Research Article

Feasibility, reliability, and validity of ambulatory cognitive tests in fibromyalgia and matched controls

Thomas R. Valentine and Anna L. Kratz

Department of Physical Medicine and Rehabilitation, University of Michigan, Ann Arbor, MI 48109, USA

Abstract

Objective: This observational study examined the feasibility, reliability, and validity of repeated ambulatory cognitive tests in fibromyalgia (FM). **Method:** Adults with FM ($n = 50$) and matched controls ($n = 50$) completed lab-based neuropsychological tests (NIH Toolbox) followed by eight days of smartphone-based ambulatory testing of processing speed (symbol search) and working memory (dot memory) five times daily. Feasibility was assessed based on response rates. Reliability was evaluated using overall average between-person reliabilities for the full assessment period and by determining the number of assessment days necessary to attain reliabilities of >.80 and >.90. To assess convergent validity, correlations were calculated between ambulatory test scores and NIH Toolbox scores. Test performance was contrasted between the FM and non-FM groups to examine known-groups validity. **Results:** Average rates of response to the ambulatory cognitive tests were 89.5% in FM and 90.0% in non-FM. Overall average between-person reliabilities were ≥.96. In FM, between-person reliability exceeded .90 after two days for symbol search and three days for dot memory. Symbol search scores correlated with NIH Toolbox processing speed scores in both groups, though there were no significant group differences in symbol search performance. Dot memory scores correlated with NIH Toolbox working memory scores in both groups. FM participants exhibited worse dot memory performance than did non-FM participants. **Conclusions:** Repeated ambulatory tests of processing speed and working memory demonstrate feasibility and reliability in FM, though evidence for construct validity is mixed. The findings demonstrate promise for future research and clinical applications of this approach to assessing cognition in FM.

Keywords: chronic pain; neuropsychological tests; ecological momentary assessment; psychometrics; cognitive symptoms; working memory

(Received 27 April 2022; final revision 15 December 2022; accepted 29 December 2022; First Published online 10 February 2023)

Introduction

Fibromyalgia (FM) is a chronic musculoskeletal disorder that affects at least 5 million US adults (Lawrence et al., 2008). In addition to widespread pain, fatigue, and unrefreshed sleep, cognitive difficulties are a common and impairing symptom of FM. In this area of research, a distinction is made between subjective cognitive difficulties (i.e., "fibrofog"), assessed via self-report, and objective cognitive difficulties, measured with neuropsychological tests. The evidence for subjective cognitive difficulties in FM is clear: approximately 70% of individuals endorse fibrofog (Katz et al., 2004), including issues with memory/learning, attention/concentration, processing speed, and executive functioning, and individuals rate fibrofog among their most troubling symptoms (Arnold et al., 2008; Bennett et al., 2007). In contrast, data regarding *objective* cognitive difficulties in FM are inconsistent. Some studies have corroborated self-reports of cognitive difficulties using neuropsychological tests, finding diminished performance across multiple cognitive domains in FM relative to non-FM controls, including processing speed, attention, learning and memory, working memory, and executive functions (Bell et al., 2018; Wu et al., 2018). Other studies have shown discrepancies between

subjective and objective cognitive functioning in FM (Suhr, 2003; Walitt et al., 2016) and limited/focal or no decrements in cognitive test performance (Grace et al., 1999; Kim et al., 2012; Landrø et al. 1997; Lee et al., 2010; Miró et al., 2011; Park et al. 2001; Suhr, 2003; Walitt et al., 2011; Walitt et al. 2008; Walteros et al., 2011).

A significant limitation of existing research on objective cognitive difficulties in FM that might explain divergent findings is the use of lab-/clinic-based neuropsychological tests administered at a single time point. Studies employing this approach lack ecological validity in that the lab/clinic setting does not resemble the real-world environment in which individuals perform cognitively demanding tasks (Sbordone, 1996; Spooner & Pachana, 2006). This issue is particularly salient when studying FM, as individuals with this condition exhibit higher susceptibility to distraction (Bell et al., 2018; Teodoro et al., 2018) and hypersensitivity to sensory stimuli (Carrillo-de-la-Peña et al., 2006; Harte et al., 2016; Hollins et al., 2009; Kosek et al., 1996; Lombion et al., 2009; Lorenz et al., 1996; et al., 2012; McDermid et al., 1996; Petzke et al., 2003) – factors that may contribute to objective cognitive difficulties in daily life but that are mitigated in a controlled testing environment. Further, subtle cognitive changes in FM may elude

traditional neuropsychological testing. Administering these tests at a single time point may fail to capture intraindividual variations in cognitive performance that indicate poor cognitive functioning (Ram et al., 2005; West et al., 2002) and risk for future cognitive decline (Bielak et al., 2010; Bielak et al., 2010). Performance on a single occasion is also susceptible to situational factors, such as a poor night's sleep and psychosocial and environmental stressors, whereas repeated testing over time captures an individual's average performance.

Ambulatory cognitive testing offers an alternative to standard neuropsychological testing in which objective cognitive functioning can be assessed in the lived environment and on repeated occasions, thereby increasing ecological validity and measurement reliability. Sliwinski et al. (2018) found that smartphone-based ambulatory tests of processing speed (i.e., a symbol search test) and working memory (i.e., a dot memory test) demonstrated high reliability and validity in a nonclinical sample of community adults. Whether these tests would exhibit similar psychometric properties in FM is unknown. Ambulatory cognitive tests that are feasible, reliable, and valid for administration in this population could be used in future studies of the nature, severity, and impact of cognitive difficulties, as well as to examine effects of interventions on cognitive functioning in daily life.

The goal of this observational study was to determine the feasibility, reliability, and validity of repeated ambulatory tests of processing speed and working memory in adults with FM. To determine psychometric considerations specific to FM, analyses were performed separately for individuals with FM and non-FM controls. First, we evaluated participant compliance to the testing protocol – a key determinant of feasibility – by examining average daily and within-day response rates. Second, we aggregated across test sessions to determine average between-person reliabilities, the number of testing sessions necessary to obtain high reliability, and the stability of reliabilities over time. Finally, we evaluated construct validity by examining the convergence of ambulatory cognitive tests with validated lab-based neuropsychological tests and by contrasting test scores in the FM and non-FM groups.

Method

Participants

Eligible participants were ≥ 18 years old and conversationally fluent and able to read in English at a minimum 6th grade level. Exclusion criteria were: (1) self-reported comorbid neurologic disorder, learning disorder, or cognitive impairment; (2) current or ≥ five-year history of alcohol/recreational drug dependence; (3) hearing or vision impairment that would preclude cognitive testing; (4) diagnosis of untreated obstructive sleep apnea; (5) atypical sleep/wake pattern (e.g., night-shift work). Fulfilling 2016 American College of Rheumatology diagnostic criteria for FM (Wolfe et al., 2016) was inclusionary for participants in the FM group and exclusionary for participants in the control group. Participants in the control group were age-, sex-, and education-matched to participants in the FM group.

Procedures

Study procedures were approved by the Medical Institutional Review Board of the University of Michigan. The research was completed in accordance with the Helsinki Declaration. Volunteers were recruited from January 2018 through August 2018 via patient registries, community groups, fliers in health centers and community settings, and advertisements on a university-based recruitment website. Volunteers were phone-screened for eligibility. Eligible participants provided written informed consent.

Study participation involved an in-person baseline visit followed by home monitoring consisting of ambulatory cognitive testing embedded within an EMA protocol. The 90-minute baseline visit included completing a battery of self-report measures and standardized cognitive tests and receiving training in home monitoring procedures (e.g., data collection device usage). Home monitoring commenced following the baseline visit and lasted eight days, in line with typical ecological momentary assessment protocols (de Vries et al., 2021). Ambulatory cognitive tests were administered using a study-specific application installed on a ZTE Axon 7 mini smartphone (5.2" display; 1080 × 1920 pixels). Participants underwent a training session in which they were shown how to use the smartphone app and practiced until they were able to demonstrate mastery of the app. Participants were instructed to keep the smartphone with them at all times except when it was important they not be disturbed by the audible alert. The time of baseline visit completion determined the number of test sessions administered on day 1. On all subsequent days, participants completed five daily cognitive test sessions, consistent with typical EMA protocols (de Vries et al., 2021) and prior study using these tests (Sliwinski et al., 2018). The first of the daily cognitive test sessions was initiated by the participant upon waking. The four test sessions that followed were prompted via audible smartphone alerts that were programmed on a quasi-random schedule based on each individual's typical waking time. Between-prompt intervals ranged 3 to 4.5 hours. After the home monitoring period, participants returned the smartphones to the laboratory via postage-paid return boxes. They were compensated up to $175 for their participation.

Measures

Baseline sociodemographic characteristics. Age, gender, race, ethnicity, years of education, and employment status were self-reported.

Baseline cognitive tests. At baseline, four cognitive tests were administered using the NIH Toolbox (Gershon et al., 2010) iPad application (Brearly et al., 2019). On each test, performance was indicated by either T-scores (mean $\pm$ SD = 50 $\pm$ 10) corrected for age, sex, race, ethnicity, and education, or uncorrected standard scores (mean $\pm$ SD = 100 $\pm$ 15). Higher scores reflected better performance.

(1) The pattern comparison test assessed processing speed. Participants were asked to identify whether pairs of visual stimuli were identical. They were allotted 85 seconds to complete as many trials as possible. (2) The list sorting test assessed working memory. Participants were presented with series of visual and auditory stimuli (e.g., animals) which they were asked to recall in sequence based on a specific dimension (e.g., size). (3) The flanker test assessed attention and inhibitory control. Participants were asked to focus their attention on a target stimulus (i.e., an arrow) while distractor stimuli (i.e., arrows faced toward or away from the target stimulus) flanked the target. On each trial, participants were asked to indicate the direction of the target stimulus. (4) The dimensional change card sort test assessed cognitive flexibility and attention. On each trial, participants were presented with a visual stimulus which they were asked to match to a target stimulus according

to alternating criteria (i.e., shape or color) as indicated by a cue word on the screen.

Ambulatory cognitive tests. Two brief cognitive tests were administered five times daily using a study-specific smartphone application (Sliwinski et al., 2018).

(1) The symbol search test assessed processing speed. Each test session consisted of sixteen trials. In each trial, participants were asked to indicate by touch which of two symbol pairs at the bottom of the screen exactly matched one of four symbol pairs at the top of the screen. Seventy-five percent of trials included a lure stimulus in which one of the two symbols within a pair at the bottom of the screen matched a symbol at the top of the screen, but the pair itself did not match. The trial ended when the participant made a selection. Variables of interest included the mean, median, and standard deviation of reaction times (milliseconds) for each session. As poor task effort produces invalid scores on neuropsychological tests, sessions with <70% accuracy were excluded from analyses. This cut point is consistent with the procedures used when validating these measures in nonclinical community adults and permits excluding scores likely produced by rote responding (i.e., indiscriminate selection), wherein around 50% accuracy is expected, or intentional poor performance ("faking bad"), wherein less than 50% accuracy is expected.

(2) The dot memory test assessed working memory. During each test session, participants completed four trials, each consisting of encoding, distraction, and retrieval phases. During the encoding phase, participants were allotted three seconds to memorize the location of three red dots appearing on a 5×5 square grid. The grid then disappeared, and the four-second distraction phase commenced, during which the participants were instructed to touch all F's in an array of E's. Finally, during the retrieval phase, participants were shown an empty 5×5 square grid and were asked to touch the squares corresponding with the locations of the three red dots as presented during the initial encoding phase. The trial ended when the participant pressed "Done." The Euclidian distance – the collective distance of the three dots from their correct locations – was calculated for each trial. Variables of interest included the mean, maximum, and standard deviation of the Euclidian distances for each session.

Data analyses

Analyses were performed using IBM SPSS Statistics (V26). Descriptive statistics were generated for sociodemographic characteristics and study variables. To evaluate feasibility, average daily and within-day response rates were calculated separately for participants in the FM and non-FM groups for the full home monitoring period. Average daily response rates were calculated for each day (i.e., days 1, 2, 3, 4, 5, 6, 7, and 8) by dividing the total number of test sessions completed by the total number of possible test sessions during the respective day. Average within-day response rates were calculated for each time point (i.e., 1, 2, 3, 4, and 5) by dividing the total number of test sessions completed by the total number of possible test sessions during the respective time point. To evaluate reliability, unconditional means multilevel models were generated for each ambulatory cognitive test in each group. The between-person and within-person variances were used to calculate intraclass correlations (ICC), wherein

ICC = between-person variance/(between-person variance + within-person variance). Between-person reliabilities for the full home monitoring period were calculated as between-person reliability = between-person variance/(between-person variance + within-person variance/n), wherein n referred to the average number of completed sessions. Additional unconditional means multilevel models were run using test scores from assessment periods of increasing duration (e.g., day 1 alone, days 1–2, days 1–3, etc.) to determine the number of days of assessment necessary to obtain between-person reliabilities of >.80 and >.90 for each ambulatory cognitive test in each group. Separate models were also run for each day ($n = 8$) to determine the stability of between-person reliabilities. In preparation for validity analyses, person-averaged variables for symbol search performance (mean, median, and SD of reaction times) and dot memory performance (mean, maximum, and SD of the Euclidian distances) were created by averaging each participant's respective test session scores across the home monitoring period. To evaluate known-groups validity, the person-averaged ambulatory cognitive test scores and NIH Toolbox T-test scores were compared between the FM and non-FM groups using independent samples t-tests. Known-groups validity would be supported by significant group differences in performance on the ambulatory tests that parallel group differences in performance on the NIH Toolbox tests. To evaluate convergent validity, Pearson's correlations were examined between the person-averaged ambulatory cognitive test scores and the NIH Toolbox standard scores in the FM and non-FM groups. Higher correlations between the ambulatory symbol search test and NIH Toolbox pattern comparison test of processing speed, as well as between the ambulatory dot memory test and NIH Toolbox list sorting test of working memory, would evidence higher convergent validity. Both NIH Toolbox T-scores and uncorrected standard scores were used for describing test performance in the FM and non-FM groups and for evaluating known-groups validity. Only NIH Toolbox uncorrected standard scores were used in correlational analyses assessing convergent validity, to optimize comparison with the ambulatory test scores, for which demographically adjusted test scores are not available.

Results

Sample characteristics. One-hundred individuals participated, including 50 with FM and 50 non-FM controls. Participants were 45.1 years old on average (range = 18–73 years), were predominantly female (88%), and had an average of 15.7 years of education. They were White (81%), Black/African American (13%), Asian (3%), or multiracial (3%), and 94% were not Hispanic or Latino/a. Participants with FM reported a higher rate of unemployment (FM = 40%, non-FM = 22%; $X^2(1) = 5.88$, $p = .02$).

Feasibility of ambulatory cognitive tests. The FM group provided data for 89.5% of ambulatory cognitive test sessions and the non-FM group provided data for 90.0% of sessions. Average daily response rates ranged 84.2% to 94.4% for the FM group and 85.6% to 94.2% for the non-FM group (Figure 1). Average within-day response rates ranged 85.8% to 94.9% for the FM group and 87.2% to 94.2% for the non-FM group (Figure 2).

Reliability of symbol search test. Between-person differences accounted for 65% of the total variance in symbol search test performance in the FM group and 61% in the non-FM group (Table 1). Overall average between-person reliabilities were .98 for both groups. For the FM group, between-person reliability

Table 1. Between-person reliabilities of individual and aggregated ambulatory cognitive test scores in individuals with FM ($n = 50$) and non-FM controls ($n = 50$)

|	Symbol search		Dot memory	
	FM ($n = 50$)	Non-FM ($n = 50$)	FM ($n = 50$)	Non-FM ($n = 50$)
Between-person variance	558155.09	382737.41	0.70	0.43
Within-person variance	295912.62	245378.73	0.63	0.55
Intraclass correlation	0.65	0.61	0.53	0.44
Reliability for day 1	0.88	0.91	0.41	0.18
Reliability for days 1-2	0.94	0.95	0.85	0.80
Reliability for days 1-3	0.96	0.96	0.92	0.89
Reliability for days 1-4	0.97	0.97	0.94	0.92
Reliability for days 1-8	0.98	0.98	0.97	0.96

Note. FM = fibromyalgia.
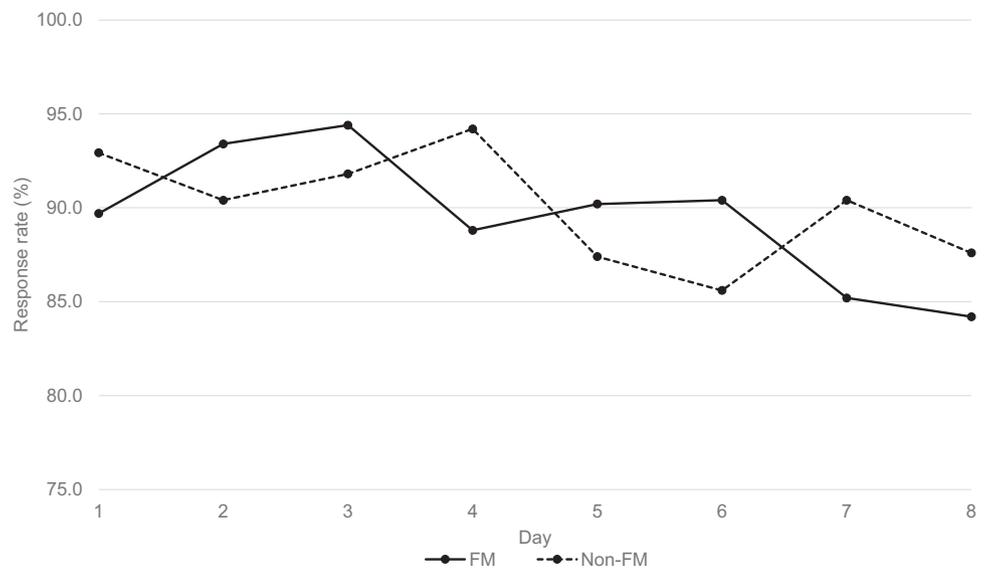
Figure 1. Average daily response rates (%) for ambulatory cognitive tests in individuals with FM ($n = 50$) and non-FM controls ($n = 50$).
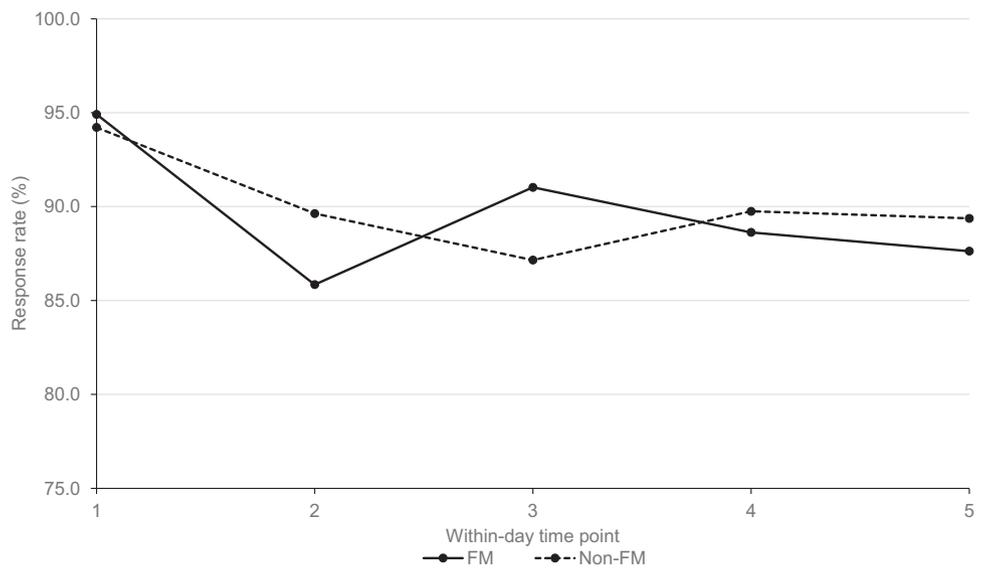

Figure 2. Average within-day response rates (%) for ambulatory cognitive tests in individuals with FM ($n = 50$) and non-FM controls ($n = 50$).

exceeded .80 after one day of assessment (reliability = .88) and .90 after two days (reliability = .94). For the non-FM group, between-person reliability exceeded both thresholds after one day of assessment (reliability = .91). The reliabilities of daily average scores were stable across the full assessment period, ranging .88 to .94 for the FM group and .90 to .93 for the non-FM group (Figure 3).

Reliability of dot memory test. Between-person differences accounted for 53% of the total variance in symbol search test

Table 2. Ambulatory and NIH Toolbox cognitive test scores in individuals with FM (n=50) and non-FM controls (n = 50)

	FM (n = 50)	Non-FM (n = 50)	t-test	p
Ambulatory cognitive tests				
Symbol search (msec)				
Mean response time	2493.81 (755.75)	2311.34 (625.15)	1.32	.19
Median response time	2304.99 (719.31)	2131.99 (582.34)	1.32	.19
SD of response times	1042.64 (342.86)	934.01 (314.87)	1.65	.10
Dot memory (Euclidian distance)				
Mean error score	1.58 (.85)	1.07 (.67)	3.31	.001
Maximum error score	3.00 (1.03)	2.31 (1.04)	3.34	.001
SD of error score	1.20 (.29)	.99 (.37)	3.12	.002
NIH Toolbox cognitive tests				
Pattern comparison (T-score)	49.76 (16.21)	57.36 (14.44)	−2.48	.02
List sorting (T-score)	49.34 (10.66)	53.18 (8.32)	−2.01	<.05
Flanker test (T-score)	39.98 (9.50)	43.78 (8.17)	−2.14	.03
Dimensional change card sort (T-score)	46.38 (11.94)	54.76 (13.20)	−3.33	.001
Pattern comparison (uncorrected standard score)	103.00 (21.85)	110.18 (17.97)	−1.79	.08
List sorting (uncorrected standard score)	102.60 (12.47)	106.46 (9.99)	−1.71	.09
Flanker test (uncorrected standard score)	95.76 (10.41)	99.12 (7.47)	−1.85	.07
Dimensional change card sort (uncorrected standard score)	100.62 (10.11)	106.66 (6.78)	−3.51	<.001

Note. FM = fibromyalgia. Values are mean (SD) unless indicated otherwise.
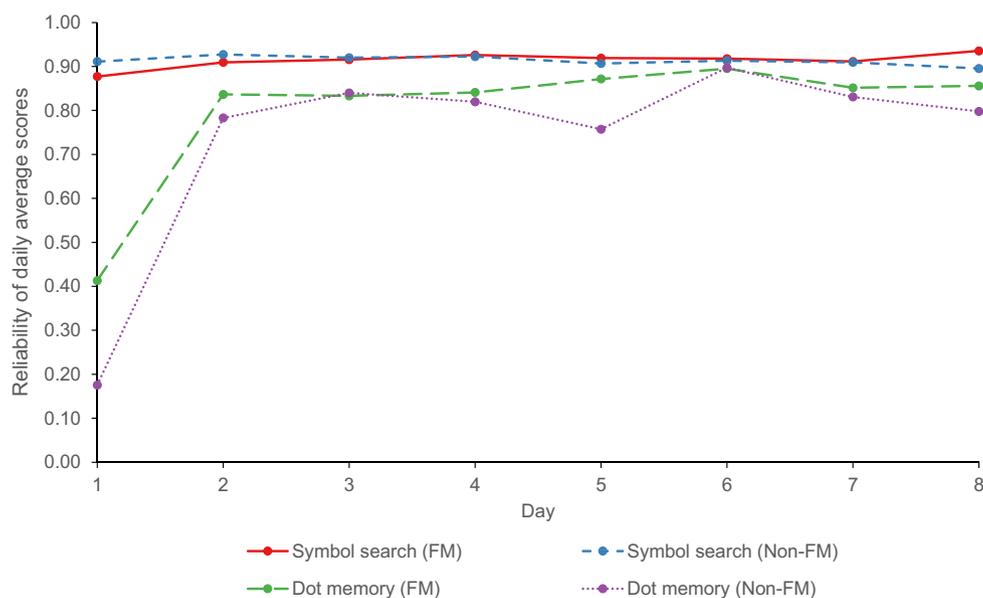

Figure 3. Between-person reliabilities of mean ambulatory cognitive test scores obtained on each study day for individuals with FM (n = 50) and non-FM controls (n = 50).

performance in the FM group and 44% in the non-FM group (Table 1). Overall average between-person reliabilities were .97 for the FM group and .96 for the non-FM group. For the FM group, between-person reliability exceeded .80 after two days of assessment (reliability = .85) and .90 after three days (reliability = .92). For the non-FM group, between-person reliability exceeded .80 after two days of assessment (reliability = .80) and .90 after four days (reliability = .92). The reliabilities of daily average scores were low on day 1 (reliability in FM group = .41; reliability in non-FM group = .18) but were markedly higher and stable for days 2 through 8, ranging .83 to .90 for the FM group and .76 to .90 for the non-FM group (Figure 3).

Validity of symbol search test. The FM and non-FM groups did not significantly differ in performance on the symbol search test, neither as measured by aggregate mean reaction time ($t(98) = 1.32$, $p = .19$), median reaction time ($t(98) = 1.32$, $p = .19$), nor SD of reaction time ($t(98) = 1.65$, $p = .10$) (Table 2). In contrast, on the NIH Toolbox pattern comparison test of processing speed, the FM group exhibited significantly worse

T-scores than did the non-FM control group ($t(98) = −2.48$, $p = .02$), although group differences in uncorrected standard scores were not statistically significant ($p = .08$). Within the FM group, worse symbol search test performance was correlated with worse performance on all four NIH Toolbox tests, including pattern comparison ($r = −.66$, $p < .001$), list sorting ($r = −.45$, $p = .001$), flanker ($r = −.51$, $p < .001$), and dimensional change card sort ($r = −.51$, $p < .001$) (Table 3). Likewise, within the non-FM group, worse symbol search test performance was correlated with worse performance on all four NIH Toolbox tests, including pattern comparison ($r = −.56$, $p < .001$), list sorting ($r = −.43$, $p = .002$), flanker ($r = −.43$, $p = .002$), and dimensional change card sort ($r = −.33$, $p = .02$) (Table 4).

Validity of dot memory test. The FM group performed significantly worse than the non-FM control group on the dot memory test, as measured by aggregate mean error score ($t(98) = 3.31$, $p = .001$), maximum error score ($t(98) = 3.34$, $p = .001$), and SD of error score ($t(98) = 3.12$, $p = .002$) (Table 2). Similarly, on the NIH Toolbox list sorting test of working memory, the

Table 3. Correlations between ambulatory and NIH Toolbox cognitive test scores in individuals with FM ($n = 50$)

	1	2	3	4	5	6
Ambulatory cognitive tests						
1. Symbol search (mean reaction time, msec)[a]	–					
2. Dot memory (mean error score, Euclidian distance)[a]	.51**	–				
NIH Toolbox cognitive tests						
3. NIH Toolbox pattern comparison (uncorrected standard score)	−.66**	−.37**	–			
4. NIH Toolbox list sorting (uncorrected standard score)	−.45**	−.31*	.43**	–		
5. NIH Toolbox flanker (uncorrected standard score)	−.51**	−.39**	.70**	.23	–	
6. NIH Toolbox dimensional change card sort (uncorrected standard score)	−.51**	−.37**	.65**	.41**	.72**	–

Note. FM = fibromyalgia. [a]person-averaged. *$p < .05$; **$p < .01$.

Table 4. Correlations between ambulatory and NIH Toolbox cognitive test scores in non-FM controls ($n = 50$)

	1	2	3	4	5	6
Ambulatory cognitive tests						
1. Symbol search (mean reaction time, msec)[a]	–					
2. Dot memory (mean error score, Euclidian distance)[a]	.29*	–				
NIH Toolbox cognitive tests						
3. NIH Toolbox pattern comparison (uncorrected standard score)	−.56**	−.07	–			
4. NIH Toolbox list sorting (uncorrected standard score)	−.43**	−.35*	.29*	–		
5. NIH Toolbox flanker (uncorrected standard score)	−.43**	−.08	.67**	.24	–	
6. NIH Toolbox dimensional change card sort (uncorrected standard score)	−.33*	.02	.55**	−.02	.37**	–

Note. FM = fibromyalgia. [a]person-averaged. *$p < .05$; **$p < .01$.

FM group exhibited significantly worse *T*-scores than did the non-FM control group ($t(98) = −2.01$, $p < .05$), although group differences in uncorrected standard scores were not statistically significant ($p = .09$). Within the FM group, worse dot memory test performance was correlated with worse performance on all four NIH Toolbox tests: pattern comparison ($r = −.37$, $p = .009$), list sorting ($r = −.31$, $p = .03$), flanker ($r = −.39$, $p = .005$), and dimensional change card sort ($r = −.37$, $p = .008$) (Table 3). Within the non-FM group, worse dot memory test performance was correlated only with worse performance on the NIH Toolbox list sorting test ($r = −.35$, $p = .01$) (Table 4).

Discussion

This study is the first to test the psychometric properties of ambulatory cognitive tests as administered in people with FM. The findings demonstrate the feasibility of repeated assessment of processing speed and working memory in the lived environment using smartphone-based symbol search and dot memory tests, with FM and non-FM participants both completing 90% of test sessions over the course of an eight-day assessment period. The high degree of within-person variability in performance on the ambulatory cognitive tests, evidenced by the ICC values, supports the notion that repeated testing may be necessary to adequately assess cognitive functioning. As indicated by overall average between-person reliabilities of ≥.96, the ambulatory tests produced highly reliable scores in both groups, and levels >.90 were attained in as few as two days of repeat testing in the FM group. Intercorrelations among scores on the ambulatory cognitive tests and in-lab neuropsychological tests, and FM vs. non-FM group differences in average scores, provided mixed evidence for construct validity.

Participant compliance to the testing protocol was examined as an indicator of feasibility. Participants were asked to complete two smartphone-based cognitive tests five times per day – upon waking and at quasi-random intervals thereafter – for eight days. Average daily response rates in both FM and non-FM groups were high

across the assessment period, ranging 84% to 94%, and were robust to differences in time of day, with average within-day response rates ranging 86% to 95%. Similarly high response rates have been found in prior studies that have employed ambulatory cognitive testing in adult lifespan and older adult samples (Cerino et al., 2021; Sliwinski et al., 2018; Yang et al., 2021), suggesting that this method is feasible across diverse groups, including, as demonstrated here, those with chronic illness and high symptom burden. It is notable that this level of compliance was observed in the context of an EMA protocol in which the participants additionally provided ratings of FM symptoms at the time they completed the cognitive tests (Kratz et al., 2020; Kratz et al., 2020; Whibley et al., 2022). Thus, these data support the feasibility of administering brief, repeated tests of cognitive performance in FM and show that ambulatory cognitive testing can be embedded in a standard EMA protocol that participants will adhere to with relatively high compliance rates.

Overall average between-person reliabilities for both the ambulatory symbol search and dot memory tests were very high across the FM and non-FM groups and were similar to overall average reliabilities obtained by Sliwinski et al. (2018) using a longer 14-day protocol in a nonclinical community adult sample. Importantly, very few days of repeat testing were needed to attain high reliability in this study; in the FM group, values exceeded .90 after just two days of testing with the symbol search test and after three days of testing with dot memory. That reliable scores can be swiftly attained with such few assessments demonstrates promise for use of ambulatory cognitive testing as an alternative to repeat in-lab/clinic neuropsychological testing, potentially mitigating several logistical challenges associated with the latter (e.g., demands on lab/clinic time, space, and personnel). Selection of the duration (eight days) and density (five times daily) of the data collection protocol was motivated by the primary study questions, which aimed to examine momentary associations between self-report phenomena (symptoms, perceived cognitive functioning) and cognitive test performance. Determinations about the duration and density of ambulatory cognitive performance data

collection should be made by identifying the protocol that presents the lowest possible participant burden while also providing data that address the primary research/clinical question(s). The reliability data from this study suggest that a briefer and potentially less dense protocol than that used here may be sufficient to provide a reliable assay of cognitive performance in future FM studies. For example, shorter data collection periods than that used here may be appropriate to detect between-group differences in change in cognitive functioning with FM treatment; however, longer and/or denser protocols may be needed depending on research/clinical objectives, such as in observational studies that aim to detect subtle intraindividual change in cognitive functioning.

Construct validity was assessed by examining correspondence between performance on the ambulatory cognitive tests and on in-lab NIH Toolbox measures and via comparing scores in the FM and non-FM groups. Regarding the symbol search task, the strongest associations were, as expected, with scores on the NIH Toolbox test of processing speed (i.e., pattern comparison), with correlations of moderate to strong magnitude observed in both the FM and non-FM controls. These data support the convergent validity of the symbol search test. However, it is notable that significant correlations of low to moderate magnitude were also observed between the symbol search test and the other NIH Toolbox tests (i.e., list sorting, flanker, and dimensional change card sort). This raises the possibility that a common factor, such as fatigue or inattention, may have impacted performance across tests.

To our surprise, known-groups analysis showed no significant group differences in average symbol search performance, despite participants with FM producing lower NIH Toolbox pattern comparison test T-scores than did non-FM controls. Although there was correlational evidence for convergence between the ambulatory and in-lab tests of processing speed, the ambulatory symbol search test used here may be less sensitive than the NIH Toolbox pattern comparison test in detecting group differences in performance. This is particularly likely if processing speed deficits are modest, as has been shown to be the case in a meta-analysis of cognitive performance in FM (Bell et al., 2018). As task difficulty is an important determinant of between-person variability in scores, it is possible that a more challenging version of the test, such as one with more complex stimuli (e.g., higher degree of similarity between stimuli in the search group, higher number of symbol pairs to consider), would have better distinguished between the groups. Alternatively, the presence of significant FM vs. non-FM differences in NIH Toolbox test performance, but lack of significant group differences in performance on the ambulatory test, could be due to the influence of FM symptoms. FM-related fatigue and attentional difficulties might impact performance on the NIH Toolbox tests to a greater degree given the tests' longer duration, thus contributing to larger FM vs. non-FM differences in performance.

There was also mixed evidence for the validity of the dot memory test. Group differences were consistent with expectations: participants with FM, relative to non-FM controls, produced significantly worse dot memory scores and T-scores on the NIH Toolbox measure of working memory (i.e., list sorting). However, within both the FM and non-FM groups, performance on the ambulatory test was only modestly correlated with the NIH Toolbox test, and within the FM group, similar magnitude correlations were observed between dot memory performance and scores on the NIH Toolbox tests of processing speed (i.e., pattern comparison), attention and inhibitory control (i.e., flanker),

and cognitive flexibility and attention (i.e., dimensional change card sort). Associations between dot memory scores and performance on tests of processing speed and attention are not surprising; slowed processing speed constrains performance in other cognitive domains (Salthouse, 1996), and attention is critical for the selection and maintenance of information in working memory (Awh et al., 2006; Oberauer, 2019).

The availability of psychometrically sound ambulatory cognitive tests has implications for assessment of cognitive functioning in FM and other groups. Ambulatory tests offer a means of assessing cognitive performance in the lived environment and on repeated occasions, therefore addressing key limitations and criticisms of traditional neuropsychological testing, including a lack of ecological validity and measurement reliability. Our own work and that of others have demonstrated clear clinical value in the advantages of ambulatory testing. We previously showed that momentary changes in processing speed in individuals with FM, as assessed by repeat ambulatory testing, correspond with momentary subjective reports of cognitive functioning (Kratz et al., 2020), giving credence to perceptions of cognitive difficulties in the daily lives of individuals with FM. In a study of healthy older adults (Allard et al., 2014), performance on repeat ambulatory cognitive testing, but not on traditional neuropsychological tests, was significantly correlated with hippocampal volume, suggesting that the former may detect subtle cognitive deficits that traditional neuropsychological testing might miss. Moreover, engagement in intellectually stimulating activities, including reading and completing crossword puzzles, was shown to precede improved performance on ambulatory cognitive tests in the subsequent three-hour period, demonstrating the utility of this measurement approach for examining real-time, dynamic associations between daily life activities or behaviors and cognitive functioning.

Study limitations and future research directions are considered. The present investigation focused on select cognitive domains – processing speed and working memory – though ambulatory tests of other cognitive domains impacted by FM also merit study. Regarding the breadth of psychometric evaluation, this study examined feasibility as indicated by compliance, between-person reliability, and construct validity. As work continues in this area, additional qualities merit consideration. The evaluation of feasibility may be expanded to include sociodemographic and clinical characteristics that impact compliance and data quality. Data regarding attitudes, perceptions, and acceptability could inform methodological improvements, and examining sensitivity to change and correlation with neuroanatomical, neurophysiological, and functional outcomes could improve clinical utility. Though the study's use of established neuropsychological measures to evaluate construct validity is a strength, it should be acknowledged that factors not considered here (e.g., environmental distractions, test anxiety) may have differentially impacted performance in the in-lab and naturalistic settings, thus affecting the measure convergence. We are unable to determine with these data the extent to which ambulatory cognitive test scores relate to performance on everyday cognitive tasks. This is an important area for future inquiry, as establishing correspondence between ambulatory test performance and real-world functioning will improve the clinical and research utility of these measures. There are several possible explanations for the differential patterns of correlation between the ambulatory tests and NIH Toolbox tests in the FM vs. non-FM groups, including the small sample size or restricted range of cognitive performance, particularly in non-FM controls. Future studies employing larger, more diverse samples with greater

variability in cognitive functioning are needed to replicate the findings, as well as to provide demographically adjusted normative data. Additionally, as the NIH Toolbox tests were only administered at baseline, we are unable to examine how changes in performance with repeat ambulatory testing compare with changes in performance with serial administration of the NIH Toolbox measures, including examining how practice effects might differ across these types of measures. Finally, test sessions with < 70% accuracy were excluded from analyses. Although this was done to reduce potential bias from data resulting from poor effort, it is possible that informative data were inadvertently excluded. As ambulatory methods continue to be refined, an important area for future investigation will be evaluating methods for assessing and controlling for suboptimal examinee effort.

In sum, the findings show promise for use of ambulatory approaches to assessing cognitive performance in people with FM. Administering brief, repeated smartphone-based cognitive tests in the everyday lives of individuals with FM is feasible and can produce reliable scores with few measurement occasions. Further development of ambulatory cognitive tests that are feasible, reliable, and valid remains an important area for future research in FM and other clinical populations.

Conflicts of interest. None.

References

Allard, M., Husky, M., Catheline, G., Pelletier, A., Dilharreguy, B., Amieva, H., Pérès, K., Foubert-Samier, A., Dartigues, J.-F., & Swendsen, J. (2014). Mobile technologies in the early detection of cognitive decline. *PLoS ONE*, 9, e112197. https://doi.org/10.1371/journal.pone.0112197

Arnold, L. M., Crofford, L. J., Mease, P. J., Burgess, S. M., Palmer, S. C., Abetz, L., & Martin, S. A. (2008). Patient perspectives on the impact of fibromyalgia. *Patient Education and Counseling*, 73, 114–120. https://doi.org/10.1016/j.pec.2008.06.005

Awh, E., Vogel, E. K., & Oh, S. -H. (2006). Interactions between attention and working memory. *Neuroscience*, 139, 201–208. https://doi.org/10.1016/j.neuroscience.2005.08.023

Bell, T., Trost, Z., Buelow, M. T., Clay, O., Younger, J., Moore, D., & Crowe, M. (2018). Meta-analysis of cognitive performance in fibromyalgia. *Journal of Clinical and Experimental Neuropsychology*, 40, 698–714. https://doi.org/10.1080/13803395.2017.1422699

Bennett, R. M., Jones, J., Turk, D. C., Russell, I. J., & Matallana, L. (2007). An internet survey of 2,596 people with fibromyalgia. *BMC Musculoskeletal Disorders*, 8, 27. https://doi.org/10.1186/1471-2474-8-27

Bielak, A. A. M., Hultsch, D. F., Strauss, E., Macdonald, S. W. S., & Hunter, M. A. (2010). Intraindividual variability in reaction time predicts cognitive outcomes 5 years later. *Neuropsychology*, 24, 731–741. https://doi.org/10.1037/a0019802

Bielak, A. A. M., Hultsch, D. F., Strauss, E., MacDonald, S. W. S., & Hunter, M. A. (2010). Intraindividual variability is related to cognitive change in older adults: Evidence for within-person coupling. *Psychology and Aging*, 25, 575–586. https://doi.org/10.1037/a0019503

Brearly, T. W., Rowland, J. A., Martindale, S. L., Shura, R. D., Curry, D., & Taber, K. H. (2019). Comparability of iPad and web-based NIH Toolbox Cognitive Battery administration in veterans. *Archives of Clinical Neuropsychology*, 34, 524–530. https://doi.org/10.1093/arclin/acy070

Carrillo-de-la-Peña, M. T., Vallet, M., Pérez, M. I., & Gómez-Perretta, C. (2006). Intensity dependence of auditory-evoked cortical potentials in fibromyalgia patients: A test of the generalized hypervigilance hypothesis. *The Journal of Pain*, 7, 480–487. https://doi.org/10.1016/j.jpain.2006.01.452

Cerino, E. S., Katz, M. J., Wang, C., Qin, J., Gao, Q., Hyun, J., Hakun, J. G., Roque, N. A., Derby, C. A., Lipton, R. B., & Sliwinski, M. J. (2021). Variability in cognitive performance on mobile devices is sensitive to mild cognitive impairment: Results from the Einstein Aging Study. *Frontiers in Digital Health*, 3, 758031. https://doi.org/10.3389/fdgth.2021.758031

de Vries, L. P., Baselmans, B. M. L., & Bartels, M. (2021). Smartphone-based ecological momentary assessment of well-being: A systematic review and recommendations for future studies. *Journal of Happiness Studies*, 22, 2361–2408. https://doi.org/10.1007/s10902-020-00324-7

Gershon, R. C., Cella, D., Fox, N. A., Havlik, R. J., Hendrie, H. C., & Wagster, M. V. (2010). Assessment of neurological and behavioural function: The NIH Toolbox. *The Lancet Neurology*, 9, 138–139. https://doi.org/10.1016/S1474-4422(09)70335-7

Grace, G. M., Nielson, W. R., Hopkins, M., & Berg, M. A. (1999). Concentration and memory deficits in patients with fibromyalgia syndrome. *Journal of Clinical and Experimental Neuropsychology*, 21, 477–487. https://doi.org/10.1076/jcen.21.4.477.876

Harte, S. E., Ichesco, E., Hampson, J. P., Peltier, S. J., Schmidt-Wilcke, T., Clauw, D. J., & Harris, R. E. (2016). Pharmacologic attenuation of cross-modal sensory augmentation within the chronic pain insula. *Pain*, 157, 1933–1945. https://doi.org/10.1097/j.pain.0000000000000593

Hollins, M., Harper, D., Gallagher, S., Owings, E. W., Lim, P. F., Miller, V., Siddiqi, M. Q., & Maixner, W. (2009). Perceived intensity and unpleasantness of cutaneous and auditory stimuli: An evaluation of the generalized hypervigilance hypothesis. *Pain*, 141, 215–221. https://doi.org/10.1016/j.pain.2008.10.003

Katz, R. S., Heard, A. R., Mills, M., & Leavitt, F. (2004). The prevalence and clinical impact of reported cognitive difficulties (fibrofog) in patients with rheumatic disease with and without fibromyalgia. *Journal of Clinical Rheumatology: Practical Reports on Rheumatic & Musculoskeletal Diseases*, 10, 53–58. https://doi.org/10.1097/01.rhu.0000120895.20623.9f

Kim, S. -H., Kim, S. -H., Kim, S. -K., Nam, E. J., Han, S. W., & Lee, S. J. (2012). Spatial versus verbal memory impairments in patients with fibromyalgia. *Rheumatology International*, 32, 1135–1142. https://doi.org/10.1007/s00296-010-1762-1

Kosek, E., Ekholm, J., & Hansson, P. (1996). Sensory dysfunction in fibromyalgia patients with implications for pathogenic mechanisms. *Pain*, 68, 375–383. https://doi.org/10.1016/s0304-3959(96)03188-0

Kratz, A. L., Whibley, D., Kim, S., Sliwinski, M., Clauw, D., & Williams, D. A. (2020). Fibrofog in daily life: An examination of ambulatory subjective and objective cognitive function in fibromyalgia. *Arthritis Care & Research*, 72, 1669–1677. https://doi.org/10.1002/acr.24089

Kratz, A. L., Whibley, D., Kim, S., Williams, D. A., Clauw, D. J., & Sliwinski, M. (2020). The role of environmental distractions in the experience of fibrofog in real-world settings. *ACR Open Rheumatology*, 2, 214–221. https://doi.org/10.1002/acr2.11130

Landrø, N. I., Stiles, T. C., & Sletvold, H. (1997). Memory functioning in patients with primary fibromyalgia and major depression and healthy controls. *Journal of Psychosomatic Research*, 42, 297–306. https://doi.org/10.1016/s0022-3999(96)00301-7

Lawrence, R. C., Felson, D. T., Helmick, C. G., Arnold, L. M., Choi, H., Deyo, R. A., Gabriel, S., Hirsch, R., Hochberg, M. C., Hunder, G. G., Jordan, J. M., Katz, J. N., Kremers, H. M., Wolfe, F., & National Arthritis Data Workgroup. (2008). Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part II. *Arthritis and Rheumatism*, 58, 26–35. https://doi.org/10.1002/art.23176

Lee, D. M., Pendleton, N., Tajar, A., O'Neill, T. W., O'Connor, D. B., Bartfai, G., Boonen, S., Casanueva, F. F., Finn, J. D., Forti, G., Giwercman, A., Han, T. S., Huhtaniemi, I. T., Kula, K., Lean, M. E. J., Punab, M., Silman, A. J., Vanderschueren, D., Moseley, C. M., Wu, F. C. W., McBeth, J., & EMAS study group. (2010). Chronic widespread pain is associated with slower cognitive processing speed in middle-aged and older European men. *Pain*, 151, 30–36. https://doi.org/10.1016/j.pain.2010.04.024

Lombion, S., Comte, A., Tatu, L., Brand, G., Moulin, T., & Millot, J.-L. (2009). Patterns of cerebral activation during olfactory and trigeminal stimulations. *Human Brain Mapping*, *30*, 821–828. https://doi.org/10.1002/hbm.20548

Lorenz, J., Grasedyck, K., & Bromm, B. (1996). Middle and long latency somatosensory evoked potentials after painful laser stimulation in patients with fibromyalgia syndrome. *Electroencephalography and Clinical Neurophysiology*, *100*, 165–168. https://doi.org/10.1016/0013-4694(95)00259-6

Lötsch, J., Kraetsch, H.-G., Wendler, J., & Hummel, T. (2012). Self-ratings of higher olfactory acuity contrast with reduced olfactory test results of fibromyalgia patients. *International Journal of Psychophysiology*, *86*, 182–186. https://doi.org/10.1016/j.ijpsycho.2012.09.003

McDermid, A. J., Rollman, G. B., & McCain, G. A. (1996). Generalized hypervigilance in fibromyalgia: Evidence of perceptual amplification. *Pain*, *66*, 133–144. https://doi.org/10.1016/0304-3959(96)03059-x

Miró, E., Lupiáñez, J., Hita, E., Martínez, M. P., Sánchez, A. I., & Buela-Casal, G. (2011). Attentional deficits in fibromyalgia and its relationships with pain, emotional distress and sleep dysfunction complaints. *Psychology & Health*, *26*, 765–780. https://doi.org/10.1080/08870446.2010.493611

Oberauer, K. (2019). Working memory and attention – A conceptual analysis and review. *Journal of Cognition*, *2*, 36. https://doi.org/10.5334/joc.58

Park, D. C., Glass, J. M., Minear, M., & Crofford, L. J. (2001). Cognitive function in fibromyalgia patients. *Arthritis and Rheumatism*, *44*, 2125–2133. https://doi.org/10.1002/1529-0131(200109)44:9 < 2125::AID-ART365>3.0.CO;2-1

Petzke, F., Clauw, D. J., Ambrose, K., Khine, A., & Gracely, R. H. (2003). Increased pain sensitivity in fibromyalgia: Effects of stimulus type and mode of presentation. *Pain*, *105*, 403–413. https://doi.org/10.1016/S0304-3959(03)00204-5

Ram, N., Rabbitt, P., Stollery, B., & Nesselroade, J. R. (2005). Cognitive performance inconsistency: Intraindividual change and variability. *Psychology and Aging*, *20*, 623–633. https://doi.org/10.1037/0882-7974.20.4.623

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*, 403–428. https://doi.org/10.1037/0033-295x.103.3.403

Sbordone, R. J. (1996). Ecological validity: Some critical issues for the neuropsychologist. In R. J. Sbordone & C. J. Long (Eds.), *Ecological validity of neuropsychological testing* (pp. 15–41). Gr Press/St Lucie Press, Inc.

Sliwinski, M. J., Mogle, J. A., Hyun, J., Munoz, E., Smyth, J. M., & Lipton, R. B. (2018). Reliability and validity of ambulatory cognitive assessments. *Assessment*, *25*, 14–30. https://doi.org/10.1177/1073191116643164

Spooner, D. M., & Pachana, N. A. (2006). Ecological validity in neuropsychological assessment: A case for greater consideration in research with neurologically intact populations. *Archives of Clinical Neuropsychology*, *21*, 327–337. https://doi.org/10.1016/j.acn.2006.04.004

Suhr, J. A. (2003). Neuropsychological impairment in fibromyalgia: Relation to depression, fatigue, and pain. *Journal of Psychosomatic Research*, *55*, 321–329. https://doi.org/10.1016/s0022-3999(02)00628-1

Teodoro, T., Edwards, M. J., & Isaacs, J. D. (2018). A unifying theory for cognitive abnormalities in functional neurological disorders, fibromyalgia and chronic fatigue syndrome: Systematic review. *Journal of Neurology, Neurosurgery, and Psychiatry*, *89*, 1308–1319. https://doi.org/10.1136/jnnp-2017-317823

Walitt, B., Ćeko, M., Khatiwada, M., Gracely, J. L., Rayhan, R., VanMeter, J. W., & Gracely, R. H. (2016). Characterizing "fibrofog": Subjective appraisal, objective performance, and task-related brain activity during a working memory task. *NeuroImage: Clinical*, *11*, 173–180. https://doi.org/10.1016/j.nicl.2016.01.021

Walitt, B., Fitzcharles, M.-A., Hassett, A. L., Katz, R. S., Häuser, W., & Wolfe, F. (2011). The longitudinal outcome of fibromyalgia: A study of 1555 patients. *The Journal of Rheumatology*, *38*, 2238–2246. https://doi.org/10.3899/jrheum.110026

Walitt, B., Roebuck-Spencer, T., Bleiberg, J., Foster, G., & Weinstein, A. (2008). Automated neuropsychiatric measurements of information processing in fibromyalgia. *Rheumatology International*, *28*, 561–566. https://doi.org/10.1007/s00296-007-0487-2

Walteros, C., Sánchez-Navarro, J. P., Muñoz, M. A., Martínez-Selva, J. M., Chialvo, D., & Montoya, P. (2011). Altered associative learning and emotional decision making in fibromyalgia. *Journal of Psychosomatic Research*, *70*, 294–301. https://doi.org/10.1016/j.jpsychores.2010.07.013

West, R., Murphy, K. J., Armilio, M. L., Craik, F. I. M., & Stuss, D. T. (2002). Lapses of intention and performance variability reveal age-related increases in fluctuations of executive control. *Brain and Cognition*, *49*, 402–419. https://doi.org/10.1006/brcg.2001.1507

Whibley, D., Williams, D. A., Clauw, D. J., Sliwinski, M. J., & Kratz, A. L. (2022). Within-day rhythms of pain and cognitive function in people with and without fibromyalgia: Synchronous or syncopated? *Pain*, *163*, 474–482. https://doi.org/10.1097/j.pain.0000000000002370

Wolfe, F., Clauw, D. J., Fitzcharles, M.-A., Goldenberg, D. L., Häuser, W., Katz, R. L., Mease, P. J., Russell, A. S., Russell, I. J., & Walitt, B. (2016). 2016 Revisions to the 2010/2011 fibromyalgia diagnostic criteria. *Seminars in Arthritis and Rheumatism*, *46*, 319–329. https://doi.org/10.1016/j.semarthrit.2016.08.012

Wu, Y.-L., Huang, C.-J., Fang, S.-C., Ko, L.-H., & Tsai, P.-S. (2018). Cognitive impairment in fibromyalgia: A meta-analysis of case–control studies. *Psychosomatic Medicine*, *80*, 432–438. https://doi.org/10.1097/PSY.0000000000000575

Yang, C.-H., Hakun, J. G., Roque, N., Sliwinski, M. J., & Conroy, D. E. (2021). Mindful walking and cognition in older adults: A proof of concept study using in-lab and ambulatory cognitive measures. *Preventive Medicine Reports*, *23*, 101490. https://doi.org/10.1016/j.pmedr.2021.101490