UC San Diego

UC San Diego Previously Published Works

Title

Remote Ecological Momentary Testing of Learning and Memory in Adults With Serious Mental Illness.

Permalink

https://escholarship.org/uc/item/83143481

Journal

Schizophrenia Bulletin, 47(3)

Authors

Parrish, Emma
Kamarsu, Snigdha
Harvey, Philip
et al.

Publication Date

2021-04-29

DOI

10.1093/schbul/sbaa172

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Schizophrenia Bulletin vol. 47 no. 3 pp. 740–750, 2021
doi:10.1093/schbul/sbaa172
Advance Access publication 21 November 2020

Remote Ecological Momentary Testing of Learning and Memory in Adults With Serious Mental Illness

Emma M. Parrish[1], Snigdha Kamarsu[2], Philip D. Harvey[3,4], Amy Pinkham[5], Colin A. Depp[*,2,6,7], and Raeanne C. Moore[2,7]

[1]San Diego State University/University of California San Diego Joint Doctoral Program in Clinical Psychology, San Diego, CA; [2]Stein Institute for Research on Aging, Department of Psychiatry, University of California San Diego, 9500 Gilman Drive La Jolla, San Diego, CA 92093-0664; [3]Miller School of Medicine, University of Miami, Miami, FL; [4]Research Service, Bruce W. Carter VA Medical Center, Miami, FL; [5]Department of Psychology, School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX; [6]Department of Psychology, Veterans Affairs San Diego Healthcare System, San Diego, CA; [7]These authors are co-senior authors.

*To whom correspondence should be addressed; tel: 858 822 4251, fax: 858 534 5475, e-mail: cdepp@ucsd.edu

Smartphone-based ecological mobile cognitive tests (EMCTs) can measure cognitive abilities in the real world, complementing traditional neuropsychological assessments. We evaluated the validity of an EMCT of recognition memory designed for use with people with serious mental illness, as well as relevant contextual influences on performance. Participants with schizophrenia (SZ), schizoaffective disorder, and bipolar disorder (BD) completed in-lab assessments of memory (Hopkins Verbal Learning Test, HVLT), other cognitive abilities, functional capacity, and symptoms, followed by 30 days of EMCTs during which they completed our Mobile Variable Difficulty List Memory Test (VLMT) once every other day (3 trials per session). List length on the VLMT altered between 6, 12, and 18 items. On average, participants completed 75.3% of EMCTs. Overall performance on VLMT 12 and 18 items was positively correlated with HVLT ($\rho = 0.52$, $P < .001$). People with BD performed better on the VLMT than people with SZ. Intraindividual variability on the VLMT was more specifically associated with HVLT than nonmemory tests and not associated with symptoms. Performance during experienced distraction, low effort, and out of the home location was reduced yet still correlated with the in-lab HVLT. The VLMT converged with in-lab memory assessment, demonstrating variability within person and by different contexts. Ambulatory cognitive testing on participants' personal mobile devices offers more a cost-effective and "ecologically valid" measurement of real-world cognitive performance.

Key words: schizophrenia/schizoaffective disorder/bipolar disorder/remote cognitive testing/memory/ecological momentary assessment

Introduction

Ecological momentary cognitive tests (EMCTs) are emerging tools for testing cognition in the real world while also providing the opportunity to evaluate intraindividual cognitive variability when intensively repeated.[1] EMCTs do not replace traditional neuropsychological testing but rather complement traditional testing.[1] Furthermore, EMCTs incorporate patient-reported outcomes (ecological momentary assessment [EMA] survey responses) with cognitive tasks and can be used with traditional assessment methods. These assessments can also help researchers and clinicians understand how other time-varying and contextual factors covary with short-term changes in cognition[2] and allows for the examination of metadata associated with the task (eg, completion time).[3] There is an increased need for validated EMCTs for remote assessment, which is further highlighted by the COVID-19 pandemic requiring physical distancing restrictions.[4] However, few studies examine the validity of these tasks or dynamic influences on task performance.

Despite the potential of utilizing EMCTs to evaluate real-time cognitive functioning, there is little feasibility or validation data to support their use, particularly in serious mental illnesses (SMI).[5] In addition to evaluating the influences on adherence, validation includes comparison against gold-standard measures of the same construct, distal constructs, and variability across patient groups that are expected to differ. Given that testing occurs in the naturalistic environment, understanding the influences of effort and social context on validity is also necessary.

This report focuses on the validation of an EMCT of learning and recognition memory called the Mobile Variable Difficulty List Memory Test (VLMT) administered via smartphone-based EMA over 30 days. Impairments in episodic memory functions of learning, recall, and recognition memory are common in individuals with SMI[6–9] and can significantly impact daily

© The Author(s) 2020. Published by Oxford University Press on behalf of the Maryland Psychiatric Research Center.
All rights reserved. For permissions, please email: journals.permissions@oup.com

functioning.[10,11] Due to the difficulty of scoring free speech recall on a mobile device, the VLMT was developed as a test of recognition memory in naturalistic environments. People with schizophrenia (SZ) spectrum illness or bipolar disorder (BD) were presented with a word list of varying lengths (6, 12, or 18 words) over 30 days. The aims of this study were to: (1) examine adherence to the VLMT in a sample of participants with SMI (SZ, schizoaffective disorder, and BD); (2) assess the intraindividual variability of performance and magnitude of within-session learning effects and between-session practice effects and compare these results across word list lengths; (3) examine VLMT performance by diagnosis evaluating whether effect sizes of comparisons of participants with SZ spectrum disorders and BD were consistent with in-lab measures; (4) estimate the correlations of aggregate means scores and intraindividual variability on the VLMT with demographic variables and laboratory-based "gold-standard" measures of global and domain-specific cognitive functioning and to understand the impact of context (eg, location and self-reported effort) on VLMT performance; and (5) examine the relationships of the VLMT with symptom severity and performance-based assessments of functional capacity and compare the strength of the associations between the mobile test and the laboratory-based memory testing and functional capacity.

Methods

Participants

Participants for this study are part of an ongoing project investigating participants' awareness of their own cognitive abilities (ie, introspective accuracy). The target sample size for this larger study is 450 and, for this analysis, we used a subsample of 168 participants gathered between November 2018 and March 2020. Participants were outpatients recruited from the University of California San Diego (UCSD), The University of Texas at Dallas, and the University of Miami, through online advertisements, flyers, and outpatient clinics. See supplemental material for full inclusion and exclusion criteria.

Diagnoses were determined using the Mini International Neuropsychiatric Interview[12] and the psychosis module of the Structured Clinical Interview for *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition,[13] complemented by medical record reviews and consensus meetings with site investigators. Raters were trained on administration and scoring through videotape and practice interviews and were determined to be reliable after reaching acceptable interrater reliability (intraclass correlation coefficient > 0.80).

The initial sample included 172 participants; however, 3 individuals were excluded as they did not contribute any VLMT data. One additional participant was excluded after data collection because investigators learned that this participant was not fluent in English. All other participants who contributed 1 or more VLMTs were included, leaving a total of 168 participants. Seven participants had missing data for some of the VLMT word list lengths or time points (eg, did not complete VLMT list length 18) noted when applicable. The study was approved by each site's institutional review board and all participants provided written informed consent.

Smartphone Procedure

Eligible participants were given a 15-minute training session on operating a Samsung Galaxy S8 lab-supplied smartphone and in completing the EMCT sessions. Participants were also given an operating manual, which detailed how to use the study smartphone and how to respond to EMCT sessions. During this session, participants completed an example survey and the VLMT task with the assistance of a study examiner. Participants selected time slots for the survey notifications, with a minimum 2-hour increment between each survey. Most participants chose the surveys to come in the morning, afternoon, and night. Surveys were scheduled to appear randomly within the requested time slot. This platform was delivered through a collaboration between our research team and Play Power Labs, LLC. The resulting platform is now a subsidiary of Play Power called NeuroUX: https://www.getneuroux.com/. Once the survey was received, the link remained active for 1 hour. Survey data were not stored on the device but was sent to an encrypted, Health Insurance Portability and Accountability Act-compliant, cloud storage location in Amazon Web Services.[14] Participants did not need to have a WiFi connection to send data to the cloud as their study-provided smartphone had a data plan. This system allowed researchers to access participant data in real time and monitor their progress daily.

30-Day Mobile Data Collection

Beginning the next day, participants were sent text notifications to complete the EMA surveys assessing mood, daily functioning, and symptoms 3 times daily for 30 days. Participants were asked about their environmental factors during each survey with questions such as "What are you doing?," "Where are you right now?," and "Who is with you at this moment?" Once every other day, one randomly selected survey also included the VLMT. If participants missed more than 3 surveys in a row, study staff contacted them to address any difficulties. Staff also performed routine check-in calls once every 2 weeks to encourage adherence to the survey protocols. Participants received $0.88 for each EMA survey completed (75 possible) and $2.25 for each completed EMA

survey accompanied by the VLMT task (15 possible), resulting in a maximum compensation amount of $100.

Mobile Variable Difficulty List Memory Test

The words used in the VLMT were originally derived for the Mobile Verbal Learning Test (mVLT), which consists of 12 semantically unrelated words.[15] For the mVLT, author R.C.M. created 15 different word lists using the SUBTLEX(US) database (http://www.lexique.org/?page_id=241), which contains word frequencies for 50 million words based on subtitles of English (US) movies and TV series.[16] Profanities and proper nouns were eliminated, and 6, 12, or 18 words were selected from this database based on a set of predefined parameters. These included word length min/max, part of speech, and minimum threshold for word frequency. The parameters excluded plural forms (ie, nouns ending in "s" where it is not preceded by "I" or "u" or nouns ending in "ae") and verb tenses (ie, any verbs ending in "ed" or "ing"). Five lists from the original mVLT were used. The foils were created using the same procedure as described above, and the foils for each list were matched to the target words on the frequency of use in the English language.

Task Description

See figure 1 for screenshots of the VLMT. During each VLMT administration, participants were presented with 3 trials where the list was shown for 30 seconds each. List lengths varied between 6, 12, or 18 items and were distributed across the 15 days so that each list length was used 5 times. The 6-item list was included intentionally to be easy to examine attention and effort. Immediately following each exposure to the list, participants were shown target and recognition foil words one-by-one and asked

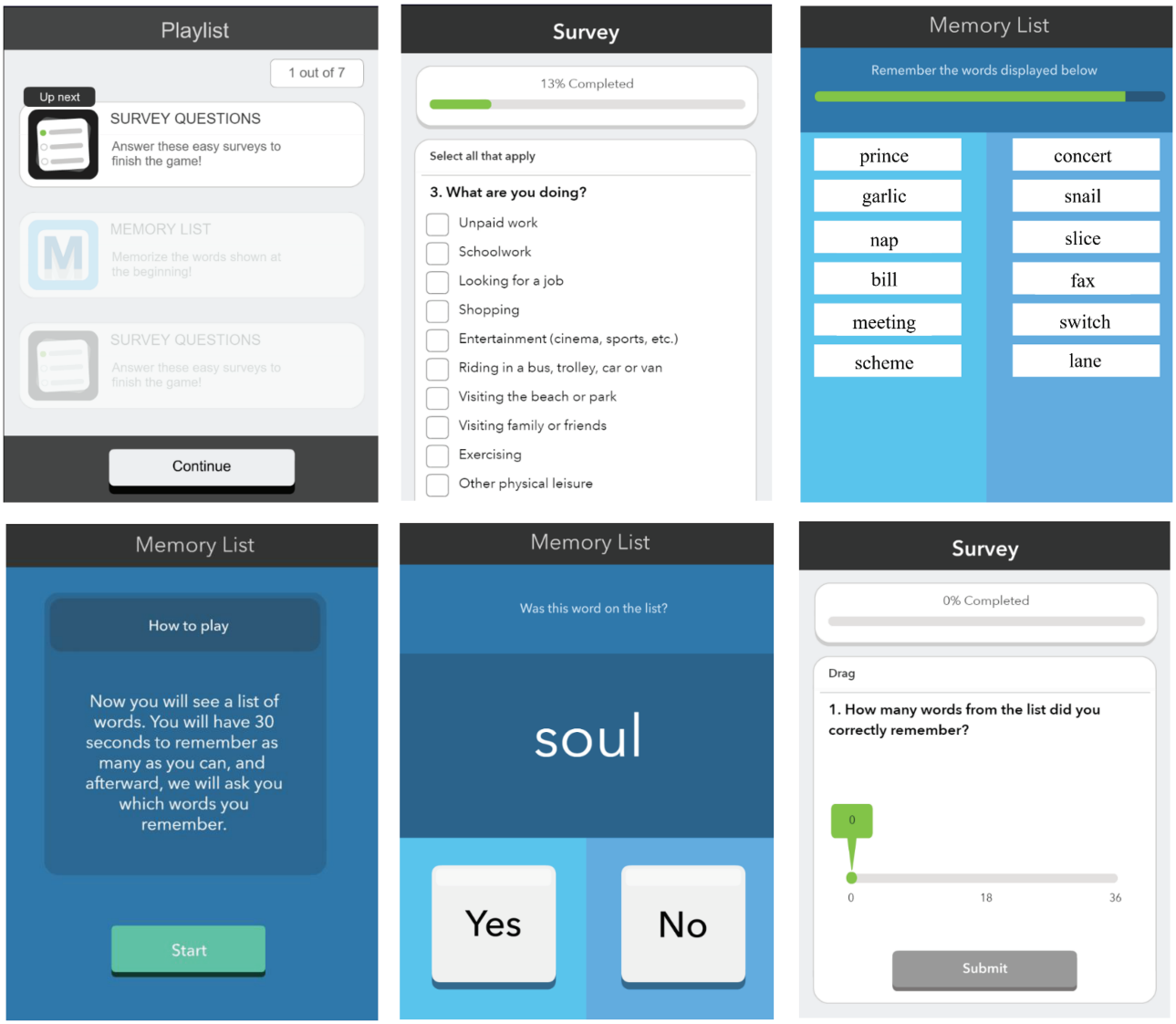

Figure 1. Screenshots of test. For test security purposes, these are sample words and not an actual word list from the VLMT.

to indicate whether or not the word appeared on the list. This forced-choice recognition task matched the number of target words to the recognition foils for all list lengths. Dependent variables were the percentage of correct identifications of target words per trial, percentage of correct rejections of foil words, and an overall score combining the percentage of correct identifications of targets and correct rejections of foils. In embedded validity tests, after each VLMT administration, participants self-reported if they were distracted or interrupted, as well as their self-reported level of effort on a 1–7 Likert scale. Completion time was calculated as the amount of time it took the participant to complete each trial. A standardized *Z* score was calculated for each of the completion time variables.

Follow-Up In-Lab Measures

Within a week of their last survey, participants returned to the lab and completed a series of assessments. Cognition was assessed with a selection of tests from the Measurement and Treatment Research to Improve Cognition in Schizophrenia Consensus Cognitive Battery (MCCB).[17] Because the VLMT produced raw scores, raw scores of each of the MCCB tasks were used as comparators. Verbal learning and recall were assessed using the Hopkins Verbal Learning Test (HVLT),[18] an MCCB subtest. The HVLT includes three learning trials of 12 semantically related words, and number correct per trial is summed to create a total recall score of the 3 initial trials. The MCCB did not include either delayed recall or delayed recognition, so the HVLT variable of interest was total recall. Processing speed was assessed using the Brief Assessment of Cognition in Schizophrenia Symbol Coding subtest [17,19] and the Trail Making Test A (TMT).[20] Working memory was assessed using the Letter Number Span (LNS).[21] Verbal fluency was assessed using the Animal Fluency test.[17]

Functional capacity was assessed using the UCSD Performance-Based Skills Assessment-Brief (UPSA-B)[22] where participants demonstrated financial and communication skills through lifelike scenarios. The total score combined the correct responses on the financial and communication sections of the assessment.

Participants were assessed for severity of positive, negative, general, and disorganized symptoms through the Positive and Negative Syndrome Scale (PANSS).[23] Depressive symptoms were assessed using the Montgomery–Asberg Depression Rating Scale (MADRS),[24] and manic symptoms were assessed using the Young Mania Rating Scale (YMRS).[25]

Statistical Analyses

Analyses focused on the 12 and 18 item lists, as the intentionally easy 6-item list was used as a measure of effort (hereafter, VLMT12/18 when results are the same for 2 list lengths). Adherence was calculated as the percentage of surveys completed by the total number possible (15). A generalized linear model was performed across the 3 trials of the VLMT. To assess between-session practice effects, a linear mixed model was used, with modeling day as a continuous factor. Greenhouse–Geisser correction was used to account for the violation of sphericity. We next examined between-group differences on the VLMT between people with BD and SZ by using an independent samples *t*-test. To analyze the relationship between the VLMT12/18 and other measures, correlations were performed between the VLMT12/18 and variables of interest. Then, an *R* to *Z* transformation was performed to compare correlations of the VLMT12/18 and variables of interest to correlations of the HVLT and other variables.[26] We performed a multiple regression analysis predicting VLMT12/18, including all cognitive variables (HVLT, Wide Range Achievement Test [WRAT], Trails, Symbol Coding, LNS, and Animal Fluency) to investigate the unique relationship of the HVLT to the VLMT. Mean square of successive difference (MSSD), or the sum of consecutive observation differences squared divided by the number of observations minus 1, was calculated as a measure of within-person variability, and correlations of variability with other measures were examined.

We used linear mixed models to evaluate the impact of response time and context and self-reported effort on VLMT performance. Finally, we removed participants who performed poorly on the 6-item list (aggregate score 2 SD below the mean) and performed analyses as above. Analyses were performed using IBM SPSS Statistics v.26, with the exception of calculating MSSD (calculated using R v.3.6.0).

Results

Demographic and clinical characteristics are presented in table 1. Fifty-eight percent of the participants had a diagnosis of SZ or schizoaffective disorder, while 42% of the participants had a diagnosis of BD. Regarding diagnostic differences on the VLMT, people with BD performed better on the VLMT percentage overall correct (M = 81.8, SD = 9.0) than people with SZ (M = 78.7, SD = 10.2, $t(159) = -1.86$, $P = .048$, Cohen's $d = 0.32$, 87.3% overlap). The magnitude of this differences was less than of the HVLT, in which people with BD performed better (M = 24.0, SD = 6.4) than people with SZ spectrum disorders (M = 20.8, SD = 5.5; $t(159) = -3.43$, $P = .001$, Cohen's $d = 0.52$, 79.5% overlap).

Adherence

Average adherence to the VLMT for the whole sample was 75.3% (SD = 23.1%) or an average of 11.3 (SD = 3.5) responses out of 15 testing opportunities; adherence did

Table 1. Demographic characteristics

	Schizophrenia spectrum (N = 98)	Bipolar disorder (N = 70)	
	Mean (SD) or (%)		T or X^2; P-value
Age—mean (SD); range	42.0 (10.2); 18–65	39.1 (11.9); 18–65	1.68 (P = .095)
Gender—% female	47 (48%)	49 (70%)	8.10 (P = .004)
Education in years—mean (SD)	12.5 (2.4)	14.3 (2.6)	−4.51 (P < .001)**
Race			
% Caucasian	32.7%	58.6%	17.54 (P = .007)
% African American	55.1%	25.7%	
% Other	12.2%	15.7%	
Ethnicity			
% Hispanic	23.5%	30%	0.901 (P = .343)
Vocational activity			
Unemployed	68.4%	57.2%	20.45 (P = .009)**
Part-time student	3.0%	0%	
Full-time student	2.1%	5.7%	
Part-time employment	21.4%	17.1%	
Full-time employment	5.1%	20%	
Clinical and cognitive variables			
WRAT-3 standard score—mean (SD)	95.5 (11.7)	101.9 (11.7)	−3.50 (P = .001)**
UPSA-B total score—mean (SD)	68.3 (15.0)	75.7 (12.5)	−3.31 (P = .001)**
PANSS positive symptoms—mean (SD); range	17.2 (4.6); 7–49	13.5 (4.8); 7–49	5.02 (P < .001)**
PANSS negative symptoms—mean (SD); range	13.6 (4.2); 7–49	10.7 (2.6); 7–49	5.50 (P < .001)**
MADRS total—mean (SD); range	9.4 (10.2); 0–60	12.8 (11.5); 0–60	−1.98 (P = .050)
YMRS total—mean (SD); range	0.9 (3.2); 0–22	2.6 (4.8); 0–19	−2.57 (P = .012)*
HVLT trial 1 score—mean (SD); range	5.4 (1.8); 0–12	6.3 (2.3); 0–12	−2.83 (P = .006)**
HVLT trial 2 score—mean (SD); Range	7.2 (2.1); 0–12	8.6 (2.3); 0–12	−3.92 (P < .001)**
HVLT trial 3 score—mean (SD); Range	8.2 (2.1); 0–12	9.1 (2.4); 0–12	−2.48 (P = .014)*
HVLT total score—mean (SD); Range	20.8 (5.5); 0–36	24.0 (6.4); 0–36	−3.43 (P = .001)**
VLMT percentage of adherence—mean (SD)	74.7 (23.6)	76.1 (22.5)	−0.39 (P = .699)
VLMT percentage of overall correct—mean (SD)	80.5 (10.5)[a]	83.3 (9.2)[b]	−1.77 (P = .079)

Note. UPSA-B, University of California San Diego Performance-Based Skills Assessment-Brief; PANSS, Positive and Negative Symptoms Scale; HVLT, Hopkins Verbal Learning Test; VLMT, Mobile Variable Difficulty List Memory Test; MADRS, Montgomery Asberg Depression Scale; YMRS, Young Mania Rating Scale; WRAT-3, Wide Range Achievement Test 3.
[a]N = 92.
[b]N = 68.
*P < .05; **P < .01.

not differ by diagnostic group, $t(166) = -0.39$, $P = .70$. Two participants completed only 1 session, whereas 26 participants completed all 15 sessions. Adherence was not correlated with HVLT performance, VLMT overall performance, years of education, premorbid IQ, PANSS, MADRS, YMRS, or any other MCCB subtest (Ps > .05). Adherence was correlated with age, with older participants having better adherence ($\rho = 0.174$, $P = .02$), and correlated with the UPSA, where individuals with better functional capacity had better VLMT adherence ($\rho = 0.164$, $P = .04$). See supplementary table 1 for all correlations.

Descriptive Information, Within-Session Learning, and Practice Effects

Percentage correct (target and foil) for all time points and word list lengths is presented in table 2. Generally, participant performance on the VLMT was high. The overall percentage correct for participants for trials 1, 2, and 3 aggregated across list lengths 6, 12, and 18 was 91.9% (SD = 8.7), 79.3% (SD = 12.3), and 72.8% (SD = 12.2), respectively. Averaging all 3 trials and all 3 word list lengths, overall percentage correct was 81.6% (SD = 10.0). There were no group diagnostic differences in variability as measured by MSSD for any list length (Ps > .219). See

supplementary figure 1 exemplar plots. When combining Z scores of completion time for VLMT12/18, there was a small but significant positive effect for overall performance, $F(1, 1230) = 21.28$, estimate = 0.02, SE = 0.00, $t = 4.61$.

Within-session learning effects were present for the percentage of target words correct, correct rejections of foils, and overall correct for all lengths of the word list (see table 2 for full data). While participants improved over the 3 trials for target and overall performance, the percentage of correctly identified foil words decreased across trials 1, 2, and 3 for all 3 word list lengths (see table 2). Over the 30 days of exposure, there was no significant effect of day on overall percentage correct of the 12-item list (estimate = −0.04, SE = 0.07, $t = -0.60$, $P = .55$) or the 18-item list (estimate = −0.12, SE = 0.07, $t = -1.68$, $P = .09$). The rate of change between the 2 tests was different from trial 1 to trial 3 such that participants improved a mean of 23.4% on the HVLT and improved a mean of 9.4% on the VLMT 12-item list, $t(156) = 9.78$, $P < .001$.

Correlations With Lab-Based Measures of Cognition, Symptoms, and Functional Capacity

All versions of the VLMT were positively correlated with the HVLT: the 6 item ($\rho = 0.44$, $P < .001$), 12 item ($\rho = 0.49$, $P < .001$), and 18 item ($\rho = 0.42$, $P < .001$). The overall percentage correct across the VLMT12/18 was highly correlated with the HVLT overall score ($\rho = 0.52$, $P < .001$; see supplementary figure 2), as was the overall percentage correct for trial 1 of the VLMT12/18 list (see table 3). The overall percentage correct for the 12-item list of the VLMT task was positively correlated with years of education but did not correlate with age. The VLMT overall percentage correct for the 12-item list also positively correlated with most lab-based MCCB subtests and negatively correlated with the TMT (higher score indicates worse performance). Correlations with the VLMT and other MCCB tasks were stronger for the percentage of foils correctly rejected than the percentage of target words correctly identified. Comparing strengths of association, the VLMT overall percentage correct for VLMT12/18 and all other measures were statistically similar to correlations of the HVLT to all other measures, with the exception of age. The VLMT was less strongly correlated with Animal Fluency performance than the HVLT, $Z = 2.64$, $P = .004$. Similarly, the VLMT was less strongly correlated with age than the HVLT, $Z = -2.15$, $P = .02$.

A multiple regression model predicting VLMT12/18, including all cognitive variables, was statistically significant, $F(6, 146) = 12.01$, $P < .001$, and the full model explained 33% of the variance in VLMT. We found that HVLT was the only independent significant predictor ($B = 0.678$, SE = 0.172, $t = 3.95$, $P < .001$).

Variability (MSSD) of the VLMT12/18 negatively correlated the HVLT overall score ($\rho = -0.173$, $P = .029$) such that people with worse HVLT overall performance had more intravariability in their VLMT scores. Greater variability of the VLMT12/18 was also negatively correlated with higher premorbid IQ (WRAT-3; $\rho = -0.172$, $P = .027$), but variability was not correlated with age, years of education, other cognitive tasks, symptoms, or UPSA-B (all Ps > .368).

Context of Test Performance

On average during the completion of the VLMT12/18, participants reported being away during 31.5% and being with others during 51.9%. Being away from home (estimate = −2.81%, SE = 0.88, $t = 3.19$, $P = .001$, Cohen's $d = 0.09$, 96.4% overlap) had a significant but small negative effect on performance, but being with others did not have a significant effect on performance (estimate = 0.55, SE = 0.83, $t = -0.665$, $P = .506$, Cohen's $d = 0.02$, 99.2% overlap). Convergent validity between performance on the VLMT12/18 and the HVLT did not significantly differ using an R to Z transformation when with others or away.

Embedded Validity Tests

On average during the VLMT12/18, participants reported distractions during 37.0% of VLMT assessments and interruptions during 24.7%. Presence of self-reported distraction (estimate = −3.77%, SE = 0.84, $t = 4.471$, $P < .001$, Cohen's $d = 0.11$, 95.6% overlap) and interruptions (estimate = −6.27%, SE = 0.91, $t = 6.87$, $P < .001$, Cohen's $d = 0.18$, 92.8% overlap) had a significant but small negative effect on overall performance on the 12- and 18-item word lists. Overall, performance on the VLMT12/18 lists did not significantly differ in the correlation with the HVLT when using an R to Z transformation when distracted or interrupted. Participants reported a mean self-reported effort of 5.32 out of 7 (SD = 1.15) for the VLMT12/18 lists. Self-reported effort also had a small but significant effect on overall performance on the VLMT12/18 lists (estimate = 1.07, SE = 0.25, $t = 4.30$, $P < .001$, Cohen's $d = 0.12$, 95.2% overlap). Finally, removing participants who had low effort on the 6-item list did not substantially impact the convergent validity of performance on the VLMT12/18 with the HVLT (see supplementary material).

Discussion

This study provides evidence for the validity of an EMCT learning and recognition task among persons with SMI. Adherence was 75% and was uncorrelated with most predictors, indicating that the task can be used by a wide constituency. Performance parametrically declined with longer word lists, as expected. The VLMT overall percentage correct score had an overall high correlation with the HVLT and a gold-standard

Table 2. Descriptive statistics of the Mobile Variable Difficulty List Memory Test

	Trial 1 *M*% (SD)	Trial 2 *M*% (SD)	Trial 3 *M*% (SD)	Total *M*% (SD)	Total d' *M*% (SD)	MSSD	Score change across trials?
6-item percentage correct (*N* = 164)							
Correct (target)	87.42 (13.63)	91.10 (10.51)	92.59 (10.92)[a]	90.39 (10.63)[a]	83.6 (17.9)	181.78	$F(1.7, 272.2) = 29.3$, $P < .001$**
Correct (foil)	94.68 (11.29)	92.86 (11.42)	91.16 (15.06)	92.90 (11.51)			$F(1.7, 281.9) = 11.7$, $P < .001$**
Overall	91.05 (9.94)	92.10 (8.78)	92.45 (9.46)[a]	91.92 (8.74)[a]			$F(1.6, 263.5) = 4.3$, $P = .021$*
12-item percentage correct (*N* = 164)							
Correct (target)	74.71 (14.09)	82.63 (12.96)	84.35 (13.26)	80.56 (12.21)	74.0 (17.8)	186.13	$F(1.7, 273.8) = 91.4$, $P < .001$**
Correct (foil)	80.57 (20.28)	76.79 (20.67)	75.75 (21.42)	77.70 (19.72)			$F(1.7, 275.1) = 16.1$, $P < .001$**
Overall	77.64 (13.20)	79.92 (12.70)	80.38 (13.01)	79.32 (12.29)			$F(1.7, 276.9) = 13.8$, $P < .001$**
18-item percent correct (*N* = 166)							
Correct (target)	69.38 (14.62)	78.89 (14.09)	80.81 (14.98)	76.36 (13.02)	69.3 (18.2)	130.96	$F(1.4, 239.0) = 97.2$, $P < .001$**
Correct (foil)	71.82 (21.80)	67.79 (23.94)	67.15 (25.39)	68.92 (22.81)			$F(1.7, 281.7) = 16.2$, $P < .001$**
Overall	70.60 (11.27)	73.51 (13.24)	74.15 (14.25)	72.75 (12.24)			$F(1.6, 269.1) = 21.3$, $P < .001$**
Percentage correct overall (*N* = 161)	80.04 (9.82)	82.15 (10.27)	82.61 (10.93)[b]	81.64 (10.04)[b]	63.0 (20.1)	326.96	$F(1.5, 241.1) = 30.4$, $P < .001$**

Note. MSSD, mean square of successive difference.
[a]*N* = 163.
[b]*N* = 160.
*$P < .05$; **$P < .01$.

Table 3. Spearman correlations with Mobile Variable Difficulty List Memory Test of word lengths 12 and 18 trial 1 percentage of correct, related measure, and demographic variables ($N = 161$[a])

	HVLT overall score; ρ (*P*)	Target, 12 items; ρ (*P*)	Distractors, 12 items; ρ (*P*)	Total, 12 items; ρ (*P*)	Target, 18 items; ρ (*P*)	Distractors, 18 items; ρ (*P*)	Total, 18 items; ρ (*P*)	Target, 12 and 18 items; ρ (*P*)	Distractors, 12 and 18 items; ρ (*P*)	Total, 12 and 18 items; ρ (*P*)
HVLT total score (total recalled)	—	0.255 (*P* = .001)*	0.516 (*P* < .001)**	0.493 (*P* < .001)**	0.143 (*P* = .143)	0.412 (*P* < .001)**	0.441 (*P* < .001)**	0.212 (*P* < .008)*	0.524 (*P* < .001)**	0.521 (*P* < .001)**
Age	−0.180 (*P* = .022)*	0.029 (*P* = .712)	−0.080 (*P* = .308)	−0.016 (*P* = .843)	0.049 (*P* = .534)	−0.058 (*P* = .458)	0.019 (*P* = .813)	0.045 (*P* = .567)	−0.086 (*P* = .275)	−0.010 (*P* = .898)
Education	0.467 (*P* < .001)**	0.132 (*P* = .092)	0.372 (*P* < .001)**	0.318 (*P* < .001)**	0.043 (*P* = .585)	0.348 (*P* < .001)**	0.351 (*P* < .001)**	0.090 (*P* = .254)	0.399 (*P* < .001)**	0.360 (*P* < .001)**
WRAT-3	0.511 (*P* < .001)**	0.223 (*P* = .004)*	0.431 (*P* < .001)**	0.434 (*P* < .001)**	−0.027 (*P* = .730)	0.374 (*P* < .001)**	0.342 (*P* < .001)**	0.110 (*P* = .165)	0.4364 (*P* < .001)**	0.412 (*P* < .001)**
Trails A	−0.351 (*P* < .001)**	−0.061 (*P* = .448)	−0.212 (*P* = .008)*	−0.218 (*P* = .006)*	−0.125 (*P* = .116)	−0.137 (*P* = .086)	−0.190 (*P* = .016)*	−0.102 (*P*=.208)	−0.189 (*P* = .019)*	−0.2232 (*P* = .005)**
Digit Symbol Coding	0.413 (*P* < .001)**	0.111 (*P* = .167)	0.361 (*P* < .001)**	0.329 (*P* < .001)**	0.037 (*P* = .643)	0.321 (*P* < .001)**	0.298 (*P* < .001)**	0.078 (*P* = .332)	0.385 (*P* < .001)**	0.349 (*P* < .001)**
Letter Number Span	0.476 (*P* < .001)**	0.169 (*P* = .034)*	0.417 (*P* < .001)**	0.394 (*P* < .001)**	0.028 (*P* = .730)	0.350 (*P* < .001)**	0.327 (*P* < .001)**	0.116 (*P* = .149)	0.418 (*P* < .001)**	0.395 (*P* < .001)**
Animal Fluency	0.470 (*P* < .001)**	0.097 (*P* = .226)	0.248 (*P* = .002)*	0.251 (*P* = .002)*	0.046 (*P* = .570)	0.001 (*P* = .001)*	0.246 (*P* = .002)*	0.060 (*P* = .462)	0.292 (*P* < .001)**	0.282 (*P* < .001)**
UPSA-B	0.500 (*P* < .001)**	0.158 (*P* = .048)*	0.440 (*P* < .001)**	0.411 (*P* < .001)**	−0.053 (*P* = .507)	0.400 (*P* < .001)**	0.359 (*P* < .001)**	0.058 (*P* = .476)	0.459 (*P* < .001)**	0.412 (*P* < .001)**
PANSS positive symptoms	−0.229 (*P* = .003)**	0.027 (*P* = .736)	−0.150 (*P* = .055)	−0.136 (*P* = .082)	−0.027 (*P* = .732)	−0.141 (*P* = .070)	−0.192 (*P* = .013)*	0.007 (*P* = .930)	−0.159 (*P* = .043)*	−0.164 (*P* = .0384)*
PANSS negative symptoms	−0.268 (*P* = .001)**	−0.092 (*P* = .242)	−0.167 (*P* = .032)*	−0.200 (*P* = .010)*	−0.072 (*P* = .359)	−0.103 (*P* = .188)	−0.108 (*P* = .166)	−0.073 (*P* = .358)	−0.154 (*P* = .051)	−0.173 (*P* = .027)*
MADRS	0.062 (*P* = .435)	0.086 (*P* = .272)	−0.132 (*P* = .091)	0.131 (*P* = .094)	−0.023 (*P* = .769)	0.150 (*P* = .054)	0.135 (*P* = .083)	0.048 (*P* = .545)	0.154 (*P* = .051)	0.133 (*P* = .090)
YMRS	0.036 (*P* = .653)	0.139 (*P* = .076)	0.011 (*P* = .886)	0.067 (*P* = .396)	0.186 (*P* = .016)*	−0.004 (*P* = .956)	0.066 (*P* = .397)	0.200 (*P* = .011)*	0.005 (*P* = .947)	0.094 (*P* = .235)

Note. HVLT, Hopkins Verbal Learning Test; VLMT, Mobile Variable Difficulty List Memory Test; WRAT-3, Wide Range Achievement Test 3; UPSA-B, University of California San Diego Performance-Based Skills Assessment-Brief; PANSS, Positive and Negative Syndrome Scale; MADRS, Montgomery–Asberg Depression Rating Scale; YMRS, Young Mania Rating Scale.

[a]*N*s range from 153 to 161 depending on available data. $N = 154$ was used for all *R* to *Z* transformations.

*$P < .05$; **$P < .01$.

assessment of verbal memory, and the correlations of the VLMT scores with other measures were not significantly different from that of the HVLT and those same variables, with the exception of Animal Fluency and age. A multiple regression analysis also suggested that the HVLT has a unique association with the VLMT overall score, despite that both the VLMT and the HVLT are correlated with other cognitive measures. We found a ceiling effect on trial length 6, but this trial length was included intentionally to induce a ceiling effect such that detection of response bias in self-assessment of memory performance could be another aim of the study. Unlike mean performance, intraindividual variability in VLMT was related to HVLT but not to other cognitive tests or psychotic symptoms, which may make variability a useful target in future memory interventions. Furthermore, the effect of poor effort, distraction, and interruption on performance was significant but did not greatly impact the convergent validity of performance. Overall, these findings support that the VLMT could help to complement or extend traditional neuropsychological testing.

Measuring serial learning and memory on a smartphone is associated with a number of challenges, including that many current paradigms require manually scoring verbal responses[15,27]; thus, recognition memory was selected as the target of the VLMT as this approach allowed for automated scoring. A strength of the smartphone approach is that it allows for the examination of metadata, which can measure other cognitive functions, index effort, provide a marker to indicate if a different individual is taking the task, and allow for real-time feedback by the researcher or clinician if desired.

VLMT scores were higher than might be expected as the task is designed to be used on a smartphone in an individual's environment. The distinction between recognition and recall can likely explain many of the patterns observed in our data. Specifically, the high performance of participants on the VLMT task (80.4% for SZ and 83.3% for BD overall) yet poorer performance on the HVLT (57.7% for SZ, 66.7% for BD) is not inconsistent with literature detailing performance on the full HVLT measure.[28] Additionally, our results are consistent with a case-control meta-analysis finding that recall is more impacted compared to recognition memory in SZ.[9] Despite these differences between tasks, performance on the VLMT and HVLT was still strongly correlated.

Over the 3 trials of the VLMT, participants had increases in false alarm errors. This likely implicates the source-monitoring deficit phenomenon in SZ and bipolar illness.[29,30] The same recognition foils were presented for each of the 3 trials and, after multiple exposures, it seems possible that identifying the source of presentation becomes challenging. This problem could likely be solved by having novel sets of recognition foils on each of the learning trial presentations, although the phenomena did not seem to affect target-item recognition, which showed systematic learning curves. Thus, for the purpose of this validation study, we primarily examined trial 1 VLMT performance.

Context, as measured by our embedded validity tests, may meaningfully impact EMCT performance, providing valuable information about performance in the real world.[2] Overall, we found a significant but small effect of distractions, interruptions, and away from the home location on performance. These findings provide data to inform a key concern that naturalistic variation may contribute to noise in EMCTs estimating cognition. We found that aggregated performance during experienced distractions and interruption was still associated with HVLT, suggesting that performance under suboptimal conditions may not greatly impact external validity. Future studies could examine how symptoms and individual vulnerabilities influence variability in cognition by context.

Limitations of this study include that the sample was of stable outpatients, and findings may not generalize to more acutely ill people. Another limitation of EMCTs is that it is difficult to know if "cheating" occurs (eg, had a friend complete a test for them) or to objectively measure attention or fatigue. Future EMCTs could employ a psychomotor vigilance task to further characterize the state of the test respondent.[31] Additionally, we did not have a healthy comparison group, control for the effect of medications, or remove individuals who had potentially noncredible scores on the MCCB. Improvements to the VLMT could include adding trial-by-trial interference, delayed recognition, emotional and nonverbal versions of these memory tasks, and, when technology becomes further developed, using speech recognition for verbal recall paradigms.[32]

In summary, the VLMT and EMCTs may be a useful tool, supplementing in-lab testing, to test cognition in naturalistic environments. The VLMT could complete longitudinal research on life-course cognitive development.[33] The VLMT may also extend cognitive testing to populations that are difficult to reach and allow clinicians and researchers to evaluate within-person variability as a dimension of impairment. Finally, the VLMT and other EMCTs could ultimately provide tools for clinical monitoring for short-term cognitive change.

Supplementary Material

Supplementary material is available at *Schizophrenia Bulletin*.

Supplemental Figure 1. Individual Variability

Supplemental Figure 2. Correlation of VLMT Overall Percentage Correct (12 and 18 List Length) with HVLT Overall Score

Acknowledgments

R.C.M. is a co-founder of KeyWise AI, Inc., and a consultant for NeuroUX. P.D.H. has received consulting fees or travel reimbursements from Acadia Pharma, Alkermes, Bio Excel, Boehringer Ingelheim, Minerva Pharma, Otsuka Pharma, Regeneron Pharma, Roche Pharma, and Sunovion Pharma during the past year. He receives royalties from the Brief Assessment of Cognition in Schizophrenia. He is chief scientific officer of i-Function, Inc. He had a research grant from Takeda and from the Stanley Medical Research Foundation. None of these companies provided any information to the authors that is not in the public domain. No other authors have conflicts of interest to report.

E.M.P. performed data analysis and took primary responsibility for writing this manuscript. S.K. collected data, conducted the literature review, assisted with data analysis, wrote the methods section, and helped write and edit the manuscript. C.A.D., a co-PI of the grant, performed data analysis with E.M.P., provided feedback throughout the process, and helped write and edit the manuscript. He assisted in the development of the mobile cognitive task and oversaw data collection at his respective site. P.D.H., a co-PI of the grant, provided feedback throughout the process and helped write and edit the manuscript. He assisted in the development of the mobile cognitive task and oversaw data collection at his respective site. A.P., PI of the grant, provided feedback throughout the process and helped write and edit the manuscript. She assisted in the development of the mobile cognitive task and oversaw data collection at her respective site. R.C.M., a co-investigator of the grant, developed the mobile cognitive task, provided feedback through the process, and helped write and edit the manuscript.

Funding

This work was supported by the National Institute of Mental Health (grant numbers R01 MH112620 to A.P; R21 MH116104 to R.C.M.; and T32 MH019934 to E.M.P).

References

1. Moore RC, Swendsen J, Depp CA. Applications for self-administered mobile cognitive assessments in clinical research: a systematic review. *Int J Methods Psychiatr Res.* 2017;26(4):e1562.
2. Weizenbaum E, Torous J, Fulford D. Cognition in context: understanding the everyday predictors of cognitive performance in a new era of measurement. *JMIR Mhealth Uhealth.* 2020;8(7):e14328.
3. Koo BM, Vizer LM. Mobile technology for cognitive assessment of older adults: a scoping review. *Innov Aging.* 2019;3(1):igy038.
4. Torous J, Keshavan M. COVID-19, mobile health and serious mental illness [published online ahead of print]. *Schizophr Res.* 2020. doi: 10.1016/j.schres.2020.04.013.
5. Charalambous AP, Pye A, Yeung WK, et al. Tools for app- and web-based self-testing of cognitive impairment: systematic search and evaluation. *J Med Internet Res.* 2020;22(1):e14551.
6. Martínez-Arán A, Vieta E, Reinares M, et al. Cognitive function across manic or hypomanic, depressed, and euthymic states in bipolar disorder. *Am J Psychiatry.* 2004;161(2):262–270.
7. Summers M, Papadopoulou K, Bruno S, Cipolotti L, Ron MA. Bipolar I and bipolar II disorder: cognition and emotion processing. *Psychol Med.* 2006;36(12):1799–1809.
8. Gold JM, Randolph C, Carpenter CJ, Goldberg TE, Weinberger DR. Forms of memory failure in schizophrenia. *J Abnorm Psychol.* 1992;101(3):487–494.
9. Grimes KM, Zanjani A, Zakzanis KK. Memory impairment and the mediating role of task difficulty in patients with schizophrenia. *Psychiatry Clin Neurosci.* 2017;71(9):600–611.
10. Bowie CR, Harvey PD. Cognitive deficits and functional outcome in schizophrenia. *Neuropsychiatr Dis Treat.* 2006;2(4):531–536.
11. Depp CA, Mausbach BT, Harmell AL, et al. Meta-analysis of the association between cognitive abilities and everyday functioning in bipolar disorder. *Bipolar Disord.* 2012;14(3):217–226.
12. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry.* 1998;59(suppl 20):22–33; quiz 34–57.
13. First MB, Williams JBW, Karg RS, Spitzer RL. *Structured Clinical Interview for DSM-5—Research Version (SCID-5 for DSM-5, research version; SCID-5-RV).* Arlington VA: American Psychiatric Association; 2015.
14. Amazon Web Services. Amazon S3 as the Data Lake Storage Platform. 2020. https://docs.aws.amazon.com/whitepapers/latest/building-data-lakes/amazon-s3-data-lake-storage-platform.html. Accessed August 3, 2020.
15. Moore RC, Paolillo EW, Sundermann EE, et al. Validation of the newly developed mobile verbal learning test (mVLT) in older adults with and without HIV infection [published online ahead of print]. *Int J Methods Psychiatr Res*. 2020:e1859. doi: 10.1002/mpr.1859.
16. Brysbaert M, New B. Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav Res Methods.* 2009;41(4):977–990.
17. Nuechterlein KH, Green MF, Kern RS, et al. The MATRICS Consensus Cognitive Battery, part 1: test selection, reliability, and validity. *Am J Psychiatry.* 2008;165(2):203–213.
18. Brandt J, Benedict RHB. *The Hopkins Verbal Learning Test—Revised.* Odessa, FL: Psychological Assessment Resources; 2001.
19. Keefe RS, Goldberg TE, Harvey PD, Gold JM, Poe MP, Coughenour L. The Brief Assessment of Cognition in Schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery. *Schizophr Res.* 2004;68(2-3):283–297.
20. Tombaugh TN. Trail Making Test A and B: normative data stratified by age and education. *Arch Clin Neuropsychol.* 2004;19(2):203–214.

21. Gold JM, Carpenter C, Randolph C, Goldberg TE, Weinberger DR. Auditory working memory and Wisconsin card sorting test performance in schizophrenia. *Arch Gen Psychiatry.* 1997;54(2):159–165.
22. Mausbach BT, Harvey PD, Goldman SR, Jeste DV, Patterson TL. Development of a brief scale of everyday functioning in persons with serious mental illness. *Schizophr Bull.* 2007;33(6):1364–1372.
23. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull.* 1987;13(2):261–276.
24. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry.* 1979;134:382–389.
25. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. *Br J Psychiatry.* 1978;133:429–435.
26. Lenhard W, Lenhard A. Hypothesis tests for comparing correlations. Psychometricia. 2014 https://www.psychometrica.de/correlation.html. Accessed April 16, 2020.
27. Schweitzer P, Husky M, Allard M, et al. Feasibility and validity of mobile cognitive testing in the investigation of age-related cognitive decline. *Int J Methods Psychiatr Res.* 2017;26(3):e1521.
28. Schretlen DJ, Cascella NG, Meyer SM, et al. Neuropsychological functioning in bipolar disorder and schizophrenia. *Biol Psychiatry.* 2007;62(2):179–186.
29. Harvey PD, Docherty NM, Serper MR, Rasmussen M. Cognitive deficits and thought disorder: II. An 8-month followup study. *Schizophr Bull.* 1990;16(1):147–156.
30. Keefe RS, Arnold MC, Bayen UJ, Harvey PD. Source monitoring deficits in patients with schizophrenia: a multinomial modelling analysis. *Psychol Med.* 1999;29(4):903–914.
31. Price E, Moore G, Galway L, Linden M. Validation of a smartphone-based approach to in situ cognitive fatigue assessment. *JMIR Mhealth Uhealth.* 2017;5(8):e125.
32. Holmlund TB, Chandler C, Foltz PW, et al. Applying speech technologies to assess verbal memory in patients with serious mental illness. *NPJ Digit Med.* 2020;3:33.
33. Kilciksiz CM, Keefe R, Benoit J, Öngür D, Torous J. Verbal memory measurement towards digital perspectives in first-episode psychosis: a review. *Schizophr Res Cogn.* 2020;21:100177.