved performance.

Table 3. Primary and secondary scores and indices for VMAT (n = 199).

	MIN	MAX	M(words)	SD	Median	Variance
Primary Scores/Indices						
Trial 1	1	13	7.09	2.03		
Trial 1 to 5	21	72	51.94	8.70		
List B (interference trial)	0	13	6.60	2.16		
Short Delay Free Recall	5	15	10.48	2.42		
Short Delay Cued Recall	5	15			11	5.83
Long Delay Free Recall	5	15	10.94	2.50		
Long Delay Cued Recall	4	15			12	5.87
Recognition Discriminability[1]	0.73	4			3.46	0.53
Secondary Scores/ Indices						
Learning Slope [T1 to T5] [2]					1.20	0.39
Semantic Clustering [3]						
Trial 1					1.14	1.88
Trial 2					2.43	4.23
Trial 3					2.86	6.18
Trial 4					3.29	7.63
Trial 5					3.43	8.43
Short Delay Free Recall					3.14	6.20
Long Delay Free Recall					4.43	7.49
Average					2.86	3.62
Serial Clustering [4]						
Trial 1					0.53	0.95
Trial 2					0.53	1.68
Trial 3					0.53	1.94
Trial 4					1.13	2.96
Trial 5					1.13	3.18
Short Delay Free Recall					0.40	1.70
Long Delay Free Recall					0.33	1.61
					0.75	0.79
Serial Position Effects[5]						
Primacy					29.82	22.09
Middle					29.31	23.68
Recency					40.74	27.69
Raw Recognition Score					44	6.47
Recognition Hit Rates					0.91	0.02
Recognition False Positives					0.01	0.00

[1]This refers to d prime (d'), or the ability to endorse all 15 target items and reject all 30 distractors. The VMAT d' can range from a +4.00 [15 hits, 0 false positives] to −4.00 [0 hits, 30 false positives]. [2] Improvement in number of words recalled from trials 1 to 5; [3] Recalling words according to their semantic categories; [4] Recalling words in the order set by the test; [5] Recalling words from beginning, end, or middle of the stimulus list.

Norms

Since age was a consistent predictor of scores, we produced age-based regression norms for trial 1, total trials 1 to 5, short delay free, and long delay free trials.[5]

Table 4. Test-retest reliability for primary and process measures of list A.

			Test (n = 199)				Re-test (n = 56)							
Measures	Rho	P	M	SD	Median	Variance	M	SD	Median	Variance	Mdiff	SD	90% CI	95% CI
Primary Measures														
Trial 1	0.03	0.80	7.64	1.65	8.00	2.71	10.48	2.22	10.00	4.95	2.84	2.60	−1.44, 7.11	−2.26, 7.93
Total 1–5	0.56	0.00	55.68	7.43	55.00	55.24	63.05	7.53	63.50	56.67	7.38	6.49	−3.29, 18.04	−5.34, 20.09
Short Delay Free Recall	0.60	0.00	11.46	2.11	11.50	4.44	13.09	1.85	13.00	3.43	1.63	1.77	−1.29, 4.54	−1.85, 5.1
Short Delay Cued Recall	0.65	0.00	11.73	2.03	12.00	4.13	13.07	1.79	14.00	3.19	1.34	1.63	−1.35, 4.02	−1.86, 4.54
Long Delay Free Recall	0.64	0.00	11.88	2.02	12.00	4.08	13.38	1.94	14.00	3.77	1.50	1.77	−1.41, 4.41	−1.97, 4.97
Long Delay Cued Recall	0.56	0.00	12.20	2.01	12.50	4.05	13.30	1.76	14.00	3.09	1.11	1.60	−1.53, 3.74	−2.04, 4.25
Recognition Discriminability	0.46	0.00	3.46	0.53	3.49	0.28	3.68	0.46	4.00	0.21	0.22	0.53	−0.65, 1.1	−0.82, 1.27
Process Measures														
Learning Slope [T1 to T5]	−0.18	0.19	1.23	0.50	1.30	0.25	0.79	0.43	0.80	0.19	−0.44	0.72	−1.62, 0.74	−1.85, 0.97
List B	0.47	0.00	7.24	2.19	8.00	4.81	7.32	2.36	7.00	5.57	0.07	2.36	−3.82, 3.96	−4.56, 4.71
Semantic Clustering														
T1 to T5 Average	0.05	0.73	3.30	1.90	2.71	3.61	5.84	2.55	5.99	6.51	2.54	3.08	−2.53, 7.61	−3.5, 8.58
Serial Clustering														
T1 to T5 Average	−0.09	0.50	0.97	1.08	0.77	1.18	1.02	1.09	0.76	1.20	0.05	1.52	−2.45, 2.56	−2.93, 3.04
Raw Recognition Score	0.40	0.00	43.73	1.38	44.00	1.91	44.30	1.04	45.00	1.09	0.57	1.39	−1.71, 2.85	−2.15, 3.29
Intrusions														
Trial 1	0.39	0.00	0.25	0.74	0.00	0.55	0.30	0.54	0.00	0.29	0.05	0.75	−1.18, 1.29	−1.41, 1.52
Short Delay Free Recall	0.10	0.46	0.14	0.35	0.00	0.12	0.32	0.54	0.00	0.29	0.18	0.61	−0.82, 1.18	−1.01, 1.37
Short Delay Cued Recall	0.18	0.18	0.23	0.47	0.00	0.22	0.38	0.68	0.00	0.46	0.14	0.72	−1.05, 1.33	−1.28, 1.56
Long Delay Free Recall	−0.06	0.69	0.25	0.44	0.00	0.19	0.30	0.54	0.00	0.29	0.05	0.70	−1.1, 1.2	−1.32, 1.42
Long Delay Cued Recall	0.35	0.01	0.32	0.51	0.00	0.26	0.32	0.58	0.00	0.33	0.00	0.57	−0.94, 0.94	−1.12, 1.12
Repetitions														
Trial 1	0.32	0.02	0.18	0.39	0.00	0.15	0.34	0.67	0.00	0.54	0.25	0.72	−0.93, 1.43	−1.16, 1.66
Short Delay Free Recall	0.07	0.60	0.59	0.87	0.00	0.76	0.07	0.26	0.00	0.45	−0.25	1.07	−2, 1.5	−2.34, 1.84
Short Delay Cued Recall*														
Long Delay Free Recall	0.03	0.82	0.21	0.41	0.00	0.17	0.16	0.42	0.00	0.17	−0.05	0.59	−1.02, 0.91	−1.2, 1.09
Long Delay Cued Recall	−0.04	0.79	0.04	0.19	0.00	0.04	0.04	0.19	0.00	0.04	0.00	0.27	−0.44, 0.44	−0.53, 0.53

*Cannot be computed because at least one of the variables is constant.

A conversion table of raw scores to standard scores based on age is available from the authors. Standard Z-scores can also be calculated using the following formula: Standard Z-score = (Obtained Raw Score – Expected Score)/RMSE. Table 5 provides the information needed to calculate the expected score for each age and score type.

General discussion

In response to growing demand, we constructed a new test of verbal learning using qualitative and quantitative methodologies in a language used by more than 370 million people worldwide.

Results from study 1 ascertained that the breadth of the VMAT content and stimuli were culturally and linguistically appropriate, at least in the Arab country of Lebanon. Importantly, by generating stimuli from the ground up, instead of translating/adapting them, we circumvented several possible biases. For example, most English tests of verbal memory use the categories of animals, food, and clothing, and none used stationery items. However, in our qualitative-quantitative analysis, we found that stimuli related to stationery fit all the inclusion criteria for word lists. Additionally, the methodology allowed us to select words with balanced variations between MSA and colloquial Arabic. Although we attempted to maximize this balance, future studies would still need to test whether the selected items function similarly in other Arab countries.

In Study 2, we tested the VMAT on a large sample of Lebanese participants from the community and clinical settings. Scores decreased as expected with age, consistent with findings on age-related memory decline (Zelinski & Burnight, 1997). Scores were able to discriminate between those without and with known memory impairment (MS patients), so the latter scored significantly lower than healthy controls on all primary measures. This supports the test's ability to discriminate between groups. In further analysis, it would be important to discern a memory profile for patients with MS, even though the evidence regarding the specific memory deficits in acquisition versus retrieval are still mixed (Lafosse et al., 2013). Future studies could also identify whether the VMAT shows specific clinical and preclinical indicators for MCI, AD, and other memory disorders and whether the VMAT can differentiate memory patterns between cortical and subcortical dementias.

Finally, scores showed acceptable consistency across time, despite practice effects among healthy controls. Such test-retest reliability is crucial, it allows test-users to judge whether the test can be used to track change across time. The stability coefficients of VMAT's primary measures of *short* and *long* recall are similar to those obtained on the original CVLT-II, and follow-up studies (Woods

Table 5. Data and formulas to produce standard scores for VMAT primary measures.

	Regression-based parameters		Percentile ranks[4]										
Measure[1]	*Predicted score calculation*	*RMSE[2]*	*1*	*2*	*5*	*16*	*25*	*50*	*75*	*84*	*95*	*98*	*99*
	*Y intercept + (age B*age)*												
Trial 1	7.606+(−0.027* age)												
1.988	2	2	3	5	5	7	8	9	10	11	12		
Total 1–5	47.936+(−0.152* age)												
8.106	30	34	37	42	45	52	58	60	65	68	69		
Short Delay Free Recall	10.443+(−0.044* age)												
2.289		5	6	7	8	10	12	13	14	14	15		
Short Delay Cued Recall*	-		5	5	6	7	8	10	12	13	14	14	15
Long Delay Free Recall	10.583+(−0.047* age)	2.335		5	6	8	9	11	12	13	14	15	15
Long Delay Cued Recall*	-		5	5	6	8	9	11	12	13	14	15	15
Recognition Discriminability*	-		1.32	1.45	1.8	2.39	2.58	3.29	3.53	3.7	3.91	3.96	3.98

[1]Each type of score (ex. Trial 1), has its unique predicted score equation and RMSE value, that should be used for Z-score computation.
[2]RMSE: Root Mean Square Error.
[4]Scores corresponding to each percentile rank were rounded to the nearest whole number for all measures (exception: recognition discriminability).
*MRAs assumptions were violated, and regression results were not used.

et al., 2006), and the HVLT-R (Benedict et al., 1998). Therefore, the VMAT is similar to other tests in this function and can be used in longitudinal study designs and clinical encounters to detect cognitive change across time.

The clinical usefulness of the VMAT can be evaluated by examining how well scores measure the intended constructs. The initial attention span is reflected in one's ability to recall the words from the first time, and this is reflected in scores obtained on Trial 1. On average, our sample encoded seven words on the first attempt (±2). This is consistent with findings in similar tests like the HVLT-R (Benedict et al., 1998), but slightly higher than those in the CVLT-II (Woods et al., 2006). This means one can make valid interpretations of the initial attention span using VMAT Trial 1 data.

The ability to retrieve information learned only a few minutes ago, with and without cues, is another cognitive ability crucial for diagnoses, treatment planning, and outcome research. This is reflected in scores on *short delay free* and *cued* recall. On average, our sample freely remembered about 10 words (±2), consistent with the learning rate of the CVLT-II (Woods et al., 2006). However, contrary to previous findings, providing cues did not improve performance significantly. This is likely because our participants were more effective than other samples in spontaneously clustering the information semantically even during the free recall, especially that most of the individuals received a university education. For example, while our sample clustered words semantically at an average of 3.67 words (median 2.86), healthy participants in the CVLT-II clustered 0.73 to 1.58 words. Therefore, our sample had little room for improvement in recall when semantic cues were provided.

Long-term memory of information learned about 20 minutes prior, is measured through scores on *long delay-free and cued recall*. Our sample recalled an average of about 11 words (±2), with a modest improvement after cues were given. However, in the long delay free recall conditions, participants recalled more words overall, therefore the slight improvement when cues were given led to a ceiling effect in the long delay cued recall condition (11.1 + 2.42 = 14.5). This phenomenon is not unusual, as other tests also show extreme kurtosis on some learning measures (Benedict et al., 1998).

Finally, a useful outcome tool can ascertain with some confidence whether the change observed in a test-taker is statistically significant. To calculate whether the observed change is beyond expected, the reader can calculate the Reliable Change Index, using obtained raw scores, and the confidence intervals, provided in Table 4 (Ogles et al., 1996). If the raw difference between test scores obtained (T2 – T1), falls within the confidence interval, then the change is unlikely to be statistically significant. This information, along with additional data regarding the test-takers fatigue, motivation, and other variables, can be used to ascertain improvement or regression in patients. Further studies and analyses can provide more data on the temporal stability of specific scores within clinical populations. Like any test, the VMAT results should be interpreted by a competent professional, in junction with other sources of data from tests that measure multiple cognitive domains, as well as history, imaging, and collateral reports.

Our study is not without limitations, and interpretations on the basis of the VMAT scores should be in light of them. Primarily, while our sample was geographically diverse (within Lebanon), it is not proportionate or representative to the national demographics of age. Also, most of the participants enrolled in the study were young and with high educational attainment (common finding among this population group (UNSECO, 2020)). The little variability in educational level in this sample might have masked the effect of education on the outcomes of the regression models. Other studies conducted in the Arab region found a positive association between education and verbal memory test scores (Azdad et al., 2017; Poreh et al., 2012; Rahmani et al., 2017). Therefore, the data is considered a prelude for future national norms, and interpretations based on them should be made with caution. Second, the matched sample with diagnosed MS is assumed to have verbal learning and memory deficits consistent with their medical presentation, but this assumption was not tested formally using well established cognitive tests. Therefore, this MS sample is a proxy for possible, but not verified verbal memory impairment. Additionally, we did not test whether the VMAT scores converge with another established memory test in Arabic, as we are not aware of another test in Arabic that has strong evidence of validity in this population.

Future directions

Although current data can be used as an informal reference in clinical settings, future studies are ongoing to develop norms based on a larger representative samples of age, education, and gender for healthy populations in Lebanon and other Arab countries. Despite the adequate test-retest reliability of the current versions, and its tolerance by patients with neurological illness, it would still be beneficial to construct an alternate version and a briefer version can be used for screening purposes and research settings.

Notes

1. "Mushroom" was not perceived as a vegetable by participants, and was replaced by "radish"; "stapler" was unfamiliar to participants because it has different pronunciations, and was replaced by "stamp"; "pigeon: was perceived as having a sexual connotation, so it was replaced by another type of bird (bulbul); "Nai" in recognition was considered unfamiliar and was replaced by "bell" (both words are distractors as an unrelated category).
2. The raw d′ score reflects the absolute difference in SD units between the examinee's hit rate (signal) and false-positive rate (noise). Hit rates were computed using the following formula: [total # of hits + 0.5]/[total # of targets + 1], and false-alarm rates were computed using the following formula [total # of false alarms + 0.5]/[total # of distracters + 1].
3. For semantic clustering, EXP was = $\frac{[(r-1)(m-1)]}{NL-1}$ = $\frac{[(r-1)(3-1)]}{15-1}$ = (r − 1)/7, and for serial clustering EXP was = (r − 1)/NL = (r − 1)/15.
4. In the short delay cued recall trial residuals were not equal across the regression line, while in the long delay cued recall scores, and recognition discriminability, assumptions of homoscedasticity and normality were violated.
5. Since age and sex, predicted only one score, we made a judgment to not include gender-based norms.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Novartis, MENACTRIMS funded by SANOFI-AVENTIS GROUPE (DMCC Branch) (NUR.HD1.05), representative office of Sanofi - Aventis Groupe SA.

ORCID

Pia Zeinoun http://orcid.org/0000-0002-9888-1533
Samia J. Khoury http://orcid.org/0000-0003-3198-6063
Hala Darwish http://orcid.org/0000-0001-5744-6466

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). John Wiley & Sons.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Azdad, A., Benabdeljlil, M., El Haimer, A., Boutbibe, F., & Faris, M. E. A. (2017). Adaptation and standardization of the free and cued selective reminding test to the Moroccan population. *Journal of the Neurological Sciences*, *381* (Suppl), 318. https://doi.org/10.1016/j.jns.2017.08.902

Benedict, R. H., Schretlen, D., Groninger, L., & Brandt, J. (1998). Hopkins verbal learning test–revised: Normative data and analysis of inter-form and test-retest reliability. *The Clinical Neuropsychologist*, *12*(1), 43–55. https://doi.org/10.1076/clin.12.1.43.1726

Carson, N., Leach, L., & Murphy, K. J. (2018). A re-examination of montreal cognitive assessment (MoCA) cutoff scores. *International Journal of Geriatric Psychiatry*, *33*(2), 379–388. https://doi.org/10.1002/gps.4756

Central Administration of Statistics (Lebanon). (2010). *Multiple indicators cluster survey 2009*. http://cas.gov.lb/images/Mics3/CAS_MICS3_survey_2009.pdf

Chiaravalloti, N. D., & DeLuca, J. (2008). Cognitive impairment in multiple sclerosis. *The Lancet Neurology*, *7*(12), 1139–1151. https://doi.org/10.1016/S1474-4422(08)70259-X

Corwin, J. (1994). On measuring discrimination and response bias: Unequal numbers of targets and distractors and two classes of distractors. *Neuropsychology*, *8*(1), 110. https://doi.org/10.1037/0894-4105.8.1.110

Daouk-Öyry, L., Zeinoun, P., Choueiri, L., & van de Vijver, F. J. (2016). Integrating global and local perspectives in psycholexical studies: A GloCal approach. *Journal of Research in Personality*, *62*, 19–28. https://doi.org/10.1016/j.jrp.2016.02.008

Davis, K. L., Price, C. C., Kaplan, E., & Libon, D. J. (2002). Error analysis of the nine-word California verbal learning test (CVLT-9) among older adults with and without dementia. *The Clinical Neuropsychologist*, *16*(1), 81–89. https://doi.org/10.1076/clin.16.1.81.8330

Delis, D. C., Kramer, J. H., Kaplan, E., Ober, B. A. (2000). California verbal learning test (2nd ed.). Psychological Corporation, TX.

Fares, S., Dirani, J., & Darwish, H. (2019). Arabic validation of the hopkins symptom checklist-25 (HSCL) in a Lebanese sample of adults and older adults. *Current Psychology*, (1–8). https://doi.org/10.1007/s12144-019-00240-x

Fasfous, A. F., Al-Joudi, H. F., Puente, A. E., & Perez-Garcia, M. (2017). Neuropsychological measures in the Arab world: A systematic review. *Neuropsychology Review*, *27* (2), 158–173. https://doi.org/10.1007/s11065-017-9347-3

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2008). *G* Power Version 3.1. 2 [computer software]*. Uiversitat Kiel.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2004). *Adapting educational and psychological tests for cross-cultural assessment*. Psychology Press.

International Organization for Migration. (2016). Regional Office for the Middle East and North Africa. *Migration to, from and in the Middle East and North Africa Data Snapshot*. https://www.iom.int/sites/default/files/country/mena/Migration-in-the-Middle-East-and-North-Africa_Data%20Sheet_August2016.pdf

Kristjansson, E. A., Desrochers, A., & Zumbo, B. (2003). Translating and adapting measurement instruments for cross-linguistic and cross-cultural research: A guide for practitioners. *CJNR (Canadian Journal of Nursing Research)*, *35*(2), 127–142. https://pubmed.ncbi.nlm.nih.gov/12908201/

Lafosse, J. M., Mitchell, S. M., Corboy, J. R., & Filley, C. M. (2013). The nature of verbal memory impairment in multiple sclerosis: A list-learning and meta-analytic study. *Journal of the International Neuropsychological Society*, *19*(9), 995–1008. https://doi.org/10.1017/S1355617713000957

Lewis, M. P., Simons, G. F., & Fennig, C. D. (2013). *Ethnologue: Languages of the world*. SIL International.

Lezak, M. D. (2004). *Neuropsychological assessment*. Oxford University Press.

Mahfoud, Z., Kobeissi, L., Peters, T., Araya, R., Ghantous, Z., & Khoury, B. (2013). The Arabic validation of Hopkins Symptoms Checklist-25 against MINI in a disadvantaged suburb of Beirut, Lebanon. *The International Journal of Educational and Psychological Assessment (TIJEPA)*, *13*(1), 17–33. https://research-information.bris.ac.uk/en/publications/the-arabic-validation-of-the-hopkins-symptoms-checklist-25-agains

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Ogles, B. M., Lambert, M. J., & Masters, K. S. (1996). *Assessing outcome in clinical practice*. Allyn & Bacon.

Oosterhuis, H. E., van der Ark, L. A., & Sijtsma, K. (2016). Sample size requirements for traditional and regression-based norms. *Assessment*, *23*(2), 191–202. https://doi.org/10.1177/1073191115580638

Poreh, A., Sultan, A., & Levin, J. (2012). The Rey Auditory Verbal Learning Test: Normative data for the Arabic-speaking population and analysis of the differential influence of demographic variables. *Psychology & Neuroscience*, *5*(1), 57–61. https://doi.org/10.3922/j.psns.2012.1.08

Rahman, T. T. A., & El Gaafary, M. M. (2009). Montreal cognitive assessment Arabic version: Reliability and validity prevalence of mild cognitive impairment among elderly attending geriatric clubs in Cairo. *Geriatrics & Gerontology International*, *9*(1), 54–61. https://doi.org/10.1111/j.1447-0594.2008.00509.x

Rahmani, F., Haghshenas, H., Mehrabanpour, A., Mani, A., & Mahmoodi, M. (2017). Shiraz Verbal Learning Test (SVLT): Normative data for neurologically intact speakers of Persian. *Archives of Clinical Neuropsychology*, *32*(5), 598–609. https://doi.org/10.1093/arclin/acx030

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93–115). Psychology Press.

Stegen, S., Stepanov, I., Cookfair, D., Schwartz, E., Hojnacki, D., Weinstock-Guttman, B., & Benedict, R. H. (2010). Validity of the California verbal learning test–II in multiple sclerosis. *The Clinical Neuropsychologist*, *24*(2), 189–202. https://doi.org/10.1080/13854040903266910

United Nations Educational, Scientific and Cultural Organization (UNSECO). (2020). *Education and literacy: Lebanon*. Retrieved May 6, 2020, from http://uis.unesco.org/country/LB

Urbaniak, G., & Plous, S. (2013). *Research randomizer* (Version 4.0) [Computer software]. Research Randomizer. http://www.randomizer.org/

Van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–63). Psychology Press.

Woods, S. P., Delis, D. C., Scott, J. C., Kramer, J. H., & Holdnack, J. A. (2006). The California Verbal Learning Test–second edition: Test-retest reliability, practice effects, and reliable change indices for the standard and alternate forms. *Archives of Clinical Neuropsychology*, *21*(5), 413–420. https://doi.org/10.1016/j.acn.2006.06.002

Zeinoun, P., Iliescu, D., & El Hakim, R. (2020). *Psychological tests in Arabic: A systematic review of methods for adaptation and validity and guidelines for future test development*. Manuscript submitted for publication.

Zelinski, E. M., & Burnight, K. P. (1997). Sixteen-year longitudinal and time lag changes in memory and cognition in older adults. *Psychology and Aging*, *12*(3), 503. https://doi.org/10.1037/0882-7974.12.3.503