Unsupervised high-frequency smartphone-based cognitive assessments are reliable, valid, and feasible in older adults at risk for Alzheimer disease

Jessica Nicosia[1], **Andrew J. Aschenbrenner**[1], **David A. Balota**[2], **Martin J. Sliwinski**[3], **Marisol Tahan**[1], **Sarah Adams**[1], **Sarah S. Stout**[1], **Hannah Wilks**[1], **Brian A. Gordon**[2,5], **Tammie L.S. Benzinger**[5], **Anne M. Fagan**[1], **Chengjie Xiong**[1,4], **Randall J. Bateman**[1], **John C. Morris**[1], **Jason Hassenstab**[1,2]

[1]Charles F. and Joanne Knight Alzheimer Disease Research Center, Department of Neurology, Washington University School of Medicine, St. Louis, MO, USA

[2]Department of Psychological & Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA

[3]Department of Human Development and Family Studies, The Pennsylvania State University, University Park, PA, USA

[4]Division of Biostatistics, Washington University School of Medicine, St. Louis, MO, USA

[5]Department of Radiology, Washington University School of Medicine, St. Louis, MO, USA

Abstract

Objective: Smartphones have the potential for capturing subtle changes in cognition that characterize preclinical Alzheimer disease (AD) in older adults. The Ambulatory Research in Cognition (ARC) smartphone application is based on principles from ecological momentary assessment (EMA) and administers brief tests of associative memory, processing speed, and working memory up to 4 times per day over 7 consecutive days. ARC was designed to be administered unsupervised using participants' personal devices in their everyday environments.

Methods: We evaluated the reliability and validity of ARC in a sample of 268 cognitively normal older adults (ages 65–97) and 22 individuals with very mild dementia (ages 61–88). Participants completed at least one 7-day cycle of ARC testing and conventional cognitive assessments; most also completed cerebrospinal fluid, amyloid and tau PET, and structural MRI studies.

Results: First, ARC tasks were reliable as between-person reliability across the 7-day cycle and test-retest reliabilities at 6-month and 1-year follow-ups all exceeded 0.85. Second, ARC demonstrated construct validity as evidenced by correlations with conventional cognitive measures

Corresponding Author: Jason Hassenstab, PhD, Associate Professor, Departments of Neurology and Psychological & Brain Sciences, Washington University in St. Louis, 4488 Forest Park Avenue, Suite 320, St. Louis, MO 63108, +1 314 747 4032, hassenstabj@wustl.edu.

(r = 0.53 between composite scores). Third, ARC measures correlated with AD biomarker burden at baseline to a similar degree as conventional cognitive measures. Finally, the intensive 7-day cycle indicated that ARC was feasible (86.50% approached chose to enroll), well tolerated (80.42% adherence, 4.83% dropout), and was rated favorably by older adult participants.

Conclusions: Overall, the results suggest that ARC is reliable and valid and represents a feasible tool for assessing cognitive changes associated with the earliest stages of AD.

Keywords

digital biomarkers; mobile testing; preclinical Alzheimer disease; ecological momentary assessment

Introduction

There have been remarkable developments in fluid and neuroimaging biomarkers that track the progression of Alzheimer disease (AD). AD biomarkers can identify pathological changes in amyloid and tau that occur well before symptom onset (Barthélemy et al., 2020; Bateman et al., 2012; Price et al., 2009; Sperling et al., 2011). Despite these developments, advances in the measurement of cognitive decline—the essence of the disease phenotype —have lagged behind. Secondary prevention trials targeting abnormal biomarker levels in preclinical (pre-symptomatic) AD are determined to be successful if they stop or slow cognitive decline (Edgar et al., 2019; Food and Drug Administration, 2018). Because the declines in cognition that occur in preclinical AD are subtle, capturing declines, slowing of declines, or improvements requires reliable cognitive tests that are sensitive to AD pathological processes. However, standard cognitive assessment tools used in AD studies include classic neuropsychological tests that were originally designed to detect overt cognitive impairments or measure facets of intelligence (Sheehan, 2012; Woodford & George, 2007; Weintraub et al., 2009) and often place heavy burden on participants. This poses a critical hurdle for randomized controlled trials (RCTs) examining therapeutics in preclinical and early-stage symptomatic AD populations. Measures with sub-optimal reliability require larger sample sizes to detect cognitive benefits, particularly when the expected effects are subtle (Dodge et al., 2015).

Advances in smartphone technology have allowed researchers to embed brief cognitive measures into ecological momentary assessments (EMA). EMA methods investigate psychological states and behaviors as they occur in natural environments (Shiffman, Stone & Hufford, 2008; Smyth & Stone, 2003; Sliwinski et al., 2018). EMA is defined by several features: (1) data is collected as participants go about their daily lives; (2) assessments are randomly sampled across various occasions to characterize an individual's average performance on a given variable of interest; and (3) participants perform multiple short assessments to capture behavioral changes over time and across different situations (Sliwinski et al., 2018).

Although traditional laboratory/clinical settings afford precise control over the testing environment, this is not representative of everyday cognitive functioning (Sliwinski et al., 2018). The use of smartphone EMAs in cognitive research can assuage ecological validity

concerns as participants perform assessments as they go about their daily lives. Additionally, repeated assessments can improve upon the reliability of conventional measures because they are not collected in just one testing session that may be influenced by variability in participants' day-to-day stress and mood, amongst other factors (Sliwinski et al., 2018). In individuals with neurodegenerative disorders, cognitive performance can vary with time of day (Wilks et al., 2021), and day-to-day variability can be exaggerated (Matar et al., 2020), further exacerbating the impact of conventional measures' low reliability. With EMA, aggregation across repeated measurements ameliorates effects of within-person variability and improves reliability by estimating average functioning (Shiffman et al., 2008; Sliwinski, 2008; Sliwinski et al, 2018). Although ambulatory cognitive testing is not necessarily a replacement of gold-standard in-person cognitive testing, smartphone EMAs provide snapshots of cognition that may reveal unique patterns that cannot be captured with conventional testing.

Smartphone-based assessments may offer a more practical and logistically plausible solution for large-scale studies and clinical trials of AD. Allowing individuals to participate in research studies unsupervised, in familiar environments, and using their own devices can increase engagement, reduce experimenter effects (e.g. demand characteristics, "white coat" testing effects), bolster sample size and diversity, and make participation more accessible and inclusive for individuals who may otherwise be unable to come into the laboratory or clinic. Indeed, interest in smartphone studies is growing and several studies have demonstrated the feasibility and validity of smartphone-based assessments for use in older adults and individuals with preclinical AD (Güsten et al., 2021; Hassenstab et al., 2020; Lancaster et al., 2020; Mackin et al., 2018; Öhman et al., 2021; Papp et al., 2021; Nicosia et al., 2021; Wilks et al., 2021), as well as the potential for high-frequency in-home monitoring to substantially increase the statistical power of therapeutic trials (Dodge et al., 2015).

The purpose of the present study was to evaluate the reliability, validity, and feasibility of unsupervised, high-frequency cognitive testing using participants' personal smartphones. Tasks assessed associate memory, processing speed, and working memory in older adults and individuals with preclinical and early symptomatic AD. If the Ambulatory Research in Cognition smartphone application (ARC) is a reliable, valid, and feasible measure, ARC should: (1) demonstrate high between-subjects and retest reliability; (2) have construct validity (indexed by correlations with correlations with conventional cognitive measures); (3) demonstrate sensitivity to age and AD-related biomarkers; and (4) be well-tolerated by older adults regardless of technology familiarity.

Methods

Participants:

We recruited participants enrolled in ongoing studies of aging and dementia at the Charles F. and Joanne Knight Alzheimer Disease Research Center (Knight ADRC) at Washington University School of Medicine in St. Louis. ARC was designed to be sensitive to subtle changes in cognition in participants at risk for, or in the earliest stages, of AD, thus enrollment in the ARC study was limited to those with a Clinical Dementia Rating® (CDR®; Morris, 1993) of 0 (cognitively normal) or 0.5 (very mild dementia). In-person

enrollment began in February of 2020 and was halted in March 2020 due to the SARS-CoV-2 (COVID-19) pandemic. Therefore, beginning April 2020, the majority of participants were enrolled remotely. All participants provided informed consent, and all procedures were approved by the Human Research Protections Office at Washington University in St. Louis and the research was conducted in accordance with the Helsinki Declaration.

Clinical Assessment.

Clinical status was determined with the CDR which uses a 5-point scale to characterize six domains of cognitive and functional performance (memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care) that are applicable to AD and other dementias (Morris, 1993). CDR scores are determined through semi-structured interviews with the participant and an informant (i.e., family member or friend). A CDR score of 0 indicates cognitive normality, 0.5 = very mild dementia, 1 = mild dementia, 2 = moderate dementia, and 3 = severe dementia.

Conventional Cognitive Assessments.

Conventional cognitive measures included measures of verbal fluency (Animals, Vegetables, Verbal Fluency), episodic memory (Wechsler Memory Scale Paired Associates Recall, Free and Cued Selective Reminding Test (FCSRT) Free Recall, Craft Story 21 immediate and delayed recall), language (the Multilingual Naming Test; MINT), processing speed (Number Span Forward, Number Symbol Test[1]), and working memory (Number Span Backwards; see Hassenstab et al., 2016 and Weintraub et al., 2018 for additional information). A global composite similar to the Preclinincal Alzheimer's Cognitive Composite (PACC; Donohue et al., 2014; Papp et al., 2017) was created by averaging the standardized scores from FCSRT free recall, Animal naming total score, Craft Story 21 delayed recall, and the total correct score from the Number Symbol test such that higher scores indicated better performance (Weintraub et al., 2009).

Ambulatory Research in Cognition (ARC) Application.

The ARC smartphone application is based on principles from ecological momentary assessment (EMA) and administers brief tests of associative memory, processing speed, and working memory up to 4 times per day over 7 consecutive days. Sampling frequency and duration were chosen based on reliability, validity, and effect size estimates reported in Sliwinski et al. (2018). ARC is programmed to run on major operating system (OS) versions (currently iOS 12.0+ and Android OS 8.0+) on iOS and Android devices. Participants were encouraged to use their personal smartphones as long as minimum technical requirements were met. Individuals interested in participating who did not own a smartphone or whose smartphone did not meet our criteria were supplied a device (either iOS or Android) for the duration of the study. Device exclusion criteria included software issues, limited phone storage, physical damage, battery problems, or poor responsivity. A trained study coordinator (M.T.) provided participants with detailed instructions regarding the ARC

---

[1]A computerized task developed and validated at the Knight ADRC that assesses similar constructs as the Wechsler Digit Symbol Substitution task.

application, and additional guidance on smartphone basics (including device setup and operation) was given to participants who were less familiar with smartphones. Throughout the study, the study coordinator provided extensive support for participants via phone, videoconferencing, email, and text messaging. Participants are reimbursed at a rate of $0.50 per completed assessment session. To incentivize participation consistency, participants receive bonus payments for completing all 4 sessions any given day ($1.00 per occurrence, max of $7.00), completing at least 2 assessments per day for 7 days ($6.00), and completing at least 21 assessments over 7 days ($5.00). The maximum compensation possible for one 7-day assessment visit was $32.00.

ARC assessment notifications were administered pseudorandomly throughout the participant's self-reported awake hours, with at least two hours between each testing session. For example, if a participant reported waking up at 7am and going to bed at 10pm, they would receive four test session notifications between 7am and 10pm, separated by at least two hours (see Figure 1, top). The ARC cognitive tasks, Grids, Prices, and Symbols (see Figure 1, bottom), were administered in a random order during each session.

Grids is a spatial working memory task in which high resolution images of three common objects (key, smartphone, and pen) are displayed on a 5×5 grid, and participants are asked to remember the locations of the items. After encoding the locations of each item, participants perform a distractor task (identify Fs in grid of Es) before moving to the retrieval phase. At retrieval, participants are asked to tap the locations where the items were shown[2]. Participants perform two trials during each test session (lasting approximately 30–40s) and, across sessions, stimuli are placed at random locations to protect against retest effects. Scores reflect a Euclidean distance estimate, agnostic to item, such that a higher score indicates retrieval placement farther away from the encoded locations (i.e., higher score indicates worse performance; Sliwinski et al., 2018).

Prices is an associate memory task with a learning and recognition phase. In the learning phase, participants are shown 10 item-price pairs for 3s per pair and asked to remember the items and their corresponding prices. Items were common shopping items (food and household supplies), and the prices were randomly assigned 3-digit prices containing no repeated digits and no more than two sequential digits. In the recognition phase, participants were presented with two prices and asked to choose which was shown with the item during the learning phase. The price choices were separated by at least $3.00 to avoid ceiling and floor effects (Hassenstab et al., 2018). To protect against retest and interference effects, 40 items, chosen without replacement, are never repeated within the same day, and item-price pairs are never re-presented over the 28 sessions. Trials last approximately 60s and scores reflect the proportion of recognition trial errors such that higher scores indicate worse performance.

---

[2]Two versions of the Grids task are included in the present analyses which differed slightly in their retrieval phase instructions. In the original version, participants were asked to tap the locations of the items from encoding. In the new version, participants are shown the items from encoding one at a time and asked to tap the location of that item from encoding. We used a scoring procedure that was agnostic to item such that scores reflect the shortest Euclidian distance between participants' taps at retrieval and the encoded locations regardless of which item they were placing. Nevertheless, to test whether participants' scores differed across versions, several *t*-tests were run to determine if this change in task administration did not dramatically affect participants' performance. Participants' scores for the old and new versions did not significantly differ at visit 1, $p = .07$, or visit 2, $p = .14$.

Symbols is a processing speed measure based on a task used by Sliwinski et al. (2018). Participants are shown three randomly assigned pairs of abstract shapes and asked to determine as quickly as possible which of two pairs match one of the three target pairs. To protect against retest effects, item pairs are randomly assigned for each session. Participants complete 12 trials during each session, lasting approximately 20–60s (duration varied based on participants' response times). Scores reflect response times (RTs) on correct trials such that higher scores indicate worse performance. An "ARC composite score" was created in two steps. Z-scores for each task were calculated by subtracting raw scores from the cohort's mean score and dividing by the cohort's standard deviation. The z-scores were then averaged together to form the ARC composite score. Similar to the individual measures, a higher ARC composite score indicated worse performance.

Feasibility and Tolerability Measures.

Technology familiarity was assessed with a novel measure described in Nicosia et al. (2021). Briefly, the assessment combined objective measurements of technology knowledge (technology-related icon recognition) and self-reported ratings of (1) the frequency with which they perform certain smartphone tasks; and (2) how difficult it would be for them to perform various technology-related tasks. For the purposes of this study, we report participants' technology icon recognition, average frequency of smartphone task performance, and average difficulty performing technology-related tasks (for more details see Nicosia et al., 2021).

ARC user experience was assessed with a 10-question survey using a 5-point Likert scale to rate aspects of user experience regarding installation, test instructions, frequency of testing, and overall tolerability. Objective measures of feasibility and tolerability included ARC adherence and drop-out rates. Adherence was defined as the number of completed test sessions divided by the total number of assessment sessions (i.e., a participant who completed 21 of 28 sessions would have a 75% adherence rate).

Cerebrospinal Fluid Collection and Processing.

Most participants underwent lumbar puncture (LP) to collect cerebrospinal fluid (CSF) following overnight fasting. Participants at the Knight ADRC undergo LP approximately every three years, however, CSF collection was postponed in March 2020 due to the pandemic, eliminating the possibility of acquiring more recent samples. Therefore, we limited the use of CSF data to those collected within five years of ARC testing (see Table 1; collected on average 2.64 +/− 1.11 years from the first ARC assessment). Twenty to thirty mL of CSF was collected in a 50 mL polypropylene tube via gravity drip using an atraumatic Sprotte 22-gauge spinal needle. CSF was kept on ice and centrifuged at low speed within 2 hours of collection. CSF was then transferred to another 50 mL tube. CSF was aliquoted at 500 μL into polypropylene tubes and stored at −80°C as previously described (Fagan et al., 2006). Prior to analysis, samples were brought to room temperature per manufacturer instructions. Samples were vortexed and transferred to polystyrene cuvettes for analysis. Concentrations of Aβ40, Aβ42, total tau (tTau), and tau phosphorylated at threonine 181 (pTau) were measured by chemiluminescent enzyme immunoassay using a fully automated platform (LUMIPULSE G1200, Fujirebio, Malvern,

PA) according to manufacturer's specifications. A single lot of reagents were used for all samples.

Neuroimaging:

Neuroimaging data was required to be collected within 5 years of ARC (see Table 1; Amyloid PET mean 2.59 +/− 1.04 years, Tau PET mean 2.50 +/− 0.96, and MRI mean 2.55 +/− 1.05 years from the first ARC assessment). Briefly, MRI data were acquired on 3T Siemens scanners and processed using Freesurfer (Fischl, van der Kouwe et al., 2004) to derive regional volumes and thicknesses. Volumes were adjusted for total intracranial volume (ICV) (see Raz et al., 2008) and a summary thickness composite was calculated (Singh et al., 2006).

Amyloid PET imaging was performed with either florbetapir (18F-AV-45) or Pittsburgh Compound B (PiB). Data were processed with an in-house pipeline using regions of interest derived from FreeSurfer (ttps://github.com/ysu001/PUP; Su et al., 2013). A summary standardized uptake value ratios (SUVR) measure was converted to the Centiloid scale (Su et al., 2018 & 2019) in order to combine PiB and florbetapir data. Tau PET imaging with flortaucipir (18F-AV-1451) was summarized using the average SUVRs of the of bilateral entorhinal cortex, amygdala, inferior temporal lobe, and lateral occipital cortex (Mishra et al., 2017). SUVRs used a cerebellar cortex reference and were partial volume corrected.

Statistical Analyses.

Statistical analyses were completed using R (v4.1.0). To characterize the reliability of ARC, descriptive statistics were examined for all ARC and conventional measures. Correlations were used to examine whether ARC captured age-related cognitive declines comparable to conventional cognitive measures. ARC test-retest reliability was assessed based on participants who completed follow-up testing ~6 months ("visit 2"; on average 6.07 +/− 1.23 months between assessments) and ~1 year later ("visit 3"; on average 11.84 +/− 0.84 months between assessments). Pearson correlation coefficients with an r of 0.80 to 0.90 were considered "good" reliability (Price et al., 2015). Intraclass correlations (ICC), which show how strongly units within the same group resemble each other, were computed to examine test-retest reliability and between-person reliability such that ICCs between 0.75 and 0.90 indicate "good" reliability (Bruton et al., 2000). ARC and conventional cognitive measure correlations were used to examine construct validity. Finally, feasibility and tolerability were assessed by examining (1) adherence and drop-out rates; (2) correlations between technology familiarity measures and ARC performance and; (3) descriptive statistics from an ARC user experience survey.

Results

Participant Characteristics.

Of the 316 participants who completed at least one ARC session, 26 were removed due to either low-quality data or unacceptable rates of missing data (>75% missingness) resulting in a sample size of 290 participants (268 CDR 0s and 22 CDR 0.5s) ranging 61 to 97 years of age. As shown in Table 1, all three ARC tasks showed good discrimination between

CDR 0 and CDR 0.5 participants[3]. Additionally, ARC performance, as indexed by the ARC composite score, did not differ as a function of gender, $t(181.46) = 0.63$, $p = 0.53$, or race, $t(28.096) = 1.92$, $p = 0.06$, and was modestly associated with education, $r = -0.18$, $p = 0.01$.

Descriptive Statistics.

Table 2 shows the descriptive statistics for the ARC and conventional cognitive measures as well as adherence and drop-out rates. The t-tests comparing ARC task performance of CDR 0 and 0.5 individuals (significant ts 2.12–3.52) were comparable to comparisons with conventional cognitive measures (significant ts 2.17–4.96). CDR 0s and 0.5s in this sample did not differ on Number Span Forward, Number Span Backward, or the MINT. Adherence and drop-out rates did not differ as a function of CDR status (ts < 0.38).

Between-Person Reliability.

As mentioned above, aggregation of EMA scores across sessions boosts reliability compared to conventional "one-shot" approaches (Shiffman et al., 2008). Unconditional multilevel mixed models using restricted maximum likelihood were employed for each ARC task to compute between-person reliability scores (Raykov & Marcoulides, 2006; Sliwinski et al., 2018). The reliabilities of scores aggregated across ARC sessions were quite high: 0.81 for Prices, 0.90 for Grids, and 0.98 for Symbols (see Table 3). These reliabilities are based on 21 (75%) sessions of ARC assessments, which reflects the average number of sessions participants completed.

Next, we conducted follow-up analyses to determine how many sessions would be required to obtain reliabilities of aggregated scores that ranged from 0.80–0.90. Following Sliwinksi et al. (2018), we fit a series of unconditional multilevel mixed models and calculated reliabilities. These results indicated that 19 sessions (or ~ 5 days) of Prices, 9 sessions (or ~ 2 days) of Grids, and 2 sessions (or ~1 day) of Symbols are required to attain reliabilities greater than 0.80 (see Table 3 and Figure 2).

Test-Retest Reliability.

As of manuscript preparation, a subset of participants also completed testing ~6 months (N = 185) and ~1 year (N = 83) after their initial visit. Figure 3 displays test-retest reliability for the 6-month and 1-year follow-ups for the individual tasks and ARC composite score. ARC demonstrated high test-retest reliability for individual ARC tasks as well as the ARC composite score at both follow-ups (all ICCs > 0.85). Considering retest effects (Table 4), there were small but significant improvements from visit 1 to visit 2 on Prices, Symbols, and the ARC Composite, but not on Grids. There were no practice effects evident between visits 2 and 3, suggesting that practice effects diminish after completion of the first testing cycle. A detailed analysis of practice effects will be considered in future studies.

Construct Validity.

As shown in Figure 4 (right), the ARC composite score was correlated with the global composite score created from the conventional measures ($r = -0.53$; this was also the case

---

[3]See Supplemental Table 1 for information on intraindividual variability for the three ARC tasks.

in the CDR 0 sample, $r = -0.47$), indicating good construct validity. Additionally, Figure 4 (left) displays correlations between ARC and conventional cognitive measures (raw scores), and the top row shows the correlations with age. ARC tasks showed similar correlations with age as the conventional cognitive measures and exhibited convergent validity such that measures were correlated within the same domains. Note that correlations between the conventional and ARC measures are negative because higher scores on the ARC tasks indicate *worse* performance, whereas higher scores on the conventional cognitive measures indicate *better* performance (except for the Trailmaking Test Parts A & B), thus the negative correlations displayed in Figure 4 (left) are in the hypothesized direction. Specifically, the Prices task was correlated with conventional memory measures (WMS Associates Recall: $r = -0.24$, FCSRT free recall: $r = -0.32$, Craft Story immediate recall: $r = -0.22$, Craft Story delayed recall: $r = -0.27$), the Grids task was correlated with all of the conventional cognitive measures (r's $= -0.15$ to -0.36), and the Symbols task was correlated with all the conventional cognitive measures but particularly the fluency tasks and the Number Symbol test (Category Fluency Animals: -0.36, Category Fluency Vegetables: -0.40, Verbal Fluency: -0.36, Number Symbol test: -0.57).

Criterion Validity.

Criterion validity of ARC was examined by comparing ARC and global composite score correlations with AD biomarkers. As shown in Figure 4 (right), the ARC composite score was correlated in the predicted directions with all AD biomarkers. All correlations remained significant after controlling for age, rs > 0.20, ps < 0.02, except for the relationships with the neurodegeneration and tauopathy measures, ps > 0.18. We also examined correlations between the ARC composite score and AD biomarkers with only CDR 0 participants. Correlations in the cognitively normal subsample (CDR 0 individuals) were weaker than in the full sample (see Supplemental Materials Figure 1), as expected, but were consistent with the magnitude of values seen in other studies which have explored such relationships (for example, see Papp et al., 2021 amongst others). Additionally, the correlations were comparable to, though slightly weaker than, correlations between the global composite score and AD biomarkers. Specifically, Fisher's z test indicated that, compared to the global composite score, all correlations with the ARC composite score were not significantly different except for the correlations with CSF pTau:Aβ42 ($z = -1.96$, $p = 0.049$), Hippocampal Volume ($z = -1.99$, $p = 0.045$), and PET Tau ($z = -2.20$, $p = 0.03$), which were only marginally to slightly weaker. There were no significant differences in correlations between AD biomarkers and the two composite scores in the CDR 0 subsample.

Feasibility and Tolerability.

Of the 290 participants included in the present analyses, a subset (N = 220) completed the technology familiarity survey. Figure 5 displays the correlations amongst age, adherence, the technology familiarity measures, and ARC performance. Greater technology-related icon recognition was associated with better performance on Grids ($r = -0.16$) and Symbols ($r = -0.14$), but not on Prices ($r = -0.02$). Self-reported frequency performing smartphone tasks was unrelated to ARC performance, but perceived difficulty performing technology tasks was related to worse performance on all ARC measures (r's 0.17–0.24). Adherence was correlated with performance on all three ARC measures (though only weakly for Prices),

and the ARC composite score, such that participants who completed more sessions tended to perform better on ARC.

A subset of participants (N = 228) also completed a user experience survey after their first ARC visit[4]. As shown in Figure 6, participants reported an overall positive experience with the ARC application, and most reported that they preferred ARC over conventional assessments. Participants reported little difficulty installing the ARC app, were generally unconcerned about privacy, and that completing 2 weeks of ARC testing per year would not be difficult.

Finally, as shown in Tables 1 and 2, adherence rates were quite high at 81% and 79% for CDR 0 and 0.5 participants, respectively. Drop-out rates were low for both groups as well - 4.9% for CDR 0s and 4.5% for CDR 0.5s. The high adherence and low drop-out rates suggest that ARC was well tolerated by older adults, even those with very mild dementia.

Discussion

The present study demonstrates that EMA cognitive assessments conducted on individuals' personal smartphones can be reliable, sensitive to age and AD biomarkers, and are well-tolerated by older adults regardless of technology experience. There were several main findings: First, between-person reliability of the ARC tasks across the 7-day protocol all exceeded 0.85. Second, individual ARC tasks and the ARC composite score showed exceptionally good test-retest reliabilities at 6-month and 1-year follow-ups (ICCs > 0.85). Third, both the individual ARC tasks and the ARC composite score were correlated with conventional measures of the same domain (r's = −0.22 to −0.57). The composite scores from ARC and conventional measures were also highly correlated (r = −0.53). Fourth, the ARC composite score showed similar validity to the global composite in predicting AD biomarkers. Finally, both cognitively normal older adults and individuals with very mild AD successfully participated in the ARC study remotely, without supervision, and had extremely low drop-out rates. Overall, the results of the present study suggest that high-frequency smartphone-based assessments are promising tools for assessing cognition in clinical studies of aging and neurodegenerative diseases.

Although classic neuropsychological tests, such as episodic memory and executive functioning tests, are regarded as the most sensitive to AD pathology, they were not designed for frequent assessment and can have poor reliability (Calamia et al., 2013). Using measures with suboptimal reliability can impact statistical power and necessitate larger sample sizes or increased measurement frequency. Our results suggest that a high-frequency EMA approach to cognitive assessments may help overcome these challenges. When averaged across sessions, all three ARC tests had excellent between-subject reliability (r's > 0.85), consistent with Sliwinski et al. (2018). The results also demonstrated that good between-person reliabilities can be achieved with < 7 days of assessments (averaging across

---

[4]Because participants completed the user experience survey voluntarily and a subset of 62 participants (21.38%) chose not to complete the survey, it is possible that the survey results may be influenced by selection bias. To test this possibility, we examined ARC task performance and adherence as a function of whether participants completed the user experience survey. These analyses indicated that there were no significant differences in either ARC task performance, ps > 0.24, or adherence, p = 0.82.

5 days produced reliabilities > 0.80 for all ARC tasks). The Symbols test achieved excellent reliability in just 3–5 sessions, which is remarkable considering that each session requires ~30–40s to complete. Although conventional cognitive measures would also receive a boost in reliability if averaged across repeated assessments, it is impractical and burdensome to assess participants at a frequency sufficient to overcome suboptimal reliability. Using an EMA smartphone protocol, researchers can efficiently obtain repeated measurements to boost reliability.

Test-retest reliability studies in AD samples have indicated "adequate" to "excellent" reliability (e.g., Benedict et al., 1998; Woods et al., 2006) over intervals ranging from several days to several weeks apart. However, cohort studies are typically conducted annually, and yield lower reliability estimates. Specifically, test-retest correlations for delayed memory tests, a cornerstone of AD clinical trials (Bateman et al., 2017; Donohue et al., 2014; Langbaum et al., 2014; Ritchie et al., 2016), can be particularly unsatisfactory, with reliabilities ranging from 0.50–0.75 (Calamia et al., 2013; Dikmen et al., 1999; Lo et al., 2012). The increased reliability demonstrated by high-frequency assessments like ARC could substantially reduce sample sizes needed in AD prevention RCTs (Dodge et al., 2015).

ARC demonstrated exceptionally high test-retest reliability for the individual ARC tasks and the ARC composite score at 6-month and 1-year follow-ups (all ICCs > 0.85). The Symbols test demonstrated exceptionally high test-retest reliability exceeding its paper and pencil equivalents (i.e., Wechsler Digit Symbol Substitution test and the Symbol-Digit Modalities test which typically have good test-retest reliabilities; Calamia et al., 2013; Pereira et al. 2015). Test-retest reliability for the Prices test, was also good but trailed behind the Symbols and Grids tests. Relatedly, a version of the Prices test demonstrated good validity and reliability in a recent EMA study of older adults (Thompson et al, 2021), but was also rated the most difficult and the least enjoyable of three cognitive tasks, reflecting the challenges of designing repeatable episodic memory measures that are reliable, feasible, and tolerable.

Our results also support the construct and predictive validity of ARC. ARC tasks exhibited convergent validity as evidenced by correlations with conventional cognitive measures (r's −0.22 to −0.57). Similarly, the ARC composite score was correlated with the global composite score ($r = −0.53$). Albeit smaller than anticipated, the correlations observed here were comparable, if not stronger, than correlations observed in other digital assessment studies including the Cambridge Neuropsychological Test Automated Battery (CANTAB; rs 0.14 to 0.39; Dorociac et al., 2021; Gills et al., 2019; Smith et al., 2013). Additionally, the individual ARC tasks and the ARC composite score showed comparable correlations with age as the conventional measures and global composite score. Given well-known associations between age and cognitive performance, these relationships provide evidence that ARC is a valid measure of cognitive aging.

ARC also demonstrated good predictive validity when assessing sensitivity to AD biomarkers. Worse ARC performance was associated with reduced cortical thickness and hippocampal volume (r's = −0.18 and −0.19, respectively), and increased levels of amyloid and tau (as indexed by both PET and CSF measures; r's = 0.11 to 0.29). These relationships were comparable, though smaller in magnitude, to AD biomarker

correlations with conventional measures suggesting that ARC captures biomarker burden similarly to conventional measures. Correlations in the cognitively normal subsample (CDR 0 individuals) were on par with other studies which have examined such relationships (Braak et al., 1991; Papp et al., 2021; Snitz et al., 2020; Van Strien et al., 2009).

Evaluation of feasibility and tolerability of a smartphone application for use in older adults is critical, and especially so for applications like ARC that require unsupervised daily interactions. Overall, adherence was excellent at 80.42%, exceeding that seen in many remote studies (Pratap et al., 2020) and similar to rates observed in other cognitive EMA studies (Sliwinski et al., 2018). A common concern regarding technology use in older adults is that of technology familiarity. Our results demonstrate that greater technology knowledge was associated with better processing speed and visual working memory task performance, but not memory performance. Interestingly, self-reported frequency of smartphone interactions was not related to ARC performance, but those who reported more difficulty interacting with technology tended to perform worse on all ARC measures. However, when the familiarity assessment results were compared to conventional cognitive measures (see Supplemental Materials Figure 2), similar patterns emerged even on non-technology-related measures like story recall, number span, confrontation naming, and verbal fluency, suggesting that difficulty with technology may also reflect, to some extent, overall cognitive ability[5]. Finally, considering the high adherence rates, and the overall favorable ratings from the user experience survey, it appears that with adequate instruction and support, older adults are capable and motivated participants in smartphone studies of cognition.

Limitations and Future Considerations.

The findings of this study should be considered in light of several limitations which may be addressed in future studies. First, although the benefits of EMA smartphone studies are clear, it can be unclear whether participants are fully engaging with the assigned tasks. To address this, participants are asked at the end of each session whether they were interrupted during the session. In the analyses presented here, sessions where participants reported being interrupted were removed. Similarly, many ambulatory assessments are limited when researchers do not collect additional contextual information. Participants were asked a battery of environmental questions at the end of each session, and future studies will investigate the impact of these factors on participants' performance. Second, as noted in the Methods, if an individual did not have a device which met study criteria, they were supplied a device. Since it is possible this could have introduced bias, several follow-up analyses were run to test for differences in age, technology familiarity, and ARC performance/adherence. As shown in Supplementary Materials Table 2, even though individuals who were supplied with a device were slightly older and less familiar with technology, there were no differences in CDR, ARC task performance, adherence, or AD biomarkers. Third, it is important to note that the Prices task lagged behind the Symbols and Grids tasks in terms of participants' performance and the between-subjects reliability (possibly due to the difficulty and task

---

[5]We explored the extent to which "overall cognitive ability," as indexed by the conventional composite score, may be associated with ARC adherence. As shown in Supplemental Materials Figure 3, individuals who performed better on the conventional measures also showed better ARC adherence.

demands). Nevertheless, the Prices task showed good reliability and was correlated with age and conventional memory measures. Finally, Knight ADRC participants consist of highly educated and primarily White older adults motivated to engage in extensive imaging and fluid biomarker studies. Future work is needed to determine the feasibility of ARC in more diverse populations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

Barthélemy NR, Li Y, Joseph-Mathurin N, Gordon BA, Hassenstab J, Benzinger TL, ... & McDade E. (2020). A soluble phosphorylated tau signature links tau, amyloid and the evolution of stages of dominantly inherited Alzheimer's disease. Nature medicine, 26(3), 398–407.

Bateman RJ, Benzinger TL, Berry S, Clifford DB, Duggan C, Fagan AM, Fanning K, Farlow MR, Hassenstab J, McDade EM, Mills S, Paumier K, Quintana M, Salloway SP, Santacruz A, Schneider LS, Wang G, & Xiong C (2017). The DIAN-TU Next Generation Alzheimer's prevention trial: Adaptive design and disease progression model. Alzheimer's & Dementia, 13(1), 8–19.

Benedict RHB, Schretlen D, Groninger L, & Brandt J (1998). Hopkins Verbal Learning Test? Revised: Normative Data and Analysis of Inter-Form and Test-Retest Reliability. The Clinical Neuropsychologist (Neuropsychology, Development and Cognition: Section D), 12(1), 43–55.

Braak H, & Braak E (1991). Neuropathological stageing of Alzheimer-related changes. Acta neuropathologica, 82(4), 239–259. [PubMed: 1759558]

Bruton A, Conway JH, & Holgate ST (2000). Reliability: what is it, and how is it measured?. Physiotherapy, 86(2), 94–99.

Calamia M, Markon K, & Tranel D (2012). Scoring Higher the Second Time Around: Meta-Analyses of Practice Effects in Neuropsychological Assessment. The Clinical Neuropsychologist (Neuropsychology, Development and Cognition: Section D), 26(4), 543–570.

Calamia M, Markon K, & Tranel D (2013). The Robust Reliability of Neuropsychological Measures: Meta-Analyses of Test–Retest Correlations. The Clinical Neuropsychologist (Neuropsychology, Development and Cognition: Section D), 27(7), 1077–1105.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, ... & Killiany RJ (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage, 31(3), 968–980. [PubMed: 16530430]

Dikmen SS, Heaton RK, Grant I, & Temkin NR (1999). Test–retest reliability and practice effects of expanded Halstead–Reitan Neuropsychological Test Battery. Journal of the International Neuropsychological Society, 5(4), 346–356. [PubMed: 10349297]

Dodge HH, Zhu J, Mattek NC, Austin D, Kornfeld J, & Kaye JA (2015). Use of high-frequency in-home monitoring data may reduce sample sizes needed in clinical trials. PLoS One, 10(9), e0138095. [PubMed: 26379170]

Donohue MC, Sperling RA, Salmon DP, Rentz DM, Raman R, Thomas RG, Weiner M, & Aisen PS (2014). The Preclinical Alzheimer Cognitive Composite: Measuring Amyloid-Related Decline. JAMA Neurology, 71(8), 961–970. [PubMed: 24886908]

Dorociak KE, Mattek N, Lee J, Leese MI, Bouranis N, Imtiaz D, ... & Hughes AM. (2021). The Survey for Memory, Attention, and Reaction Time (SMART): Development and Validation of a Brief Web-Based Measure of Cognition for Older Adults. Gerontology, 1–13.

Edgar CJ, Vradenburg G, & Hassenstab J (2019). The 2018 Revised FDA guidance for early Alzheimer's disease: establishing the meaningfulness of treatment effects. The journal of prevention of Alzheimer's disease, 6(4), 223–227.

Fagan AM, Mintun MA, Mach RH, Lee S-Y, Dence CS, Shah AR, LaRossa GN, Spinner ML, Klunk WE, Mathis CA, DeKosky ST, Morris JC, & Holtzman DM (2006). Inverse relation between in vivo amyloid imaging load and cerebrospinal fluid Aβ 42in humans. Annals of Neurology, 59(3), 512–519. [PubMed: 16372280]

Fischl B, & Dale AM (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proceedings of the National Academy of Sciences, 97(20), 11050–11055.

Fischl B, Van Der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, ... & Dale AM. (2004). Automatically parcellating the human cerebral cortex. Cerebral cortex, 14(1), 11–22. [PubMed: 14654453]

Food and Drug Administration. (2018). Early Alzheimer's Disease: Developing Drugs for Treatment: Guidance for Industry. Food and Drug Administration.

Gills JL, Glenn JM, Madero EN, Bott NT, & Gray M (2019). Validation of a digitally delivered visual paired comparison task: reliability and convergent validity with established cognitive tests. GeroScience, 41(4), 441–454. [PubMed: 31463649]

Güsten J, Ziegler G, Düzel E, & Berron D (2021). Age impairs mnemonic discrimination of objects more than scenes: A web-based, large-scale approach across the lifespan. Cortex, 137, 138–148. [PubMed: 33611227]

Hassenstab J, Aschenbrenner AJ, Balota DA, McDade E, Lim YY, Fagan AM, ... & Dominantly Inherited Alzheimer Network. (2020). Remote cognitive assessment approaches in the Dominantly Inherited Alzheimer Network (DIAN) Using digital technology to drive clinical innovation in brain-behavior relationships: A new era in neuropsychology. Alzheimer's & Dementia, 16, e038144.

Hassenstab J, Chasse R, Grabow P, Benzinger TLS, Fagan AM, Xiong C, Jasielec M, Grant E, & Morris JC (2016). Certified normal: Alzheimer's disease biomarkers and normative estimates of cognitive functioning. Neurobiology of Aging, 43, 23–33. [PubMed: 27255812]

Hassenstab J, Nicosia J, LaRose M, Aschenbrenner AJ, Gordon BA, Benzinger TL, ... & Morris JC. (2021). Is comprehensiveness critical? Comparing short and long format cognitive assessments in preclinical Alzheimer disease. Alzheimer's Research & Therapy, 13(1), 1–14.

Lancaster C, Koychev I, Blane J, Chinner A, Chatham C, Taylor K, & Hinds C (2020). Gallery Game: Smartphone-based assessment of long-term memory in adults at risk of Alzheimer's disease. Journal of clinical and experimental neuropsychology, 42(4), 329–343. [PubMed: 31973659]

Langbaum JB, Hendrix SB, Ayutyanont N, Chen K, Fleisher AS, Shah RC, Barnes LL, Bennett DA, Tariot PN, & Reiman EM (2014). An empirically derived composite cognitive test score with improved power to track and evaluate treatments for preclinical Alzheimer's disease. Alzheimer's & Dementia, 10(6), 666–674.

Lo AHY, Humphreys M, Byrne GJ, & Pachana NA (2012). Test-retest reliability and practice effects of the Wechsler Memory Scale-III. Journal of Neuropsychology, 6(2), 212–231. [PubMed: 22257421]

Mackin RS, Insel PS, Truran D, Finley S, Flenniken D, Nosheny R, ... & Weiner MW. (2018). Unsupervised online neuropsychological test performance for individuals with mild cognitive impairment and dementia: results from the Brain Health Registry. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 10, 573–582.

Matar E, Shine JM, Halliday GM, & Lewis SJ (2020). Cognitive fluctuations in Lewy body dementia: towards a pathophysiological framework. Brain, 143(1), 31–46. [PubMed: 31612904]

Mishra S, Gordon BA, Su Y, Christensen J, Friedrichsen K, Jackson K, ... & Benzinger TL. (2017). AV-1451 PET imaging of tau pathology in preclinical Alzheimer disease: defining a summary measure. Neuroimage, 161, 171–178. [PubMed: 28756238]

Morris JC (1991). The clinical dementia rating (cdr): Current version and. Young, 41, 1588–1592.

Nicosia J, Aschenbrenner AJ, Adams S, Tahan M, Stout SH, Wilks H, … Hassenstab J (2021, December 1). Bridging the technological divide: Stigmas and challenges with technology in clinical studies of older adults. 10.31234/osf.io/yeufs

Öhman F, Hassenstab J, Berron D, Schöll M, & Papp KV (2021). Current advances in digital cognitive assessment for preclinical Alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 13(1), e12217.

Papp KV, Rentz DM, Orlovsky I, Sperling RA, & Mormino EC (2017). Optimizing the preclinical Alzheimer's cognitive composite with semantic processing: The PACC5. Alzheimer's & Dementia: Translational Research & Clinical Interventions, 3(4), 668–677.

Papp KV, Samaroo A, Chou HC, Buckley R, Schneider OR, Hsieh S, ... & Amariglio RE (2021). Unsupervised mobile cognitive testing for use in preclinical Alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 13(1), e12243.

Pereira DR, Costa P, & Cerqueira JJ (2015). Repeated assessment and practice effects of the written symbol digit modalities test using a short inter-test interval. Archives of Clinical Neuropsychology, 30(5), 424–434. [PubMed: 25994156]

Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, ... & Omberg L. (2020). Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ digital medicine, 3(1), 1–10. [PubMed: 31934645]

Price PC, Jhangiani RS, & Chiang ICA (2015). Reliability and validity of measurement. Research methods in psychology.

Price JL, McKeel DW Jr, Buckles VD, Roe CM, Xiong C, Grundman M, ... & Morris JC. (2009). Neuropathology of nondemented aging: presumptive evidence for preclinical Alzheimer disease. Neurobiology of aging, 30(7), 1026–1036. [PubMed: 19376612]

Raykov T, & Marcoulides GA (2006). On multilevel model reliability estimation from the perspective of structural equation modeling. Structural Equation Modeling, 13(1), 130–141.

Raz N, Lindenberger U, Ghisletta P, Rodrigue KM, Kennedy KM, & Acker JD (2008). Neuroanatomical correlates of fluid intelligence in healthy adults and persons with vascular risk factors. Cerebral cortex, 18(3), 718–726. [PubMed: 17615248]

Ritchie K, Ropacki M, Albala B, Harrison J, Kaye J, Kramer J, ... & Ritchie CW. (2017). Recommended cognitive outcomes in preclinical Alzheimer's disease: Consensus statement from the European Prevention of Alzheimer's Dementia project. Alzheimer's & Dementia, 13(2), 186–195.

Sheehan B (2012). Assessment scales in dementia. Therapeutic advances in neurological disorders, 5(6), 349–358. [PubMed: 23139705]

Shiffman S, Stone AA, & Hufford MR (2008). Ecological momentary assessment. Annu. Rev. Clin. Psychol, 4, 1–32. [PubMed: 18509902]

Singh V, Chertkow H, Lerch JP, Evans AC, Dorr AE, & Kabani NJ (2006). Spatial patterns of cortical thinning in mild cognitive impairment and Alzheimer's disease. Brain, 129(11), 2885–2893. [PubMed: 17008332]

Sliwinski MJ (2008). Measurement-burst designs for social health research. Social and Personality Psychology Compass, 2(1), 245–261.

Sliwinski MJ, Mogle JA, Hyun J, Munoz E, Smyth JM, & Lipton RB (2018). Reliability and validity of ambulatory cognitive assessments. Assessment, 25(1), 14–30. [PubMed: 27084835]

Smith PJ, Need AC, Cirulli ET, Chiba-Falek O, & Attix DK (2013). A comparison of the Cambridge Automated Neuropsychological Test Battery (CANTAB) with "traditional" neuropsychological testing instruments. Journal of clinical and experimental neuropsychology, 35(3), 319–328. [PubMed: 23444947]

Smyth JM, & Stone AA (2003). Ecological momentary assessment research in behavioral medicine. Journal of Happiness studies, 4(1), 35–52.

Snitz BE, Tudorascu DL, Yu Z, Campbell E, Lopresti BJ, Laymon CM, ... & Cohen AD. (2020). Associations between NIH Toolbox Cognition Battery and in vivo brain amyloid and tau pathology in non-demented older adults. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 12(1), e12018.

Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, ... & Phelps CH. (2011). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & dementia, 7(3), 280–292.

Su Y, D'Angelo GM, Vlassenko AG, Zhou G, Snyder AZ, Marcus DS, ... & Benzinger TL. (2013). Quantitative analysis of PiB-PET with freesurfer ROIs. PloS one, 8(11), e73377. [PubMed: 24223109]

Su Y, Flores S, Hornbeck RC, Speidel B, Vlassenko AG, Gordon BA, ... & Benzinger TL. (2018). Utilizing the Centiloid scale in cross-sectional and longitudinal PiB PET studies. NeuroImage: Clinical, 19, 406–416. [PubMed: 30035025]

Su Y, Flores S, Wang G, Hornbeck RC, Speidel B, Joseph-Mathurin N, ... & Benzinger TL (2019). Comparison of Pittsburgh compound B and florbetapir in cross-sectional and longitudinal studies. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 11(1), 180–190.

Thompson L, Harrington K, Roque N, Strenger J, Correia S, Jones R, Salloway S, & Sliwinski M (in press). A highly feasible, reliable, and fully remote protocol for mobile app-based cognitive assessment in cognitively healthy older adults. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring.

Van Strien NM, Cappaert NLM, & Witter MP (2009). The anatomy of memory: an interactive overview of the parahippocampal–hippocampal network. Nature reviews neuroscience, 10(4), 272–282. [PubMed: 19300446]

Weintraub S, Besser L, Dodge HH, Teylan M, Ferris S, Goldstein FC, ... & Morris JC. (2018). Version 3 of the Alzheimer Disease Centers' neuropsychological test battery in the Uniform Data Set (UDS). Alzheimer disease and associated disorders, 32(1), 10. [PubMed: 29240561]

Weintraub S, Salmon D, Mercaldo N, Ferris S, Graff-Radford NR, Chui H, ... & Morris JC. (2009). The Alzheimer's disease centers' uniform data set (UDS): The neuropsychological test battery. Alzheimer disease and associated disorders, 23(2), 91. [PubMed: 19474567]

Wilks HM, Aschenbrenner AJ, Gordon BA, Balota DA, Fagan AM, Musiek E, Balls-Berry J, Benzinger LS, T.L.S., Cruchaga C, Morris JC, Hassenstab J (2021). Sharper in the Morning: Cognitive Time of Day Effects Revealed with High-Frequency Smartphone Testing. Journal of Clinical and Experimental Neuropsychology. 10.1080/13803395.2021.2009447

Woodford HJ, & George J (2007). Cognitive assessment in the elderly: a review of clinical methods. QJM: An International Journal of Medicine, 100(8), 469–484. [PubMed: 17566006]

Woods S, Delis D, Scott J, Kramer J, & Holdnack J (2006). The California Verbal Learning Test – second edition: Test-retest reliability, practice effects, and reliable change indices for the standard and alternate forms. Archives of Clinical Neuropsychology, 21(5), 413–420. [PubMed: 16843636]
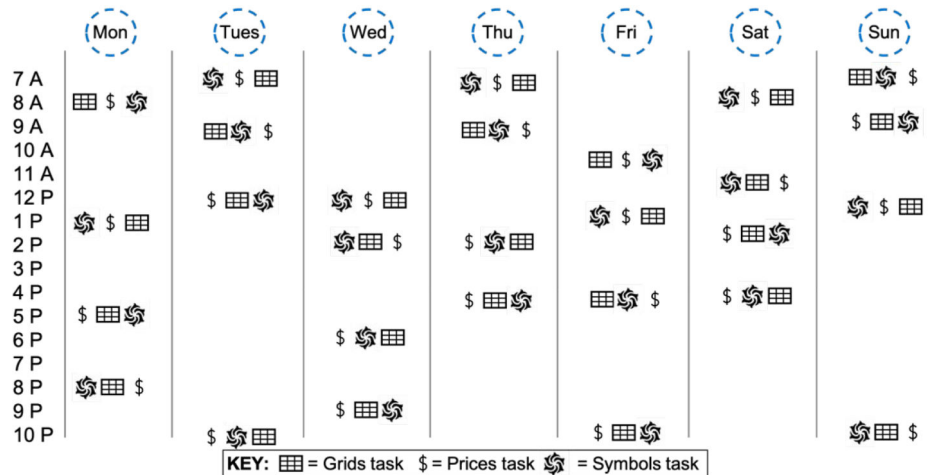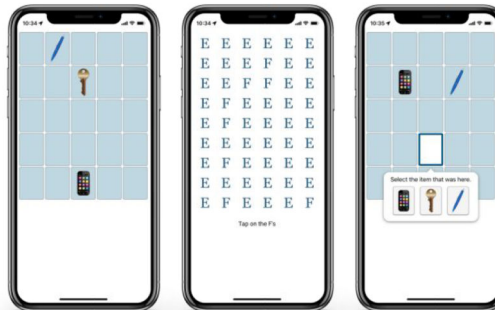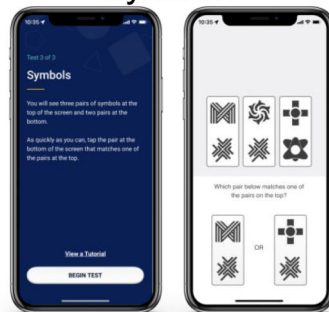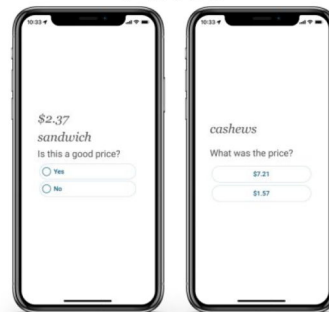
Figure 1. ARC Design and Cognitive Tasks

Note. Top demonstrates if a participant reported waking up at 7am and going to bed at 10pm, they would receive four test session notifications between 7am and 10pm, separated by at least two hours. The ARC cognitive tasks, Grids, Prices, and Symbols are displayed on the bottom.
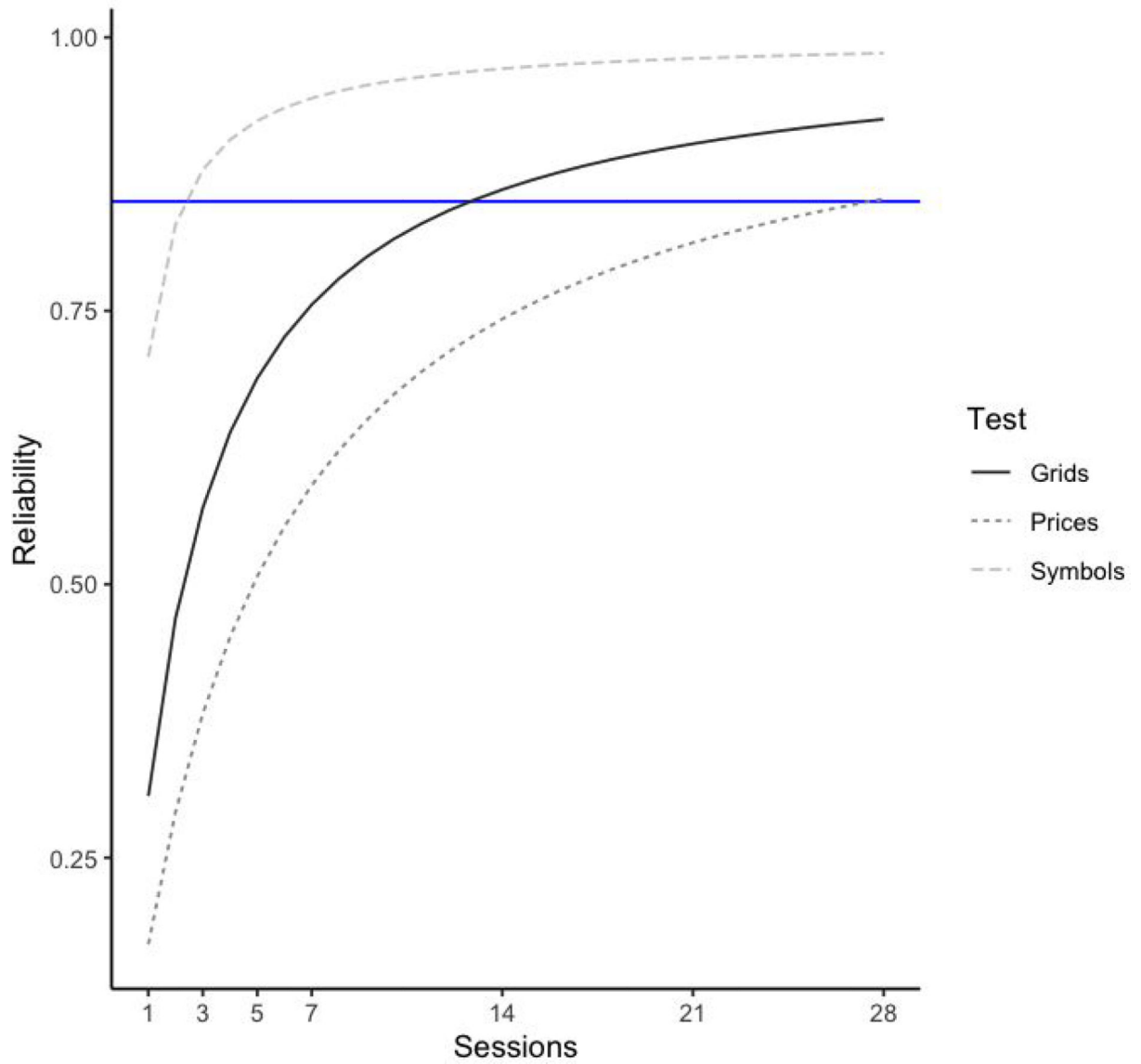
Figure 2. Between-Person Reliabilities for ARC Tasks
Note. Between-person reliabilities for each ARC cognitive task. Following Sliwinksi et al. (2018), a series of unconditional multilevel mixed models were fit to determine how many sessions would be required to obtain good reliability. Blue line indicates 0.85 reliability threshold.
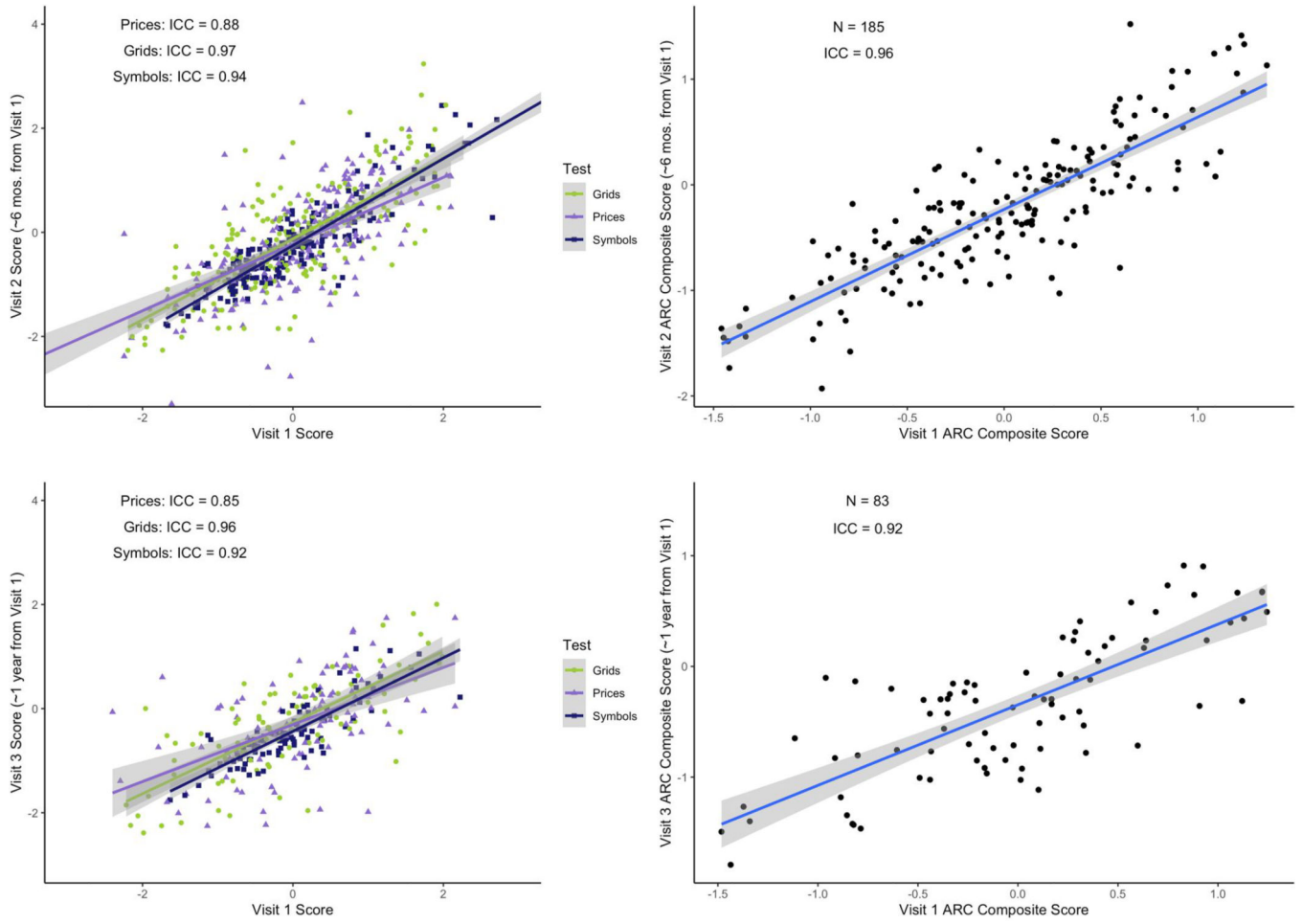
Figure 3. ARC Test-Retest Reliabilities at 6 Month (top) and 1 Year (bottom) Follow-Up
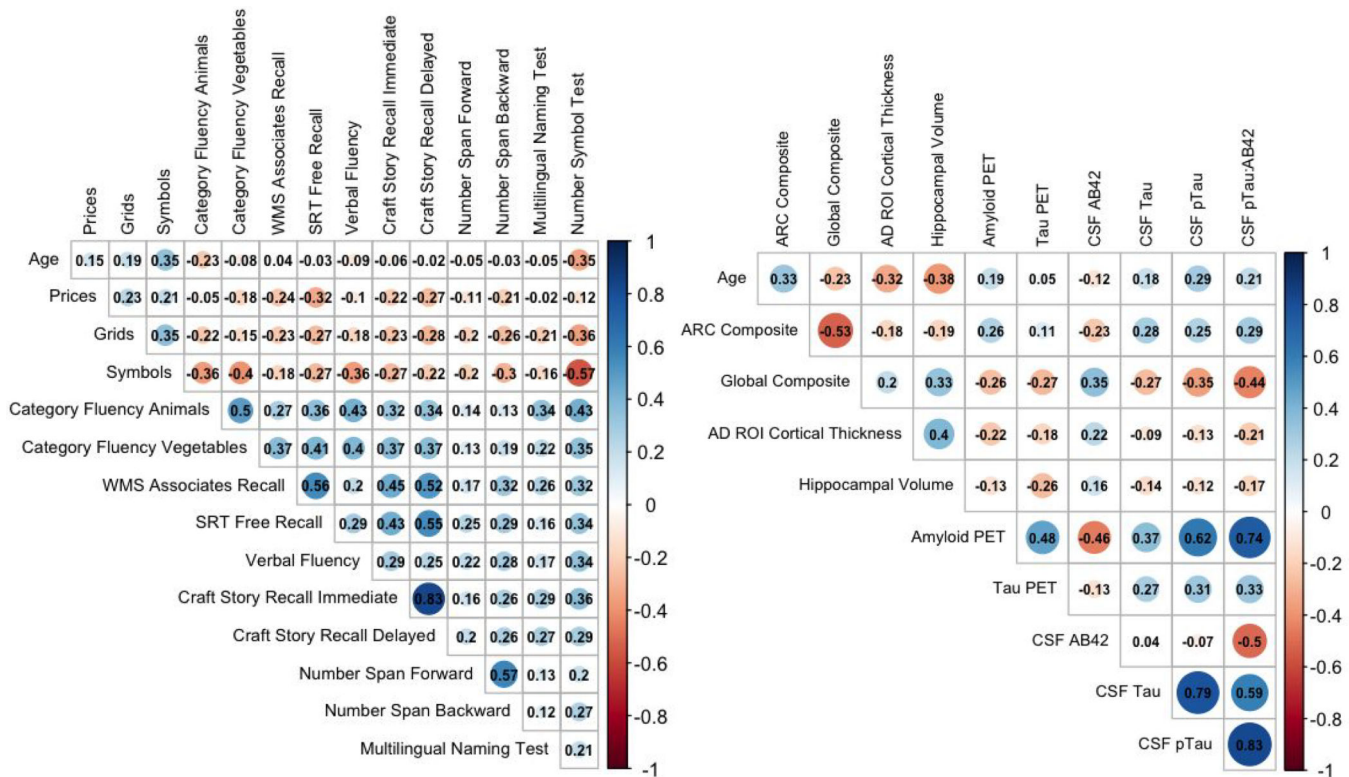
Figure 4. ARC, Conventional, and AD Biomarker Correlations

Note. Correlations amongst ARC and conventional measures (raw scores) shown on the left (N = 282). Correlations of the ARC composite score (higher = worse) and global composite score (higher = better), and AD-related biomarkers are shown on the right (Ns = 146 for CSF measures, 212 for amyloid PET, 173 for tau PET, 175 for AD ROI cortical thickness, and 290 for hippocampal volume). Significant correlations ($p < 0.05$) are displayed with colored circles, non-significant correlations are blank. Because in-clinic and ARC measures have opposing directionality, the negative correlations amongst the conventional and ARC measures are in the hypothesized direction.
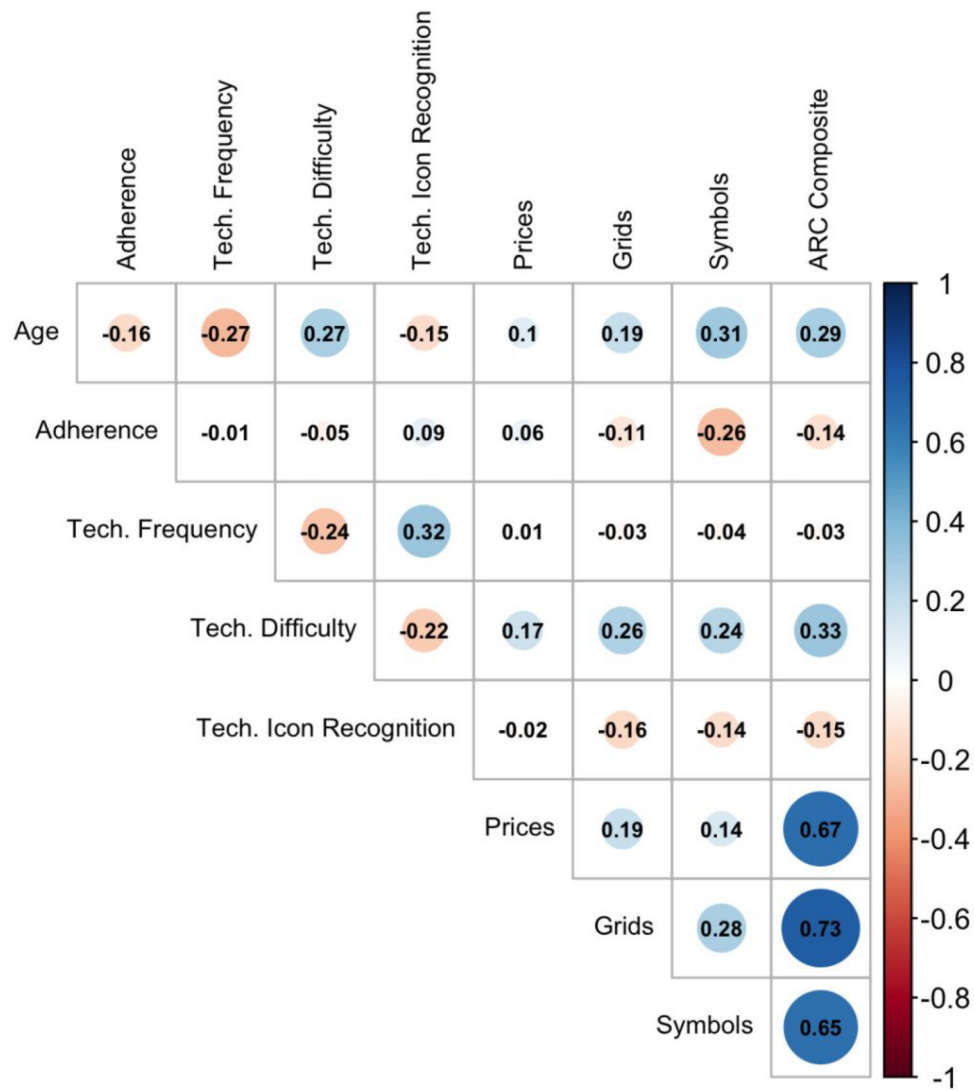
Figure 5. Age, Technology Familiarity, and ARC Performance Correlations
Note. Of the 290 participants included in the present analyses, 220 completed the technology familarity survey (see Nicosia et al., 2021) which assessed the frequency with which participants perform smartphone-related tasks, how difficult participants find various technology-related tasks, and how well participants could recognize technology-related icons. Significant correlations ($p < 0.05$) are displayed with colored circles whereas non-significant relationships are blank.
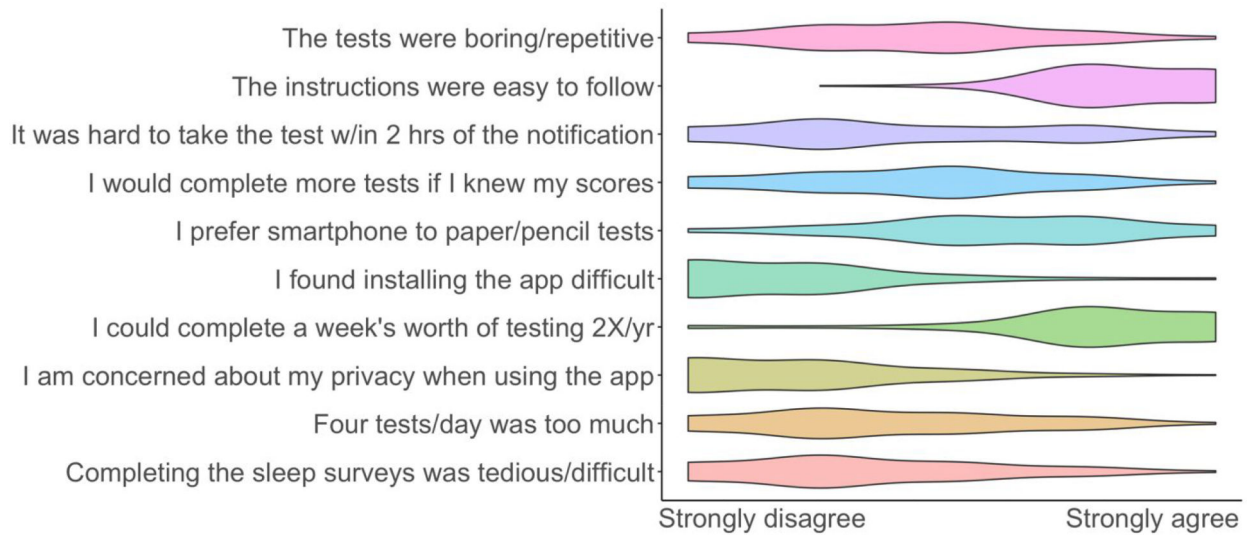
Figure 6. ARC User Experience Survey Results

Note. Of the 290 participants included in the present analyses, 228 completed the ARC user experience survey which assessed participants attitudes towards their experience with the ARC application after their first week using it.

Table 1.

Demographic Data

	CDR 0, N = 268[1]	CDR 0.5, N = 22[1]	p-value[2]
Age	76.6 (5.7)	77.0 (6.1)	0.76
Gender (% Female)[3]	148 (55%)	6 (27%)	0.021
Race[3]			0.10
Black	45 (17%)	0 (0%)	
Other	2 (0.8%)	0 (0%)	
White	219 (82%)	22 (100%)	
Education	16 (2)	17 (2)	0.74
APOE Status (% Positive)	82 (31%)	14 (67%)	0.002
Grids	0.71 (0.27)	0.93 (0.28)	0.002
Prices	0.25 (0.06)	0.29 (0.05)	0.002
Symbols	3.24 (0.95)	3.85 (1.32)	0.045
Adherence	81% (18%)	79% (20%)	0.71
Drop-out	13 (4.9%)	1 (4.5%)	0.99
	CDR 0, N = 134	**CDR 0.5**, N = 12	**p-value**
CSF Aβ42	970 (400)	533 (222)	<0.001
CSF Tau	347 (183)	474 (197)	0.052
CSF pTau:Aβ42	0.06 (0.06)	0.13 (0.07)	0.006
	CDR 0, N = 202	**CDR 0.5**, N = 10	**p-value**
Amyloid PET (Centiloid)	18 (27)	49 (43)	0.053
	CDR 0, N = 165	**CDR 0.5**, N = 8	**p-value**
AD ROI Tau PET (Standardized)	1.20 (0.15)	1.43 (0.42)	0.17
	CDR 0, N = 165	**CDR 0.5**, N = 10	**p-value**
AD ROI Cortical Thickness (mm)	2.57 (0.10)	2.46 (0.14)	0.043
Hippocampal Volume (mm^3)	7,790 (912)	6,983 (1,022)	0.035

[1] Mean (SD); n (%)

[2] Welch Two Sample t-test; Pearson's Chi-squared test

[3] Gender and race were self-reported

Table 2.

Descriptive Statistics At ARC Baseline

	N	Mean	SD	Range	Skew	Kurtosis	CDR 0 vs. 0.5 (t)
Age	290	76.61	5.768	36.07	0.446	0.224	0.31
Prices	290	0.249	0.062	0.4	−0.995	2.055	3.51 **
Grids	289	0.731	0.272	1.438	0.153	−0.284	3.52 **
Symbols	290	3.287	0.989	6.313	1.746	4.723	2.12 *
Adherence	290	80.42	18.19	71.43	−0.995	0.119	0.38
Drop-out	290	4.83	0.215	1	4.193	15.64	0.06
Category Fluency Animals	282	20.14	5.432	28	0.189	−0.519	3.90 ***
Category Fluency Vegetables	282	13.87	4.122	26	0.374	0.543	4.96 ***
WMS Associates Recall	282	14.91	3.571	16	−0.418	−0.581	3.10 **
FCSRT Free Recall	242	31.31	6.451	46	−0.847	1.612	4.86 ***
Verbal Fluency (Letters)	283	28.25	8.201	43	0.145	−0.073	3.36 **
Craft Story Recall Immediate	282	17.36	3.559	20	−0.524	0.144	2.17 *
Craft Story Recall Delayed	282	16.74	4.141	25	−1.229	2.951	2.65 *
Number Span Forward	283	8.445	2.308	12	−0.023	−0.399	1.18
Number Span Backward	282	7.355	2.192	12	0.408	0.481	1.88
Multilingual Naming Test	243	30.25	1.9	14	−2.015	7.426	0.96
Number Symbol Test	245	38.49	7.15	39	−0.182	0.147	2.19 **

Note.

*
indicates *p*-value < 0.05

**
indicates *p*-value < 0.01

***
indicates *p*-value < 0.001.

Table 3.

ARC Reliabilities for Individual Tasks

Sessions	Symbols	Grids	Prices
1	0.71	0.31	0.17
3	0.88	0.57	0.38
5	0.92	0.69	0.51
7	0.94	0.76	0.59
14	0.97	0.86	0.74
21	0.98	0.90	0.81
28	0.99	0.93	0.85

Note. ARC participants received 4 sessions/day for 7 day

Table 4.

ARC Test-Retest

| Measure | Visit 1 vs. Visit 2 | | | Visit 2 vs. Visit 3 | | |
	1, N = 185	2, N = 185[1]	p-value	2, N = 83[1]	3, N = 83[1]	p-value[2]
Prices	0.25 (0.06)	0.23 (0.06)	0.002	0.23 (0.06)	0.23 (0.06)	0.80
Grids	0.69 (0.27)	0.65 (0.27)	0.11	0.62 (0.30)	0.61 (0.26)	0.73
Symbols	3.09 (0.71)	2.88 (0.70)	0.004	2.82 (0.62)	2.67 (0.54)	0.080
ARC Composite	0.19 (0.69)	−0.08 (0.72)	<0.001	−0.15 (0.73)	−0.23 (0.67)	0.46

[1]Mean (SD)

[2]Welch Two Sample t-test

Note. Values represent participants mean score for that visit, values in parentheses represent standard deviations. Significant *p*-values indicate the presence of a practice effect