Journal of Psychiatric Research 114 (2019) 170–177

Contents lists available at ScienceDirect

Journal of Psychiatric Research

journal homepage: www.elsevier.com/locate/jpsychires

Development of VM-REACT: Verbal memory RecAll computerized test

Sharon Naparstek[a,b], Dawlat El-Said[a,b], Michelle L. Eisenberg[a,b], Joshua T. Jordan[a,c], Ruth O'Hara[a,b,1], Amit Etkin[a,b,*,1]

[a] Department of Psychiatry and Behavioral Sciences and Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, 94304, USA
[b] Sierra Pacific Mental Illness, Research, Education, and Clinical Center (MIRECC), Veterans Affairs Palo Alto Healthcare System, Palo Alto, CA, 94304, USA
[c] Department of Psychiatry, University of California, San Francisco, USA

ARTICLE INFO

Keywords:
Neuropsychological assessments
Verbal memory
Computerized tests
Aging

ABSTRACT

When tracking the progression of neuropsychiatric or neurodegenerative diseases, assessment tools that enable repeated measures of cognition and require little examiner burden are increasingly important to develop. In the current study, we describe the development of the VM-REACT (Verbal Memory REcAll Computerized Test), which assesses verbal memory recall abilities using a computerized, automated version. Four different list versions of the test were applied on a cohort of 798 healthy adults (ages 20–80). Recall and learning scores were computed and compared to existing gender- and age-matched published norms for a similar paper-and-pencil test. Performance was similar to existing age-matched norms for all but the two oldest age groups. These adults (ages 60–80) outperformed their age-matched norms. Processing speed, initiation speed, and number of recall errors are also reported for each age group. Our findings suggest that VM-REACT can be utilized to study verbal memory abilities in a standardized and time efficient manner, and thus holds great promise for assessment in the 21st century.

1. Introduction

Tracking changes in mental and cognitive states has long been a goal of neuropsychological assessments. Tracking memory abilities is key when working with young children with learning disabilities, with elderly patients, when measuring the progression of neurodegenerative diseases, and when working with patients following brain injury. Cognitive (dys)functions in general, and memory (dys)functions more specifically, are increasingly recognized as core symptoms that cut across multiple psychiatric disorders (Etkin et al., 2013; Etkin et al., 2013; Weiser et al., 2004). Attempts to discover the neural underpinnings of psychiatric disorders have led to a broader interest in the utility of neuropsychological tests with these populations, across the lifespan. For example, it has been proposed that memory deficits that appear during the first episode of major depression might assist in early identification and intervention of future episodes (Lee et al., 2012). In Post-Traumatic Stress Disorder (PTSD), poor capacity for verbal memory prior to treatment predicted reduced clinical gains with treatment (Parslow and Jorm, 2007; Scott et al., 2017; Wild and Gur, 2008).

Multiple tasks have been developed for standardized assessment of verbal memory. Many of these measures are widely used, highly structured and have published norms (i.e., the Wechsler Memory Scale, WMS). Ultimately, however, traditional paper and pencil formats are impractical for the repeated assessment of cognitive abilities, including memory function, in large patient or research cohorts, particularly when performed remotely. The promise of computerized versions of tests has been recognized for over thirty years by the American Psychological Association (Schoenfeldt, 1989, American Psychological Association, 1986), and more recently by both the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology (Bauer et al., 2012). These organizations recognize the utility and potential of computerized tests while stressing their reliability and validity, ease of administration and unbiased interpretations (see also Butcher et al., 2000; Noyes and Garland, 2008). From a clinical perspective, computerized tests minimize examiner effects in administration and scoring (Wiens and Bryan, 1994) and maximize standardized administration; often include additional measures that cannot be obtained otherwise, such as inspection time or reaction time (Parsey and Schmitter-Edgecombe, 2013); and have automatic scoring algorithms that minimize scoring errors. From a research perspective, computerized tests can be administered to a large number of

* Corresponding author. 401 Quarry Road, Stanford, CA, 94305.
E-mail address: amitetkin@stanford.edu (A. Etkin).
[1] These authors contributed equally as senior authors.

https://doi.org/10.1016/j.jpsychires.2019.04.023
Received 5 November 2018; Received in revised form 23 April 2019; Accepted 25 April 2019

Table 1
Computerized verbal memory tests.

Test	Key processes	Measures obtained	Alternate forms
Computerized neuropsychological scan/Penn word memory test (Gur et al., 2001, Gur et al., 1993)	Verbal recognition of words: Visual presentation of 20 target words, followed by 2 recognition trials of 40 words each (targets + distractors).	Number of words correctly recognized in the immediate and delayed.	NA
IntegNeuro (Paul et al., 2005)	Free verbal recall: auditory presentation of 12 words followed by a verbal recall trial through a voice recording system. The list is repeated 4 times followed by an interference list, immediate and delayed recall and a recognition trial. Verbal responses are manually scored by the examiner.	Number of words correctly recalled across the four learning trials, the immediate recall trial and the delayed recall trial, and the total number of correctly identified words on the recognition trial.	NA
WebNeuro (Silverstein et al., 2007)	Verbal recognition: visual presentation of 12 words followed by a recognition trial. The list is repeated 4 times and includes a delayed recognition trial.	Number of words correctly recognized across the four learning trials and the delayed trial.	NA
MicroCog (Elwood, 2001; Powell, 1993)	Story recognition test: Visual presentation of 2 stories followed by multiple choice questions on each story, including immediate and delayed recognition.	Immediate and delayed recognition.	NA
Neurobehavioral evaluation system (NES) (Arcia and Otto, 1992)	Verbal recognition test: visual presentation of 9 pairs of words followed by a recognition/matching task.	Number of words correctly recognized.	NA
NeuroTrax Mindstreams (Dwolatzky et al., 2003; Schweiger et al., 2003)	Verbal recognition: visual presentation of 10 pairs of words followed by a recognition/matching trial. The list is repeated 4 times followed by delayed recognition phase.	Number of words correctly recognized in each trial.	Available
CANTAB Verbal Recognition Memory (VRM) (Robins et al., 1994)	Verbal recall and recognition: visual presentation of 12 words followed by a free recall and force-choice recognition test (immediate and delayed). Verbal responses for recall are manually scored by the examiner.	Number of words correctly recalled, number of correct and incorrect responses for the immediate and delayed recognition.	Available
The Computerized Neuropsychological Test Battery (CNTB) (Veroff et al., 1991)	Visual presentation of 15 words followed by free recall. Verbal responses for recall are manually entered by the examiner.	Number of words correctly recalled, number and type of errors (intrusions, preservations).	NA

participants; are optimal for those who may be geographically remote from the laboratory or clinical site; administration can be carried out without the involvement of a clinician; and integrated results are usually available as soon as the test is completed. Even when patients or research participants can participate in-person, these time-locked procedures can be easily paired with the collection of physiological or neural measures. Indeed, standard task-based fMRI procedures have incorporated a range of cognitive measures to examine which brain neurocircuits subserve which cognitive functions (Gur et al., 2010). However, most of these tasks are not normed, and due to the constraints of the scanning procedure itself rely on motor responses which favor recognition memory over delayed recall. That said, over the past years, several computerized batteries have been developed, which include measures of verbal learning and memory (see Table 1).

As can be seen in Table 1, although there are several available memory tests, two major aspects are lacking. First, whereas traditional neuropsychological assessment relies strongly on ***recall*** abilities, most computerized tests (aside from the IntegNeuro and CANTAB) asses ***recognition*** alone (Wild et al., 2008). Assessment of recognition is highly valuable if it is interpreted in comparison to free recall, as it is then possible to determine whether the origin of the deficit is in retrieval or in acquisition (Elwood, 2001). However, when free recall is not assessed, recognition has less utility. Second, in the two computerized batteries that includes a verbal recall component, this is achieved by the verbal responses being manually scored by the examiner. Finally, most computerized tests (aside from Neurotrax and CANTAB) do not include alternate versions or forms for repeated measures. Repeated measures of cognitive abilities, and memory abilities specifically, are an important factor not only when monitoring the course of disease or deterioration (such as a neurodegenerative disease, i.e., Ewers et al., 2012), but also when assessing improvement or effectiveness of therapy. Importantly, both healthy individuals and those who suffer from cognitive deficits show a practice effect in repeated testing, namely, better performance when a test is repeated twice. The use of alternate forms or versions, especially within the memory domain, has shown to attenuate the practice effect and thus should be considered (Knight et al., 2007).

The current study therefore has four aims: 1) Create a computerized verbal memory test that addresses the limitations of previous tests by assessing recall abilities via the computer without the need for an assessor to record these responses; 2) Include alternate forms for repeated measures over time; 3) Provide data on a large number of participants from different age groups; and 4) Provide data for additional measures of performance that cannot be obtained in traditional paper-and-pencil tests, namely, measures of speed of processing.

The verbal memory test described here, the Verbal Memory REcAll Computerized Test (VM-REACT), is a computerized adaptation of the Rey Auditory Verbal Learning Test (RAVLT). In the traditional in-person administration, a list of 15 unrelated nouns (list A) is presented auditorily by the examiner five times, each followed by a free, verbal recall test. Following these five learning trials, a new list of 15 nouns is presented (interference list) followed by free recall. Following the presentation of the interference list, there is an immediate delay recall test of the first list and another delayed recall test after 20–30 min. The last two recall tests do not include a presentation of the list. In the traditional version, following the delayed recall trial, the examinee is asked to recognize the words from list A embedded in a list of 50 words or within a story. The English version has alternate forms enabling repeated measures. Test scores include information about acquisition, learning rate, interference and retention.

2. Methods

2.1. Memory test

The VM-REACT (Verbal Memory REcAll Computerized Test) is a computerized adaptation of the RAVLT described above. Importantly, in order to enable computerized implementation to the RAVLT two main adaptations were made: words are presented visually rather than auditorily, and recall is tested by typing of the words by participants

instead of verbal responses. Visual presentations are widely used in many computerized tests including all but one of the tests in Table 1 (i.e., IntegNeuro). When tested for cross-model reliability, traditional auditory tests and computer-based visual tests show medium to high reliability (Gur et al., 2001) suggesting the visual presentation is a valid one. Manual typing of individual word responses is not used in other tests. However, this possible modification is widely discussed in the literature including specific quality assurance (QA) measures (Schlegel and Gilliland, 2007).

The test was designed, deployed and administered via Inquisit Millisecond Software (https://www.millisecond.com). The Inquisit web platform ensures millisecond accuracy by prompting the participant to locally download the application. Thus, although the task was completed on the participant's local computer, the application prevents differences in reaction times and test timing due to varying computer hardware or internet speed. Additionally, the application takes over the computer, preventing multi-tasking.

Test instructions were carefully refined to resemble delivery of the in-person administration. Instructions were gradually presented (i.e., line by line) and participants were instructed to press the spacebar to move to the next line. During presentation, there was a 500 ms (ms) time lock, where key presses were disabled preventing the participant from rushing through the instructions prior to the test and allowing for the test to be entirely self-guided. In every trial, participants were presented with 15 words that appeared sequentially on the computer screen. Each word appeared for a total of 1000 ms with a 750 ms interval between word presentations. Stimulus duration and interstimulus interval were determined based on the investigator-administered, auditory version. Following each list presentation participants were presented with a screen including 20 text boxes and were instructed to type as many words as they can remember, in any order that comes to mind. To be similar to in-person administration, only one textbox was available at a time, with the other 19 textboxes greyed out. Participants were instructed to press the 'enter' key once they were ready to type the next word, which triggered the text box with their previous response to turn black and hide their response. Once participants completed recalling all the words, they were instructed to press 'enter' twice, and were then prompted with an additional instructions page asking whether they wanted to move on to the next trial or continue typing in the words they remembered. This additional step ensured that a list that contains no recalled words is a true response and not due to accidently pressing 'enter' and moving on to the next trial. This same procedure (a 15-word presentation followed by a free recall trial) was repeated 5 times, utilizing the same set of instructions and guidelines. Then, a new list, the interference list, was presented followed by a free recall trial. Following the interference list recall, participants were asked to recall the words from the first list again. Finally, after a 20–25-min delay (during which participants performed an unrelated attention test) participants were asked to report the words they recall from the first list.

2.2. Participants

Eight hundred and eighteen adults participated in the study. Participants were recruited via Amazon's Mechanical Turk (MTurk) platform and completed the test online. Recruitment was carried out in several launches between January 21, 2017 and December 19, 2017. Due to previous reports on the effect of age on memory task performance, participants were recruited specifically based on defined age groups to ensure adequate sampling across the lifespan. The number of females and males were also balanced within each age range (for detailed description see supplementary material).

Of the original 818 participants that completed the test, 20 were excluded from the analysis due to: a history of brain injury (11 participants), missing information on the demographic survey (5 participants) or negative values in recall trials, possibly due to a timing issue (4 participants). Demographics of the remining 798 appear in Table 2.

Table 2
Demographic characteristics of the sample.

Age Range	Gender	Age (mean)	Age (std)	N
20-29	Male	25.35	2.81	82
	Female	25.37	2.8	54
	Both	25.36	2.8	136
30-39	Male	33.7	2.65	109
	Female	34.28	2.88	122
	Both	34	2.8	231
40-49	Male	43.92	2.93	62
	Female	44.47	2.81	73
	Both	44.21	2.87	135
50-59	Male	54.82	3	65
	Female	55.41	2.92	80
	Both	55.14	2.96	145
60-69	Male	64.04	2.92	52
	Female	63.45	2.86	75
	Both	63.7	2.9	127
70-79	Male	71.92	1.89	13
	Female	72.18	2.6	11
	Both	72	2.2	24

2.3. Procedure

Interested participants meeting the criteria mentioned above were transferred to the test's custom launch page, prompted to locally install the Inquisit application and then press the start button. Once launched, participants were introduced to a set of demographic questions followed by the memory test. Each participant was assigned one of 4 alternate test forms (Geffen et al., 1994; Lezak et al., 2012) see supplementary material). Following the learning, interference and immediate recall trials, participants were presented with a filler task testing sustained attention (the Attention Network Test- Revised, Fan et al., 2009) that lasted between 20 and 25 min. When the attention task was over, a delayed recall trial was presented. Upon completion of the memory test, participants were asked to complete two additional components: 1) Typing speed test: participants were presented with 3 sentences and asked to type these sentences exactly as they were presented. Average typing time for each sentence was measured. 2) A set of questionnaires assessing emotional well-being. Upon completion of the study, participants were paid $10. Additional information regarding recruitment and procedure appear in the supplementary material.

3. Results

Each individual's responses were automatically scored (for detailed description see supplementary material). Fifteen different measures were derived from the test including raw scores and composite scores (Vakil et al., 2010; Vakil and Blachstein, 1997). Normed data including means and standard deviations for each gender and age range are reported separately for recall scores on each trial (Tables 3a and 3b composite scores (Tables 4a and 4b) as well as errors and processing speed (Tables 6a and 6b, Fig. 1). Additional analyses including structural analysis of the test, analysis of alternate test forms and the effects of time spent on the computer over performance appear in the supplementary material.

3.1. Recall and composite scores

We first examined the effects of age range and gender on the learning curve, which was calculated as the number of words recalled from Trial 1 thru Trial 5. A three-way analysis of variance (ANOVA) was carried out with gender (male, female), age range (20–29, 30–39, 40–49, 50–59, 60–69 and 70–79) and trial (1 through 5) as independent variables and number of words correctly recalled as the dependent variable. There was a main effect of trial F (4, 3144) = 849, p = .00, η^2 = 0.52. The linear trend was significant F (1, 786) = 1754, p = .00,

Table 3a
Males: Means and standard deviations of the recall scores.

Trial	20-29		30-39		40-49		50-59		60-69		70-79	
	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev
T1	6.2	3.03	6.61	2.74	6.44	2.55	7.08	2.67	7.19	2.77	6	1.96
T2	9.27	3.27	9.61	2.74	9.66	2.44	10.26	3.01	10	3.12	9.38	2.5
T3	10.88	3.33	11.42	2.71	11.15	2.35	11.51	2.72	11.27	2.6	11.31	2.78
T4	11.57	2.74	11.88	2.47	12.06	2.37	12.09	2.42	12.02	2.45	11	2.58
T5	11.79	3.28	12.6	2.59	13.06	1.82	12.63	2.29	12.56	2.44	11.85	2.44
List B	6.79	3.16	7.02	2.99	7	2.75	7.34	2.95	7.31	3.39	6.85	2.54
T6	9.37	4.49	9.43	4.77	11.29	3.36	10.52	4	10.04	4.72	9.85	3.65
T7	9.15	4.33	10.65	3.58	11.08	3.1	11.23	3.24	11.42	3.01	9.54	4.14

Table 3b
Females: Means and standard deviations of the recall scores.

Trial	20-29		30-39		40-49		50-59		60-69		70-79	
	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev
T1	6.94	2.74	7	3.01	7.15	3.07	7.35	2.72	7.41	2.65	5.91	2.47
T2	9.83	2.81	9.96	2.81	9.93	2.91	10.13	2.38	9.73	2.62	9.36	3.61
T3	11.15	2.89	10.93	3.17	11.86	2.24	11.68	2.52	11.36	2.63	10.18	2.96
T4	11.76	3.03	11.77	2.78	12.1	2.35	12.38	2.24	12.23	2.21	9.91	3.18
T5	12.5	2.86	12.51	2.76	12.85	2.25	13	1.93	12.29	2.68	10.91	2.88
List B	6.89	2.94	7.21	3.28	7.14	2.91	7.13	3.42	6.4	2.99	5.55	1.92
T6	11.26	3.52	10.02	4.13	10.53	4.42	10.11	4.47	10.55	3.68	9.27	3.58
T7	11.37	3.26	10.61	3.72	10.82	3.71	11.25	3.1	11.17	3.09	9.45	3.96

$\eta^2 = 0.7$, suggesting the number of words recalled increased from trial to trial, as expected. None of the other main effects or interactions were significant.

Previous studies have consistently reported an effect of age on performance in the test. Thus, a series of ANOVAs with age range (20–29, 30–39, 40–49, 50–59, 60–69, 70–79) as the independent variable, were carried out. Only three variables were modified by age: best learning trial (Trial 5) F (5, 792) = 3, p = .01, η^2 = 0.02, the total learning composite score F (5, 792) = 2.2, p = .04, η^2 = 0.01, and the delayed recall (Trial 7) F (5, 792) = 3, p = .01, η^2 = 0.02. Post hoc analysis of these effects revealed all three originated from a significant difference in performance between the 70–79 age group and the 60–69 age group. Namely, the oldest age group recalled fewer words in the best learning trial (Trial 5) and following a delay (Trial 7), in addition to displaying a lower ability to accumulate information over time (total learning) (F (1, 786) = 5, p = .02, η^2 = 0.00; F (1, 792) = 5, p = .02, η^2 = 0.00; and F (1, 792) = 3.9, p = .04, η^2 = 0.00; for Trial 5, Trial 7 and total learning, respectively).

To test the comparability of our current results to traditional paper and pencil versions, the mean scores of the most robust measures in each age group were compared to existing meta-norms (Schmidt, 1996). These measures include: the initial span (Trial 1), the best learning trial (Trial 5), delayed recall (Trial 7) and total learning composite score. Cohen's d values calculated as the difference between the means divided by the pooled standard deviations for each of these measures are presented in Table 5.

As can be seen in Table 5, there is a difference between the younger and older age groups. Within the younger age groups (20–59) Cohen's d values are mostly within the small to moderate range suggesting recall scores on trials 1, 5 and 7 and the total learning composite score in our sample are very similar to the norms previously published. However, in the two oldest age groups (60–69 and 70 to 79), the differences between our sample and previous published norms are quite high (Cohen's d values are in the large range). Importantly, since the comparison was carried out by subtracting the mean of the current sample from the age-equivalent norms, the negative values represent an interesting trend. Namely, in the current sample, the older groups (60–79) exceeded the performance of their age-equivalent norms in Trials 1, 5, 7 and in the total learning scores.

3.2. Error and processing speed

Mean number of errors are reported in Tables 6a and 6b for each gender and each age group. To examine the effect of gender and age range over errors a three-way ANOVA with gender (male, female), age range (20–29, 30–39, 40–49, 50–59, 60–69, and 70–79) and trial (Trial 1 through 5) as independent variables, and mean number of errors as the dependent variable. There was a main effect for trial F (4,

Table 4a
Males: means and standard deviations for composite scores.

Score	20-29		30-39		40-49		50-59		60-69		70-79	
	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev
TL	49.71	12.91	52.12	11.3	52.37	9.55	53.57	11.32	53.04	11.59	49.54	10.74
cTL	18.73	9.89	19.09	8.05	20.19	7.42	18.18	8.25	17.08	7.39	19.54	6.6
Rate	5.6	3.33	5.99	2.7	6.63	2.26	5.55	2.56	5.37	2.32	5.85	2.38
Pro	-0.6	3.23	-0.41	2.66	-0.56	2.8	-0.26	2.87	-0.12	2.87	-0.85	2.91
Retro	2.43	4.36	3.17	4.55	1.77	2.8	2.11	3.23	2.52	3.73	2	2.2
Reten	2.65	4.33	1.94	2.96	1.98	2.17	1.4	1.98	1.13	2.01	2.31	2.63

Table 4b
Females: means and standard deviations for composite scores.

Score	20-29		30-39		40-49		50-59		60-69		70-79	
	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev
TL	52.19	11.96	52.16	12.1	53.89	10.62	54.53	9.4	53.03	10.87	46.27	12.88
cTL	17.46	10.25	17.16	11.15	18.14	9.29	17.78	9.44	15.96	8.39	16.73	12.81
Rate	5.56	3.18	5.51	3.23	5.7	2.84	5.65	3	4.88	2.7	5	2.65
Pro	0.06	2.6	-0.21	2.99	0.01	2.81	0.23	3.23	1.01	2.34	0.36	1.43
Retro	1.24	2.73	2.49	3.58	2.32	3.9	2.89	3.88	1.75	3.3	1.64	1.43
Reten	1.13	1.99	1.89	2.88	2.03	3.07	1.75	2.5	1.12	2.48	1.45	1.86

Note: TL = total learning, the sum of words recalled over first five learning trials. This measure reflects the ability to accumulate information over time; cTL = corrected total learning, the total of five learning trials minus five times the number of words recalled in Trial 1. Represents an uncontaminated estimate of the individual's learning over time; Rate = learning rate, the difference between Trial 5 and Trial 1. Reflects the learning slope or learning process that is not affected by immediate learning; Pro = proactive interference, the difference between Trial 1 and List B. Represents the effect of previously learned information over the ability to acquire new information; Retro = retroactive interference, the difference between Trial 5 and Trial 6 (immediate recall) and represents the effect of new information over recall of previously learned material; Reten = retention, the difference between Trial 5 and Trial 7 and reflects the amount of information remembered over time.

Table 5
Cohen's d value for comparison between scores on VM-REACT and previously published meta norms.

Measure	20-29	30-39	40-49	50-59	60-69	70-90
T1	0.24	−0.06	−0.11	−0.47	−0.67	−0.27
T5	0.38	0.07	−0.34	−0.35	−0.46	−0.46
T7 (Delayed)	0.44	0.16	−0.25	−0.42	−0.82	−0.94
TL	0.69	0.16	−0.23	−0.71	−1.04	−1.34

Note: TL = total learning. Cohen's d, calculated as the difference between the means divided by the pooled SD. Values are interpreted as following: 0.2 represents a small difference, 0.5 represents a moderate difference and 0.8 represents a large difference between the means.

3144) = 850, $p = .00$, $\eta^2 = 0.52$, with a significant linear trend F (1, 786) = 1754, $p = .00$, $\eta^2 = 0.7$, suggesting the number of errors decreased from trial to trial. Neither one of the other main effects or interactions were significant.

Finally, in 638 eligible participants that completed the typing test, we examined the effects of gender and age over processing speed on two measures: 1) Average typing time-the mean reaction time averaged across typing of three sentences; and 2) Initiation time-calculated as the average time from the presentation of the recall textbox to the first key press of the first letter in a word, averaged across all words typed in all trials for each subject. As can be seen in Fig. 1, age significantly affected both typing time, F (5, 626) = 10, $p = .00$, $\eta^2 = 0.07$, and initiation time F (5, 626) = 5.6, $p = .00$, $\eta^2 = 0.04$. The linear trends were significant for both measures suggesting both typing, and initiation time increases gradually with age thus indicating slower processing speed F (1, 626) = 37, $p = .00$, $\eta^2 = 0.06$ and F (1, 626) = 19, $p = .00$, $\eta^2 = 0.03$ for typing and initiation, respectively. For typing time only, there was a significant effect of gender, with females typing faster than males (mean typing times were 16.3 s and 19.8 s for females and males, respectively) F (1, 626) = 17, $p = .00$, $\eta^2 = 0.03$. Regression models were employed to determine whether typing time and gender significantly predicted participants' recall scores. The results indicated that neither typing time nor gender significantly predicted immediate or delayed recall ($R^2 = 0.002$, $p = .45$ and $R^2 = 0.005$ $p = .2$, for immediate and delayed recall, respectively).

4. Discussion

The current study presents the development of a new computerized self-guided, visual version of a free recall verbal memory test. This test was designed to address the caveats that exist in previous tests (i.e. recall instead of recognition, be self-guided, have alternate forms, as well as additional measures of cognitive performance such as speed), while retaining the utility of prior tests. We then compared performance of a large sample of 798 participants in six different age groups, to existing norms, for both recall and composites scores. We also include data for errors and reaction time measures. To the best of our knowledge, this is the first most comprehensive report of such findings, using a novel computerized, self-guided, free recall verbal memory test.

First and most importantly, all eligible participants completed the test. Such low dropout rates suggest the current computerized version of the test maximizes engagement and can thus be used with both younger and older adults. Importantly, computer proficiency, as examined by the mean time spent on computer daily, did not affect performance on the task, suggesting the task can be used with participants with different levels of computer skills.

Overall, analysis of recall scores suggests that the current computerized version yielded similar scores compared to previously published meta-norms for auditory, paper and pencil, verbal memory tests. Across age and gender, participants showed a learning effect (both in accuracy and error rates) over the first five trials, followed by a decrease in performance for the interference list, and then increase in performance

Table 6a
Males: Means and standard deviations for the number of errors.

Trial	20-29		30-39		40-49		50-59		60-69		70-79	
	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev
T1	8.8	3.03	8.39	2.74	8.56	2.55	7.92	2.67	7.81	2.77	9	1.96
T2	5.73	3.27	5.39	2.74	5.34	2.44	4.74	3.01	5	3.12	5.62	2.5
T3	4.12	3.33	3.58	2.71	3.85	2.35	3.49	2.72	3.73	2.6	3.69	2.78
T4	3.43	2.74	3.12	2.47	2.94	2.37	2.91	2.42	2.98	2.45	4	2.58
T5	3.21	3.28	2.4	2.59	1.94	1.82	2.37	2.29	2.44	2.44	3.15	2.44
List B	8.21	3.16	7.98	2.99	8	2.75	7.66	2.95	7.69	3.39	8.15	2.54
T6	5.63	4.49	5.57	4.77	3.71	3.36	4.48	4	4.96	4.72	5.15	3.65
T7	5.85	4.33	4.35	3.58	3.92	3.1	3.77	3.24	3.58	3.01	5.46	4.14

Table 6b
Females: Means and standard deviations for the number of errors.

Trial	20-29		30-39		40-49		50-59		60-69		70-79	
	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev	Mean	stdev
T1	8.06	2.74	8	3.01	7.85	3.07	7.65	2.72	7.59	2.65	9.09	2.47
T2	5.17	2.81	5.04	2.81	5.07	2.91	4.88	2.38	5.27	2.62	5.64	3.61
T3	3.85	2.89	4.07	3.17	3.14	2.24	3.33	2.52	3.64	2.63	4.82	2.96
T4	3.24	3.03	3.23	2.78	2.9	2.35	2.63	2.24	2.77	2.21	5.09	3.18
T5	2.5	2.86	2.49	2.76	2.15	2.25	2	1.93	2.71	2.68	4.09	2.88
List B	8.11	2.94	7.79	3.28	7.86	2.91	7.88	3.42	8.6	2.99	9.45	1.92
T6	3.74	3.52	4.98	4.13	4.47	4.42	4.89	4.47	4.45	3.68	5.73	3.58
T7	3.63	3.26	4.39	3.72	4.18	3.71	3.75	3.1	3.83	3.09	5.55	3.96

for the immediate and delayed recall scores. Importantly, this was not affected by list version, suggesting that the four versions used lead to similar results. Given the importance of repeated memory testing over time in many neurological and psychiatric conditions, the value of such alternate versions lies in the ability to asses an individual repeatedly over time. In the current study alternate test forms were examined between subjects, thus, future studies should address this question with the appropriate within-subject design. Factor analysis of our computerized test yielded two or four unique factors depending on the input (whether all raw scores or a sub-set of interest). Two of these factors resemble the ones previously reported by Vakil and Blachstein (1993) and may be interpreted as reflecting the processes of acquisition and retrieval suggesting the current computerized version taps into similar cognitive processes.

Performance on many verbal recall tests is affected by age. However, some measures are more sensitive to age compared to others. Most studies suggest that the number of words recalled in learning, immediate and delayed recall trials decline with age (Dunlosky and Salthouse, 1996; Vakil and Blachstein, 1997). In our sample, age affected only three measures: the number of recalled words in the best learning trial (Trial 5), the number of recalled words in the delayed trial (Trial 7), and the ability to accumulate information over time (total learning composite score). For all three measures, the effect sizes were generally small and originated from a difference between the oldest age group (70–79) and the second-to-oldest age group (60–69) and not necessarily from a general overall decrease over age. Thus, the age effects in our current data should be interpreted with caution. Why did we fail to fully replicate previously published effects of age? When compared to previously published meta-norms, our older adults seem to outperform their age matched norms in these three measures (Trial 5, Trial 7 and total learning). Taken together it seems that the older age groups in our sample performed the task similarly to the younger adults in the sample, leading to a lack of an age effect in this study, and a big difference between the norms for the older adults in this study and those from previous studies (Vakil and Blachstein, 1997; Dunlosky and Salthouse, 1996). One possible explanation for this lies in the nature of the online task. Although MTurk interface is user-friendly and simple to handle, participating as a worker on this platform necessitates fluent use of computers and the internet. It has been suggested that cognitive engaging tasks, and computer use in particular, enhances cognitive abilities in elder population and might even act as a protective of aging over cognition (Vaportzis et al., 2017; Wagner et al., 2010). Thus, it is possible that our older adults sample have higher cognitive abilities, and as a result perform like young adults in the memory task. Alternatively, it has been shown that use of computer and web-technologies are more prevalent among people with higher cognitive abilities suggesting a possible selection bias in our sample (Czaja et al., 2006). Specifically, that healthy and well-educated, community-dwelling older adults were more likely to approach and participate in this online study. Finally, the sample size for the oldest age group is significantly smaller compared to other age groups and to previously published data. Thus, it could be that the lack of the age effect is a result of the small sample size for this specific age group. This calls for an additional investigation of the performance of older adults in the task.

In the current study, we fail to find gender differences in task performance, other than in typing speed. Whereas some studies report female outperformance (Gur et al., 2010; Vakil and Blachstein, 1997) others fail to replicate such an effect (Kurylo et al., 2001). In a review of 24 studies of verbal recall performance, Loftus and colleagues report no sex differences in 1/3 of the studies indicating mixed results across investigations (Loftus et al., 1987). More recently, Weber and colleagues (Weber et al., 2014, 2017) studied the effects of living condition and education opportunities over gender differences in cognitive performance. Interestingly, they found that women's outperformance in an episodic memory task was attenuated by birth cohort (less prominent for older adults) and by regional development (less prominent in less developed areas). Another study (Horne, 2007) compared performance of school aged boys and girls between paper-and-pencil and computerized tests. The authors reported sex differences in students' performance were evident only for paper-and-pencil but not for computerized tests. The lack of gender differences were suggested to result either from differences between the students or from the tests themselves (i.e., computerized tests might be more objective and less susceptible to gender bias). Interestingly, we did find an effect of gender on typing speed. Specifically, we found that on average, females typed faster than males. Whereas one previous study reported faster typing speed for males compared to females (Yang and Cho, 2012), another study failed to find gender differences and concluded that typing speed alone was not useable in identifying the gender of the user (Tsimperidis et al., 2018). Note, that the study reporting faster typing speed for males included an in-person administration whereas the one failing to report such differences was administered online. It is possible that the in-person administration led females to slower typing speed due to a “stereotype threat”, namely, the technical nature of the task might have been enough to trigger a negative stereotype in the female participants and lead to slower responses (Steele et al., 2002). These effects might have been attenuated and even reversed when tests are carried online, where there is no experimenter-participant encounter, which might also lead to a gender bias in performance (Chapman et al., 2018). However, since neither processing speed nor gender were predictive of memory abilities, the above differences favoring females cannot affirm or refute gender differences in memory performance. Taken together it seems that the lack of gender differences in our study could result from either one of the factors suggested above including effects of generation; gender equity or living conditions; a lack of gender differences on computerized measures; or selection bias in terms of who responds to a study using self-administered computerized cognitive test. Future work might include additional measures such as socio-economic status, and computer anxiety when assessing cognitive abilities and gender differences.

The use of a computerized platform enabled us to extract additional measures of performance. Specifically, timing or speed measures. We found an effect of age over both timing measures. It has been suggested that the age-related decline in memory performance can be explained

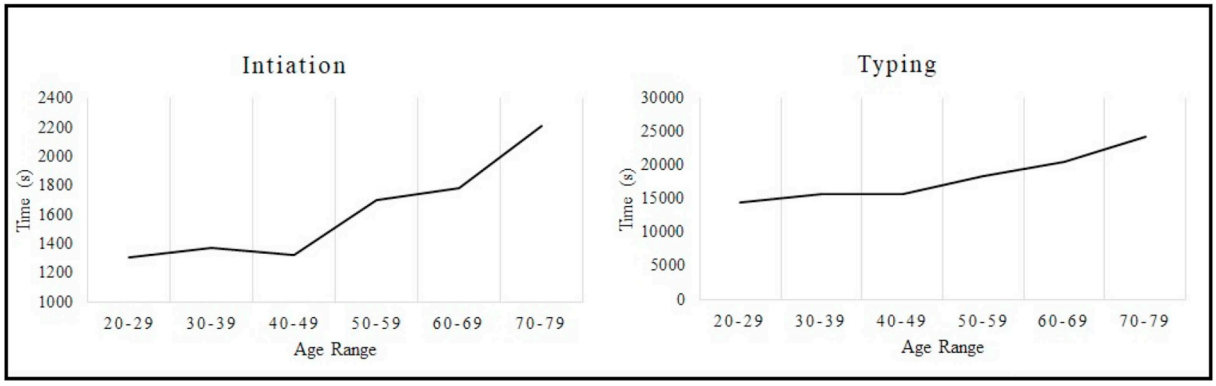

Fig. 1. The effect of age over two timing measures: mean initiation (left) and typing time (right).

by the decline in speed of processing (Dunlosky and Salthouse, 1996). Specifically, it has been shown that age affected the process of data acquisition and that deficits in data acquisition correlated with measures of processing speed. Although the current study did not include a measure of simple processing speed, we examined two measure that reflect processing speed: average typing time and average initiation time. Both measures showed an effect of age whereas older age leads to an increase in average typing and initiation. However, these age-related effects did not interact with any of the learning measures, nor did they predict recall scores. Thus, it seems that our current data does not support the processing speed/memory decline explanation since our older adults did show the expected decline in processing speed however, memory deficits were not as pronounced as previously published results.

Data collection in the study was carried out via Amazon's Mechanical Turk (MTurk) platform. This online workforce is gaining attention from behavioral and computer sciences due to its potential for having a large pool of participants who are constantly available to complete different research studies on demand (Chandler et al., 2014). Previous work suggests that MTurk workers perform comparable to college samples in many of the tasks and surveys including cognitive tasks (Crump et al., 2013); and appear to be truthful and reliable (Mason and Suri, 2012). However, monitoring study participants online imposes limitations and issues that need to be considered. First, MTurk workers tend to be younger and overeducated compared to the general population, and generalizability may be limited (Berinsky et al., 2012). Thus, a selection bias may exist that affected our results, as discussed above for both age and gender. Second, in this platform, workers select which work they would like to perform and where and when it will be completed. Chandler and colleagues (Chandler et al., 2014) describe several potential problems that may arise from this and affect the data. Of importance to the current study are findings on worker attention. In their survey, 18% of workers reported completing the task while being engaged in another activity such as watching TV, listening to music or chatting online (Chandler et al., 2014). In the current study, several steps were taken to try and minimize such issues. For example, the task was made available only for high-rated workers that have a history of adequate task completion. Also, the use of Inquisit platform was chosen, among other reasons, due to its ability to take over the computer, and prevent online multi-tasking. These steps could not eliminate the possibility that some of the participants used a paper-and-pencil while performing the task rather than relying on recall alone. However, this did not seem to be the case in our investigation, since most participants did not outperform their age-matched norms.

The current study has several limitations. First, whereas our current results suggest performance in this test was comparable to existing meta-norms for adults between ages 20–59, performance was not compared between the two versions (self-administered computerized and investigator-administered auditory). Future studies should validate this version using empirical psychometric validation procedures. Second, whereas the test offers a unique measure of free recall abilities, it lacks a measure of memory recognition. Importantly, if an examinee shows low recall abilities, it is impossible to tell whether the deficits originate from a problem in retention or retrieval of verbal information. Adding a measure of recognition is the most commonly used way to differentiate between the two and is also part of the traditional test, either in the form of a word list or in the form of a story. Future versions of this test should include a measure of recognition as well. Third, although the test was examined on a large number of participants, our oldest age group has a relatively small N (24 participants). As previously mentioned, the small number of participants might have led to the lack of age effects in some of the expected measures as well as differences between current data and previously published meta-norms. In order to make sure this test can be applied with older participants future studies should examine the utility of the test in a larger group of elderly participants. Finally, this new computerized version necessitated a few modifications in stimulus presentation and response configuration (Noyes and Garland, 2008). Whereas visual presentation is widely used and accepted (Gur et al., 2010), this is the first test to employ manual typing of recalled responses to assure complete self-administration of the test. As human-computer interface continues to develop, automatic voice recording might substitute manual typing, possibly providing richer information from the test including: vocal tone, speech tempo and speech patterns (Butcher et al., 2000).

To conclude, the VM-REACT offers an easy to administer, self-guided computerized measure of immediate and delayed verbal recall. Recall and composite scores can be extracted as well as measures of processing speed which are highly important. Performance of younger adults (20–59) was highly similar to traditional pencil and paper norms suggesting the test can be utilized within these populations. Performance of the two oldest age groups should be addressed in further studies due to a small number of participants.

Conflicts of interest

AE owns equity in Mindstrong Health and Akili Interactive.

Funding

This work was supported by a grant from Cohen Veterans Bioscience. SN is supported by the Israel Science Foundation (ISF grant 98/17). RO, AE and MLE are funded by the Sierra-Pacific Mental Illness Research, Education and Clinical Center (MIRECC) at the Palo Alto VA.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpsychires.2019.04.023.

References

American Psychological Association, 1986. Committee on Professional Standards, American Psychological Association. Board of Scientific Affairs. Committee on Psychological Tests, Assessment, 1986. Guidelines for Computer-Based Tests and Interpretations. The Association.

Arcia, E., Otto, D.A., 1992. Reliability of selected tests from the neurobehavioral evaluation system. Neurotoxicol. Teratol. 14, 103–110.

Bauer, R.M., Iverson, G.L., Cernich, A.N., Binder, L.M., Ruff, R.M., Naugle, R.I., 2012. Computerized neuropsychological assessment devices: joint position paper of the American Academy of clinical Neuropsychology and the national Academy of Neuropsychology. Clin. Neuropsychol. 26, 177–196.

Berinsky, A.J., Huber, G.A., Lenz, G.S., 2012. Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. Polit. Anal. 20, 351–368.

Butcher, J.N., Perry, J.N., Atlis, M.M., 2000. Validity and utility of computer-based test interpretation. Psychol. Assess. 12, 6–18.

Chandler, J., Mueller, P., Paolacci, G., 2014. Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. Behav. Res. Methods 46, 112–130.

Chapman, C.D., Benedict, C., Schiöth, H.B., 2018. Experimenter gender and replicability in science. Science advances 4, e1701427.

Crump, M.J., McDonnell, J.V., Gureckis, T.M., 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PLoS One 8, e57410.

Czaja, S.J., Charness, N., Fisk, A.D., Hertzog, C., Nair, S.N., Rogers, W.A., Sharit, J., 2006. Factors predicting the use of technology: findings from the center for research and education on aging and technology enhancement (CREATE). Psychol. Aging 21, 333–352.

Dunlosky, J., Salthouse, T.A., 1996. A decomposition of age-related differences in multitrial free recall. Aging Neuropsychol. Cognit. 3, 2–14.

Dwolatzky, T., Whitehead, V., Doniger, G.M., Simon, E.S., Schweiger, A., Jaffe, D., Chertkow, H., 2003. Validity of a novel computerized cognitive battery for mild cognitive impairment. BMC Geriatr. 3, 4.

Elwood, R.W., 2001. MicroCog: assessment of cognitive functioning. Neuropsychol. Rev. 11, 89–100.

Etkin, A., Gyurak, A., O'Hara, R., 2013. A neurobiological approach to the cognitive deficits of psychiatric disorders. Dialogues Clin. Neurosci. 15, 419–429.

Ewers, M., Walsh, C., Trojanowski, J.Q., Shaw, L.M., Petersen, R.C., Jack Jr., C.R., Feldman, H.H., Bokde, A.L., Alexander, G.E., Scheltens, P., Vellas, B., Dubois, B., Weiner, M., Hampel, H., 2012. North American Alzheimer's Disease Neuroimaging Initiative (ADNI). Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. Neurobiol. Aging 33, 1203–1214.

Fan, J., Gu, X., Guise, K.G., Liu, X., Fossella, J., Wang, H., Posner, M.I., 2009. Testing the behavioral interaction and integration of attentional networks. Brain Cogn. 70, 209–220.

Geffen, G.M., Butterworth, P., Geffen, L.B., 1994. Test-retest reliability of a new form of the auditory verbal learning test (AVLT). Arch. Clin. Neuropsychol. 9, 303–316.

Gur, R.C., Jaggi, J.L., Ragland, J.D., Resnick, S.M., Shtasel, D., Muenz, L., Gur, R.E., 1993. Effects of memory processing on regional brain activation: cerebral blood flow in normal subjects. Int. J. Neurosci. 72, 31–44.

Gur, R.C., Ragland, J.D., Moberg, P.J., Turner, T.H., Bilker, W.B., Kohler, C., Siegel, S.J., Gur, R.E., 2001. Computerized neurocognitive scanning:: I. Methodology and validation in healthy people. Neuropsychopharmacology 25, 766–776.

Gur, R.C., Richard, J., Hughett, P., Calkins, M.E., Macy, L., Bilker, W.B., Brensinger, C., Gur, R.E., 2010. A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: standardization and initial construct validation. J. Neurosci. Methods 187, 254–262.

Horne, J., 2007. Gender differences in computerised and conventional educational tests. J. Comput. Assist. Learn. 23, 47–55.

Knight, R.G., McMahon, J., Skeaff, C.M., Green, T.J., 2007. Reliable Change Index scores for persons over the age of 65 tested on alternate forms of the Rey AVLT. Arch. Clin. Neuropsychol. 22, 513–518.

Kurylo, M., Temple, R.O., Elliott, T.R., Crawford, D., 2001. Rey Auditory Verbal Learning Test (AVLT) performance in individuals with recent-onset spinal cord injury. Rehabil. Psychol. 46, 247–261.

Lee, R.S., Hermens, D.F., Porter, M.A., Redoblado-Hodge, M.A., 2012. A meta-analysis of cognitive deficits in first-episode major depressive disorder. J. Affect. Disord. 140, 113–124.

Lezak, M., Howieson, D., Loring, D., 2012. Neuropsychological Assessment, fifth ed. Oxford University Press, Oxford, New York ISBN 10, 9780195395525.

Loftus, E.F., Banaji, M.R., Schooler, J.W., Foster, R.A., 1987. Who remembers what? Gender differences in memory. Mich. Q. Rev. 26, 64–85.

Mason, W., Suri, S., 2012. Conducting behavioral research on Amazon's Mechanical Turk. Behav. Res. Methods 44, 1–23.

Noyes, J.M., Garland, K.J., 2008. Computer-vs. paper-based tasks: are they equivalent? Ergonomics 51, 1352–1375.

Parsey, C.M., Schmitter-Edgecombe, M., 2013. Applications of technology in neuropsychological assessment. Clin. Neuropsychol. 27, 1328–1361.

Parslow, R.A., Jorm, A.F., 2007. Pretrauma and posttrauma neurocognitive functioning and PTSD symptoms in a community sample of young adults. Am. J. Psychiatry 164, 509–515.

Paul, R.H., Lawrence, J., Williams, L.M., Richard, C.C., Cooper, N., Gordon, E., 2005. Preliminary validity of “integneuroTM”: a new computerized battery of neurocognitive tests. Int. J. Neurosci 115, 1549–1567.

Powell, D.H., 1993. MicroCog: Assessment of Cognitive Functioning: Manual. Psychological Corporation.

Schlegel, R.E., Gilliland, K., 2007. Development and quality assurance of computer-based assessment batteries. Arch. Clin. Neuropsychol. 22, S49–S61.

Schmidt, M., 1996. Rey Auditory Verbal Learning Test (RAVLT). Western Psychological Services, Los Angeles, CA.

Schoenfeldt, L.F., 1989. Guidelines for computer-based psychological tests and interpretations. Comput. Hum. Behav. 5, 13–21.

Schweiger, A., Doniger, G., Dwolatzky, T., Jaffe, D., Simon, E., 2003. Reliability of a novel computerized neuropsychological battery for mild cognitive impairment. Acta Neuropsychologica 1, 407–413.

Scott, J.C., Harb, G., Brownlow, J.A., Greene, J., Gur, R.C., Ross, R.J., 2017. Verbal memory functioning moderates psychotherapy treatment response for PTSD-Related nightmares. Behav. Res. Ther. 91, 24–32.

Silverstein, S.M., Berten, S., Olson, P., Paul, R., Williams, L.M., Cooper, N., Gordon, E., 2007. Development and validation of a World-Wide-Web-based neurocognitive assessment battery: WebNeuro. Behav. Res. Methods 39, 940–949.

Steele, C.M., Spencer, S.J., Aronson, J., 2002. Contending with group image: the psychology of stereotype and social identity threat. In: Anonymous Advances in Experimental Social Psychology. Elsevier, pp. 379–440.

Tsimperidis, I., Arampatzis, A., Karakos, A., 2018. Keystroke dynamics features for gender recognition. Digit. Invest. 24, 4–10.

Vakil, E., Blachstein, H., 1997. Rey AVLT: developmental norms for adults and the sensitivity of different memory measures to age. Clin. Neuropsychol. 11, 356–369.

Vakil, E., Blachstein, H., 1993. Rey auditory-verbal learning test: structure analysis. J. Clin. Psychol. 49, 883–890.

Vakil, E., Greenstein, Y., Blachstein, H., 2010. Normative data for composite scores for children and adults derived from the Rey Auditory Verbal Learning Test. Clin. Neuropsychol. 24, 662–677.

Vaportzis, E., Martin, M., Gow, A.J., 2017. A tablet for healthy ageing: the effect of a tablet computer training intervention on cognitive abilities in older adults. Am. J. Geriatr. Psychiatry 25, 841–851.

Wagner, N., Hassanein, K., Head, M., 2010. Computer use by older adults: a multi-disciplinary review. Comput. Hum. Behav. 26, 870–882.

Weber, D., Dekhtyar, S., Herlitz, A., 2017. The Flynn effect in Europe–Effects of sex and region. Intelligence 60, 39–45.

Weber, D., Skirbekk, V., Freund, I., Herlitz, A., 2014. The changing face of cognitive gender differences in Europe. Proc. Natl. Acad. Sci. U.S.A. 111, 11673–11678.

Weiser, M., Reichenberg, A., Rabinowitz, J., Knobler, H., Lubin, G., Yazvitzky, R., Nahon, D., Gur, R., Davidson, M., 2004. Cognitive performance of male adolescents is lower than controls across psychiatric disorders: a population-based study. Acta Psychiatr. Scand. 110, 471–475.

Wiens, A.N., Bryan, J.E., 1994. Assessment strategies. In: Anonymous Advanced Abnormal Psychology. Springer, pp. 19–44.

Wild, J., Gur, R.C., 2008. Verbal memory and treatment response in post-traumatic stress disorder. Br. J. Psychiatry 193, 254–255.

Wild, K., Howieson, D., Webbe, F., Seelye, A., Kaye, J., 2008. Status of computerized cognitive testing in aging: a systematic review. Alzheimer's Dementia 4, 428–437.

Yang, J., Cho, C., 2012. Comparison of posture and muscle control pattern between male and female computer users with musculoskeletal symptoms. Appl. Ergon. 43, 785–791.