Published in final edited form as:

Assessment. 2018 January ; 25(1): 14–30. doi:10.1177/1073191116643164.

Reliability and Validity of Ambulatory Cognitive Assessments

Martin J. Sliwinski¹, Jacqueline A. Mogle², Jinshil Hyun¹, Elizabeth Munoz¹, Joshua M. Smyth³, and Richard B. Lipton⁴

¹Department of Human Development and Family Studies and Center for Healthy Aging, Pennsylvania State University

²College of Nursing and Center for Healthy Aging, Pennsylvania State University

³Department of Biobehavioral Health, Pennsylvania State University

⁴Department of Neurology, Department of Psychiatry and Behavioral Sciences, Department of Epidemiology and Population Health, Albert Einstein College of Medicine

Abstract

Mobile technologies are increasingly used to measure cognitive function outside of traditional clinic and laboratory settings. Although ambulatory assessments of cognitive function conducted in people's natural environments offer potential advantages over traditional assessment approaches, the psychometrics of cognitive assessment procedures have been understudied. We evaluated the reliability and construct validity of ambulatory assessments of working memory and perceptual speed administered via smartphones as part of an ecological momentary assessment (EMA) protocol in a diverse adult sample (N=219). Results indicated excellent between-person reliability (.97) for average scores, and evidence of reliable within-person variability across measurement occasions (.41–.53). The ambulatory tasks also exhibited construct validity, as evidenced by their loadings on working memory and perceptual speed factors defined by the in-lab assessments. Our findings demonstrate that averaging across brief cognitive assessments made in uncontrolled naturalistic settings provide measurements that are comparable in reliability to assessments made in controlled laboratory environments.

Keywords

ambulatory assessment; smartphone; ecological validity; cognition

Traditional approaches to measuring cognitive function have relied on testing procedures administered by trained technicians, in laboratory or clinical settings, and usually on a single occasion. The artificial nature of standard testing environments and unmeasured sources of within-person variability can negatively impact the ecological validity and reliability of standard approaches to cognitive assessment (Allard et al., 2014; Timmers et al., 2014). Health and social scientists have addressed similar concerns by using ecological momentary assessment (EMA) methods to study people's physical symptoms, psychological states, and

behaviors in real-time and in people's natural environment (Shiffman, Stone, & Hufford, 2008; Smyth & Stone, 2003). EMA is a particular approach to ambulatory assessment defined by several key features (Shiffman et al., 2008). First, data collection occurs in real-world environments as participants go about their daily activities, which is the "ecological" aspect of EMA. Second, assessments emphasize an individual's current state, which is the "momentary" aspect of EMA. Third, EMA involves the strategic sampling of moments for assessment. For example, random sampling would facilitate characterization of a person's "average" or typical level on a variable of interest (e.g., memory performance) through representative sampling of occasions, situations, and settings. And fourth, EMA involves multiple assessments which can elucidate how behaviors arise and vary over time and different situations.

Recent advances in mobile technology have enabled researchers to embed objective assessments of cognitive function into studies that use EMA methods. Important questions remain, however, regarding the reliability and validity of cognitive assessments administered using mobile technology in uncontrolled and naturalistic settings. The purpose of the present study is to evaluate the reliability and construct validity of a brief battery of cognitive tests administered on smartphones as part of an ecological momentary assessment (EMA) study.

The Rationale for Ambulatory Cognitive Assessment

A primary motivation for ambulatory assessment is to enhance ecological validity. Ecological validity is especially important for the field of neuropsychology, which has become increasingly focused on understanding the relationship between assessment results and real-world cognitive function (Spooner & Pachana, 2006). Cognitive scientists and neuropsychologists have, for the most part, been concerned with the *verisimilitude* of cognitive assessments, which refers to the ecological validity of stimuli, responses, and other task demands (Spooner & Pachana, 2006), and *veridicality*, which refers to the predictive relationship between cognitive assessment and activities of daily living. Ecological validity, however, reflects more than face validity (verisimilitude) and external validity (veridicality). Ecological validity also, and perhaps most importantly, pertains to questions of which settings of daily life are represented in a particular study design and data collection procedure (Fahrenberg, Myrtek, Pawlik, & Perez, 2007). Relatively little attention has been paid to this important facet of ecological validity which concerns the nature of the settings in which measurements of cognitive function take place (Timmers et al., 2014).

Traditional approaches to cognitive assessment take place in physical and social environments that are fundamentally dissimilar to the environments in which people perform cognitively demanding tasks in their daily lives. Although laboratory and clinical settings provide the benefit of experimental control over the testing environment, this control provides no assurance that cognitive performance observed in the laboratory is similar to how people would perform in the real world. A classic example of ecological validity concerns is "white coat hypertension", which refers to the observation that some people have higher blood pressure readings when taken in clinical settings compared to when measurements are taken in their natural environment. As regards cognition, performance measured in clinic or laboratory research settings may – for example – be either elevated

through social facilitation (e.g., Strauss, 2002) or hampered by unintended evaluative and stereotype threat (e.g., Schmader, Johns, & Forbes, 2008). Regardless of the specific differences, the clinical and laboratory settings in which researchers typically measure cognitive abilities are very unlike the situations in which people typically *use* those abilities.

The use of mobile technology in cognitive research can address ecological validity concerns through *ambulatory* assessment of cognitive function in peoples' natural environment as they go about their daily lives. Ambulatory assessment can also improve the reliability of measuring individual differences in cognitive function. Traditional approaches to studying cognition typically rely on measurements obtained on single testing occasion. Such "single-shot" assessments are influenced by both random and systematic within-person variability (e.g., sleep quality, recent exposure to stressors) that is difficult or impossible to control. Aggregating across repeated ambulatory measurements "cancels out" the effects of within-person variability and can improve measurement precision and reliability by estimating an individual's average level of functioning (Shiffman et al., 2008; Sliwinski, 2008).

In addition to addressing ecological validity and reliability concerns, ambulatory methods also permit frequent cognitive assessments that can be made in conjunction with other measurements of a person's environment and internal states. Repeated and proper sampling of situations can aid researchers in more accurately characterizing a person's average level of cognitive performance over the variety of naturalistic contexts in their daily lives (Brunswik, 1943) as well as in exploring dynamic relationships between cognitive performance and other time-varying psychological, social, and biological processes.

There are only a few examples describing the use of ambulatory cognitive assessments in EMA designs. In an early study Shiffman and colleagues compared the cognitive performance of regular and light smokers under naturally occurring conditions of abstinence and smoking by administering repeated cognitive tests via a handheld computer (Shiffman, Paty, Gnys, Kassel, & Elash, 1995). In a later study, they used cognitive assessments embedded in an EMA design to examine the time-course and natural history of the effects of nicotine withdrawal on cognitive function (Shiffman et al., 2006). Another study used EMA to demonstrate the acute effects of alcohol consumption at ecologically relevant doses on ambulatory working memory performance (Tiplady, Oshinowo, Thomson, & Drummond, 2009). In a study of older adults, Allard and colleagues (2014) found that ambulatory semantic memory was higher during the hours following engagement in specific intellectual activities such as reading and word puzzles compared to other activities. A recent study combined EMA, ambulatory cognitive assessment and passive physiological monitoring to demonstrate that age moderates the effects of cardiovascular activation on working memory performance (Riediger et al., 2014).

These examples describe how researchers have used naturally occurring behaviors, contexts, and psychological and biological states to predict ambulatory cognitive function. Other applications have used ambulatory cognitive assessments as sensitive indicators of individual differences and predictors of proximal psychological states and behaviors. For example, Allard and colleagues (2014) found that aggregated scores on an ambulatory semantic memory test were more sensitive indicators of individual differences in hippocampal volume

than scores from the same memory test administered by conventional means (single-shot, in the lab). Waters and colleagues have used modified versions of the standard and emotional Stroop task to obtain ambulatory measurements of attentional bias in individuals being treated for substance dependency. They found that reliable Stroop effects could be obtained using mobile-based administration, and that elevated attentional bias, measured on a momentary basis (Waters & Li, 2008), may predict imminent temptation episodes among recovering addicts (Waters, Marhe, & Franken, 2012).

Apart from the studies reviewed above, there have been relatively few studies that have used mobile technology (e.g., smartphones) to assess cognitive function via repeated, momentary assessments. There is also a lack of evidence regarding the reliability and validity of scores obtained by brief cognitive assessments administered by smartphones in uncontrolled, naturalistic settings. One notable exception is a recent study by Dirk and Schmiedek (in press), in which they described the reliability and validity of smartphone based ambulatory assessments of working memory in 3rd and 4th grade students. They found that that scores averaged across the one month period exhibited very high between-person reliability (>.99) but also that there was evidence of systematic within-person variability across measurement occasions.

The Present Study

The purpose of the present study is to evaluate the reliability and construct validity of a brief battery of ambulatory cognitive tests administered on smartphones. The ambulatory tests were designed to measure two distinct and fundamental cognitive constructs: perceptual speed and working memory. Perceptual speed, which reflects the ability to quickly and accurately compare different types of stimuli (e.g., numbers, symbols, patterns), is important for understanding child and adolescent cognitive development (Rucklidge & Tannock, 2002; Stenneken et al., 2011), cognitive aging (Salthouse, 2000), and is also predictive of real world outcomes, such as job performance (Ackerman & Beier, 2007) and vehicular crash risk (Owsley et al., 1998). Working memory ability, which reflects the capacity to maintain information in active memory while simultaneously performing distracting or interfering activities (Just & Carpenter, 1992), is also predictive of developmental, aging, and real world outcomes (Alloway & Alloway, 2010; Schneider-Garces et al., 2010). One perceptual speed and two distinct working memory tasks were embedded in a signal contingent EMA protocol, which involved prompting individuals to complete a brief assessment 5 times per day for 14 days using study provided smartphones. Each assessment included some self-report questions (e.g., current mood, location, recent activity), followed by the cognitive tests.

Given the demands on participants imposed by an EMA design, each assessment, and therefore, each cognitive task must be kept very brief. For example, the ambulatory speed test used in the current study consisted of just 12 trials at each assessment, compared to laboratory assessments of speeded performance which may include 100 trials or more. We designed each of the three ambulatory cognitive tasks to be completed in less than one minute. The necessary brevity of ambulatory cognitive tests coupled with their unsupervised administration in uncontrolled, naturalistic settings raises legitimate questions about

measurement reliability. That is, can scores obtained from brief ambulatory assessments capture systematic (i.e., reliable) sources of variance in cognitive performance?

The precise type of variance that counts as systematic (or error), however, depends on the purpose for which EMA data are being used. One way to use EMA data involves aggregating across measurement occasions in order to characterize between-person individual differences (Shiffman et al., 2008). For this purpose, only systematic between-person differences are of interest, and all sources of within-person variation (i.e., across occasions, across trials within occasions) are considered error. Another use of EMA data involves analyses of within-person variability in order to identify contextual associations and time-ordered effects across measurement occasions. For this purpose, within-person variance across occasions is of primary interest, and variability within measurement occasions (e.g., across trials) is considered error. Thus, one goal of the present study is to examine whether and to what extent scores obtained on brief, repeated ambulatory assessments capture systematic between-person and within-person (across occasion) variability in cognitive performance.

The second goal of the present study is to determine whether scores obtained from ambulatory cognitive tests measure the same constructs (i.e., perceptual speed and working memory) that are commonly studied in laboratory settings. This goal is especially important for comparing results obtained through ambulatory methods with those obtained through more traditional assessment approaches. Thus, we examined the pattern of correlations between cognitive tasks administered on smartphones in naturalistic settings and a battery of tasks administered in a traditional controlled laboratory setting that are valid indicators of the perceptual speed and working memory.

Method

The current investigation is part of a larger study, The Effects of Stress on Cognitive Aging, Physiology and Emotion (ESCAPE). The ESCAPE project is a prospective measurement burst study (Nesselroade, 1991; Sliwinski, 2008) that tracks variability and change in daily experiences of stress and cognitive function across four 14-day EMA measurement bursts (Scott et al., 2015). The present study reports data from the smartphone-based and laboratory cognitive tests from the baseline measurement burst.

Participants

Participants were 219 adults (34% men, 66% women) aged 25–65 years ($Mean = 46.99$, $SD = 10.74$) recruited using systematic probability sampling of New York City Registered Voter Lists for the zip code 10475 (Bronx, NY). Nine percent of participants self-identified as Non-Hispanic White, 63% as Non-Hispanic Black, 17% as Hispanic White, 6% as Hispanic Black, <1% as Asian, and 4% as other. Most of the sample had completed high school or some college (50%) or had a college degree (45%). During the data collection period (2011–2013), 51% of the sample was working, 12% were retired, 26% were unemployed looking for working, and 10% were unemployed not looking for work. The sample displayed marked diversity in income as well: 5% reported an annual income <\$4,999, 17% between \$5,000 – \$19,999, 24% between \$20,000 – \$39,999, 20% between \$40,000 – \$59,999, 12% between

\$60,000 – \$79,999, 7% between \$80,000 – \$99,999, 8% between \$100,000 – \$149,999, and 1% greater than \$150,000 annually; 7% declined to report income. Married persons made up 31% of the sample; 9% were not married but were living with someone, 17% were divorced or separated, 34% were never married, 2% were widowed, and 6% described their marital status as other. The sample slightly over-represents women (66% of the sample compared to 58% of the population), but is representative of the area from which it is obtained with respect to age, race and income.

Procedure

During recruitment, introductory letters were mailed to individuals from a sampling frame (obtained from the Registered Voter Lists) and a research assistant phoned to establish eligibility, and enroll and consent interested persons. Eligibility criteria included aged 25–65, ambulatory, fluent in English, without visual impairment that would interfere ability to use the study smartphone, resident of Bronx County. Participants were mailed paper survey batteries assessing demographic and individual difference characteristics which they complete at home and brought to their first lab visit. At the first visit to the research offices, participants completed 1.5 hours of training on the use of study smartphones to complete surveys.

Beginning the next day after their first visit, participants completed 2 days of the EMA protocol then returned to the lab for a second visit to complete a battery of cognitive tests and to determine compliance with the EMA protocol. The EMA protocol involved carrying specially programmed study smartphones that were programmed to beep at quasi-random times to alert participants to complete 5 ambulatory assessments per day. The scheduled interval between beeps was between 2 to 3 hours, the average time between scheduled beeps was 2 hours and 33 minutes. Participants who completed 80% of the EMA protocol were invited to the 14 day EMA study. At the end of 14 days, participants returned phones to the lab. Participants were generally highly adherent to the study protocol, responding to an average of 85% surveys; most (79%) survey responses were completed within 10 minutes of the beep prompt. Participants who satisfactorily completed the entire study protocol could receive up to \$160.

Measures

In-Lab Cognitive Assessments—Participants were individually tested by a trained technician in a quiet testing room in a single session. In addition to the tests of perceptual speed and working memory, tests of fluid and crystallized intelligence, and associative learning were also administered. Because of its focus on reliability and construct validity of ambulatory perceptual speed and working memory tasks, the present analyses used data from the in-lab perceptual speed and working memory tests. We also examined the correlation of fluid intelligence with in-lab and ambulatory cognitive tests because of its established association with speed and working memory.

Fluid Intelligence: The Ravens Progressive Matrices (Raven, 2000) was administered as a measure of fluid intelligence. Items in the Ravens Progressive Matrices tests consisted of a

series of geometrical shapes and participants were asked to guess what the next shape should be among multiple options.

Perceptual speed: Tasks consisted of a symbol match, a letter match, and a number match task. Each task required participants to make speeded perceptual comparisons among items present in a visual display. The dependent measure for each of the perceptual speed tasks was the median reaction time for correct trials.

Symbol match: Participants saw six symbol pairs at the top of the screen and were presented with a comparison symbol pair at the bottom of the screen. They had to decide as quickly as possible whether the target pair at the bottom of the screen exactly matched one of the pairs at the top of the screen. Participants completed 70 trials of this task.

Letter match: Two letter strings were presented and participants identified whether the strings were the same or different as quickly as possible. Strings were 3, 6, or 9 characters in length and mismatched strings varied by one character. Participants completed 70 trials with one break at trial 35.

Number match: Participants saw two number strings and indicated whether the strings were the same or different as quickly as possible. Strings were 3, 6, or 9 items in length and participants completed 70 trials with a break midway through the test.

Working memory: Working memory tasks consisted of three tasks that previous work has demonstrated are valid indicators of working memory capacity: operation span, counting span (Conway, Cowan, Bunting, Theriault, & Minkoff, 2002), and a backward letter span task (Waters & Caplan, 2003).

Counting span: The counting span task was administered following procedures described in Engle et al. (1999). Participants were instructed to memorize the number of targets in a series of displays that included dark blue circles (targets) and dark blue squares (color distractors) and light blue circles (shape distractors). Each display was initiated by the tester. Participants counted the number of dark blue circles aloud and repeated the digit corresponding to the final count. The number of targets per display varied from two to six, with three trials of each. After two to six displays, a recall cue was presented, at which point participants reported the number of targets in each of the previous displays, in the serial order in which they occurred. The dependent measure was the total number of target counts recalled in the correct order.

Operation span: In the operation span, participants verified equations aloud while trying to remember letters. As with the counting span, after a series of number of equations and letters was presented, participants were prompted to recall all the letters from that series. The number of equation and letter pairs per series varied from two to five with three series of each length presented in a fixed random order. The dependent measure was the total number of letters recalled in correct order.

Backward span: Participants saw a series of letters presented one at a time for 1 second each. At the end of each series participants recalled all of the letters they saw in reverse order. The number of letters in each series varied from 3–8 and participants attempted 2 trials at each length for a total of 12 trials. The dependent measure was the total number of the items for trials that were recalled in the correct order.

Ambulatory Cognitive Assessments—Each ambulatory assessment consisted of a brief survey that assessed affective state and stress, as well as physical/social activities at the time of the prompt. After completing each survey, participants performed three ambulatory cognitive tasks in the following fixed order: symbol search, dot memory and n-back. The ambulatory cognitive tasks were administered on Droid X which has a 4.3" display (480 × 854 pixels) and a 60 Hz refresh rate, and response times were recorded in milliseconds.

Symbol search: On each trial of the symbol search task, participants saw a row of three symbol pairs at the top of the screen and were presented with two symbol pairs at the bottom of the screen. Stimuli were presented until a response was provided there was an interval of 200 msec. between each response and the following stimulus. Participants decided, as quickly as possible, which of the two pairs presented at the bottom of the screen was among the pairs at the top of the screen (see Figure 1). Participants completed 12 trials of this task. The dependent variable was median response time of correct trials. Because this task requires speeded comparisons similar to standard laboratory tests, we reasoned it would be a viable indicator of perceptual speed.

Dot memory: Each trial of the *dot memory* task consisted of 3 phases: encoding, distraction, and retrieval (see Figure 2). During the encoding phase, the participant was asked to remember the location three red dots appear on a 5 × 5 grid. After a 3 second study period, the grid was removed and the distraction phase began, during which the participant was required to locate and touch F's among an array of E's. After performing the distraction task for 8 seconds, and empty 5 × 5 grid reappeared on the screen and participants were then prompted to recall the locations of the 3 dots presented initially and press a button labeled "Done" after entering their responses to complete the trial. Participants completed 2 trials (encoding, distractor, retrieval) with a 1 second delay between trials. The dependent variable was an error score with partial credit given based on the deviation from the correct positions. If all dots were recalled in their correct location the participant received a score of zero. In the case of one or more retrieval errors, Euclidean distance of the location of the incorrect dot to the correct grid location was calculated, with higher scores indicating less accurate placement and poorer performance (Siedlecki, 2007).

The rationale for our use of this task as an indicator of working memory has both an empirical and theoretical basis. Previous research (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001) has demonstrated that a similar dot memory task loaded on a factor representing working memory. The authors of this study reasoned that the spatial dot memory task placed high demands on controlled attention—a hallmark of working memory tasks. Indeed, individual differences in working memory capacity arise "in situations where information needs to be actively maintained or when a controlled search of memory is required" (Unsworth & Engle, 2007, p. 123). The ambulatory dot memory task satisfies this

requirement by using an interference task to prevent rehearsal and produce interference with encoded locations, which creates the demand for active maintenance and controlled retrieval of previously encoded location during the recall phase.

N-Back: Participants saw a series of 3 standard playing cards slide from one box on the right of the screen to the second box on left of the screen. In the practice phase, cards were facing up and participants were asked to determine if the *target* card in the rightmost box matched the *test* card in the leftmost box. After each response there was a 500 msec. delay after which the cards shifted positions from right to left, with a new target card appearing in the rightmost box (see Figure 3). After 10 practice trials, participants started the 2-back condition in which each new target card would turn face down prior to shifting positions, requiring participants to retain the card's identity in working memory. As the cards shifted positions, participants were asked to determine if the current target card matched the face-down card they saw 2 trials back while retaining the identity of the previous target card. Feedback on response errors was provided by turning all cards face up. Participants completed twelve trials and the dependent variable was the proportion of correct responses. We selected the N-back because of its widespread use as an indicator of working memory (Schmiedek, Lövdén, & Lindenberger, 2014).

Results

Descriptive Statistics

Table 1 shows the means and standard deviations for each of the ambulatory and in-lab cognitive tasks. Because both age and fluid intelligence are related to working memory and perceptual speed, we also examined correlations of task performance with age and performance on the Raven's Matrices. The correlations between age and two of the three ambulatory tasks (dot memory and symbol search) were comparable to the age correlations with the in-lab tasks of working memory and perceptual speed. Specifically, older age was associated with higher error scores on the ambulatory dot memory task ($r = .16, p = .02$) and worse performance on the in-lab counting span ($r = -.12, p = .09$), operation span ($r = -.19, p < .01$) and backward span ($r = -.16, p = .02$) tasks. Age was also correlated with slower RTs on the ambulatory symbol search ($r = .36, p < .01$), as well as with the in-lab letter match ($r = .35, p < .01$), number match ($r = .29, p < .01$), and symbol match ($r = .35, p < .01$) tasks. Age was not significantly correlated with the accuracy on the ambulatory n-back task ($r = -.04, ns$). Higher scores on the Ravens test were associated with better performance on both the ambulatory dot memory ($r = -.42, p < .01$) and n-back tasks ($r = -.47, p < .01$) and faster response times on the ambulatory symbol search task ($r = -.50, p < .01$). These correlations were comparable in magnitude to the in-lab working memory (r s ranged from .27 to .38) and speed (r s ranged from $-.41$ to $-.43$) tasks.

Reliability Analysis

Between-person reliability—Between-person reliability refers to the proportion of total variance in scores attributable to differences between individuals. Aggregation of EMA scores may boost reliability (due to averaging) for measuring individual differences to levels that exceed typical “single-shot” measurement approaches (Shiffman et al., 2008). We used

SAS PROC MIXED to estimate multilevel mixed models (MLMs) in order to compute the between-person (BP) reliability of scores on the ambulatory cognitive tests using the following formula (Raykov & Marcoulides, 2006):

$$BP\ Reliability = \frac{Var(BP)}{Var(BP) + \frac{Var(WP)}{n}} \quad (1)$$

where Var(BP) is the total variance in scores that is between-persons, Var(WP) is the total variance in scores that is within-persons, and n is the number of assessments. To determine the reliability of a test based on a single measurement, n would equal 1, which defines one type of intraclass correlation (ICC). The ICC is the expected correlation between two randomly sampled measurements from the same person, and hence reflects the stability of measurements. To determine the reliability of aggregated scores based on the average of multiple measurements, n would be set equal to the number of observations on which the average is based.

Table 2 shows the between- and within-person variance and resulting reliabilities produced by fitting unconditional MLMs using restricted maximum likelihood to each of the ambulatory tasks. The ICCs were 0.39 for the dot memory task, 0.54 for the symbol search task, and 0.59 for the N-back. The ICCs indicate that between 39% and 59% of the variance in performance across the three ambulatory tasks was between-persons. The reliabilities of average scores aggregated across all the ambulatory assessments were exceptionally high: .97 for the dot memory task, and .99 for both the symbol search and N-Back tasks. These reliabilities are based on scores that reflect the average of 56 (80%) of the ambulatory assessments, which was the approximate average number of assessments participants completed in this study.

We next conducted a series of follow-up analyses to determine how many days of ambulatory assessment would be required to obtain reliabilities of aggregated scores that exceeded .80 and .90. To do this we fit a sequence of unconditional MLMs and calculated reliabilities of average scores using only observations from the first study day, from the first two study days, and from the first three study days. The bottom of Table 2 displays these results, indicating that 2 days of ambulatory assessments (10 total assessments) are required to attain a reliability >.80 for the dot memory task, and 3 days of assessments (15 total assessments) are required for aggregated scores to attain a reliability of .90. Averaging across a single day of 5 assessments produced scores with reliabilities .90 for the Symbol Search and N-back.

Because our sample contained people who varied widely in age (25–65), and because age is related to cognitive function, we next conducted a sensitivity analysis to examine whether this age heterogeneity influenced the reliability results. First, we recalculated the ICCs and reliability of the average scores after statistically partialling for age. The age partial ICC for the symbol search task (.50) was slightly lower compared to the unconditional ICC reported in Table 2. The ICCs for the dot memory and the N-back task were not changed by statistically partialling for age. The reliabilities for the average scores aggregated across the

14 days of EMA remained above .96 for all three tasks after partialling for age. Second, we categorized individuals into 4 equal width age bins (age 25–34, 35–44, 45–54, 55+) and conducted an age-stratified analysis to determine whether the tests exhibited similar patterns of between-person and within-person variability across ages. Table 3 indicates the ICCs exhibited an increasing trend for older ages, particularly for the symbol search and N-back tasks. Inspection of the variance components from the stratified analyses indicated that this was primarily due to more between-person heterogeneity among older compared to younger participants. For all age strata, reliability exceeded .85 for scores aggregated across 3 days, and .95 for scores aggregated across 14 days.

Next, we examined whether reliability remained constant across the 2-week ambulatory assessment period. Separate unconditional MLMs were fit to the data from the 5 ambulatory assessments made on each of the 14 study days. Then, reliabilities of the average scores for each day were calculated for each of the 3 ambulatory tasks, and plotted in Figure 4. Reliabilities were very stable across the 14 days: the standard deviation of daily reliabilities was .01 for the symbol search task and .02 for both the dot memory and N-back tasks. Thus, increased practice with the tasks did not alter the reliability of aggregated scores.

Within-person reliability—Next, we examined whether there was evidence of systematic within-person variation in cognitive performance across assessments using an approach similar to other studies of within-person cognitive variability (Brose, Schmiedek, Lövdén, & Lindenberger, 2012; Schmiedek, Lövdén, & Lindenberger, 2013; Sliwinski, Smyth, Hofer, & Stawski, 2006). This examination determined whether within-person differences in cognitive performance from one occasion to the next reflect systematic variance by decomposing within-person variation into two sources: across occasions and within occasions. Because our interest is in across-occasion variability, this source of variability is considered *systematic* and within-person variability within-occasions (i.e., across trials) is considered *error*¹. The logic for quantifying the proportion of systematic within-person variance across assessment occasions is analogous to estimation of between-person reliability shown in equation 1 above:

$$WP\ Reliability = \frac{Var(WP_{occasion})}{\left(Var(WP_{occasion}) + \frac{Var(WP_{trial})}{i} \right)} \quad (2)$$

where $Var(WP_{occasion})$ is the within-person variance in scores *across assessment occasions*, $Var(WP_{trial})$ is the within-person variance *across trials* (within-occasions) of scores, and i is the number of trials. The symbol search task consists of $i=12$ trials, and the dot memory task consists of $i=2$ trials. As the n-back task consist of 12 binary trials (1= correct, 0=incorrect), we followed a procedure used by Brose and colleagues (Brose et al., 2012) and calculated the proportion correct for odd and even trials and treated $i=2$ (scores from odd and even

¹Referring to within-occasions (i.e., across trial) variance as “error” in this context does *not* imply that it reflects only noise. In fact, a large body of research has shown that trial-to-trial variability on a single measurement occasion reflects systematic variance related variables such as psychological stress (Sliwinski et al., 2006), aging (Hultsch, MacDonald, Hunter, Levy-Bencheton, & Strauss, 2000), and brain damage (MacDonald, Li, & Bäckman, 2009). The designation of whether a source of variability is considered as systematic or error depends upon whether it is the focus of study.

trials) for this task. We used SAS VARCOMP to quantify systematic within-person reliability by partitioning the within-person variance into variability across assessment occasions and variability across trials within occasions for each of the 3 ambulatory cognitive tests as described by Cranford and colleagues (see Cranford et al., 2006, pp.925 formula 5). Results indicated within-person reliabilities of .53 for the symbol search task, .50 for the dot memory task, and .41 for the n-back task.

We next conducted a sensitivity analyses to determine whether within-person reliabilities were similar across age. Table 4 shows that the within-person reliabilities for the symbol search were constant across age strata, but that the within-person reliabilities for both the dot memory and n-back tasks were slightly lower for the older age strata. Finally, we examined whether within-person reliabilities remained constant across the duration of the study or whether reliabilities changed with repeated practiced on the ambulatory tasks. To accomplish this, we estimated within-person reliabilities separately for each of the 14 days of the study, based on the 5 daily assessments obtained on each of those days. As this approach treats each study day separately, the resulting reliabilities (displayed in Figure 5) only reflect systematic sources of within-person variance that transpires *within a given day*, and do not reflect systematic sources of within-person variability that transpire *across days*. Therefore these study day specific reliabilities will tend to be somewhat lower than reliabilities calculated from the entire data series, which reflect both within-day and across day sources of systematic within-person variance. Inspection of Figure 5 indicates that the within-person reliabilities for the symbol search and dot memory tasks were relatively stable across study days: the standard deviation of within-person reliabilities was .04 for both the symbol search task and dot memory tasks. The within-person reliabilities for the N-back were somewhat more variable ($SD=.07$). Inspection of the data indicated that the within-person reliabilities were much more variable for the first several days compared to the remaining days: excluding the data from the first three days reduced the SD of within-person reliabilities for the n-back to .04.

Validity Analysis

Construct validity of the ambulatory tasks was assessed by first examining the pattern of correlations among the ambulatory and in-lab tasks, and then by conducting a confirmatory factor analysis (CFA) to assess formally the fit of our hypothesized measurement model. Table 5 shows the intercorrelations among all the ambulatory and in-lab cognitive tasks. The ambulatory symbol search task was correlated strongly with all of the in-lab speed tasks (r .61 to .74). The ambulatory dot memory task correlated significantly with in-lab working memory tasks (r $-.39$ to $-.45$), which were similar in magnitude to the correlations between the in-lab backward span task and operation span ($r = .42$) and counting span tasks ($r = .48$). The correlation between the operation and counting span tasks was higher ($r = .60$) likely because of shared method variance (i.e., they are both complex span tasks). The ambulatory n-back task exhibited significant correlations with in-lab working memory tasks (r .24 to .36) that were similar in magnitude to its correlations with in-lab speed tasks (r $-.29$ to $-.34$). These patterns of correlations are not surprising, given that the task demands of the n-back involve aspects of working memory (e.g., memory updating), but also speeded perceptual comparisons.

We next fit a confirmatory factor analysis (CFA) using MPLUS (Version 7.2) software to provide a formal evaluation of our hypothesized measurement model. The model specified two correlated latent variables: 1) a working memory factor, indicated by 3 in-lab tasks (operation span, counting span, backward span) and two ambulatory tasks (dot memory and n-back); and 2) a perceptual speed factor, indicated by 3 in-lab tasks (number, letter and symbol match) and one ambulatory task (symbol search). The initial model had unacceptable fit ($\chi^2[26]=94.94$, $df=26$, $p<.001$; CFI=.925, RMSEA=.110 [90% CI= 0.087, 0.134], SRMR=.057), and examination of modification indices suggested that model fit would be improved by adding covariances between pairs of indicators that shared common method variance. Specifically, we added a covariance between the operation and counting span variables (because they are both complex span tasks) and between the in-lab symbol match and the ambulatory symbol search (because they both shared a common stimulus set). The resulting model, shown in Figure 6, had acceptable fit ($\chi^2[24]=45.82$, $p=.005$; CFI=.976, RMSEA=.063 [90% CI=.035, .092], SRMR=.045. Standardized regression weights were all significant and $>|.56|$ for the working memory factor and $>|.66|$ for the speed factor. The correlation between the two factors ($-.53$) is consistent with speed-working memory correlations in the literature (Verhaeghen & Salthouse, 1997). Adding cross-factor loadings did not improve model fit, indicating good discriminant validity.

We conducted two follow-up analyses to examine the sensitivity of these results to the effects of age. That is, because our sample varies considerably in age and because age is related to cognitive performance, is possible that the correlations among the tests are driven by age effects. Therefore, we re-ran the CFA partialling for age to determine the extent to which the factor loadings may have been driven by shared age differences in memory and speed. Thus, we fit a CFA identical to the model depicted in Figure 6, but with the addition of regressing each indicator variable on age, so that the factor loadings reflected effects residualized for individual differences in age. The residualized factor loadings on the working memory factor (dot memory= $-.68$, n-back = $.57$, counting span= $.59$, operation span = $.54$, and backward span= $.66$) and the speed factor (symbol search= $.57$, symbol match= $.78$, letter match= $.83$, number match= $.87$) remained high, statistically significant and were relatively unchanged compared to their loadings from the unadjusted model (see Figure 6). The fit for the age-partial CFA was also acceptable: ($\chi^2[23]=43.24$, $p<.01$; CFI=.979, RMSEA=.063 [90% CI= .033, .092], SRMR=.040). These results indicate that the pattern of correlations between the ambulatory and in-lab cognitive tests is not a by-product of the age heterogeneity of the sample.

Next, we examined whether the CFA factor model applied to the younger and older members of the sample. Accordingly, we split the sample into two subsamples at the median age (age 48) and tested for factorial age invariance between the younger and older subsamples. We compared the fit of a sequence of two-group confirmatory factor models that stipulated configural (same factors structure), metric (equal factor loadings), scalar (equal loadings, equal intercepts), and strict (equal loadings, equal intercepts, equal residual variances) invariance between the younger and older subsamples (Yoon & Millsap, 2007). Table 6 displays the results from these models. Model fit was not significantly worse for increasingly stringent invariance models and strict invariance obtained between the younger and older subsamples.

Discussion

Findings from the current study contribute to the literature by demonstrating both the reliability and construct validity of brief ambulatory cognitive assessments conducted by smartphones in people's natural environments. Specifically, there were three main findings: (1) between-person reliability of ambulatory cognitive test scores averaged across the 14 day EMA protocol were .97; (2) within-person reliabilities (.41–.53) indicated the ambulatory tests captured systematic fluctuations in memory and speeded performance across occasions; (3) performance on the ambulatory tests of memory and speed was correlated with in-lab cognitive assessments of the same constructs. In addition, supplementary analyses indicated that these findings were consistent across duration of the study protocol (e.g., robust to practice effects) and invariant across age. We now discuss the implications of these findings and how they extend previous work on ambulatory cognitive assessments, as well as some important limitations and directions for future research.

Reliability of Ambulatory Cognitive Assessments

Between-person reliability—Our findings describe two different types of reliability: between-person and within-person reliability. As regards between-person reliability, all three of the ambulatory cognitive tests produced highly reliable average scores (.97), consistent with reliability values obtained from ambulatory working memory tests administered to school-aged children (Dirk & Schmiedek, in press). The results also demonstrated that good between-person reliabilities can be obtained with fewer than 14 days of ambulatory assessments. For example, averaging across assessments from a single day produced a reliability of .75 for the dot memory task, and reliabilities .90 for the symbol search and n-back tasks. Averaging across the 3 days produces scores with reliabilities of .90 for all the ambulatory cognitive tasks. The boost to reliabilities obtained by averaging across repeated assessments would also apply to cognitive tests repeatedly administered in more conventional laboratory or clinic settings. Obtaining many repeated measurements in a controlled laboratory setting that are distributed across a relatively narrow interval (e.g., within a day or week) would, however, present a number of logistical challenges, including personnel, space and time constraints. Using mobile technology, such as smartphones, to obtain repeated assessments can increase the feasibility of conducting intensive longitudinal studies of cognitive function.

It is important to recognize, however, that these reliability values are not a property of a test *per se*, but depend upon a specific assessment procedure. The reliability values provided for standard neuropsychological tests presume that those tests are administered by a trained tester and according to prescribed procedures. Similarly, the reliability values provided in Table 2 presume a certain set of procedures, which included pre-training of the participants on the tasks, as well as the specific assessment protocol which distributed five measurements across multiples times of the day for a two week period.

The reliability values reported in Table 2 could be used for planning purposes to determine the required number of assessments required to obtain reliable cognitive scores; however, researchers must also consider the frequency and the time period across which measurements are distributed. This has significant implications for planning assessment

protocols because administering 20 trials of the dot memory task, for example, on a single assessment occasion may not result in equivalent reliability as scores based on the same number trials distributed across multiple assessments. Similarly, distributing 20 trials across five assessments in a single day may produce different results than distributing those 20 trials across a week. Scores based on the average of multiple temporally distributed assessments may boost reliability because they “average out” the effects of state-based influences, such as fatigue and stress, as well as other sources of natural variation that could impact performance on any given measurement occasion (Allard et al., 2014). There is still relatively little research on the psychometric properties of ambulatory cognitive assessments, and insufficient data to make clear cut recommendations as to whether researchers should favor a few lengthy assessment (with many trials) or multiple brief and temporally distributed cognitive assessments.

Within-person reliability—The issue of planning the temporal spacing of assessments becomes even more complex if one is interested in studying within-person variability in cognitive performance. Decisions about how often, how frequently, and over what time period to measure cognitive function should take into account the cadence of the underlying time-varying processes that may drive fluctuations in function (Ram & Gerstorf, 2009; Sliwinski, 2011; Sliwinski, Almeida, Smyth, & Stawski, 2009). For example, if most of the systematic within-person cognitive variability occurs over a relative fast time scale (e.g., across hours), then a given ambulatory cognitive test may exhibit lower within-person reliability if assessments are made only once per day compared to if assessments were made multiple times per day (Schmiedek et al., 2013). Our results showed that the proportion of systematic within-person variance that transpired *across* occasions (separated by a few hours) relative to within-person variability transpiring *within* occasions varied between .41 and .53, depending upon the task.

Very little work has examined within-person reliabilities of cognitive assessments. A few studies have done so in adult samples using measurements made in laboratory settings (Brose, Lövdén, & Schmiedek, 2014; Sliwinski et al., 2006) and in children assessed in the classroom and in their homes (Dirk and Schmiedek, in press). The present results extend this research in two ways. First, our study examined the reliability of within-person fluctuations in ambulatory cognitive performance among adults (ages 25–65). And second, the ambulatory assessments were made at unpredictable (i.e., pseudo-random) times and took place in random locations and situations. Thus, the present study demonstrates the feasibility of embedding cognitive tests in EMA protocols for studying the systematic effects of time-varying and contextual influences on cognitive function among adults measured in naturalistic settings.

That said, the within-person reliabilities (.41–.53) were lower than the between-person reliabilities (>.95 for all tasks). For comparison, previous laboratory based intensive repeated measurement studies have reported within-person reliabilities in the range of .50–.70 for response time tasks (Brose et al., 2014; Sliwinski et al., 2006) and in the range of .15–.36 for accuracy measures (Brose et al., 2014). In an ambulatory study, Dirk and Schmiedek (in press) reported within-person reliabilities for working memory tasks (accuracy scores) in the range of .58–.70 for 3rd and 4th grade students assessed in the classroom and in their homes.

Although the within-person reliabilities of the present set of brief ambulatory cognitive tasks are comparable to previously reported values from other tests, these values raise the question of whether the within-person reliabilities of cognitive tests are sufficient to detect the effects of contextual and time-varying variables, such as stress, mood or fatigue, on performance.

Indeed, a number of studies have reported significant effects of time-varying variables on dependent measures with low to modest within-person reliability. For example, the negative affect scale used in the National Study of Daily Experiences (NSDE), which is part of the Midlife in the United States study (MIDUS Study), has a within-person reliability of only .55 (Charles, Piazza, Mogle, Sliwinski, & Almeida, 2013), yet there are dozens of publications from this data set showing time-varying influences on the negative affect that were detectable. Other studies reporting low to modest (.18–.70) within-person reliabilities of cognitive measures have been able to detect time-varying effects mood, motivation and stress (Broese et al., 2014, 2012; Sliwinski et al., 2006). However, the effects involving lower reliability measures will be attenuated compared to effects involving measures with higher reliability. This attenuation decreases statistical power for detecting an expected effect for a given sample size and design. Researchers planning future studies using ambulatory cognitive assessments should power their studies to detect within-person effects that are much smaller than the expected effect sizes. Lower within-person reliabilities, however, may be tolerable for intensive repeated measures studies (e.g. daily diary and EMA designs) because these designs have many observations which can partially offset the effects of unreliability.

One obvious approach to increasing within-person reliability is to increase the number of trials made at each assessment. According to equation 2, the within-person reliability of the dot memory task could theoretically be increased to .80 if the number of trials were increased from 2 to 8. Doing so, however, would increase the assessment time on each occasion, adding to participant burden and possibly having an adverse effect on compliance. Moreover, increasing the length of cognitive tests administered in an ambulatory setting might increase the likelihood of an external distraction disrupting performance. Finally, certain time-varying influences, such as stress, could increase the trial-to-trial level variability on a given occasion (Sliwinski et al., 2006), effectively reducing within-person reliability of across occasion measurements. The systematic study of procedures for ambulatory cognitive assessment is in its very early stages and we know very little about how to design cognitive tests that have optimal psychometric properties to detect systematic within-person effects, particularly across short time scales. Additional work is required to determine how to optimize within-person reliability of ambulatory cognitive assessment procedures.

Validity of Ambulatory Cognitive Assessments

Two findings support the construct validity of the ambulatory working memory and perceptual speed tasks used in this study. First, the ambulatory dot-memory and symbol search tasks exhibited similar correlations with age and fluid intelligence as the in-lab working memory and perceptual speed tasks, respectively. The ambulatory N-back task also correlated with fluid intelligence; however, it was not significantly correlated with age. The

lack of association between the N-back and age likely resulted from the low difficulty level of 2-back procedures (compared to, for example, 3-back tasks), and is consistent with other studies that have found no age differences on 2-back tasks administered in a laboratory setting (Schmiedek, Li, & Lindenberger, 2009; Wild-Wall, Falkenstein, & Gajewski, 2011). In addition, the ambulatory N-back task exhibited roughly equal first order correlations with working memory and perceptual speed indicators.) This may have been due to the N-back being a speeded task and it not being as not sufficiently difficult. Together, these two features of the ambulatory N-back task may have resulted in performance that was largely constrained by processing speed rather than working memory capacity because the working memory demands were relatively light. Despite its correlation with fluid intelligence and its high loading on the working memory factor, the ambulatory N-back task used in the current study did not exhibit good discriminant validity—it correlated as strongly with perceptual speed measures as it did with working memory measures. Therefore, the version of the ambulatory n-back task used in this study may not be a useful indicator of individual differences.

Second, a CFA showed that the ambulatory tasks loaded significantly and strongly on their hypothesized factors. Because these factors were also defined by standard in-lab cognitive tests, this result indicates that ambulatory working memory and speed tasks measure the same underlying constructs that researchers study in controlled testing environments. Although using different cognitive tests for the ambulatory and in-lab assessments precluded making direct comparisons between assessment modes, it also made the present demonstration of construct validity even more compelling. That is, evidence that the ambulatory and in-lab tests measure the same underlying constructs cannot be attributed to shared method variance that could have resulted by using the same procedures for ambulatory and in-lab assessments. Supplementary analyses demonstrated that evidence of construct validity was not a by-product of the age heterogeneity of the sample and was age invariant.

It is also important to note, however, that these analyses only support the between-person construct validity of the ambulatory tasks for analyses of individual differences. We were not able to examine the construct validity at the within-person level since we did not have a sufficient number of ambulatory indicators (i.e., 3) of the working memory and speed constructs. Additionally, we did not explore possible validity threats to the ambulatory tasks, such as the impact of external distractions (e.g., noise). That said, our results provide a novel demonstration of a strong pattern of correlations between ambulatory cognitive tests administered in uncontrolled settings and standard working memory and processing speed tests administered in the laboratory.

Limitations and Considerations for Future Research

Although our results contribute to the literature on ambulatory cognitive assessment, they must be considered with respect to several limitations. First, the same procedures were not used in both ambulatory and in the lab, precluding direct comparison of performance differences between ambulatory and in-lab cognitive assessments. Some people might perform systematically worse (or better) in naturalistic environments compared to laboratory

or clinical settings, perhaps due to features of their everyday environment, features of the testing environment, or their dispositions (e.g., test anxiety). In particular, ambulatory cognitive assessments might be more sensitive than in-lab assessments to contextual and time-varying influences, such as stress, physical symptoms, social activity and health behaviors.

A second limitation is that each ambulatory cognitive assessment consisted of relatively few trials per task in order to minimize participant burden. One negative consequence is that the small number of trials per occasion may have negatively impacted within-person reliability. In addition, the small number of trials precluded in-depth examination of variability in performance across faster time scales (i.e., within occasions) which other studies have shown to exist (Dirk & Schmiedek, in press). Although systematic within-person variability may operate across time scales that we did not examine (e.g., across minutes), our findings demonstrated that fluctuations in memory and speeded performance throughout the course of a day (across hours) reflects systematic sources of intraindividual variability.

A third limitation is that we did not sample a broad range of memory functioning (e.g., secondary and primary memory) in either the laboratory or ambulatory setting. Secondary memory is strongly correlated with, and is an important component of, working memory (Unsworth & Spillers, 2010). Therefore, we cannot rule out possibility that the ambulatory dot memory task is a better indicator of secondary than working memory.

A fourth consideration is that our protocol for sampling occasions prompted individuals to complete their cognitive assessments at unpredictable times during the day in order to obtain a representative sampling of situations. This type of repeated and random sampling of occasions allows ambulatory methods to characterize a person's *average* (or typical) cognitive performance over a wide variety of contexts in their daily lives with a high degree of reliability. This approach to cognitive assessment, however, differs from traditional laboratory or clinic-based approaches which seek to obtain an individual's best or *maximum* rather than their average or *typical* performance. For some purposes, such as assessing scholastic aptitude, measuring an individual maximum performance might be ideal. However, for other purposes, such as detecting subtle cognitive impairment or the influence of social context on cognitive function, a person's average performance may be more sensitive than their maximum performance.

Finally, the analyses reported do not yet demonstrate value added to conducting ambulatory vs. in-lab cognitive assessments. Additional research is required to study the potential added value by obtaining intensive repeated measurements of cognition in ecological settings. Such demonstrations might involve comparing the predictive validity of ambulatory with in-lab cognitive assessments for some important health outcome (e.g., incident Alzheimer's disease or mild cognitive impairment) or increased sensitivity to detecting cognitive change due to much higher reliability compared to single-shot assessments (Sliwinski, 2008). Other demonstrations could focus on identifying contextual and 'real-time' determinants of cognitive function (such as emotional distress or rumination) as potential targets for time-sensitive interventions delivered through mobile technology (Heron & Smyth, 2010). We hope that these findings support and motivate future research to utilize ambulatory methods

to study the influences of environmental, social, psychological, physiological and behavioral factors on cognitive function in people's natural environments.

Acknowledgments

This research was supported by National Institute of Health (NIH) grants R01 AG039409, R01 AG042595, P01 AG03949, CTSA 1UL1TR001073 from the National Center for Advancing Translational Sciences (NCATS), the Leonard and Sylvia Marx Foundation, and the Czup Foundation.

References

- Ackerman PL, Beier ME. Further explorations of perceptual speed abilities in the context of assessment methods, cognitive abilities, and individual differences during skill acquisition. *Journal of Experimental Psychology Applied*. 2007; 13(4):249–272. <http://doi.org/10.1037/1076-898X.13.4.249>. [PubMed: 18194049]
- Allard M, Husky M, Catheline G, Pelletier A, Dilharreguy B, Amieva H, Swendsen J. Mobile Technologies in the Early Detection of Cognitive Decline. *PLoS ONE*. 2014; 9(12):e112197. <http://doi.org/10.1371/journal.pone.0112197>. [PubMed: 25536290]
- Alloway TP, Alloway RG. Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*. 2010; 106(1):20–29. <http://doi.org/10.1016/j.jecp.2009.11.003>. [PubMed: 20018296]
- Brose A, Lövdén M, Schmiedek F. Daily fluctuations in positive affect positively co-vary with working memory performance. *Emotion (Washington, DC)*. 2014; 14(1):1–6. <http://doi.org/10.1037/a0035210>.
- Brose A, Schmiedek F, Lövdén M, Lindenberger U. Daily variability in working memory is coupled with negative affect: The role of attention and motivation. *Emotion*. 2012; 12(3):605–617. <http://doi.org/10.1037/a0024436>. [PubMed: 21787075]
- Brunswik E. Organismic achievement and environmental probability. *Psychological Review*. 1943; 50(3):255–272. <http://doi.org/10.1037/h0060889>.
- Charles ST, Piazza JR, Mogle J, Sliwinski MJ, Almeida DM. The wear and tear of daily stressors on mental health. *Psychological Science*. 2013; 24(5):733–741. <http://doi.org/10.1177/0956797612462222>. [PubMed: 23531486]
- Conway ARA, Cowan N, Bunting MF, Theriault DJ, Minkoff SRB. A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*. 2002; 30(2):163–183. [http://doi.org/10.1016/S0160-2896\(01\)00096-4](http://doi.org/10.1016/S0160-2896(01)00096-4).
- 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? *Personality & Social Psychology Bulletin*. 2006; 32(7):917–929. <http://doi.org/10.1177/0146167206287721>. [PubMed: 16738025]
- Engle RW, Tuholski SW, Laughlin JE, Conway A. Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*. 1999; 128(3):309–331. <http://doi.org/10.1037/0096-3445.128.3.309>. [PubMed: 10513398]
- Fahrenberg J, Myrtek M, Pawlik K, Perrez M. Ambulatory assessment—Monitoring behavior in daily life settings: A behavioral-scientific challenge for psychology. *European Journal of Psychological Assessment*. 2007; 23(4):206–213. <http://doi.org/10.1027/1015-5759.23.4.206>.
- Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology*. 2010; 15(Pt 1):1–39. <http://doi.org/10.1348/135910709X466063>. [PubMed: 19646331]
- Hultsch DF, MacDonald SW, Hunter MA, Levy-Bencheton J, Strauss E. Intraindividual variability in cognitive performance in older adults: comparison of adults with mild dementia, adults with arthritis, and healthy adults. *Neuropsychology*. 2000; 14(4):588–598. [PubMed: 11055261]
- Just MA, Carpenter PA. A capacity theory of comprehension: individual differences in working memory. *Psychological Review*. 1992; 99(1):122–149. [PubMed: 1546114]

- MacDonald SWS, Li SC, Bäckman L. Neural underpinnings of within-person variability in cognitive functioning. *Psychology and Aging*. 2009; 24(4):792–808. <http://doi.org/10.1037/a0017798>. [PubMed: 20025396]
- Miyake A, Friedman NP, Rettinger DA, Shah P, Hegarty M. How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology General*. 2001; 130(4):621–640. [PubMed: 11757872]
- Nesselroade, JR. The warp and the woof of the developmental fabric. In: Downs, RM, Liben, LS., Palermo, DS., editors. *Visions of aesthetics, the environment & development: The legacy of Joachim F Wohlwill*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.; 1991. p. 213-240.
- Owsley C, Ball K, McGwin G, Sloane ME, Roenker DL, White MF, Overley ET. Visual processing impairment and risk of motor vehicle crash among older adults. *JAMA*. 1998; 279(14):1083–1088. [PubMed: 9546567]
- Ram N, Gerstorf D. Time-structured and net intraindividual variability: Tools for examining the development of dynamic characteristics and processes. *Psychology and Aging*. 2009; 24(4):778–791. <http://doi.org/10.1037/a0017915>. [PubMed: 20025395]
- Raven J. The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology*. 2000; 41(1):1–48. <http://doi.org/10.1006/cogp.1999.0735>. [PubMed: 10945921]
- Raykov T, Marcoulides GA. On Multilevel Model Reliability Estimation From the Perspective of Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*. 2006; 13(1):130–141. http://doi.org/10.1207/s15328007sem1301_7.
- Riediger M, Wrzus C, Klipker K, Müller V, Schmiedek F, Wagner GG. Outside of the laboratory: Associations of working-memory performance with psychological and physiological arousal vary with age. *Psychology and Aging*. 2014; 29(1):103–114. <http://doi.org/10.1037/a0035766>. [PubMed: 24660800]
- Rucklidge JJ, Tannock R. Neuropsychological profiles of adolescents with ADHD: effects of reading difficulties and gender. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*. 2002; 43(8):988–1003.
- Salthouse TA. Aging and measures of processing speed. *Biological Psychology*. 2000; 54(1–3):35–54. [PubMed: 11035219]
- Schmader T, Johns M, Forbes C. An Integrated Process Model of Stereotype Threat Effects on Performance. *Psychological Review*. 2008; 115(2):336–356. <http://doi.org/10.1037/0033-295X.115.2.336>. [PubMed: 18426293]
- Schmiedek F, Li SC, Lindenberger U. Interference and facilitation in spatial working memory: age-associated differences in lure effects in the n-back paradigm. *Psychology and Aging*. 2009; 24(1): 203–210. <http://doi.org/10.1037/a0014685>. [PubMed: 19290752]
- Schmiedek F, Lövdén M, Lindenberger U. Keeping It Steady Older Adults Perform More Consistently on Cognitive Tasks Than Younger Adults. *Psychological Science*. 2013; 24(9):1747–1754. <http://doi.org/10.1177/0956797613479611>. [PubMed: 23842960]
- Schmiedek F, Lövdén M, Lindenberger U. A task is a task is a task: putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in Psychology*. 2014; 5 <http://doi.org/10.3389/fpsyg.2014.01475>.
- Schneider-Garces NJ, Gordon BA, Brumback-Peltz CR, Shin E, Lee Y, Sutton BP, Fabiani M. Span, CRUNCH, and beyond: working memory capacity and the aging brain. *Journal of Cognitive Neuroscience*. 2010; 22(4):655–669. <http://doi.org/10.1162/jocn.2009.21230>. [PubMed: 19320550]
- Scott SB, Graham-Engeland JE, Engeland CG, Smyth JM, Almeida DM, Katz MJ, Sliwinski MJ. The Effects of Stress on Cognitive Aging, Physiology and Emotion (ESCAPE) Project. *BMC Psychiatry*. 2015; 15(1) <http://doi.org/10.1186/s12888-015-0497-7>.
- Shiffman S, Patten C, Gwaltney C, Paty J, Gnys M, Kassel J, Balabanis M. Natural history of nicotine withdrawal. *Addiction (Abingdon, England)*. 2006; 101(12):1822–1832. <http://doi.org/10.1111/j.1360-0443.2006.01635.x>.
- Shiffman S, Paty JA, Gnys M, Kassel JD, Elash C. Nicotine withdrawal in chippers and regular smokers: Subjective and cognitive effects. *Health Psychology*. 1995; 14:301–309. <http://doi.org/http://dx.doi.org.ezaccess.libraries.psu.edu/10.1037/0278-6133.14.4.301>. [PubMed: 7556033]

- Shiffman S, Stone AA, Hufford MR. Ecological Momentary Assessment. *Annual Review of Clinical Psychology*. 2008; 4(1):1–32. <http://doi.org/10.1146/annurev.clinpsy.3.022806.091415>.
- Siedlecki KL. Investigating the structure and age invariance of episodic memory across the adult lifespan. *Psychology and Aging*. 2007; 22(2):251–268. <http://doi.org/10.1037/0882-7974.22.2.251>. [PubMed: 17563181]
- Sliwinski MJ. Measurement-burst designs for social health research. *Social and Personality Psychology Compass*. 2008; 2(1):245–261.
- Sliwinski, MJ. Approaches to modeling intraindividual and interindividual facets of change for developmental research. In: Fingerman, KL, Berg, CA, Smith, J., Antonucci, TC., editors. *Handbook of life-span development*. New York, NY, US: Springer Publishing Co.; 2011. p. 1-25.
- Sliwinski MJ, Almeida DM, Smyth J, Stawski RS. Intraindividual change and variability in daily stress processes: Findings from two measurement-burst diary studies. *Psychology and Aging*. 2009; 24(4):828–840. <http://doi.org/10.1037/a0017925>. [PubMed: 20025399]
- Sliwinski MJ, Smyth JM, Hofer SM, Stawski RS. Intraindividual coupling of daily stress and cognition. *Psychology and Aging*. 2006; 21(3):545–557. <http://doi.org/10.1037/0882-7974.21.3.545>. [PubMed: 16953716]
- Smyth JM, Stone AA. Ecological Momentary Assessment Research in Behavioral medicine. *Journal of Happiness Studies*. 2003; 4(1):35–52. <http://doi.org/10.1023/A:1023657221954>.
- Spooner D, Pachana N. Ecological validity in neuropsychological assessment: A case for greater consideration in research with neurologically intact populations. *Archives of Clinical Neuropsychology*. 2006; 21(4):327–337. <http://doi.org/10.1016/j.acn.2006.04.004>. [PubMed: 16769198]
- Stenneken P, Egetemeir J, Schulte-Körne G, Müller HJ, Schneider WX, Finke K. Slow perceptual processing at the core of developmental dyslexia: a parameter-based assessment of visual attention. *Neuropsychologia*. 2011; 49(12):3454–3465. <http://doi.org/10.1016/j.neuropsychologia.2011.08.021>. [PubMed: 21903119]
- Strauss B. Social facilitation in motor tasks: a review of research and theory. *Psychology of Sport and Exercise*. 2002; 3(3):237–256. [http://doi.org/10.1016/S1469-0292\(01\)00019-X](http://doi.org/10.1016/S1469-0292(01)00019-X).
- Timmers C, Maeghs A, Vestjens M, Bonnemayer C, Hamers H, Blokland A. Ambulant cognitive assessment using a smartphone. *Applied Neuropsychology: Adult*. 2014; 21:136–142. <http://doi.org/10.1080/09084282.2013.778261>. [PubMed: 24826507]
- Tiplady B, Oshinowo B, Thomson J, Drummond GB. Alcohol and cognitive function: assessment in everyday life and laboratory settings using mobile phones. *Alcoholism, Clinical and Experimental Research*. 2009; 33(12):2094–2102. <http://doi.org/10.1111/j.1530-0277.2009.01049.x>.
- Unsworth N, Engle RW. The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*. 2007; 114(1):104–132. <http://doi.org/10.1037/0033-295X.114.1.104>. [PubMed: 17227183]
- Unsworth N, Spillers GJ. Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model. *Journal of Memory and Language*. 2010; 62(4):392–406. <http://doi.org/10.1016/j.jml.2010.02.001>.
- Verhaeghen P, Salthouse TA. Meta-analyses of age–cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin*. 1997; 122(3):231–249. <http://doi.org/10.1037/0033-2909.122.3.231>. [PubMed: 9354147]
- Waters AJ, Li Y. Evaluating the utility of administering a reaction time task in an ecological momentary assessment study. *Psychopharmacology*. 2008; 197(1):25–35. <http://doi.org/10.1007/s00213-007-1006-6>. [PubMed: 18064440]
- Waters AJ, Marhe R, Franken IHA. Attentional bias to drug cues is elevated before and during temptations to use heroin and cocaine. *Psychopharmacology*. 2012; 219(3):909–921. <http://doi.org/10.1007/s00213-011-2424-z>. [PubMed: 21833505]
- Waters GS, Caplan D. The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*. 2003; 35(4):550–564. <http://doi.org/10.3758/BF03195534>.

- Wild-Wall N, Falkenstein M, Gajewski PD. Age-Related Differences in Working Memory Performance in A 2-Back Task. *Frontiers in Psychology*. 2011; 2 <http://doi.org/10.3389/fpsyg.2011.00186>.
- Yoon M, Millsap RE. Detecting Violations of Factorial Invariance Using Data-Based Specification Searches: A Monte Carlo Study. *Structural Equation Modeling: A Multidisciplinary Journal*. 2007; 14(3):435–463. <http://doi.org/10.1080/10705510701301677>.

Figure 1.
Example of symbol search test. Participants were asked to decide which of the two pairs at the bottom of the screen was among the pairs at the top of the screen as quickly as possible.

Figure 2.

Example of dot memory test. (a) The participants were asked to remember the location of three red dots. (b) After grid was removed, participants were required to touch F's among E's for 8 seconds. (c) Empty 5×5 grid reappeared on the screen and participants were prompted to recall the location of initial 3 dots.

Figure 3.

Example of N-back test. (a) In the practice session, participants were asked to determine whether the two cards matched. (b) Cards shifted position from right to left, with each new target card turning face down. (c) Participants were asked to determine if the current target card matched the face-down card they saw 2 trials back.

Figure 4.

The between-person reliability of average scores obtained on each of the days of ambulatory assessments.

Figure 5.
The within-person reliability estimated separately for each of the 14 days of ambulatory assessments.

Figure 6. Confirmatory Factor Model and standard coefficients. Ambulatory tasks (dot memory, n-back, and symbol search) are indicated by rectangles with thick borders.

Table 1

Descriptive statistics

Task	Mean	SD	Range	Skewness	Kurtosis	Correlation with age ^d	Correlation with Ravens ^{d,e}
Symbol search ^a	2.70	0.67	1.28–5.27	0.74	1.08	0.35***	–0.50***
Dot memory ^c	1.79	0.91	0.07–4.04	0.13	–0.72	0.16*	–0.42***
N-back ^b	0.82	0.13	0.48–0.99	–1.27	0.53	–0.04	0.47***
Symbol match ^a	3.37	0.68	1.8–5.66	0.47	0.19	0.36***	–0.43***
Letter match ^a	2.72	0.57	1.06–4.61	0.12	0.24	0.35***	–0.41***
Number match ^a	2.13	0.45	1.01–3.88	0.74	0.80	0.29***	–0.41***
Counting span ^b	19.12	12.20	0–60	0.81	0.50	–0.12	0.27***
Operation span ^b	19.77	8.21	0–38	0.02	–0.51	–0.19**	0.35***
Backward span ^b	25.17	8.46	1–51	0.08	0.69	–0.16*	0.38***

Note. N=218; Ambulatory tasks indicated by bold text;

^aUnit: second;

^bUnit: number of correct response;

^cUnit: Euclidean distance;

^dPearson Correlation Coefficients;

* < .05,

** < .01,

*** < .001;

^eRaven's Progressive Matrices

Table 2

Reliabilities for Individual and Aggregated Ambulatory Cognitive Test Scores

	Symbol search	Dot memory	N-back
between-person variance	.435	0.770	0.0177
within-person variance	.364	1.185	0.012
Reliability of 1 occasion (ICC)	.54	.39	.59
Reliability of Average (1 day)	.90	.75	.92
Reliability of Average (2 day)	.95	.86	.95
Reliability of Average (3 days)	.96	.90	.96
Reliability of Average (14 days)	.98	.97	.99

Note: ICC= intraclass correlation. There are five assessment occasions per day, so 1 day reflects 5 assessments, 2 days reflects 10 assessments, etc.

Table 3
Between-person Reliabilities for Ambulatory Cognitive Test Scores by Age Group

Age	Symbol Search		Dot Memory		N-Back	
	ICC	3 Day Average	14 Day Average	ICC	3 Day Average	14 Day Average
25–34	0.29	0.86	0.96	0.39	0.90	0.97
35–44	0.48	0.93	0.98	0.39	0.91	0.97
45–54	0.49	0.93	0.98	0.37	0.90	0.97
55+	0.58	0.95	0.99	0.42	0.92	0.98

Note. Values in table indicate reliabilities for a single assessment (ICCs) and average scores that are aggregated across either first 3 or entire 14 days of ambulatory assessments.

Table 4

Within-Person Reliabilities for Ambulatory Cognitive Test Scores by Age Group

Age	Symbol Search	Dot Memory	N-Back
25–34	0.57	0.54	0.46
35–44	0.58	0.50	0.46
45–54	0.47	0.50	0.35
55+	0.54	0.44	0.33

Table 5

Intercorrelations among ambulatory and in-lab cognitive tasks

	1	2	3	4	5	6	7	8
1. Symbol search	–							
2. Dot memory	0.38							
3. N-back	–0.32	–0.48						
4. Symbol match	0.74	0.33	–0.34					
5. Letter match	0.61	0.30	–0.29	0.76				
6. Number match	0.61	0.37	–0.29	0.79	0.83			
7. Counting span	–0.25	–0.39	0.25	–0.24	–0.26	–0.29		
8. Operation span	–0.26	–0.45	0.24	–0.22	–0.23	–0.26	0.60	
9. Backward span	–0.31	–0.43	0.36	–0.37	–0.34	–0.29	0.48	0.42

Note. Pearson Correlation Coefficients; All correlations were significant, $p<.01$. Ambulatory tasks are in bold text. Scores on the N-back and the three complex span tasks reflect accuracy (higher values=better performance); the dot memory task reflects an error score (higher values=poorer performance). Scores on the Symbol search and the three matching tasks reflect response time (higher values=poorer performance).

Table 6

Model Fit Statistics for Factorial Age-Invariance Models

Step	Invariance Level	df	χ^2	p	CFI	RMSEA	RMSEA 90% CI	df	χ^2	p
1	Configural	46	67.190	.022	.976	.046	.018 – .069	–	–	–
2	Metric	53	72.296	.040	.978	.041	.009 – .063	7	5.106	.647
3	Scalar	60	79.455	.047	.978	.039	.005 – .060	7	7.159	.413
4	Strict (Residual)	69	90.377	.043	.976	.038	.007 – .058	9	10.922	.281

Note. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; RMSEA 90% CI = 90% confidence interval for the RMSEA