

UC San Diego

UC San Diego Previously Published Works

Title

Feasibility and validity of ecological momentary cognitive testing among older adults with mild cognitive impairment.

Permalink

<https://escholarship.org/uc/item/9th5h5v6>

Authors

Moore, Raeanne
Ackerman, Robert
Russell, Madisen
[et al.](#)

Publication Date

2022

DOI

10.3389/fdgth.2022.946685

Peer reviewed



OPEN ACCESS

EDITED BY

Amit Baumel,
University of Haifa, Israel

REVIEWED BY

Nelson Roque,
University of Central Florida, United States
Richard Lipton,
Albert Einstein College of Medicine,
United States

*CORRESPONDENCE

Raeanne C. Moore
r6moore@health.ucsd.edu

SPECIALTY SECTION

This article was submitted to Human Factors and Digital Health, a section of the journal Frontiers in Digital Health

RECEIVED 17 May 2022

ACCEPTED 20 July 2022

PUBLISHED 05 August 2022

CITATION

Moore RC, Ackerman RA, Russell MT, Campbell LM, Depp CA, Harvey PD and Pinkham Amy E. (2022) Feasibility and validity of ecological momentary cognitive testing among older adults with mild cognitive impairment. *Front. Digit. Health* 4:946685. doi: 10.3389/fdgth.2022.946685

COPYRIGHT

© 2022 Moore, Ackerman, Russell, Campbell, Depp, Harvey and Pinkham. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Feasibility and validity of ecological momentary cognitive testing among older adults with mild cognitive impairment

Raeanne C. Moore^{1*}, Robert A. Ackerman², Madisen T. Russell², Laura M. Campbell³, Colin A. Depp^{1,4}, Philip D. Harvey^{5,6} and Amy E. Pinkham²

¹Department of Psychiatry, University of California San Diego, La Jolla, CA, United States, ²Department of Psychology, School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, United States, ³San Diego State University/University of California San Diego Joint Doctoral Program in Clinical Psychology, San Diego, CA, United States, ⁴Veterans Affairs San Diego Healthcare System, San Diego, CA, United States, ⁵Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, Miami, FL, ⁶Research Service, Bruce W. Carter VA Medical Center, Miami, FL, United States

It is critical to intervene early in the mild cognitive impairment (MCI) stage of the Alzheimer's disease trajectory, but traditional cognitive testing methods are costly, burdensome, and difficult to access. We examined adherence and validity data to a 30-day self-administered ecological momentary cognitive testing protocol among a sample of older adults with MCI and cognitively normal controls to evaluate feasibility, tolerability, and initial validity in comparison to standard neuropsychological tests. Participants included 48 participants with MCI (Mean age = 72 years, SD = 7 years) and 46 demographically-matched cognitively normal (NC) control participants (Mean age = 70 years, SD = 7 years). Participants completed traditional neuropsychological testing to determine MCI status, followed by 30 days of remote ecological momentary cognitive testing. Ecological momentary assessment (EMA) surveys were administered 3 times per day for 30 days (possible total = 90), and mobile cognitive tests were administered every other day (for a total of 15 administrations). Mobile cognitive tests included the Variable Difficulty List Memory Test (VLMT; measure of learning and memory), Memory Matrix (measure of visual working memory), and the Color Trick Test (measure of executive function). EMA and mobile cognitive test adherence, fatigue effects, mobile cognitive test performance and group differences, and psychometrics (reliability, convergent validity, ceiling effects, and practice effects) were examined. Overall mean-level adherence to the mobile cognitive tests was 85% and did not differ by MCI status. The reliability of stable between-person individual differences for the VLMT and Memory Matrix were very high. Moreover, although the reliability of within-person change for Memory Matrix was adequate, the corresponding reliability for VLMT was somewhat low. Averaged performance on the mobile cognitive tests was correlated with lab-based tests measuring the same construct. Participants with MCI performed worse than NCs on the VLMT and Color Trick Test, and there was no evidence of fatigue effects for these two tests. These findings support the feasibility and potential for ecological momentary cognitive testing to support clinical trials and for measuring cognitive changes over time in persons with increased risk for Alzheimer's disease such as those with MCI.

KEYWORDS

ecological momentary assessment, ambulatory assessment, smartphones, Alzheimer's disease, adherence, psychometrics

1 Introduction

Research that examines cognitive functioning has traditionally taken place in a lab with paper and pencil neuropsychological testing; however, there are barriers with this method, including high cost, time burden, and access to testing locations which are limited by transportation and uneven distribution in rural or remote areas. As a result, neurocognitive testing is infrequently repeated, if at all. Ecological momentary cognitive tests (EMCTs), which are brief and repeatable cognitive assessments that are self-administered *via* smartphone in participants' own environments, may be a valuable complement to traditional neuropsychological testing that can help overcome some of these barriers (1–4).

There are several advantages to EMCTs that may make them well suited for use in clinical trials. Cognition can fluctuate from day to day, which makes it difficult to determine what should be considered a real change on neuropsychological testing from one time point to another. This is particularly problematic when trying to examine improvement over time (e.g., recovery from stroke) or cognitive decline as seen in Alzheimer's disease and related dementias. Alzheimer's disease is the most common cause of dementia in older adults (5) and places significant financial and emotional burden on affected families, not to mention the financial impact on healthcare systems. Therefore, it is no surprise that there are currently hundreds of ongoing clinical trials aimed at prevention of and intervention in Alzheimer's disease and related dementias (6).

To date, pharmacological interventions have been slow to show reductions in cognitive decline, and no treatments have been able to reverse cognitive decline despite some evidence for slowing disease progression; however, many of these studies use less-than-optimal cognitive outcome measures. For example, the Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog) has been shown to have significant ceiling effects in those with normal cognition and mild cognitive impairment (MCI) and there are concerns about its ability to detect cognitive changes early in the disease course (7–9). Given that EMCTs can be given over multiple days, EMCTs may be a cost effective and time efficient method to establish a more accurate baseline for cognitive functioning and to detect person-specific changes more sensitively over time. Such procedures could also allow for dynamic titration of difficulty in order to more effectively probe variation in performance.

EMCTs can also be paired with other technologies such as ecological momentary assessment (EMA) or wearable devices (e.g., actigraphy to objectively assess physical activity and sleep). Therefore, observational studies or interventional studies can examine how mood, activities, sleep, and other fluctuating daily-life factors associate with cognition over time

without relying on retrospective recall, which is particularly relevant to persons with memory impairments (e.g., 10, 11). Utilizing EMCTs to examine cognition in a person's everyday life with different contextual variables could lead to person-specific intervention strategies (4).

Additionally, the use of EMCT may reduce the number of in-person visits, which could reduce the burdens of time and transportation, particularly for participants that live in rural areas and older adults with mobility limitations. The tradeoff is that technology familiarity may impact one's ability to engage in EMCTs and is something to be mindful of in this group. However, a study conducted in 2021 by the Pew Research Center found that 83% of those aged 50–64 own a smartphone and 61% of adults aged 65+ own a smartphone, indicating that the majority of older adults are already engaged with smartphone technology (12). To date, there have been a handful of studies by other groups utilizing smartphone-based mobile cognitive testing among cognitively normal older adults (e.g., 10, 13, 14) and older adults with MCI (e.g., 15, 16), all of which have demonstrated feasibility, good adherence, and promising initial psychometric properties for use of these tests in this population.

Despite the clear appeal of EMCT in aging research, there are some current limitations. For example, a recent systematic search and evaluation found that the majority of currently-available commercial-grade app-based tools to assess cognition lack validity data for their assessments (17). This is concerning, as an absence of validity data in these tools could lead to unreliable information about possible cognitive impairment. Therefore, we present adherence and validity data in a group of older adults with and without MCI for three NeuroUX EMCTs assessing the domains of memory and executive functioning: 1) Variable Difficulty List Memory Test (VLMT), which is a verbal list-learning test in which we administered 6-word, 12-word, and 18-word versions; 2) Memory Matrix, a visual working memory task; and 3) Color Trick Test, an executive functioning task examining inhibition using a Stroop-Type paradigm. The aims of the study were to examine the 1) adherence to the 30-day EMCT protocol, 2) fatigue effects, 3) EMCT task performance and group differences, and 4) EMCT psychometrics, including reliability, convergent validity (compared to traditional neuropsychological tests), ceiling effects, and practice effects.

2 Materials and methods

2.1 Participants

Participants were English-proficient individuals aged 50 or older who met criteria for any subtype of mild cognitive impairment (MCI) using Jak/Bondi criteria, which require performance of one standard deviation below normative

expectations on two different assessments within a single cognitive domain (i.e., memory, attention, language, executive functioning), or cognitively normal (NC) control participants. Exclusion criteria included: (1) presence or history of medical or neurological disorders that may affect brain function (e.g., stroke, epilepsy, Parkinson’s disease), (2) presence of dementia, (3) history of unconsciousness for a period greater than 15 min, (4) significant impairment of vision (e.g., blindness, glaucoma, vision uncorrectable to 20/40, color blindness) or hearing (e.g., hearing loss) that would interfere with their ability to complete the study protocol, (5) presence of intellectual disability (defined as IQ < 70), (6) current diagnosis of substance use disorder, (7) or presence or history of a psychotic disorder or bipolar disorder.

Data were collected across three sites between December 2020 and December 2021: The University of Texas at Dallas (UTD), University of California San Diego (UCSD), and University of Miami Miller School of Medicine (UM), resulting in a total of 94 participants (48 MCI, 46 NC). UTD participants were recruited from community advertisements and previous participation in aging-related research studies at the Center for Vital Longevity at UTD. UCSD participants were recruited from word of mouth and posting in the Stein Institute for Successful Aging monthly newsletter. UM participants were recruited from the clinical programs at the Miller School of Medicine Memory Disorders Center, the Florida ADRC, and through advertisements and previous study participants.

2.2 Procedures

The study was approved by each University’s respective Institutional Review Board, and all participants provided written informed consent. After a brief phone screen, participants completed a baseline visit either remotely *via* Microsoft Teams or Zoom or in-person. During the baseline visit, participants completed a neuropsychological battery. Research staff held a bachelor’s degree or higher, and were trained over the course of several weeks, within and across sites, to administer and score the neuropsychological tests accurately. Jak/Bondi diagnostic criteria for MCI were applied

to the neuropsychological test data to determine MCI status. The Jak/Bondi diagnostic criteria show a good balance of sensitivity, specificity, and reliability compared to other conventional MCI criteria (18). Study eligibility, all neuropsychological test scores, and diagnoses were reviewed by the first author (RCM). Once eligibility and group status were confirmed, staff contacted participants to set up their smartphones for the EMCT period. Participants could either complete the EMCTs using their personal smartphone or, if they requested or did not own a smartphone, they were provided with a study-owned Android smartphone. Those using study-provided smartphones were trained to operate the device and given a user manual to reduce technological issues. Participants were trained on the EMCT protocol and completed a mock EMA survey and mobile cognitive testing session to allow for technical questions and troubleshooting.

For the following 30 days, participants completed the EMCT protocol using the NeuroUX platform (19). Participants were sent text message notifications to take the EMA surveys three times per day. Every other day, participants were asked to complete the three different mobile cognitive tests (i.e., Variable Difficulty List Memory Test, Memory Matrix, Color Trick Task) of varied difficulty along with each of their EMA surveys. The mobile cognitive tests were counterbalanced throughout the EMA period by test type and difficulty level, resulting in a total of 5 easy, 5 medium, and 5 hard conditions of each of the three mobile cognitive tests (see Figure 1). To encourage EMA adherence and help troubleshoot any difficulties, researchers contacted participants if they missed more than three surveys in a row. Participants were compensated up to \$190 total for completing the baseline visit (\$50) and EMCT sessions (EMA questions only – \$0.88; EMA + mobile cognitive tests – \$2.25).

2.2.1 Remote visit task modifications

Due to evolving restrictions on in-person data collection during the height of the COVID-19 pandemic, some individuals participated in-person (*n* = 28) whereas others participated *via* remote visits (*n* = 66). For remote appointments, all tasks were completed *via* video conferencing using Microsoft Teams or Zoom meetings and required minimal modification. Participants were asked to complete the visit in a quiet environment away

Mobile Cognitive Test	Study Day (and administration order)														
	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29
VLMT	1	2	3	2	1	3	3	1	2	2	3	1	1	2	3
Color Trick	2	1	2	3	3	1	2	2	1	3	1	3	2	3	1
Memory Matrix	3	3	1	1	2	2	1	3	3	1	2	2	3	1	2

FIGURE 1 Protocol of mobile cognitive testing administration. Note. Difficulty levels are depicted as green (easy), yellow (medium), and red (hard).

from distractions (e.g., away from other individuals, powering off/silencing unrelated devices) and a screening measure was completed to ensure participants could hear the researcher well and see the PowerPoint materials on their desktop, laptop, or iPad. Researchers also asked participants to refrain from utilizing any performance aids, such as writing down stimulus items, searching for answers on the internet, or seeking help from other individuals.

Tasks that were typically administered orally (Hopkins Verbal Learning Test – Revised (HVLTR), Number Span Test: Forward) were implemented as is. Tasks that required visual presentations (Wide Range Achievement Test-4 (WRAT-4), Delis-Kaplan Executive Function System Color-Word Interference Test (D-KEFS), Brief Visuospatial Memory Test - Revised (BVMTR)) were administered *via* video call using a PowerPoint screenshare function. Prior to the baseline visit, research staff instructed participants to prepare four blank pieces of printer paper for the BVMTR task. Additionally, during the BVMTR task, after the participant completed each trial drawing, the researcher asked the participant to hold the paper in front of the camera so that a photo could be taken, then instructed them to flip the paper over and place it out-of-sight before beginning the next trial.

2.3 Measures

2.3.1 Traditional Neuropsychological measures (lab or remote administered at baseline)

To determine premorbid IQ, the Wide Range of Achievement Test 4 (WRAT-4; 20) word reading subtest was used. The Montreal Cognitive Assessment-BLIND version 7.1 (MoCA-BLIND; 21) was administered to screen for the presence of dementia using established cutoff scores. This version of the MoCA was used for participants who completed virtual visits as well as participants who completed in-person visits. To determine MCI eligibility, the following tests were administered: Hopkins Verbal Learning Test – Revised (HVLTR; 22), Brief Visuospatial Memory Test – Revised (BVMTR; 23), Oral Trail Making Test- A and B (24), Digit Span Forward (25), Verbal Fluency – Letter and Animals (25), Multilingual Naming Test (MINT; 26), Number Span Test: Forward (25), and the D-KEFS-Color Word Interference Test (27).

For validity analyses in the current study, we used non-demographically adjusted scores from the HVLTR (verbal memory), BVMTR (visual memory), Letter-Number Span (attention/working memory), and D-KEFS Color-Word Interference Test (executive function).

2.3.2 EMA surveys

Each EMA survey asks participants questions about their daily functioning, including where they are (dichotomized as

“at home” versus “away”) and who they are with (dichotomized as “alone” versus “with others”). The EMA surveys also generally queried participants’ mood, cognitive concerns, substance use, pain, and sleep as additional questions but data are not reported here.

2.3.3 Mobile cognitive tests

See **Table 1** for a list of the mobile cognitive tests, the cognitive domains assessed, completion times, and screenshots.

2.3.3.1 Mobile variable difficulty list memory test (VLMT)

The VLMT has been described and validated by Parrish et al. (2020). For this task, participants are presented with a list of words (list length varies between 6, 12, or 18) on 3 separate trials for 30 s each. Immediately following each trial, participants are shown target and distractor words one-by-one and asked to identify whether the word appeared on the list (matched number of target and distractor words presented). Each trial is scored by number of words correctly recalled or based on a percentage of correct target items (range 0%–100%).


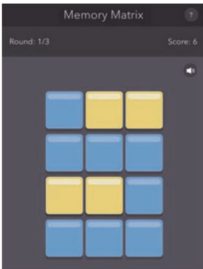


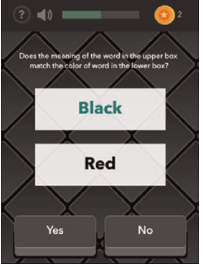
2.3.3.2 Memory matrix

During the Memory Matrix task, participants are presented with a matrix of blue tiles. A pattern of yellow tiles is then displayed, and the participant is asked to memorize the location of the yellow tiles. After 1.5 s, the yellow tiles are then switched back to blue, and the participant is asked to tap the tiles that were previously yellow. Matrix sizes are varied across administration days so that participants complete 5 days of 6-tile matrices, 5 days of 12-tile matrices, and 5 days of 18-tile matrices. Each administration also includes three trials of 9 patterns each. Participants earn 1 point for each pattern correctly recreated for a score range of 0–9 per trail and 0–27 per administration.

2.3.3.3 Color trick

The Color Trick task was modelled after the Stroop-type paradigm (Stroop, 1935). Participants completed three different conditions of this task (Meaning-to-Meaning, Meaning-to-Color, Yes-No Mechanic) divided across the 15 days of EMCTs such that each condition was administered 5 times. Each condition includes three trials of 9 items/questions for a total of 27 items per administration. Each item in each condition shows a word in an upper box of the smartphone screen and between 1 and 3 words on the lower half of the screen. The font colors and actual meanings of the upper and lower words are either the same or different colors. The first condition type is *Meaning-to-Meaning*, in which participants are presented with one word in an upper box on their screen and 2–3 word choices on the lower half of their screen and asked to select the word choice that has the same meaning as the word in the top box (e.g., matching top word “pink” with bottom word “pink”). The second condition type

TABLE 1 Mobile cognitive tests.

Mobile Cognitive Test	Cognitive Domain Assessed	Time to Complete	Screenshot of Task
Variable Difficulty List Memory Test (VLMT)	Recognition Memory	30 s for list presentation	
Memory Matrix	Visual Working Memory	Variable; 3 trials; approximately 1–2 min (Mean completion time: 1.5 min)	
Color-Trick: Meaning-to-Meaning	Executive Function	Variable; 3 trials; approximately 1.5–3 min (Mean completion time: 2.25 min)	
Color-Trick: Meaning-to-Color	Executive Function	Variable; 3 trials; approximately 2–3.5 min (Mean completion time: 2.75 min)	
Color-Trick: Yes-No Mechanic	Executive Function	Variable; 3 trials; approximately 2.5–3.5 min (Mean completion time: 3 min)	

is *Meaning-to-Color*, in which participants are presented with one word in an upper box on their screen and 2–3 word choices on the lower box of their screen and asked to select

the word choice that has the same font color as the meaning of the word in the top box (e.g., matching top word “pink” with bottom word printed in pink font). The third condition

type is *Yes-No Mechanic*, in which participants are presented with one word in an upper box on their screen and one word in a lower box on their screen, and asked, “Does the meaning of the word in the upper box match the color of the word in the lower box?” and the participant can choose either “yes” or “no.” Each trial is scored based on the number of items correct (range 0–9) and average response time for correct items.

2.4 Statistical analyses

Demographic differences between groups (MCI+ vs. NCs) and administration formats (in-person vs. remote) were assessed using independent samples *t*-tests or Chi-Square tests (χ^2) as appropriate. Adherence was calculated as the percentage of EMA surveys completed by the total number possible (90), as well as the percentage of each of the three mobile cognitive tests completed by the total number possible (15 each). Adherence differences between groups and administration formats were assessed using independent samples *t*-tests. In addition, Pearson’s *r* correlations were used to estimate relationships between adherence and demographic differences.

To further assess whether adherence changed over time, we computed missing data variables for the EMCTs that denoted whether participants skipped a test that they were scheduled to take (0 = completed test, 1 = missed test). We then estimated fatigue effects for each of the EMCTs (i.e., whether participants’ odds of missing a test was greater on later versus earlier study days) using growth-curve models specified with multilevel logistic regression model in Mplus v. 8.4 (28). Using maximum likelihood estimation, each model regressed participants’ log odds of missing a test on time (scaled such that 0 is the midpoint of the EMA period and a one-unit change corresponds to the total change in the log odds of missing a test across the EMA period), MCI status (effect coded such that $-1 = \text{NC}$ and $1 = \text{MCI}$), and the interaction of time with MCI status. Each model also included an unstructured variance-covariance matrix for the random intercepts and slopes. These specifications enabled us to estimate the average probability of missing a test across the EMA period (*via* the threshold value¹), the average fatigue effect in the sample (*via* the first-order effect of time²), whether the average log odds of missing a test across the EMA period differs between NC and MCI (*via* the first-order effect of MCI

status), and whether fatigue effects differ between NC and MCI (*via* the interaction of time and MCI status).

We next investigated participants’ average performance on the EMCTs across the EMA period. To evaluate group differences (i.e., NC vs. MCI) on EMCT performance across trials, we conducted independent samples *t*-tests.

The final sets of analyses provided additional psychometric evidence for each EMCT – namely, reliability, convergent validity, ceiling effects, and practice effects. We first calculated Intraclass Correlation Coefficients (ICCs) for each EMCT to quantify the proportion of variance in the tests attributed to trait vs. state components across the EMA period. We then used generalizability theory (see Ref. 29) to estimate the reliability of stable between-person individual differences (R_{KF}) as well as the reliability of within-person change (R_{C}) in the EMCT measures that contained multiple trials (i.e., list-learning and matrix memory). These analyses used the Minimum Norm Quadratic Unbiased Estimate (MINQ) method within SPSS v. 26 to estimate the variance components linked to the factorial combination of participant, day, and item (where only participant was treated as a random factor).

We then evaluated the convergent validity evidence for each EMCT by estimating correlations between participants’ average performance on a given EMCT and their parallel performance on a similar lab-based measure. Ceiling effects for each EMCT were subsequently evaluated by counting the number of participants who earned the maximum score consistently across the EMA period. Practice effects for each of the EMCTs (i.e., whether participants’ performance on the measures systematically changed across the course of the EMA period) were then assessed *via* growth-curve models specified with linear multilevel regression in Mplus v. 8.4 (28). Using maximum likelihood estimation with robust standard errors, each model regressed participants’ test scores on time, MCI status, and the interaction of time with MCI status (we used the same scaling for time and MCI status as our analyses investigating fatigue effects). When sufficient variability was present, we specified an unstructured variance-covariance matrix for the random intercepts and slopes. These specifications enabled us to estimate participants’ average performance on the EMCT (*via* the intercept), the average practice effect in the sample (*via* the first-order effect of time), whether average levels of performance for an EMCT differs between NC and MCI (*via* the first-order effect of MCI

0.50) of the study:

$$\pi = \frac{\exp[-(\tau) + \beta_1 X_i]}{1 + \exp[-(\tau) + \beta_1 X_i]}$$

where τ = threshold, β_1 = Slope reflecting fatigue effect, and X_i = the specific value of time.

¹Participants’ average probability of missing an EMCT item was computed as $1/(1 + \exp(\tau))$, where τ = threshold.

²In cases where there was evidence for a fatigue effect, we used the following formula to determine participants’ average probability of missing an EMCT item at the beginning (time = -0.50) and end (time =

status), and whether practice effects differ between NC and MCI (*via* the interaction of time and MCI status).

3 Results

3.1 Sample characteristics

Demographic and clinical characteristics by MCI status are displayed in **Table 2**. Groups were comparable on demographics and did not significantly differ on age, sex, race, ethnicity, or years of education. Groups were also comparable on type of phone used, with 55% of MCI participants and 62% of NCs using iPhones, while the other participants used Android devices (Chi-Square = 11.3, $p = 0.334$; **Supplementary Table S1**).

Sixty-six participants completed the lab-based neuropsychological visit remotely *via* telehealth, while 28 completed this visit in-person. There were no demographic differences for participants who completed this visit remotely versus in-person except for fewer Hispanic individuals in the in-person group ($\chi^2 = 6.4$, $p = 0.01$). Additionally, there were no significant differences in MCI status ($\chi^2 = 0.59$, $p = 0.44$) or performance on any of the neuropsychological tests based on remote vs. in-person participation (all $ps > 0.09$).

3.2 Adherence

For the whole sample, adherence to EMA surveys was 86% (SD = 15.8%; range = 24%–100%). In regard to the mobile cognitive tests, adherence to the VLMT was 84% (SD = 19.3%; range = 7%–100%), adherence to Memory Matrix was 85% (SD = 18%; range = 20%–100%), and adherence to Color Trick was 85% (SD = 17%; range = 13%–100%). Adherence to EMA surveys did not differ by diagnostic status, $t = 1.21$, $p = 0.23$, and neither did completion rates of the mobile cognitive tests (VLMT: $t = 0.83$, $p = 0.41$; Memory Matrix: $t = 1.56$, $p = 0.12$; Color Trick: $t = 0.97$, $p = 0.33$). Further, there was no difference in EMA adherence or mobile cognitive test completion rates for participants who completed the lab-visit remotely or in-person (all $ps > 0.19$). Age, education, and estimated IQ (measured by the WRAT-4) did not correlate with adherence to EMCTs nor with percentage of surveys completed at home or alone, except for a small negative correlation between years of education and completion of the Memory Matrix test. Higher adherence was positively correlated with answering more surveys when home and when alone (see **Table 3**).

3.3 Fatigue effects

Because we used varying list lengths for the VLMT, we included list length (*via* two effect-codes that treated the 18-

word list length as the reference group) and its interaction with time as covariates in the VLMT fatigue effect analyses. On average, participants' probability of missing (i.e., failing to complete) a list-learning item was 0.08 for Trial 1 (threshold = 2.40, SE = 0.23, $p < 0.001$), 0.08 for Trial 2 (threshold = 2.38, SE = 0.22, $p < 0.001$), and 0.09 for Trial 3 (threshold = 2.34, SE = 0.22, $p < 0.001$), where trials refer to trials within the same test (e.g., for the VLMT, there were three trials administered at each session). We found no evidence of a fatigue effect for Trial 1 (logit = 0.46, SE = 0.53, $p = 0.39$; OR = 1.58), Trial 2 (logit = 0.56, SE = 0.53, $p = 0.29$, OR = 1.74), or Trial 3 (logit = 0.52, SE = 0.52, $p = 0.32$, OR = 1.68). Moreover, MCI participants did not significantly differ from controls on their log odds of missing a list-learning item vs. not missing the item for Trials 1, 2, or 3 (all p 's > 0.12) or their fatigue effects for Trials 1, 2, or 3 (all p 's > 0.57).

Similar to the VLMT, participants' average probability of missing a Memory Matrix item across the EMA period was 0.08 for Trial 1 (threshold = 2.45, SE = 0.21, $p < 0.001$), 0.08 for Trial 2 (threshold = 2.43, SE = 0.21, $p < 0.001$), and 0.08 for Trial 3 (threshold = 2.42, SE = 0.21, $p < 0.001$). Unlike the VLMT, however, we found evidence of fatigue effects for the Memory Matrix items across the three trials. In particular, participants' odds of missing a Memory Matrix item vs. not missing a Memory Matrix item from the beginning to the end of the EMA period increased approximately 3.23-fold for Trial 1 (logit = 1.174, SE = 0.52, $p = 0.023$), approximately 3.47-fold for Trial 2 (logit = 1.244, SE = 0.51, $p = 0.014$), and approximately 3.42-fold for Trial 3 (logit = 1.231, SE = 0.50, $p = 0.014$). That is, whereas participants' probability of missing a Memory Matrix item was 0.05 at the beginning of the EMA period for Trials 1, 2, and 3, their probability of missing a Memory Matrix item at the end of the EMA period was 0.13 for Trials 1 and 2 and 0.14 for Trial 3. Nonetheless, MCI participants did not significantly differ from controls on their log odds of missing a Memory Matrix item vs. not missing the item for Trials 1, 2, or 3 (all p 's > 0.06) or on their fatigue effects for Trials 1, 2, or 3 (all p 's > 0.59).

Participants' average probability of missing a Color Trick item across the EMA period was 0.09 for Trial 1 (threshold = 2.285, SE = 0.19, $p < 0.001$), 0.09 for Trial 2 (threshold = 2.269, SE = 0.19, $p < 0.001$), and 0.09 for Trial 3 (threshold = 2.256, SE = 0.19, $p < 0.001$). We found no evidence of a fatigue effect for Trial 1 (logit = 0.299, SE = 0.46, $p = 0.514$, OR = 1.35), Trial 2 (logit = 0.242, SE = 0.46, $p = 0.598$, OR = 1.27), or Trial 3 (logit = 0.269, SE = 0.45, $p = 0.55$, OR = 1.31). MCI participants also did not significantly differ from controls on their log odds of missing a Color Trick item vs. not missing the item for Trials 1 to 3 (all p 's > 0.07) or on their fatigue effects for Trials 1 to 3 (all p 's > 0.20).

TABLE 2 Demographics and clinical characteristics by mild cognitive impairment (MCI) status.

	MCI (<i>n</i> = 48)	Cognitively Normal (CN) (<i>n</i> = 46)	Test-statistic ^a	<i>p</i> -value
Demographics				
Age in years, <i>M</i> (SD); range	72 (7.7); 54–85	70 (6.6); 60–87	0.96	0.34
Sex (% F)	27 (56%)	34 (73%)	3.22	0.07
Race (%)				
White	45 (94%)	41 (89%)	4.81	0.09
Black/African American	1 (2%)	5 (11%)		
More than one race	2 (4%)	0 (0%)		
Ethnicity (% Hispanic/Latino)	8 (17%)	5 (11%)	0.66	0.42
Education (years), <i>M</i> (SD)	16.1 (2.5)	16.2 (2.1)	0.26	0.80
Premorbid IQ (WRAT-4 SS), <i>M</i> (SD)	110.2 (15.1)	109.9 (12.0)	0.11	0.91
Employment status				
Retired	26 (54%)	32 (70%)	2.64	0.45
Unemployed	2 (4%)	1 (2%)		
Part-time employment or volunteer	14 (29%)	8 (17%)		
Full-time employment or volunteer	6 (13%)	5 (11%)		
Residential Status				
Independent/Financially Responsible	48 (100%)	44 (96%)	2.13	0.14
Independent/Not Financially Responsible	0 (0%)	2 (4%)		
Smartphone used for study				
Personal iPhone	27 (56%)	31 (67%)	4.36	0.11
Personal Android	17 (36%)	15 (33%)		
Study Loaned Android	4 (8%)	0 (0%)		
Remote Participation	32 (67%)	34 (74%)	0.59	0.44
Lab-Based Neuropsychological Scores^b				
Hopkins Verbal Learning Test (HVLT) – Immediate Recall	40.7 (9.9)	51.4 (10.0)	5.24	<0.001
Brief Visuospatial Memory Test-R (BVM-T-R) – Immediate Recall	50.8 (9.7)			
Letter Number Span	45.1 (8.9)	49.6 (9.3)	2.4	0.02
D-KEFS Interference	54.7 (11.4)	56.4 (10.0)	0.73	0.47
Mobile Cognitive Tests – Mean aggregated scores^c				
VLMT 6 words (% Correct)	94.5 (5.7)	95.6 (5.0)	1.04	0.30
VLMT 12 words (% Correct)	85.0 (8.5)	87.3 (6.1)	1.50	0.14
VLMT 18 words (% Correct)	76.6 (9.3)	80.8 (6.9)	2.41	0.02
Memory Matrix (Total Score)	7.3 (0.93)	7.4 (0.83)	0.97	0.33
Color Trick: Meaning-to-Meaning (Total Score)	8.2 (0.51)	8.5 (0.46)	2.13	0.04
Color Trick: Meaning-to-Color (Total Score)	8.6 (0.41)	8.7 (0.42)	1.50	0.07
Color Trick: Yes-No Mechanic (Total Score)	8.6 (0.41)	8.7 (0.28)	1.19	0.24

Note. Values are presented as mean (SD) or *n* (%).

^aT-tests for continuous variables; Chi square for dichotomous variables.

^bDemographically-adjusted *T*-Scores from lab-based neuropsychological scores are reported.

^cRaw scores are reported.

3.4 EMCT performance and group differences

Table 2 presents average mobile cognitive test performance for the MCI and NC groups across the EMA period. As expected, participants generally committed more errors on the

VLMTs when the list length was greater. Participants' performance on the Memory Matrix and Color Trick tests was also quite high. While participants with MCI scored lower on all EMCTs, they only performed significantly worse than the NC participants on the 18-word VLMT and the Color Trick: Meaning-to-Meaning task.

TABLE 3 Correlations between adherence and demographic characteristics in the whole sample ($N = 94$).

	Age	Education	Estimated IQ	% surveys completed at home	% surveys completed alone
EMA Adherence	-0.122	-0.167	-0.129	0.582**	0.286**
VLMT Adherence	-0.029	-0.023	-0.075	0.536**	0.249*
Memory Matrix Adherence	-0.158	-0.205*	-0.129	0.511**	0.274**
Color Trick Adherence	-0.117	-0.114	-0.132	0.363**	0.381**

Note. * $p < 0.05$; ** $p < 0.01$.

We also examined performance differences by phone type. In the overall sample, there were no significant performance differences based on phone type (Supplementary Table S2). When examining the effects of both phone type and group (and their interaction) on mobile cognitive test performance, no main effects were found for the VLMT 6- or 12-word list, Memory Matrix, Color Trick Meaning-to-Color, or Color Trick Yes-No Mechanic (all p 's > 0.05). Further, there were no significant interactions between phone type and group on any of the mobile cognitive tests (all p 's > 0.05). For the VLMT 18-word list, a main effect for group was observed, such that NC participants performed better than participants with MCI ($F = 6.53$, $p = 0.01$); there was no main effect for phone type ($F = 0.53$, $p = 0.47$). Lastly, there was a main effect for group on Color Trick Meaning-to-Meaning, such that MCI participants performed worse than NC participants ($F = 5.23$; $p = 0.03$), but there was no main effect for phone type ($F = 0.11$, $p = 0.74$).

3.5 EMCT psychometrics: Reliability, convergent validity, ceiling effects, and practice effects

3.5.1 Psychometric evidence for VLMT

Aggregated across trials, the Intraclass Correlation Coefficients (ICCs) for each trial length of the VLMT were 0.22, 95% CI [0.11, 0.32] for the 6-word list, 0.33, 95% CI [0.22, 0.44] for the 12-word list, and 0.32, 95% CI [0.20, 0.42] for the 18-word list. Thus, most of the variance on VLMT can be attributed to within-person differences in performance across trials. Using generalizability theory, we further found that the reliability of stable between-person individual differences in VLMT scores across list lengths and trials was quite high ($R_{KF} = 0.94$). In contrast, the reliability of within-person change across list lengths and trials was somewhat low ($R_C = 0.57$).

To examine convergent validity, we examined relationships between the VLMT with immediate recall scores from the HVLMT and BVMT (see Table 4). We examined the VLMT data in two ways: percentage correct by trial length and overall correct across all trial lengths. In the overall sample, percent of items correct on the 18-item VLMT list was

positively correlated with the HVLMT ($r = 0.33$, $p < 0.001$). The relationships between the 6- and 12-item percent correct VLMT lists were not significantly related to HVLMT performance. When looking at the overall correct data across all three list lengths, the VLMT was positively associated with HVLMT ($r = 0.26$, $p = 0.012$). When comparing the VLMT to the BVMT, percent of items correct on the 6-item VLMT list was positively correlated with the BVMT ($r = 0.27$, $p = 0.01$); 12- and 18-item VLMT lists were unrelated to the BVMT. The VLMT overall correct scores (across all three list lengths) was positively correlated with BVMT performance ($r = 0.27$, $p = 0.01$).

We next examined whether there were ceiling effects at any of the VLMT list lengths. At length 6, there was some evidence for ceiling effects such that on Trial 1, 13 (28%) NC and 15 (31%) MCI participants consistently scored 100%; on Trial 2, 23 (50%) NC and 26 (54%) MCI consistently scored 100%; and on Trial 3, 29 (63%) NC and 27 (56%) MCI participants consistently scored 100%. No ceiling effects were observed for list length 12 or 18.

Practice effects were subsequently investigated with linear mixed effect models to determine whether participants' performance on the VLMT systematically changed across the EMA period.³ Note that all effects were adjusted for list length. On average, participants recognized 10.06 out of an average of 12 words (i.e., average of 6, 12, and 18) correctly (SE = 0.08), averaging across the list lengths. Moreover, participants showed a systematic decline in the number of words they got correct for the list-learning task across the EMA period (on average, participants' total change = -0.84, SE = 0.14, $p < 0.001$). Although MCI participants ($M = 9.87$) significantly differed from controls ($M = 10.25$) on their average number of words correct across the trials ($b = -0.19$, SE = 0.08, $p = 0.015$), participants' systematic change in words correct across the EMA period was not significantly related to MCI status ($b = -0.10$, SE = 0.14, $p = 0.471$).

³Practice effects were treated as fixed effects as opposed to random given limited variability in the data set.

TABLE 4 Correlations between mobile cognitive tests and in-lab neuropsychological performance in whole sample ($N = 94$).

Mobile Cognitive Tests (Raw Scores)	Demographic Characteristics				Lab Administered Neuropsychological Tests				
	Age	Sex	Race	Education	WRAT-4	HVLT-Immediate Recall	BVMT-Immediate Recall	Letter Number Span	D-KEFS Color-Word Interference Test (time)
VLMT 6 words (% Correct)	-0.27*	0.25*	0.11	0.04	0.07	0.12	0.27**	0.23*	-0.29*
VLMT 12 words (% Correct)	-0.17	0.09	0.11	0.04	-0.03	0.13	0.09	0.07	-0.17
VLMT 18 words (% Correct)	-0.12	0.24*	0.02	0.04	-0.04	0.33**	0.17	0.03	-0.020
VLMT Overall Mean (all trials)	-0.01	0.37**	0.01	0.08	0.17	0.26**	0.27**	0.10	-0.29*
Memory Matrix (Total Score)	-0.43**	0.09	0.11	0.21*	0.04	0.20	0.17	0.38**	-0.26*
Color Trick: Meaning-to-Meaning (Total Score)	-0.12	0.24*	0.13	0.28**	0.30**	0.28**	0.32**	0.24*	-0.33**
Color Trick: Meaning-to-Color (Total Score)	-0.05	0.18	0.03	-0.25*	0.22*	0.21*	0.29**	0.18	-0.19
Color Trick: Yes-No Mechanic (Total Score)	-0.04	0.23*	0.07	0.33**	0.28**	0.21*	0.19	0.20	-0.18

Note. * $p < 0.05$; ** $p < 0.01$.

3.5.2 Psychometric evidence for the Memory Matrix task

The ICC for the average Memory Matrix score across trials was 0.07, 95% CI [0.03, 0.11], indicating that the majority of the variance on this measure can be attributed to within-person differences in performance across trials. Generalizability theory analyses further showed that the reliability of stable between-person individual differences was 0.97. The reliability of within-person change was also satisfactory, with a value of 0.72.

To assess convergent validity, we looked at associations between the Letter-Number Span and performance on Memory Matrix. Memory Matrix scores were positively and significantly correlated with Letter-Number Span ($r = 0.38$, $p < 0.001$). Relationships with demographics and the other lab-administered tests are presented in **Table 3**.

Although we did not find any evidence of a ceiling effect for Memory Matrix, we nonetheless decided to modify our analyses for the practice effects to account for the possibility of right-hand censoring in the data. Because participants' average scores on these EMCTs tended to be close to the maximum number correct, we wanted to ensure that the growth-curve analyses could accurately capture systematic changes in performance across the EMA period in spite of any measurement limitations. As such, these analyses use Mplus v. 8.4 to estimate what the scores would be if there was not an upper limit (e.g., scores can be greater than 9).

Averaging across trials, participants were estimated to get 8.43 items correct on average out of 10 ($SE = 0.12$, $p < 0.001$). Moreover, participants showed systematic change in the number of Memory Matrix items they got correct across the EMA period (on average, participants' total change = 1.75, $SE = 0.20$, $p < 0.001$). However, MCI participants did not significantly differ from NCs on either the intercepts ($b = -0.16$, $SE = 0.13$, $p = 0.22$) or the slopes ($b = -0.11$, $SE = 0.20$, $p = 0.577$). In addition, although the data suggest evidence of a practice effect, closer inspection of participants' trajectories *via* spaghetti plots suggests that participants' performance on the Memory Matrix ebbs and flows throughout the EMA period. Specifically, there appears to be a slight decrease in performance from days 1 to 13, then a marked improvement in performance from days 13 to 21, and then a slight decrease in performance from days 21 to 30.

3.5.3 Psychometric evidence for the Color Trick task

We computed ICCs for participants' accuracy on each version of the Color Trick task: Meaning-to-Meaning, ICC = 0.13, 95% CI [0.07, 0.18]; Meaning-to-Color, ICC = 0.17, 95% CI [0.10, 0.23]; and Yes-No Mechanic, ICC = 0.23, 95% CI [0.15, 0.30]), indicating that the majority of the variance on these measures can be attributed to within-person differences in performance across trials. **Table 3** presents associations between the Color Trick tasks with demographics and lab-

based assessments. As can be seen, the D-KEFS Interference Trial showed a moderate negative correlation with the Meaning-to-Meaning Color Trick task, such that faster performance on the D-KEFS was related to better performance on Meaning-to-Meaning.

We next examined whether there were ceiling effects for participants' accuracy on any of the Color Trick tasks. There was some evidence for ceiling effects, such that 5 (11%) NC and 1 (2%) MCI participants consistently scored 100% for the Meaning-to-Meaning task; 8 (17%) NC and 4 (8%) MCI participants consistently scored 100% for the Meaning-to-Color task; and 5 (11%) NC and 5 (10%) MCI participants consistently scored 100% for the Yes-No Mechanic task. To account for the possibility of right-hand censoring in the data, we adapted our practice effect analyses for the color trick tasks to be consistent with the modifications we made for the memory matrix task analyses.

For Meaning-to-Meaning trials, participants were estimated to get 9.86 items correct on average ($SE = 0.16$, $p < 0.001$). Moreover, participants showed systematic change in the number of items they got correct across the EMA period (on average, participants' total change = 2.19, $SE = 0.32$, $p < 0.001$). Although MCI participants ($M = 9.51$) significantly differed from NCs ($M = 10.21$) on their average number of items correct across the EMA period ($b = -0.35$, $SE = 0.14$, $p = 0.011$), participants' systematic change in the number of items that they got correct across the EMA period was not significantly related to MCI status ($b = -0.12$, $SE = 0.30$, $p = 0.677$).

For Meaning-to-Color trials, participants were estimated to get 11.01 items correct on average ($SE = 0.23$, $p < 0.001$). Moreover, participants showed systematic change in the number of items they got correct across the EMA period (on average, participants' total change = 1.75, $SE = 0.42$, $p < 0.001$). Similar to performance on Meaning-to-Meaning trials, MCI participants ($M = 10.62$) significantly differed from NCs ($M = 11.40$) on their average number of items correct across the EMA period ($b = -0.39$, $SE = 0.16$, $p = 0.015$). In addition, participants' systematic change in the number of items correct across the EMA period was not significantly related to MCI status ($b = 0.53$, $SE = 0.37$, $p = 0.154$).

Lastly, for Yes-No Mechanic trials, participants were estimated to get 11.05 items correct on average ($SE = 0.21$, $p < 0.001$). Participants also showed systematic change in the number of items they got correct across the EMA period (on average, participants' total change = 0.91, $SE = 0.43$, $p = 0.035$). MCI participants did not significantly differ from NC on either the intercepts ($b = -0.29$, $SE = 0.15$, $p = 0.06$) or the slopes ($b = -0.04$, $SE = 0.36$, $p = 0.902$).

4 Discussion

This study evaluated the feasibility and validity of three mobile cognitive tests among persons with and without MCI.

Adherence to this 30-day, fully remote, ecological momentary cognitive testing protocol was very good, with 86% of assigned EMA sessions completed and 84–85% of mobile cognitive testing sessions completed. In this sample of cognitively normal and cognitively impaired older adults, adherence did not differ by MCI status. Further, these findings indicate adherence does not differ by demographic characteristics. Participants who had higher adherence answered more surveys when home and alone compared to people with lower adherence.

We found mixed findings of a fatigue effect at the level of the individual tests, such that there was no evidence of a fatigue effect for the VLMT or Color Trick tests, but participants were more likely to miss Memory Matrix tests over the course of the 30-day protocol (with no difference by NC vs MCI). In another study using the VLMT and Memory Matrix test (14-day protocol in participants with bipolar disorder and control participants) we found an overall fatigue effect for the EMCT protocol, such that participants were more likely to miss a test as study day increased (no differences by diagnostic status), but we did not examine fatigue effects at the level of the individual test (30). Of note, the prior study had a more intensive protocol than the current study, with participants pinged to complete 2–3 mobile cognitive tests three times daily for 14-days. When designing EMCT protocols there is always a frequency and duration trade-off when considering participant burden and capturing outcomes of interest. Our prior work has shown that a 14-day period is sufficient to capture cognition and mood data across various contexts (e.g., 31–35), and other groups have demonstrated strong feasibility and psychometric properties for measuring cognition in as few as 7–8 days (e.g., 14, 16). In general, the 30-day EMCT protocol in this study was largely well tolerated and provides further support for the feasibility of remote, smartphone-based cognitive testing among older adults. Participants had higher rates of adherence than has been reported with other digital health apps (36), which is likely due to a combination of factors including incentives for completing each testing session, brief, gamified tests that varied in difficulty, establishment of good rapport with the study team, and a time-limited engagement with the app.

The psychometric properties of the tasks in this sample were generally good. The reliability of stable between-person individual differences for the VLMT and Memory Matrix were very high, indicating that participants' averaged scores on each mobile cognitive test across the EMA period can reliably assess differences between participants' average levels of the variables. In addition, although the reliability of within-person change (i.e., the consistency in the degree of systematic within-person change across multiple items over time) for Memory Matrix was adequate, the corresponding reliability estimate for the VLMT was not.

Of note, the reliability of within-person change would likely increase if there were more trials, but this would also increase participant burden. As hypothesized, the VLMT overall percentage correct score had an overall moderate positive correlation with the HVLMT and BVMT, demonstrating convergent validity. Further, MCI participants recognized significantly fewer words on this task than CN participants. The trajectories of word recognition did not differ by group status across the 30-day study period, but rather, on average, the participants with MCI remembered fewer words overall. In the whole sample, females performed significantly better than males on both the VLMT and HVLMT, which is consistent with the female verbal memory advantage highlighted in the Alzheimer's disease literature (e.g., 37), and further supports utility of the VLMT in people with MCI.

Also consistent with our hypotheses, Memory Matrix had a moderately positive correlation with Letter Number Span. Group differences in Memory Matrix performance were not found, although the data did demonstrate variability in performance on this task over the 30-day study period, and future work is needed to examine whether context (e.g., home vs. away from home; alone vs. with others; time of day effects) affected performance on this task. Lastly, data from the Meaning-to-Meaning condition of the Color Trick task was related to faster performance on the D-KEFS Interference Trial. The other two Color Trick conditions were not significantly related to D-KEFS performance. For the Meaning-to-Meaning and Meaning-to-Color trials, MCI participants performed significantly worse than NCs. There was some evidence for ceiling effects, especially among the NC participants, for all versions for Color Trick, and future development of this task, such as increasing the number of trials at each administration or increasing difficulty of the task, may be beneficial if this task is to be adopted in a cognitively normal sample. It is worth noting that traditional neuropsychological tests, albeit used as the "gold standard" comparison for mobile cognitive tests in this study, are limited in that they only provide a snapshot of cognitive abilities at one time point. We would not expect a high correlation between once-administered tests and averaged mobile cognitive testing performance. Additional research is needed to examine whether one testing method is superior to the other when examining clinical outcomes such as disease progression, medication effects, reversion rates, and associations with pathology.

This study is not without limitations. Our sample was largely White and highly educated, which may limit generalizability. There were significantly more women in the cognitively normal group compared to the MCI group, which could have an effect on our findings, especially given the female advantage to verbal memory. Future work is needed with larger and more representative samples to determine

whether these tests would be appropriate to detect differences based on cognitive status in randomized controlled trials. Additionally, data were collected during the COVID-19 pandemic, and we did not measure how pandemic-related factors may have influenced performance on these tasks. Another limitation that applies to all ambulatory mobile cognitive testing is that it is difficult to identify suspected cheating, such as whether the participant or someone else took the tests. Relatedly, it is difficult to assess effort on mobile cognitive tests. However, aggregating mobile cognitive test scores can reduce error associated with instances of low effort, as evidenced by the construct validity findings of our mobile cognitive tests with lab-based tests. We did observe evidence of ceiling effects on the VLMT 6-item list and the Color-Trick task in the whole sample, and these trials could possibly be adapted to be made more difficult or used as performance-validity tests in future EMCT protocols. A final limitation is that while we were able to examine differences by smartphone make (iOS vs. Android), we did not have a sufficient sample size to examine differences by smartphone model or OS version, service providers, connectivity, and screen size, all of which may impact response times. Touch sensitivity and latency can differ by up to 100 ms between difference devices, especially between newer and older devices (38, 39). In this study none of the mobile cognitive test outcomes were based on speed. In future work examining timing of responses, these smartphone differences should be examined.

In conclusion, our data add to the extant literature on self-administered mobile cognitive testing in older adults, and is one of the first studies examining an EMCT protocol in people with MCI. The tests are automatically scored, integrated with EMA surveys, and available on iOS and Android operating systems for ease of use by other investigators. Adherence to the EMCTs was high, and the psychometric data are promising. Thus, the three mobile cognitive tests in this study, and particularly the VLMT, may serve as useful tools in future clinical trials with cognition as an endpoint, especially in persons with increased risk for Alzheimer's disease such as those with MCI.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary Files](#), further inquiries can be directed to the corresponding author/s.

Ethics statement

The studies involving human participants were reviewed and approved by UCSD IRB, UTD IRB, UM IRB. The

patients/participants provided their written informed consent to participate in this study.

Author contributions

AEP, RCM, RAA, CAD, and PDH contributed to the conception, design, and obtaining funding for this study. RAA and AEP performed the statistical analysis. RCM wrote the manuscript. RAA, MTR, and LMC wrote sections of the manuscript. AEP, CAD, and PDH provided critical edits. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by grant R01 MH112620-03S1 from the National Institute of Mental Health. LMC was supported by grant F31 AG067869 from the National Institute on Aging.

Conflict of interest

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 Alkermes, Bio Excel, Boehringer Ingelheim, Karuna Pharma, Merck Pharma, Minerva Pharma, and Sunovion (DSP) Pharma in the past year. He receives royalties from the Brief Assessment of Cognition in Schizophrenia (Owned by WCG Verasci, Inc. and contained in the MCCB). He is Chief Scientific Officer of i-Function, Inc. and Scientific Consultant to EMA Wellness, Inc. No other authors report conflicts.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fdgth.2022.946685/full#supplementary-material>.

References

- Lancaster C, Koychev I, Blane J, Chinner A, Wolters L, Hinds C. Evaluating the feasibility of frequent cognitive assessment using the mezurio smartphone app: observational and interview study in adults with elevated dementia risk. *JMIR Mhealth Uhealth*. (2020) 8:e16142. doi: 10.2196/16142
- 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:e1562. doi: 10.1002/mpr.1562
- Sliwinski MJ, Mogle JA, Hyun J, Munoz E, Smyth JM, Lipton RB. Reliability and validity of ambulatory cognitive assessments. *Assessment*. (2016) 25(1):14–30. doi: 10.1177/1073191116643164
- 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:e14328. doi: 10.2196/14328
- Alzheimer's Association. 2017 Alzheimer's disease facts and figures. *Alzheimer's Dementia*. (2017) 13:325–73. doi: 10.1016/j.jalz.2017.02.001
- Aging NIO. NIA-Funded Active Alzheimer's and Related Dementias Clinical Trials and Studies. U.S. Department of Health & Human Services (2022). Available at: [https://www.nia.nih.gov/research/ongoing-AD-trials#:~:text=The%20National%20Institute%20on%20Aging,dementias%20\(AD%2FADRD\)](https://www.nia.nih.gov/research/ongoing-AD-trials#:~:text=The%20National%20Institute%20on%20Aging,dementias%20(AD%2FADRD) (Accessed May 4, 2022).
- Harvey PD, Cosentino S, Curiel R, Goldberg TE, Kaye J, Loewenstein D, et al. Performance-based and observational assessments in clinical trials across the Alzheimer's disease Spectrum. *Innov Clin Neurosci*. (2017) 14:30–9.
- Kueper JK, Speechley M, Montero-Odasso M. The Alzheimer's disease assessment scale–cognitive subscale (ADAS-cog): modifications and responsiveness in pre-dementia populations. a narrative review. *J Alzheimer's Dis*. (2018) 63:423–44. doi: 10.3233/JAD-170991
- Posner H, Curiel R, Edgar C, Hendrix S, Liu E, Loewenstein DA, et al. Outcomes assessment in clinical trials of Alzheimer's disease and its precursors: readiness for short-term and long-term clinical trial needs. *Innov Clin Neurosci*. (2017) 14:22–9.
- Weizenbaum EL, Fulford D, Torous J, Pinsky E, Kolachalama VB, Cronin-Golomb A. Smartphone-based neuropsychological assessment in Parkinson's disease: feasibility, validity, and contextually driven variability in cognition. *J Int Neuropsychol Soc*. (2021) 28(4):401–13. doi: 10.1017/S1355617721000503
- Zlatar ZZ, Campbell LM, Tang B, Gabin S, Heaton A, Higgins M, et al. Daily level association of physical activity and performance on ecological momentary cognitive tests in free-living environments: a Mobile health observational study. *JMIR Mhealth Uhealth*. (2022) 10:e33747. doi: 10.2196/33747
- Center PR. Mobile Fact Sheet: Mobile phone ownership over time (2021). Available at: <https://www.pewresearch.org/internet/fact-sheet/mobile/> (Accessed May 4, 2022).
- Schweitzer P, Husky M, Allard M, Amieva H, Peres K, Foubert-Samier A, et al. Feasibility and validity of mobile cognitive testing in the investigation of age-related cognitive decline. *Int J Methods Psychiatr Res*. (2016) 26(3):e1521. doi: 10.1002/mpr.1521
- Thompson LI, Harrington KD, Roque N, Strenger J, Correia S, Jones RN, et al. A highly feasible, reliable, and fully remote protocol for mobile app-based cognitive assessment in cognitively healthy older adults. *Alzheimers Dement*. (2022) 14:e12283. doi: 10.1002/dad2.12283
- Cerino ES, Katz MJ, Wang C, Qin J, Gao Q, Hyun J, et al. Variability in cognitive performance on Mobile devices is sensitive to mild cognitive impairment: results from the einstein aging study. *Front Digit Health*. (2021) 3:758031. doi: 10.3389/fdgth.2021.758031
- Wilks H, Aschenbrenner AJ, Gordon BA, Balota DA, Fagan AM, Musiek E, et al. Sharper in the morning: cognitive time of day effects revealed with high-frequency smartphone testing. *J Clin Exp Neuropsychol*. (2021) 43:825–37. doi: 10.1080/13803395.2021.2009447

17. Charalambous AP, Pye A, Yeung WK, Leroi I, Neil M, Thodi C, et al. Tools for app- and web-based self-testing of cognitive impairment: systematic search and evaluation. *J Med Internet Res.* (2020) 22:e14551. doi: 10.2196/14551
18. Jak A, Bondi M, Delano-Wood L, Wierenga C, Corey-Bloom J, Salmon D, et al. Quantification of five neuropsychological approaches to defining mild cognitive impairment. *Am J Geriatr Psychiatry.* (2009) 17:368–75. doi: 10.1097/JGP.0b013e31819431d5
19. Neurox (2022). Available at: <https://www.getneurox.com/> (Accessed May 4, 2022).
20. Wilkinson GS, Robertson GJ. *Wide Range Achievement Test Fourth Edition (WRAT-4) Professional Manual* Lutz, FL: Psychological Assessment Resources (2004).
21. Nasreddine Z, Phillips N, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal cognitive assessment. MoCA: A brief screening tool for mild cognitive impairment. *Am J Geriatr Psychiatry.* (2005) 53:695–9. doi: 10.1111/j.1532-5415.2005.53221.x
22. Benedict RH, Schretlen D, Groninger L, Brandt J. Hopkins verbal learning test-revised: normative data and analysis of inter-form and test-retest reliability. *Clin Neuropsychol.* (1998) 12:43–55. doi: 10.1076/clin.12.1.43.1726
23. Benedict RHB, Schretlen D, Groninger L, Dobraski M, Shpritz B. Revision of the brief visuospatial memory test: studies of normal performance, reliability, and validity. *Psychol Assess.* (1996) 8:145–53. doi: 10.1037/1040-3590.8.2.145
24. Ricker JH, Axelrod BN. Analysis of an oral paradigm for the trail making test. *Assessment.* (1994) 1:47–51. doi: 10.1177/1073191194001001007
25. Besser L, Kukull W, Knopman DS, Chui H, Galasko D, Weintraub S, et al. Version 3 of the national Alzheimer's coordinating center's uniform data set. *Alzheimer Dis Assoc Disord.* (2018) 32:351–8. doi: 10.1097/WAD.0000000000000279
26. Gollan T, Weissberger G, Runnqvist E, Montoya R, Cera C. Self-ratings of spoken language dominance: a multilingual naming test (MINT) and preliminary norms for young and aging spanish–English bilinguals. *Biling: Lang Cogn.* (2012) 15:594–615. doi: 10.1017/S1366728911000332
27. Delis DC, Kaplan E, Kramer JH. *Delis-Kaplan Executive Function System (D-KEFS)*. San Antonio: APA PsycTests (2001).
28. Muthén LK, Muthén BO. *Mplus User's Guide: eighth Edition*. Los Angeles, CA: Muthén & Muthén (1998–2017).
29. Cranford JA, Shrout PE, Iida M, Rafaeli E, Yip T, Bolger N. A procedure for evaluating sensitivity to within-person change: can mood measures in diary studies detect change reliably? *Pers Soc Psychol Bull.* (2006) 32:917–29. doi: 10.1177/0146167206287721
30. Moore RC, Parrish EM, Van Patten R, Paolillo EW, Filip T, Bomyea J, et al. *Initial Psychometric Properties of 7 NeuroUX Remote Ecological Momentary Cognitive Tests Among People with Bipolar Disorder*. JMIR mHealth Uhealth (2022). In Press.
31. Bomyea JA, Parrish EM, Paolillo EW, Filip TF, Eyler LT, Depp CA, et al. Relationships between daily mood states and real-time cognitive performance in individuals with bipolar disorder and healthy comparators: a remote ambulatory assessment study. *J Clin Exp Neuropsychol.* (2021) 43:813–24. doi: 10.1080/13803395.2021.1975656
32. Campbell LM, Paolillo EW, Heaton A, Tang B, Depp CA, Granholm E, et al. Daily activities related to Mobile cognitive performance in middle-aged and older adults: an ecological momentary cognitive assessment study. *JMIR Mhealth Uhealth.* (2020) 8:e19579. doi: 10.2196/19579
33. Jones SE, Moore RC, Pinkham AE, Depp CA, Granholm E, Harvey PD. A cross-diagnostic study of adherence to ecological momentary assessment: comparisons across study length and daily survey frequency find that early adherence is a potent predictor of study-long adherence. *Pers Med Psychiatry.* (2021) 29–30:100085. doi: 10.1016/j.pmp.2021.100085
34. Moore RC, Campbell LM, Delgadillo JD, Paolillo EW, Sundermann EE, Holden J, et al. Smartphone-Based measurement of executive function in older adults with and without HIV. *Arch Clin Neuropsychol.* (2020) 35:347–57. doi: 10.1093/arclin/acz084
35. Moore RC, Paolillo EW, Sundermann EE, Campbell LM, Delgadillo J, Heaton A, et al. Validation of the mobile verbal learning test: illustration of its use for age and disease-related cognitive deficits. *Int J Methods Psychiatr Res.* (2021) 30:e1859. doi: 10.1002/mpr.1859
36. Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. *NPJ Digit Med.* (2020) 3:21. doi: 10.1038/s41746-020-0224-8
37. Caldwell JZK, Berg JL, Cummings JL, Banks SJ. Moderating effects of sex on the impact of diagnosis and amyloid positivity on verbal memory and hippocampal volume. *Alzheimers Res Ther.* (2017) 9:72. doi: 10.1186/s13195-017-0300-8
38. Siegal J. Here's why typing on Android phones is harder than typing on an iPhone. BGR (2013a). Available at: <http://bgr.com/2013/09/20/iphone-android-touchscreen-responsiveness/> (Accessed).
39. Siegal J. Study: iPads are the most responsive tablets in the world (2013b). Available at: <http://bgr.com/2013/10/09/tablet-touch-screen-responsiveness/> (Accessed).