Routledge
Taylor & Francis Group

Aging, Neuropsychology, and Cognition

A Journal on Normal and Dysfunctional Development

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/nanc20

Naturalistic assessments in virtual reality and in real life help resolve the age-prospective memory paradox

Nathan S. Rose & Joseph M. Saito

To cite this article: Nathan S. Rose & Joseph M. Saito (16 Feb 2024): Naturalistic assessments in virtual reality and in real life help resolve the age-prospective memory paradox, Aging, Neuropsychology, and Cognition, DOI: 10.1080/13825585.2024.2315791

To link to this article: https://doi.org/10.1080/13825585.2024.2315791

Published online: 16 Feb 2024.

AGING, NEUROPSYCHOLOGY, AND COGNITION
https://doi.org/10.1080/13825585.2024.2315791

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Naturalistic assessments in virtual reality and in real life help resolve the age-prospective memory paradox

Nathan S. Rose[a] and Joseph M. Saito[b]

[a]Department of Psychology, University of Notre Dame, Notre Dame, IN, USA; [b]Department of Psychology, University of Toronto, Mississauga, ON, Canada

ABSTRACT

Cognitive aging researchers have long reported "paradoxical" age differences in prospective memory (PM), with age deficits in laboratory settings and age benefits (or no deficits) in real-world settings. We propose a theoretical account that explains this "age-PM-paradox" as a consequence of both methodological factors and developmental changes in cognitive abilities and personality traits. To test this account, young and older adults performed a series of naturalistic PM tasks in the lab and real world. Age-related PM deficits were observed in both lab-based tasks where demands were implemented using virtual reality and in-person role-playing. In contrast, older adults performed equal to or better than young adults on both real-world tasks, where demands were implemented in participants' daily lives. Consistent with our proposed account, an index of these "paradoxical" effects was partially predicted by age-related differences in working memory, vigilance, agreeableness, and neuroticism, whose predictive utility varied across task settings.

ARTICLE HISTORY
Received 15 December 2022
Accepted 5 January 2024

KEYWORDS
Prospective memory; aging; virtual reality; working memory; personality

Prospective memory (PM) refers to the cognitive processes that enable individuals to remember to do intended actions at the appropriate moment in the future. In a typical day, individuals engage in multiple behaviors that depend on PM (e.g., remembering to take medications and turn off appliances). Given the ubiquity of these cognitive processes in daily life, preserving PM into older adulthood is essential for maintaining an independent lifestyle (Crovitz & Daniel, 1984; Einstein & McDaniel, 1990; Ellis & Kvavilashvili, 2000; Hering et al., 2018; Kliegel & Martin, 2003; Terry, 1988; Woods et al., 2015).

Although a substantial body of research has been conducted on the nature and extent of age-related differences in PM performance, studies over the last 40 years have repeatedly shown paradoxical patterns across laboratory and real-world settings (for reviews, see Kliegel et al., 2016; Peter & Kliegel, 2018; Schnitzspahn et al., 2020). Specifically, while older adults typically show age-related deficits on conventional lab-based paradigms (Einstein & McDaniel, 1990; N. S. Rose et al., 2010), they often perform as well as or better than young adults on real-world tasks (Aberle et al.,

CONTACT Nathan S. Rose nrose1@nd.edu Department of Psychology, University of Notre Dame, 390 Corbett Hall, Notre Dame, IN 46556, USA

2010; Kliegel et al., 2008; Rendell & Craik, 2000; Rendell & Thomson, 1999). This pattern has been described as "one of the most important puzzles in the study of cognitive aging" (p. 3, Verhaeghen et al., 2012). Here, we propose a theoretical account of this paradox that helps explain this pattern in performance as the result of developmental changes in both cognitive abilities and motivational factors related to personality traits, which are differentially implicated in different settings between young and older adults.

The age-prospective memory paradox

In 1999, Rendell and Thomson reported one of the first formal investigations of age-related differences in PM across lab and real-world task settings. In the lab, young and older adult participants were asked to turn off a stop-clock after seven minutes and to note the time they finished a questionnaire during a word-list learning task. Here, robust patterns of age-related deficits were observed. Outside the lab, participants were asked to log times at four set times for one week, with six different regimens that varied the complexity of the time schedule and the opportunity to use external aids and conjunction cues (e.g., recoding the task of "put casserole in oven at 5:30" as "when you get home from work"). In contrast to the lab-based tasks, older adults outperformed the young adults on all conditions of the real-world tasks. It was suggested that these results confirmed the existence of an "age-prospective-memory paradox" that had been observed and reported in the literature since at least 1982 (Moscovitch, 1982; Schaffer & Poon, 1982; Rendell & Thomson, 1999; for meta-analyses, see; Henry et al., 2004; Uttl & Greene, 2008). Now, there are over 400 publications referencing this paradox.

Even though this phenomenon has been observed for over 40 years, there still is no comprehensive account of it. Whenever there are discussions of the paradox, there is often a litany of just-so stories used to explain away the findings. For example, many assume that age-related differences reflect lifestyle factors, including busyness and the number and structure of routine everyday life activities. Yet, studies directly investigating this hypothesis have shown that these variables alone do not fully explain the paradox. While perceived busyness has been associated with self-reported PM failures (Gondo et al., 2010; Martin & Park, 2003), it does not mediate the paradox (Festini et al., 2016).

Age differences in the use of reminders are another common just-so story. Most studies of PM in daily life instruct participants to avoid using reminders (e.g., alarms and notes), yet some individuals still report using some sort of external aid to help them with PM tasks. However, a large and growing literature shows that there are little-to-no age differences in the use of PM reminders and that this also cannot explain the paradox (Ihle et al., 2012; Maylor, 1990; Niedźwieńska & Barzykowski, 2012; Patton & Meit, 1993; Rendell & Craik, 2000; Rendell & Thomson, 1999; Schnitzspahn et al., 2011, 2020; West, 1988). So then, what causes paradoxical patterns of PM performance between the lab and real world?

Resolving the age-PM paradox

Below, we highlight three factors that may play a meaningful role in the age-PM paradox. The first has to do with methodological issues related to the variety of ways that PM constructs are operationalized and measured across lab and real-world settings. For example, while lab-based studies often use time-based tasks with target intentions spaced across relatively short retention intervals (e.g., 5 or 10 minutes), real-world tasks often incorporate longer delays with target intentions occurring at particular times of day instead (e.g., 5PM). Consistent with this account, we recently showed that older adults did worse than young adults on time-interval tasks performed in the lab, but outperformed young adults on time-of-day tasks performed in everyday life, even for a demanding version of the real-world tasks (Haines et al., 2020). However, in that same study, when the target intentions in the real-world task were modified to be based on time intervals rather than times of day, older adults no longer outperformed young adults (see also Bailey et al., 2010). This suggests that it is not a difference in task setting, per se, that drives paradoxical patterns of PM performance, but is instead a difference in the types of PM tasks that tend to be implemented across task settings. Outside the lab, people can convert time-of-day cues (e.g., 5PM) to event cues by associating them with events that happen at that time of day (e.g., when leaving work). By contrast, this type of recoding scheme would be much more difficult to implement in the lab when targets are based on short, arbitrary time intervals.

A related hypothesis is that conventional lab-based measures differ from real-world tasks in their ecological validity (see, e.g., Phillips et al., 2008). Does remembering to press a computer key when a specific word is presented during a lexical decision task depend on the same cognitive functions as following a medication regimen over the course of several days in everyday life? Prior work has shown that differences in ecological validity between lab-based and real-world measures cannot be resolved just by incorporating familiar elements of daily life into otherwise artificial tasks (Rendell & Craik, 2000; Rose et al., 2010). For example, Rendell and Craik (2000) developed a boardgame paradigm, *Virtual Week*, in which participants simulate five virtual days of activities and simulate performing PM tasks that are common in everyday life (e.g., taking medication at breakfast, dropping off dry cleaning after lunch), with each virtual "day" lasting about 10 minutes of actual time. Despite the everyday nature of the PM tasks embedded within Virtual Week, young adults still outperformed older adults.

One possibility is that the artificial structure and condensed nature of tasks like Virtual Week produce critical limitations on older adults' ability to translate their years of real-world experience to the task. It is reasonable to expect that performing an entire day's worth of PM intentions in 10 minutes places considerably more demands on short-term/working memory processes than performing these intentions over a more naturalistic time scale. Indeed, N. S. Rose et al. (2010) showed that age-related deficits in Virtual Week performance were largely attributable to age- and individual-differences in working memory (WM). WM is known to decline with age (Bopp & Verhaeghen, 2005; Park, 2000; Park et al., 2002; N. S. Rose et al., 2009, 2010; Rypma & D'Esposito, 2000; Salthouse, 1994; Sarter & Bruno, 2004) and predict individual differences in PM (Engle et al., 1999; Kane et al., 2007; N. S. Rose et al., 2010, 2015; Unsworth et al., 2012). Given these findings, it may be the case that lab-based tasks like Virtual Week, despite incorporating familiar elements of everyday behavior, recruit cognitive mechanisms differently from everyday PM functioning. That said, few studies have systematically assessed the same participants on multiple naturalistic PM paradigms across both lab-based and real-world contexts. As such, a more complete examination of the age-PM paradox will require an individual differences approach in which large numbers of the same young and older adults perform similar naturalistic PM tasks in the lab and real-world settings, as well as tasks of potential cognitive predictors (e.g., WM, vigilance).

Lastly, age- and individual-differences in various motivational factors may also contribute to the age-PM paradox. Some data suggest that motivation plays a direct role in PM performance via financial, social,[1] and intrinsic motivators (Aberle et al., 2010; Peter & Kliegel, 2018). Motivational factors also appear to play an indirect role via developmental differences in personality traits (Lodi-Smith et al., 2011). For example, older adults score higher than young adults on measures of conscientiousness and agreeableness on average (Helson et al., 2002). These factors relate to age- and individual-differences in time-management and procrastination (D. C. Watson, 2001), as well as associated cognitive abilities (e.g., WM, vigilance) that help facilitate PM (Hering et al., 2014). To this point, some studies have reported positive associations between conscientiousness and PM performance (Cuttler & Graf, 2007; Smith et al., 2011), but the research overall remains mixed, likely due to variability in the use of self-report and objective measures of PM performance in lab-based and real-world contexts (Uttl & Kibreab, 2011; Uttl et al., 2013). These mixed findings further underscore the need for a more comprehensive examination of the age-PM paradox that incorporates objective measures of both personality and cognitive predictors of PM performance in both lab- and real-world contexts within the same individuals.

The present study

We designed the present study with the aim of addressing the factors described in the previous section. To examine the role of ecological validity, we used four naturalistic PM paradigms that were done in lab-based or real-world settings. One was a new immersive virtual reality (VR) paradigm with both time-interval and event-based PM measures, and the other three were selected based on logistic feasibility and prior evidence of their ecological validity: the *Breakfast Task*, which involved simulating the preparation of breakfast for a large group of people with both time-interval PM intentions (e.g., flip the eggs after 3 minutes) and event-based PM intentions (e.g., turn off all appliances when done; Altgassen et al., 2015; Craik & Bialystok, 2006), the *Belongings Task*, which involved remembering to retrieve one's cell phone and return an activity-tracker at the end of the lab-based portion of the experiment (i.e., event-based PM intentions, Cuttler & Graf, 2008; N. Rose et al., 2023; Wilson et al., 1985), and the *Call-Back Task*, which involved calling the experimenter's office while at home in everyday life after relatively short time intervals (i.e., time-interval PM intentions, N. S. Rose et al., 2015). These naturalistic PM paradigms enable a high-degree of experimental control while also capturing common types of PM intentions across task settings (Altgassen et al., 2010; Einstein & McDaniel, 1990; Einstein et al., 1992).

Our motivation for the VR paradigm was based on the fact that VR has been increasingly used over the last decade to make lab-based paradigms more realistic, including those used to measure PM. However, prior efforts have largely involved two important limitations. The first is methodological. PM studies using VR have used either standard computer monitors (Debarnot et al., 2015; Gonneaud et al., 2014; Sakai et al., 2018) or head-mounted displays (Banville et al., 2010; Parsons et al., 2017). While both can present three-dimensional environments, HMDs have the added advantage of adjusting visual input in real-time with physical head movements. This has been shown to evoke patterns of neural activity that are more like real-world navigation (e.g., Taube et al., 2013). For example, rodent models using VR have suggested that only 25% of localized place cell activation is achieved with visual input alone. The remaining 75% requires synchronous proprioceptive and vestibular information from physical movement (Chen et al., 2013). Thus, VR paradigms that do not use HMDs may not fully capture the immersive, neurocognitive processes that are engaged in the real world (Debarnot et al., 2015; Jylkkä et al., 2023; O'Rear & Radvansky, 2019; Okahashi et al., 2013; Parsons & Parsons et al., 2017; Sakai et al., 2018). Moreover, even for PM studies that have used HMDs previously, many required people to sit or lie static throughout the task (Dong et al., 2018; Kalpouzos et al., 2010). The novel VR paradigm and setup used in the current study allowed full physical mobility, which provides more immersion and a closer approximation of real-world activity.

The second limitation is theoretical. While some recent studies have involved immersive PM assessments in VR (e.g., Jylkkä et al., 2023; Seesjärvi et al., 2022), to our knowledge, no study has shown that VR-based PM tasks capture age-differences in naturalistic PM. One study showed that age-differences in PM performance in an immersive VR shopping task were related to self-reports of everyday memory (Ouellet et al., 2018). While this presents promising preliminary evidence supporting the use of VR, correlating performance with a subjective rating instead of an objective assessment of real-world PM limits its validation (Uttl & Kibreab, 2011). In the present study, we perform a direct comparison between VR and multiple, objective measures of naturalistic PM performed in either lab or real-world settings.

Within the VR paradigm used here, participants performed PM tasks embedded in the commercially-available role-playing videogame, *Job Simulator* (Owlchemy Labs, Co., Austin, TX), in which players participate in futuristic, humorous simulations of different jobs such as a convenience store clerk or a short-order cook. Despite participants' lack of familiarity with the *Job Simulator* environment, we hypothesized that performing realistic occupational tasks in an immersive environment would more closely approximate real-world behavior and result in patterns of performance that correspond with those of our other validated PM measures. Specifically, age differences in *Job Simulator* performance were predicted to be closely related to age differences in performance on the Breakfast Task, given that they both required managing multiple PM intentions while performing other demanding, ongoing tasks. Although the Belongings Task and Call-Back task also required performing PM tasks, the context and demands were considerably different, especially for the Call-Back task. For example, the Call-Back task was completed at home in the participant's own environment where they may be able to reduce the overall cognitive load (i.e., WM demands) of the task by adjusting the pace and demands of other ongoing activities. For this reason, although we hoped that the immersive VR tasks would capture real world PM, we suspected that differences in context and load would reveal "paradoxical" patterns of age differences in PM that we could use to test our theoretical account of the paradox based on age- and individual-differences in cognition and personality.

To test the impact of developmental changes in personality and cognition on PM performance, we administered measures of WM, vigilance, and the "big five" personality traits. As aforementioned, age differences in PM may be partially explained by differences in WM abilities (Engle et al., 1999; Kane et al., 2007; N. S. Rose et al., 2010; Unsworth et al., 2012). Vigilance has also been associated with PM previously, particularly in conditions that require sustained attention and online monitoring (e.g., Einstein & McDaniel, 2008). However, some have suggested that tasks that can be fully accomplished via vigilance abilities lack mnemonic processing demands that most would consider to be central to the definition of PM (Brandimonte et al., 2001; Graf & Uttl, 2001; N. S. Rose et al., 2010). Finally, developmental changes in personality traits across adulthood—i.e., increased agreeableness and conscientiousness, decreased neuroticism, extraversion, and openness (Allemand et al., 2008; McCrae et al., 1999) – may relate to age differences in PM. For example, one study showed that conscientiousness predicted both lab-based and real-world PM (Cuttler & Graf, 2007), and neuroticism predicted lab-based PM in young adults (see Anderson & McDaniel, 2019, for another example of neuroticism predicting young adults' self-reported thoughts about real world PM using experience sampling methods). In the present study, we constructed hierarchical regression models using these variables as predictors of PM performance to assess their contribution to the age-PM paradox. We hypothesized that WM, but not vigilance, would predict PM performance on *Job Simulator* and Breakfast tasks, and that age differences in personality traits (particularly conscientiousness and neuroticism) would help explain any "paradoxical" differences in PM performance across settings (e.g., on the Belongings and Call-back tasks).

Method

Power analysis

Because we aimed to conduct hierarchical, multiple-regression modeling to assess the effects of age- and individual-differences in cognition and personality on age differences in PM performance across task settings, we relied on our previous study that used a similar approach (N. S. Rose et al., 2010) to determine an appropriate sample size to test in this study. N. S. Rose et al. (2010) found moderate effect sizes with regards to the prediction of age- and individual-differences in PM performance in young and older adults. A power calculation performed with an alpha level of 0.05, statistical power of 0.8, and an anticipated between-subjects effect size of $f^2 = 0.15$ revealed that we would need at least 100 participants split between the two groups to detect an effect with 8 predictors in the model (Faul et al., 2007).

Participants

Young ($N = 59$; 18–30 years; $M_{age} = 19.4$; $SD_{age} = 1.8$; $M_{edu} = 13.0$) and cognitively healthy older adult volunteers ($N = 52$; 58–83 years; $M_{age} = 70.4$; $SD_{age} = 5.1$; $M_{edu} = 16.3$) were recruited to participate in exchange for either course credit or $10/hour. The older adults were those who had attended and participated in our outreach activities at a local church and Community Learning Center (or were referred by their friends and family). There was one participant below the age of 60 and three below the age of 65. We attempted to address the wide age range in our sample by analyzing the data with correlational/regression analyses with age included as a continuous variable. To ensure that the pattern of results did not depend on including the few older adults who were below 65, we excluded these participants and reanalyzed the data. The interpretations and conclusions from the results reported below were unaffected. All participants were screened for psychiatric illness, the use of psychoactive medications, and risk factors associated with using VR (e.g., epilepsy, motion sickness; see Supplemental Table 1 for the full list of exclusionary criteria). Older adults were also screened for cognitive impairments using the Telephone Interview for Cognitive Status (TICS; Brandt et al., 1988) ($M = 38.1$, $SD = 3.3$, cutoff < 32). No participants withdrew due to cybersickness. Some data were unavailable for ten participants due to either early withdrawal (five older; did not return for second session), not following instructions (one older), or computer malfunction during a session (one young, three older). Additionally, data for the out-of-lab PM tasks were unavailable for ten participants (four young, six old) due to either researcher error in task administration (four young), because the participant did not own a cell phone (five older), or a scheduling error (one older). Each participant provided informed consent in accordance with the procedures approved by the Institutional Research Board at the University of Notre Dame (17 June 3930).

Immersive VR game

Job Simulator is a commercially-available immersive VR videogame that entails role-playing through simulations of real-world jobs and completing associated tasks (see Figure 1). Participants completed a short-order cook and a convenience store clerk simulation. These simulations included the ongoing videogame narrative as well as a set of PM tasks that were appended by the researchers (see supplemental material). A floating computer "Job-bot" provided audible instructions for what participants needed to do in the virtual environment to advance to the next in-game task. These audio instructions were delivered through a speaker in the experimental room at a comfortable listening volume for each participant.

Participants began by completing a tutorial that covered how to interact with objects and monitor time in the VR environment. Afterward, researchers removed the HMD from the participant's head and positioned them comfortably in front of a computer in the lab space. Using the computer, participants studied a collection of seven PM tasks (see Supplemental Table 2) that were to be performed during the *Job Simulator* game narrative. The studied PM tasks consisted of four event-based intentions (e.g., turn the sink on and off every three orders to clear the drain) and three time-based intentions (e.g., "drink" a cup of water every 5 minutes). The tasks were not listed by task type or by the order in which they were to be performed in the game during study. Participants were told that the tasks needed to be memorized and executed in response to the corresponding event-based or time-based cue during gameplay. Event-based cues consisted of external events intended to trigger retrieval of the PM intention while time-based cues required monitoring a clock embedded in the in-game menu of *Job Simulator*.

Table 1. # = count, prop. = proportion, acc. = accuracy, comp. = composite, sqrt = square-root-transformed, discrep. = discrepancy, [c]*t*- and p-values were corrected due to significant Levene's test for unequal Variances; d is Cohen's d effect size with Hedges correction.

Variable	Young Adults N = 59, 46% Female				Older Adults N = 51, 57% Female						
	Mean	SEM	Skew	Kurt.	Mean	SEM	Skew	Kurt.	*t*	*p*	*d*
Age (years)	19.37	0.23	3.72	20.49	70.71	0.67	0.28	−0.42			
Job Simulator Study Trials (#)	**1.99**	**0.07**	**0.16**	**−0.03**	**3.34**	**0.21**	**1.61**	**3.72**	**−6.20c**	**<.001**	**−1.27**
Job Simulator Clock Checks (#)	**45.97**	**2.63**	**0.72**	**0.18**	**24.16**	**2.51**	**1.21**	**1.82**	**5.95**	**<.001**	**1.13**
Job Simulator Event-Regular Acc. (prop.)	**0.647**	**0.025**	**−0.811**	**0.251**	**0.439**	**0.032**	**−0.132**	**−0.864**	**5.16**	**<.001**	**0.98**
Job Simulator Event-Irregular Acc. (prop.)	**0.627**	**0.028**	**−0.026**	**−0.435**	**0.373**	**0.036**	**−0.022**	**−1.114**	**5.62**	**<.001**	**1.07**
Job Simulator Time-Regular Acc. (prop.)	**0.736**	**0.033**	**−0.756**	**−0.538**	**0.364**	**0.029**	**0.302**	**0.093**	**8.37**	**<.001**	**1.59**
Job Simulator Time-Irregular Acc. (prop.)	**0.767**	**0.031**	**−1.028**	**0.967**	**0.343**	**0.04**	**0.449**	**−0.641**	**8.38c**	**<.001**	**1.61**
Job Simulator Post-test Recall Acc. (prop.)	**0.952**	**0.011**	**−3.185**	**14.331**	**0.833**	**0.02**	**−0.985**	**1.04**	**5.19c**	**<.001**	**1.02**
Breakfast Task Study Trials (#)	1.45	0.08	0.8	−0.34	1.45	0.11	1.97	4.22	0.04	0.966	0.01
Breakfast Task Planning (comp. score)	28.89	1.02	0.79	1.9	26.91	0.97	0.25	0.05	1.4	0.163	0.26
Breakfast Task Clock Checks (#)	**16.46**	**0.74**	**0.37**	**0.01**	**11.08**	**0.69**	**0.34**	**−0.62**	**5.26**	**<.001**	**1**
Breakfast Task Cooking Time Discrep. (sec)	**00:00:36**	**00:00:02**	**0.745**	**0.544**	**00:01:20**	**00:00:04**	**0.996**	**1.518**	**−8.23c**	**<.001**	**−1.61**
Breakfast Task Average Acc. (prop.)	**0.817**	**0.024**	**−1.41**	**3.218**	**0.457**	**0.03**	**0.037**	**−0.157**	**9.52**	**<.001**	**1.81**
Breakfast Task Stop Range	00:00:56	00:00:11	3.938	17.847	00:01:35	00:00:17	2.231	5.114	−1.87[c]	0.064	−0.36
Breakfast Task Table Settings (#)	**34.9**	**1.86**	**−0.19**	**−0.7**	**24.84**	**1.61**	**0.49**	**0.21**	**4.03**	**<.001**	**0.76**
Breakfast Task Appliances Left On	1.08	0.18	1.42	1.09	0.96	0.17	1.38	1.71	0.49	0.312	0.09
Breakfast Task Serving Acc. (prop.)	**0.898**	**0.04**	**−2.705**	**5.502**	**0.569**	**0.07**	**−0.286**	**−1.998**	**4.10c**	**<.001**	**0.8**
Belongings Task Acc (sqrt sum discrep.)	4.848	0.544	−0.084	−1.632	5.001	0.693	0.047	−1.724	−0.18	0.86	−0.04
Call-back Task Acc (avg. sqrt discrep.)	**9.11**	**0.85**	**0.79**	**−0.02**	**4.73**	**0.53**	**0.99**	**1.5**	**4.37c**	**<.001**	**0.8**
OSPAN (abs. score)	**48.76**	**1.91**	**−0.69**	**0.49**	**21.57**	**2.3**	**0.49**	**−1.07**	**9.19**	**<.001**	**1.74**
Psychomotor Vigilance Task (avg. sqrt sec)	0.583	0.005	0.955	1.786	0.593	0.005	0.096	−0.736	−1.37	0.175	−0.26
Open-mindedness (comp.)	46.53	1.03	−0.57	0.67	46.51	0.88	0.09	−0.89	0.01	0.991	0
Conscientiousness (comp.)	**43.27**	**1.16**	**−0.67**	**0.32**	**49.41**	**0.95**	**−0.62**	**−0.34**	**−4.01**	**<.001**	**−0.76**
Extraversion (comp.)	42.31	1.01	−0.21	−0.17	44.02	1.12	−0.15	−0.39	−1.14	0.257	−0.22
Agreeableness (comp.)	**45.78**	**0.89**	**−0.49**	**−0.05**	**49**	**0.75**	**−0.35**	**−0.21**	**−2.71**	**0.008**	**−0.52**
Neuroticism (comp.)	30.02	1.33	0.5	−0.31	27.12	1.15	0.43	0.06	1.62	0.107	0.31

Figure 1. Examples of participants completing PM tasks in the immersive *job simulator* VR game for the convenience-store clerk (*top*) and short-order chef (*bottom*) scenario.

After studying the task list for one minute, participants were asked to vocally recall as many of the tasks as they could remember without prompting. Afterward, the researcher repeated the tasks aloud to the participant and corrected any tasks that were either recalled incorrectly or not recalled. This process was repeated until participants were able to correctly recall all of the tasks aloud without any assistance from the researcher. Young adults took an average of 1.99 ($SD = 0.49$) study trials and older adults took an average of 3.31 ($SD = 1.46$) study trials, which was a significant difference, $t(109) = 6.54$, $p < .0001$. After memorizing the list of tasks, participants then watched a brief, narrated introduction video about the impending VR simulation that included an overview of the environment (e.g., the kitchen in the short-order cook simulation), including the location of several items that were to be used in some of the PM tasks (e.g., the location of the wine bottles) (see Supplemental Figure S1).

Once participants reentered the virtual environment, before beginning the simulation, they were told to press the menu button on their handheld controller to bring up the clock and to state the current time of day aloud. The researcher informed the participant that the current time of day represented the starting time of the simulation and asked them to verbally report the target times at which they were to perform the time-based tasks (see Supplemental Figure S1 and Video). For example, if the participant began the short-order cook simulation at 12:05PM, the participant was to state that they had to place the wine into the fridge at exactly 12:17PM. After reporting the target times for each of the time-based tasks, participants played through the simulation. During the VR gameplay, screen-capture software was used to record in-game footage of participants' performance to allow for offline scoring including the frequency of clock checking behavior for control analyses (see Supplemental Video for an example).

After completing the simulation, participants' retrospective memory of the PM tasks was assessed by again asking to verbally recall the PM tasks that they had memorized earlier. Tasks that were not successfully recalled on this retrospective memory check were excluded from analysis because failure to complete the task could have resulted from a failure to encode or retrieve the content of the PM intention (i.e., retrospective memory failure) rather than a failure with remembering to perform the delayed intention at the appropriate moment (i.e., a PM failure). The percentage of tasks that were excluded from analysis due to retrospective memory failures was 4% for young adults and 15% for older adults. This difference was significant, $t(105) = 21.72$, $p < .001$.One young adult and six older adults failed to recall at least 70% of the tasks. When these participants were excluded from analysis, the difference in retrospective memory between the groups still remained (96.0% vs. 88.2%, $t(100) = 5.06$, $p < .001$)[2]

Breakfast task

The Breakfast Task was administered to provide another objective measure of naturalistic PM that was performed in the lab under direct observation. The task design and analyses were based on similar paradigms described in detail elsewhere (Altgassen et al., 2015; Craik & Bialystok, 2006; Feinkohl et al., 2016; Hering et al., 2014). Participants were asked to simulate preparing a breakfast comprising both "global" and "local" goals with varying complexity involving five different foods and beverages. Each food and beverage item that was to be prepared in the simulation required a different series of steps and had different time-interval constraints associated with their preparation (e.g., flip the bacon after 4 minutes). There were 10 of these local food/beverage preparation goals in total. The four global task goals included (1) finishing all of the foods and drinks as close together in time as possible, (2) turning off all appliances after finishing meal preparation, (3) accurately arranging the food on the plate according to the instructions before serving, and (4) finishing all of the meal preparation demands in under 13 minutes. While preparing the meal, participants were asked to perform an ongoing task that involved setting as many table settings as possible at each of four stations at a square-shaped table. After the instructions for the task were learned to criterion (perfect recall), participants made a handwritten plan about how they were going to execute the task to fulfill all of the task goals. The plan was scored to assess planning ability for control analyses. The clock that participants used to monitor the time was positioned to allow for recording the frequency of clock checking behavior for control analyses. Specific details and the task instructions are provided in the Supplemental Methods.

Belongings task

A modified version of a personal belongings task (Cuttler & Graf, 2008; Wilson et al., 1985) was administered in order to measure participants' performance on naturalistic PM tasks that are or are not personally relevant in a lab-based setting under controlled experimental conditions. The researcher and participant exchanged personal items (i.e., the participant's cell/mobile phone and the researcher's activity-tracker/step-counter) before the beginning of the in-lab experiment that needed to be returned or retrieved at the end of the experiment. When the participant came into the lab for their first testing session, after providing informed consent, the researcher asked for the participant's cell phone to be kept in a secure location so that they would not be "distracted by it during their test session". The researcher also told the participant that they would attach an activity-tracker (step-counter) to the back of their clothes to track their amount of "fidgeting during the experiment" (for details, see the explicit encoding condition of N. Rose et al., 2023). The participants were explicitly instructed to remember to both ask for their cell phone and return the activity-tracker at the end of the experimental session. Upon completion of the experiment, participants were debriefed about the session and were told that they were free to leave (i.e., there was no prompt or reminder about their phone or the tracker). A confederate experimenter whom the participant had not met or seen waited outside the testing room to see if and when they stopped to remember to ask for their cell phone back or return the activity tracker. The precise location where they stopped to remember was recorded on a copy of the floor plan of the building. PM accuracy was calculated as the sum of the distance from the target location where the participant was to retrieve the intentions before leaving the lab. If the participant completely forgot, they were stopped and reminded just before they exited the building and assigned this maximum distance value. To account for skew in the data, the square root of each participant's summed distance was recorded. Histograms, skewness, and kurtosis values were examined for both the full sample and each age-group for the participants' square root transformed summed distance as well as the number of forgotten items (0, 1, or 2). The psychometric properties in terms of approximating a normal distribution with acceptable skewness and kurtosis values were best for the square root transformed summed distance.

Call-back task

The Call-back task was administered to measure naturalistic PM performance in the real world, outside of the lab (for details, see N. S. Rose et al., 2015). The participant coordinated a day between their first and second test session in which they had a two-hour window during normal business-hours when they would be at home and available to receive and make four phone calls to the lab. At the start of the two-hour window, the researcher called the participant and told them to call the lab back and leave a message with their initials at exactly 15 and 40 minutes after they hung up. For example, if, upon hanging up the phone, the time of day was 11:13AM, the target times to return a call to the lab would be 11:28 AM and 11:53AM. The researcher noted the start time. The target time to return each phone call was not mentioned to the participant. They were to calculate it on their own and remember to call as close to the target time as possible and leave a voicemail that stated the time they intended to call. The researcher explicitly instructed the participant that they were to avoid using reminders, timers, or any other form of memory "offloading" strategy during the delay. Participants were told that the purpose of the task was for them to remember on their own. PM performance was scored as the deviation from the message time-stamp relative to the target times. The process was repeated for the second hour with target times of 20 and 35 minutes after they hung up. All participants reported complying with the instructions. If a call was not returned at any point, the actual time assessed was imputed to three standard deviations longer than the median deviation time for the participant's respective group. Mean square-root transformed deviations from the target times across the four calls represented the participants' Call-back score.[3]

Psychomotor vigilance task (PVT)

The PVT was administered to measure vigilance ability. It is a reaction time measure that consists of monitoring for a change in a dot's color that occurs at unpredictable times onscreen, and responding (via keypress) as rapidly as possible (for details, N. S. Rose et al., 2010). The variable inter-trial interval, which randomly ranged from 3000 to 7000 milli-seconds (ms) in steps of 500 ms, required that the participant remained vigilant in monitoring the dot's color change for 100 trials without a break. Trials with response times (RT) less than 200 ms or greater than 1000 ms were removed, resulting in loss of 0.82% of the data. One older adult was identified as an outlier due to an excessive number of omitted trials (26%) so their data were excluded from analysis. Mean RT for the trimmed data set was calculated. Histograms, skewness, and kurtosis values were examined for both the full sample and each age group for participants' square-root-transformed mean RT, as well as their standard deviation (SD) of RT, coefficient of variation (SD/mean), and proportion of "lapses" (response times > 500 msec). The psychometric properties in terms of approximating a normal distribution with acceptable skewness and kurtosis values were best for the square-root-transformed mean RT.

Big five inventory (BFI-II)

The BFI-II was administered to measure the Big Five personality dimensions (D. Watson & Clark, 1992). It is a well-validated assessment of personality consisting of a 60-item self-reported questionnaire. Each item contained a statement that described a general characteristic (e.g., "I am someone who is outgoing, sociable"), and participants identified the extent to which they agreed it was representative of themselves on a 5-point scale, from 1 (strongly disagree) to 5 (strongly agree). Each participant's inventory was scored for the big five domain scales of personality: open-mindedness, conscientiousness, extraversion, agreeableness, and neuroticism (see Soto & John, 2017 for scoring criteria).

Operation span task (OSPAN)

The Operation Span Task (OSPAN) was administered to measure WM capacity (Turner & Engle, 1989). It is a well-validated measure of WM capacity that consists of a dual-task procedure in which participants were asked to answer arithmetic questions while trying to remember an interspersed sequence of letters. The number of letters ranged from three to seven per trial. After each trial, participants were asked to select the letters they had seen, in order of presentation, from a set of twelve possible letters. This recall test was performed by clicking the mouse on each appropriate letter. Three trials of each set size were performed in a randomized order across the entire task. The dependent measure with the best psychometric properties in this sample was the total number of letters that were recalled in correct order on correct trials (i.e., absolute score; Conway et al., 2005).

Procedure

The experiment was completed in two separate sessions, each approximately two hours in length, and spaced approximately one week apart ($M = 7.2$ days). Participants completed both VR simulations at the start of the first in-lab testing session. Afterward, they spent the remainder of the session completing the BFI-II and the PVT. The Belongings Task began at the beginning of the first session with the initial exchange of personal belongings and completed at the end of the first session with the return of the belongings. Participants performed the Call-back task at home between the first and second sessions. In the second in-lab testing session, participants first received instructions for the Breakfast task and completed the planning stage. Each participant then performed the OSPAN task before beginning the Breakfast task to provide a filled retention-interval between Breakfast task encoding and performance. The filled retention-interval was comparable between the age groups (M $(SD) = 13.5$ (2.12) and 13.6 (2.32) min for young and older adults, respectively; $t(109) = -0.31$, $p = 0.76$).[4] For further procedural details (instructions, etc.), see the Supplemental Material.

Scoring criteria

PM task performance in *Job Simulator* and the Breakfast task was measured as the absolute difference between the ideal and actual times of task performance. Given the large number of observations, these deviation measures were converted to accuracy scores for each trial based on whether performance happened within a particular time window surrounding the ideal time (10 or 60 seconds for event and time-based tasks, respectively) analogous to previous studies (Rendell & Craik, 2000; N. S. Rose et al., 2010).[5] Performance was averaged across the two *Job Simulator* simulations per task type.

Data analysis

Before we conducted correlational analyses, we first examined the psychometric properties of each task and the extent to which it reliably measured its intended construct. To do so, we computed Spearman-Brown split-half reliability coefficients for each age group, which can be found in Supplemental Table S6, along with the bivariate Pearson correlations among all the other measures above and below the diagonal for the older and young adults, respectively. Scores from the Call-back task, PVT, and Belongings task were transformed to account for non-normality in their distributions. For the Call-back task, we square-root-transformed the time discrepancy for each call and computed the average discrepancy across these transformed observations. For the PVT, we square-root-transformed the RT on each trial and computed the average RT across these transformed observations. For the Belongings task, we summed the distances traveled from the target locations and computed the square root of this summed measure. After these transformations, measures of skewness and kurtosis ranged from – 0.084 to 0.99 and – 1.724 to 1.786, respectively, which reflect approximately normal distributions (skewness <2, kurtosis < 4, Kline, 1998).

To assess the age-PM paradox, an index was calculated to provide the most appropriate, direct comparisons of age-differences in PM performance on the most similar tasks (time-based tasks with similar retention intervals) that primarily differed in that they were either performed in lab-based or real-world settings. This involved calculating the difference between an individual's performance on the time-based PM tasks during *Job Simulator* or Breakfast Task and the Call-back task. The size and direction of these differences between the same individual's performance on time-based PM in the different task-settings provided a key index of the age-PM paradox that we attempted to predict with the cognitive and personality variables in hierarchical, multiple-regression analyses. Adjusted R-squared and Bonferroni corrections were used to control for the multiple parameters and comparisons.

Results

To outline the results, we first examined age differences in PM performance across the four paradigms and the relations among the various measures of performance from those paradigms. Afterward, we examined the relations between PM performance and the cognitive and personality predictors. Finally, we conducted hierarchical regression modeling on a novel age-PM index to test the hypothesis that the age-PM paradox is partly due to there being different predictors of young and older adults' PM performance across lab and real-world settings.

Age differences in PM across task settings

The means and psychometric properties for each variable and each group are in Table 1 with the between-group *t*-, *p*-, and Cohen's d-values.

Job simulator

First, we conducted a mixed-design ANOVA to examine age differences in PM accuracy on the *Job Simulator* tasks and any potential interactions with cue type (event-based, time-based) and task regularity (regular, irregular) as within-subjects factors, and age group as a between-subjects factor. There was a main effect of age group with young adults outperforming older adults across the four task types, $F(1,109) = 108.90$, $p < 0.001$, $\eta_p^2 = 0.500$ (see Figure 2). We also observed an interaction between age-group and cue-type with larger age differences on time- based tasks (Y: $M = 0.75$, $SD = 0.21$; O: $M = 0.36$, $SD = 0.21$) than event-based tasks (Y: $M = 0.64$, $SD = 0.18$; O: $M = 0.41$, $SD = 0.18$; $F(1,109) = 13.35$, $p < 0.01$, $\eta_p^2 = 0.11$). There was neither a main effect, $F(1,109) = 1.34$, $p = 0.25$, nor any interaction with task regularity, $Fs(1,109)<2.16$, $p > 0.15$; thus, further analysis collapsed over this factor.

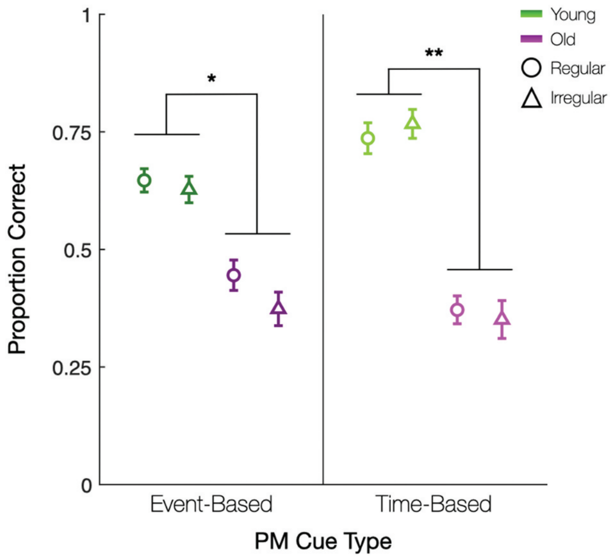

Figure 2. Mean proportion of correctly performed *job simulator* PM tasks in young and older adults. Error bars indicate standard error of the mean. *$p < 0.05$; **$p < 0.01$.

Given the cue type by age interaction, we investigated the extent to which time monitoring behavior (i.e., number of clock checks) impacted participants' ability to perform time-based PM tasks correctly. We conducted bivariate correlations on individuals' mean number of clock checks and mean time-based PM task accuracy within the young and older adult groups separately. Both young adults', $r = 0.50$, $p < 0.01$, and older adults', $r = 0.66$, $p < 0.01$, time-based PM performance was positively correlated with the number of clock checks. There was also a significant difference in the average number of clock checks between young ($M = 46.0$, SD = 20.2) and older adults ($M = 24.5$, SD = 17.9), $t(109) = 5.89$, $p < 0.01$. However, after controlling for the number of clock checks as a covariate in an ANCOVA, age differences in time-based PM accuracy remained, $F(1,108) = 49.05$, $p < 0.01$, $\eta_p^2 = 0.31$, with young adults ($M = 0.69$, $SD = 0.18$) still significantly outperforming older adults ($M = 0.43$, $SD = 0.18$). Thus, young adults outperformed older adults on all of the PM tasks in *Job Simulator*, particularly time-based measures – a pattern that was not solely explained by the observed differences in time-monitoring behavior. Correlational analyses reported below help to explain the source of this differential age deficit.

Breakfast task

Next, we conducted a mixed-design ANOVA with ten levels for the different cooking tasks (i.e., local goals) as a within-subjects factor and age group as a between-subjects factor to investigate for possible age-related differences in performance. Mean accuracies for each group and task are reported in Supplemental Figure S2. The test for between-subjects effects revealed significant age-related deficits in the proportion of PM tasks accurately performed, $F(1,109) = 90.55$, $p < 0.01$, $\eta_p^2 = 0.45$. A main effect of task was also observed, $F(9,981) = 4.58$, $p < 0.01$, $\eta_p^2 = 0.04$. However, the task by age-group interaction was not significant, $F(9,981) = 0.53$, $p = 0.85$, $\eta_p^2 = 0.01$. Therefore, subsequent analyses involving Breakfast task performance collapsed over task type.

With regards to the global task goals, independent samples t-tests revealed that, although there was no difference between young adults ($M = 1.08$, $SD = 1.42$) and older adults ($M = 0.94$, $SD = 1.19$) in the number of appliances left powered on, $t(109) = 0.57$, $p = 0.57$, young adults ($M = 0.90$, $SD = 0.30$) were significantly better than older adults ($M = 0.58$, $SD = 0.50$) at remembering to serve the food as instructed at the end of the task, $t(109) = 4.15$, $p < 0.01$, stopping the foods closer together in time ($M = 56$ sec., $SD = 90$ sec. vs. $M = 94$ sec., $SD = 122$ sec. respectively, though this difference failed to reach significance, $t(109) = -1.87$, $p = 0.06$), and finishing all tasks closer to the 13-minute cutoff time ($M = 28$ sec., $SD = 48$ sec. vs. $M = 98$ sec., $SD = 102$ sec., $t(109) = -4.714$, $p < 0.0001$). Overall, older adults showed decrements in performance on all tasks relative to young adults, except for the two global task goals of finishing cooking the food and beverages at the same time and turning off the appliances. Young and older adults' mean proportions of responses for each task type on both the *Job Simulator* and Breakfast tasks are presented in Supplemental Table 5 for the different response categories (time-based VR PM: on time <10 sec.; a little late <30 sec.; very late >30 sec.; time-based Breakfast task PM: on time <60 sec.; a little late <90 sec.; very late >90 sec).

Similar to our analysis of *Job Simulator* performance, we investigated whether performance on the Breakfast task was influenced by time monitoring behavior or planning ability. For time monitoring, we found that young adults ($M = 16.46$, $SD = 5.67$) checked the clock significantly more often than older adults ($M = 11.15$, $SD = 4.94$), $t(109) = 5.22$, $p < 0.001$. However, as with clock checking in *Job Simulator*, an ANCOVA with the number of clock checks as a covariate showed that significant age-differences remained, $F(1,108) = 55.10$, $p < 0.01$, $\eta_p^2 = 0.34$, with young adults ($M = 0.79$, $SD = 0.20$) still outperforming older adults ($M = 0.49$, $SD = 0.20$). We measured planning ability using a point-based, composite score across three sub-categories: prioritization, rule description, and specification of action (see Supplemental Methods). In contrast to time monitoring behavior, planning ability was shown to yield no significant differences between young ($M = 28.89$, $SD = 7.87$) and older adults ($M = 26.90$, $SD = 6.84$), $t(109) = 1.41$, $p = 0.16$. Thus, despite comparable planning abilities, and the advantage of having prepared breakfast many times over their lifetime, older adults still performed the Breakfast task less efficiently and accurately than young adults. This age-deficit in simulating a naturalistic PM task dovetails with the age-deficit observed in PM performance during *Job Simulator*.

Real-world PM

We conducted independent samples t-tests to assess age differences on the Belongings and Call-back tasks. For the Belongings task, young adults ($M = 0.79$, $SD = 0.27$) did *not* remember to retrieve their cell phone and return the activity tracker more often than older adults ($M = 0.71$, $SD = 0.37$), $t(99) = 1.28$, $p = 0.20$. There was also no difference between young ($M = 2.81$, $SD = 2.76$) and older adults ($M = 3.38$, $SD = 3.34$) for the distance-traveled measure, $t(99)$=-0.94, $p = 0.35$. An ANOVA with item-type (cell-phone, activity-tracker) as a within-subjects variable, and age-group as a between-subjects variable showed that both young and older adults were significantly better at remembering to retrieve their cell-phone than remembering to return the activity-tracker, $F(1,99) =$

24.51, $p < 0.01$. The item-type by age-group interaction was not significant, $F(1,99) = 0.20$, $p = 0.66$.

For the Call-back task, we found that age-related differences in performance were *reversed*, with older adults showing significantly *smaller* absolute deviation times ($M = 58$ sec., $SD = 76$ sec.) in their phone calls than young adults ($M = 192$ sec., $SD = 232$ sec.), $t(108) = 3.96$, $p < 0.01$. Thus, in contrast to the large age-related deficits observed in PM performance on both the *Job Simulator* and Breakfast tasks, older adults performed comparably or outperformed young adults on the Belongings and Call-back tasks, respectively.

Relations among lab-based and real-world PM

To assess the relations among naturalistic PM measures performed in lab-based and real-world settings, correlations among the various measures of PM performance were examined within each age group (see Figure 3). The correlation matrix displayed in Table 2 reports each bivariate Pearson correlation coefficient between the measures of the *Job Simulator* task, the Breakfast task, the Belongings task, the Call-back task, and age within each group, with the older adults below the diagonal and young adults above the diagonal. This correlational analysis is important for assessing the construct validity of

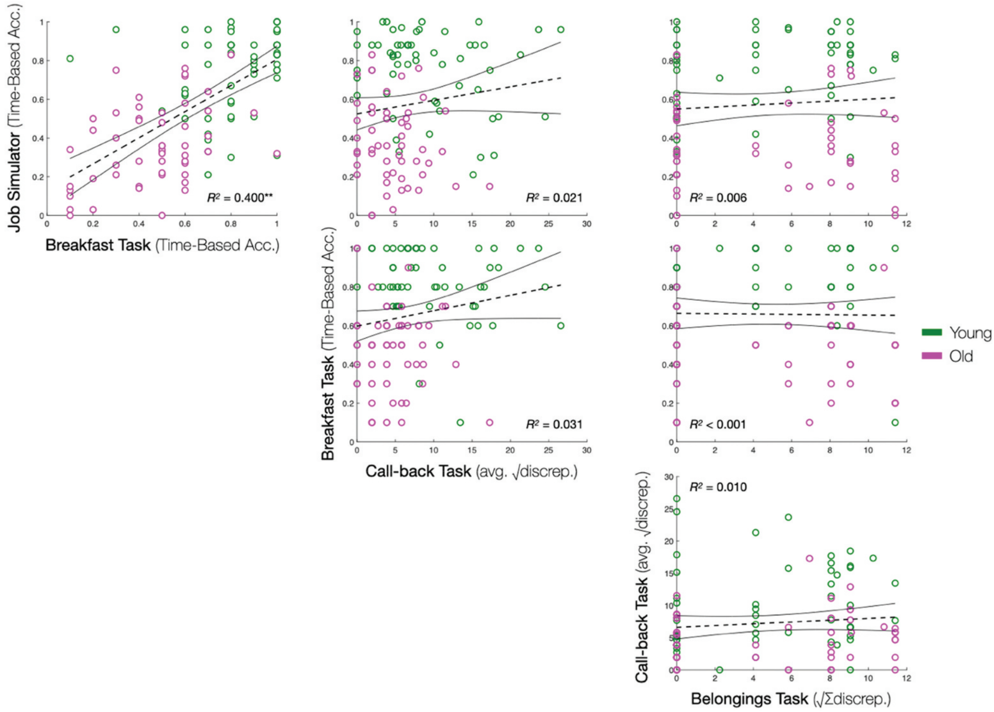

Figure 3. Performance relationships amongst lab-based and real-world PM tasks. The black dashed and solid lines indicate the best-fitting regression model and 95% confidence interval, respectively; the R^2 value indicates the proportion of variance explained across individuals in both age groups; to assess correlations within each age group, see Table 2. $^{**}p < 0.01$.

Table 2. Correlations among the measures from the PM paradigms within each age group, with young adults above the diagonal and older adults below the diagonal. Bolded values indicate $p < 0.05$; bolded and italicized values indicate $p < 0.01$. Green and red color coding indicate positive and negative correlations, respectively.

Variable		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Age	1		-0.013	0.188	***0.408***	-0.070	0.187	0.097	0.122	-0.154	0.071	-0.105	-0.059	-0.016	0.085	-0.150	-0.244	-0.033	-0.040	0.158	-0.051	-0.198	-0.089	-0.060
Job Simulator Study Trials	2	0.084		-0.050	**0.334**	-0.062	-0.053	-0.073	-0.007	0.215	0.075	-0.053	0.032	0.115	-0.092	0.067	0.125	-0.098	-0.151	0.103	0.112	-0.219	0.123	0.083
Job Simulator Clock Checks	3	**-0.279**	**-0.344**		0.234	0.194	***0.503***	***0.487***	0.192	0.074	0.021	**0.271**	-0.068	-0.118	0.018	-0.197	***-0.350***	0.034	**-0.279**	0.265	-0.024	-0.192	-0.143	-0.068
Job Simulator Total Game Time	4	***0.383***	**0.305**	-0.205		-0.248	-0.030	-0.168	-0.024	0.218	-0.127	-0.251	0.052	0.180	-0.121	0.065	-0.145	0.008	**-0.325**	-0.040	0.110	-0.037	0.117	0.142
Job Simulator Event Acc.	5	***-0.373***	***-0.383***	***0.366***	-0.256		0.119	***0.673***	***0.362***	-0.186	0.030	***0.394***	-0.063	-0.129	0.023	-0.244	**-0.273**	-0.028	-0.026	**0.289**	-0.116	-0.213	***-0.377***	***-0.326***
Job Simulator Time Acc.	6	***-0.368***	-0.249	***0.654***	-0.234	***0.410***		***0.814***	0.239	-0.021	-0.106	0.214	-0.045	**-0.307**	0.217	-0.119	-0.217	0.075	-0.039	**0.285**	-0.054	-0.196	-0.148	-0.175
Job Simulator Total Acc.	7	***-0.437***	***-0.379***	***0.609***	**-0.290**	***0.840***	***0.838***		***0.388***	-0.122	-0.061	***0.395***	-0.075	**-0.306**	0.178	-0.234	**-0.320**	0.041	-0.045	***0.384***	-0.109	**-0.274**	***-0.334***	**-0.323**
Job Simulator Post-Test Recall Acc.	8	-0.238	-0.113	***0.441***	-0.140	***0.442***	***0.538***	***0.585***		-0.266	0.155	0.084	0.056	-0.166	0.118	-0.125	-0.148	-0.099	0.249	0.255	-0.127	-0.195	-0.232	-0.199
Breakfast Task Study Trials	9	0.010	***0.372***	-0.107	0.093	-0.132	**-0.295**	-0.254	-0.197		0.023	-0.150	-0.082	0.175	-0.066	-0.018	0.030	0.221	-0.139	-0.242	0.260	0.152	0.077	0.001
Breakfast Task Planning	10	0.105	-0.093	-0.036	-0.113	-0.126	0.097	-0.018	-0.100	-0.023		-0.035	0.035	0.050	-0.201	-0.206	-0.113	-0.176	0.168	0.063	0.046	-0.114	-0.187	-0.098
Breakfast Task Clock Checks	11	-0.009	-0.222	**0.328**	0.039	0.109	0.167	0.165	-0.013	-0.128	0.013		0.094	**-0.302**	0.221	0.122	**-0.312**	0.044	0.037	0.196	0.041	-0.183	-0.071	-0.062
Breakfast Task Total Game Time	12	0.119	0.083	***-0.376***	0.022	-0.130	-0.267	-0.235	**-0.312**	0.148	**0.303**	***-0.468***		0.184	-0.158	0.172	-0.035	-0.018	0.060	0.067	0.161	-0.048	-0.106	-0.193
Breakfast Task Average Cooking Discrep.	13	0.151	0.177	***-0.421***	0.150	-0.081	**-0.344**	-0.252	**-0.300**	0.089	0.168	**-0.344**	***0.819***		***-0.882***	0.070	0.025	**0.292**	-0.217	-0.214	0.030	0.054	0.197	0.190
Breakfast Task Average Cooking Acc.	14	-0.141	-0.151	***0.487***	-0.108	0.051	***0.395***	0.265	0.199	-0.113	-0.039	***0.386***	***-0.562***	***-0.836***		-0.023	0.093	-0.191	0.203	0.102	0.101	-0.004	-0.111	-0.154
Breakfast Task Stop Range	15	0.017	0.059	**-0.294**	***0.372***	0.009	-0.271	-0.160	-0.214	0.112	0.004	0.205	0.029	**0.321**	**-0.340**		0.137	0.018	-0.109	-0.024	0.001	-0.020	**0.257**	0.115
Breakfast Task Serving Acc.	16	-0.256	0.007	0.019	-0.148	0.058	-0.045	0.005	-0.026	-0.058	-0.049	0.078	-0.032	0.174	-0.197	0.234		0.180	0.021	-0.135	0.092	0.085	0.117	0.065
Breakfast Task Appliances Left On	17	0.043	0.008	0.053	-0.221	-0.078	0.111	0.020	0.118	-0.049	-0.009	-0.154	0.083	0.032	-0.007	-0.219	-0.129		-0.111	-0.038	0.076	-0.025	0.227	0.136
Breakfast Task Table Settings	18	***-0.501***	-0.089	***0.378***	**-0.355**	***0.405***	0.251	***0.392***	***0.367***	-0.129	-0.206	0.129	***-0.480***	***-0.467***	**0.278**	-0.115	0.099	-0.084		0.003	-0.085	0.104	-0.048	-0.048
Belongings Task Distance Travelled	19	0.151	-0.114	0.025	***0.475***	0.192	-0.062	0.084	0.051	0.220	-0.183	-0.060	0.100	0.131	-0.098	0.205	-0.235	0.060	-0.184		***-0.410***	***-0.794***	-0.019	0.122
Belongings Task Cell Phone Acc.	20	-0.199	-0.042	0.252	***-0.449***	0.048	0.260	0.183	0.269	-0.050	-0.018	0.060	-0.052	-0.130	0.199	**-0.307**	0.233	0.145	0.253	***-0.597***		0.012	-0.039	-0.108
Belongings Task Activity Tracker Acc.	21	-0.110	0.259	-0.163	***-0.349***	***-0.317***	-0.038	-0.218	-0.097	-0.105	0.242	0.066	0.004	-0.022	0.007	-0.103	0.190	-0.091	0.039	***-0.898***	***0.480***		0.064	-0.027
Call-back Task Average Discrep.	22	-0.028	-0.048	-0.200	-0.124	0.020	-0.157	-0.083	**-0.343**	0.267	0.005	0.028	***0.437***	**0.339**	-0.096	0.001	0.017	-0.117	-0.198	0.148	0.038	-0.051		***0.906***
Call-back Task Transformed Discrep.	23	-0.030	-0.013	-0.218	-0.175	0.049	-0.206	-0.098	**-0.292**	0.279	-0.107	0.082	**0.284**	0.211	-0.043	0.043	0.038	-0.097	-0.047	0.122	0.048	-0.080	***0.876***	

the different measures of PM and for testing the hypothesis that the age-PM paradox is due in part to there being different predictors of young and older adults' PM performance on tasks performed in different task settings. However, given the multiple comparisons and the relatively exploratory nature of this analysis, caution must be taken with forming strong conclusions based on these results. The regression analyses reported below are based on the summary measures of PM performance with appropriate psychometric properties from both the novel *Job Simulator* paradigm and the other, previously validated PM paradigms.

First, we tested our hypothesis that PM performance on *Job Simulator* was related to PM performance on the Breakfast task, especially along time-interval PM demands (e.g., Bailey et al., 2010; Haines et al., 2020). As shown in Figure 3, time-based PM performance in *Job Simulator* was indeed predicted by time-based PM performance in the Breakfast task (average cooking time discrepancy) for both the young adults (r=-.307) and the older adults (r=-.344). Partial correlation analyses were conducted to determine whether or not this relationship persisted irrespective of variance attributable to age. In doing so, we found that the relationship between time-based PM accuracy in *Job Simulator* and the Breakfast task remained significant after controlling for the discrete variable of age group, $r = 0.31$, $p < 0.01$. Linear regressions performed within each age group also indicated that older adults' time-based PM accuracy on Job Simulator was significantly predicted by their time-based PM accuracy on the Breakfast task, $F(1,50) = 9.77$, $p < 0.01$, $R^2 = 0.16$, even when we controlled for variation in the continuous variable of age within the older adult group, $r = 0.37$, $p < 0.01$. Thus, as anticipated, individual differences in time-based PM performance in *Job Simulator* were meaningfully related to time-based performance on the Breakfast task, especially for older adults.

Overall PM performance on *Job Simulator* was also significantly predicted by the number of clock checks and posttest recall accuracy (retrospective memory). When we investigated the correlations within each age group, we found that older adults' performance was predicted by the total time to complete the *Job Simulator* task and by the number of table settings they completed on the Breakfast task. Young adults' performance was predicted by the Breakfast task clock checks, average cooking time discrepancy, serving accuracy, and by performance on both the Belongings and Call-back tasks.

Thus, there were significant relations among PM performance for both the young adults and older adults, especially between the time-based tasks on *Job Simulator* and Breakfast. Notably, time-based PM on *Job Simulator* was not significantly correlated with time-based PM from the Call-back task performed in the real-world for either group.

PM performance on the Breakfast task, in addition to its aforementioned relationship with *Job Simulator* performance, was primarily predicted by its own measures, including clock checks within each group, total time, stop range, table settings, and by the number of appliances left on for young adults, as well as mean Call-back discrepancy for older adults. Therefore, the Breakfast task showed good construct validity with PM performance on *Job Simulator* for young adults, and with PM performance on the Call-back task for older adults. Therefore, the relations between PM task paradigms differed between the groups, especially for time-based PM performed in the lab and in the real-world.

We then moved to investigate predictors of PM performance on our real-world tasks. PM performance on the Belongings task was predicted by *Job Simulator* PM performance for the young adults; however, the positive correlation (.384) means that those who had higher accuracy on *Job Simulator* had *larger* distances (i.e., worse performance) on the Belongings task for the young adults. For the older adults, the Belongings task was predicted by *Job Simulator* game time because those who took longer to complete the game had larger distances (i.e., worse performance). Therefore, PM performance on the Belongings task did not show good construct validity based on the correlations with other PM performance measures.

For the Call-back task, older adult performance was predicted by average time discrepancy and total task time on the Breakfast task, as well as *Job Simulator* posttest task recall (retrospective memory). Call-back performance for young adults was predicted by overall PM performance on *Job Simulator*. Therefore, unlike the Belongings task, PM performance on the Call-back task showed moderate construct validity for both young and older adult groups, but via different PM task paradigms.

Cognitive and personality predictors of PM

Next, we conducted a series of hierarchical multiple regression analyses to predict performance on each PM task using age and independent measures of WM, vigilance, and personality traits. Figure 4 shows the predictive relationships from each model that persisted after controlling for age. Each model was constructed using a sequential (i.e., Type 1) sum of squares approach to determine how the predictors contributed to performance after controlling for age. In predicting each measure of PM, model 1 assessed the effect of age to determine the amount of age-related variance. Then model 2 added the WM, vigilance, and personality traits to see if each predictor contributed unique variance over and above the variance explained by age. Table 3 shows the bivariate correlations between these measures.

For both event-based and time-based *Job Simulator* performance, the net contribution of WM, vigilance, and personality trait measures accounted for a significant amount of the variability (EB: $F(7,103) = 6.03$, $p < 0.01$, $R^2 = 0.29$; TB: $F(7,103) = 8.85$, $p < 0.01$, $R^2 = 0.38$). However, 8% and 17% of the residual variance in event- and time-based PM, respectively, was attributable to age alone. For event-based VR performance, WM significantly contributed to the model above and beyond the effect of age ($\beta = 0.25$, $t(110) = 2.38$, $p =$

0.02), while the other predictors, including agreeableness (β =-0.08, t(110)=-0.84, $p = 0.40$) and conscientiousness ($\beta = 0.11$, t(110) = 1.22, $p = 0.23$) did not. For time-based VR performance, agreeableness was the only predictor that significantly contributed to the model above and beyond the effect of age (β=-0.22, t(110)=-2.72, $p < 0.01$). While WM did predict time-based VR performance initially while controlling for vigilance and personality ($\beta = 0.47$, t(110) = 5.64, $p < 0.01$), the relationship was no longer significant after also controlling for age ($\beta = 0.15$, t(110) = 1.71, $p = 0.09$). Moreover, we found that the association between WM and time-based VR performance within the older adult group was not significant when controlling for age ($r = 0.17$, $p = 0.22$).

For the Breakfast task, the net contribution of our predictors was again able to account for a significant amount of the variance in Breakfast Task performance, F(7,103) = 5.47, $p <$ 0.01, $R^2 = 0.27$, with 26% of the residual variance accounted for by age. Within this model, only agreeableness was found to be significantly associated with PM performance after controlling for age (Figure 4, β=-0.28, t(110)=-3.38, $p < 0.01$). Similar to our model predicting time-based VR performance, we found a significant association between WM and Breakfast task performance initially, while controlling for vigilance and personality traits ($\beta = 0.35$, t(110) = 3.83, $p < 0.01$), but this relationship was no longer significant after also controlling for age (β=-0.04, t(110)=-0.48, $p = 0.64$).

Finally, we ran two hierarchical multiple regression analyses to determine how variability in WM, vigilance, and personality was associated with PM performance on the Belongings and Call-back tasks, respectively. In the Belongings task, which did not show an age difference in performance, the regression analysis revealed that there was no reliable association between the potential predictor variables and PM performance, F (7,93) = 0.41, $p = 0.89$, $R^2 = 0.05$. With regards to the Call-Back task, the predictors were able to account for a significant amount of the variance, F(7,102) = 3.92, $p < 0.01$, $R^2 = 0.21$, but left a residual 12.7% of the variance to be accounted for by age. Amongst the predictor variables that contributed to the Call-back task, WM (β=-0.258, t(109)=-2.40, p = 0.02) and vigilance ($\beta = 0.31$, t(109) = 3.51, $p < 0.01$) remained significant after accounting for age, with poorer Call-back performance being associated with poorer WM and vigilance (Figure 4). For a comprehensive report of the results from each hierarchical model, see Table 4.

PM-Paradox index

Following these hierarchical regression analyses for each PM task, we conducted an analysis to try and identify the source of paradoxical age-differences in PM performance across lab and real-world contexts. We hypothesized that paradoxical age differences in PM performance across task settings—i.e., the age-PM paradox – were driven by different predictor variables. We created an index of these paradoxical performance patterns by comparing time-based PM measures from the Call-back task to time-based PM measures from *Job Simulator* and the Breakfast task. We normalized (i.e., z-scored) the data from each task and computed the difference between each individual's Call-Back task performance and their performance on *Job Simulator* (i.e., JS-CB), and the Breakfast task (i.e., BF-CB), respectively. We then repeated our hierarchical regression analyses with these PM-paradox indices as outcome variables to determine the extent to which WM, vigilance, and personality traits predicted age differences in PM performance across task settings.

Importantly, repeating this approach for the JS-CB and BF-CB indices separately allowed us to assess the replicability of any observed relationships. Figure 5 shows the predictive relationships in each of these model that persisted after controlling for age. For a comprehensive report of the results from each hierarchical model, see Table 5.

For the JS-CB index, our moderators were able to account for a significant amount of the observed variance over and above age. Vigilance ($\beta = 0.15$, $t(109) = 2.10$, $p = 0.04$) and neuroticism (β =-0.16, $t(109)$=-2.06, $p = 0.04$) were found to be significantly associated with performance differences while controlling for age. WM ($\beta = 0.45$, $t(109) = 5.00$, $p <$ 0.01) was associated with performance differences while controlling for vigilance and personality traits, but was no longer significant after controlling for age ($\beta = 0.08$, $t(109) =$ 0.83, $p = 0.41$).

For the BF-CB index, the same moderators were again able to account for a significant amount of the variance, $F(7,102) = 4.77$, $p < 0.01$, $R^2 = 0.25$, but left a residual 31.8% of the variance to be accounted for by age. The roles of vigilance ($\beta = 0.21$, $t(109) = 2.92$, $p < 0.01$),

and neuroticism (β =-0.16, $t(109)$=-2.10, $p = 0.04$) were replicated in this second model, with an additional effect of agreeableness (β =-0.20, $t(109)$=-2.60, $p = 0.01$). Moreover, the significant effect of WM capacity was replicated when controlling for vigilance and personality traits ($\beta = 0.36$, $t(109) = 3.87$, $p < 0.01$) and persisted in the BF-CB model after including age ($\beta = -0.08$, $t(109)$=-0.93, $p = 0.36$). Taken together, these findings revealed that paradoxical age differences in PM performance in the lab and real-world were partially predicted by WM and vigilance, and by the personality traits of agreeableness and neuroticism.

Discussion

In the present study, we tested a theoretical account of the age-PM paradox that explains conflicting patterns of age-differences in PM performance across lab and real-world settings as a consequence of developmental changes in both cognitive and motivational factors associated with personality traits that differentially influence performance across settings. We measured age differences between the same young and older adults who performed similar naturalistic PM tasks inside and outside of the lab, including tasks in a novel, immersive VR-based paradigm. Despite the immersive, interactive nature of our in-lab VR and Breakfast task assessments of PM, and the

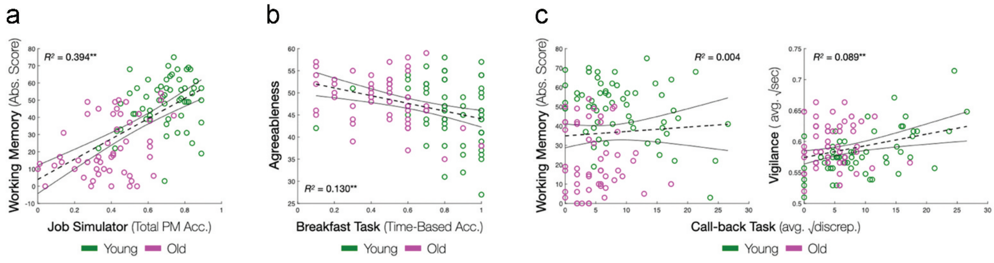

Figure 4. Predictors of (a-b) lab-based and (c) real-world PM. The black dashed and solid lines indicate the best-fitting regression models and 95% confidence intervals, respectively; R^2 values indicate the proportion of variance explained across individuals in both age groups; to assess correlations within each age group, see Table 3. *p < 0.05; **p < 0.01.

Table 3. Correlations between age, the measures of PM performance, working memory capacity, vigilance, and personality traits within each group with older adults above the diagonal and young adults below the diagonal. Bolded values indicate $p < 0.05$; bolded and italicized values indicate $p < 0.01$. Green and red color coding indicate positive and negative correlations, respectively.

Variable		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Age	1		***−0.387***	***−0.432***	***−0.482***	0.170	−0.123	−0.077	**−0.331**	−0.044	−0.154	−0.216	−0.086	**0.102**	−0.110
Job Simulator Event-Based Acc.	2	−0.070		***0.424***	***0.839***	−0.089	−0.197	0.063	**0.352**	−0.184	0.070	0.010	−0.166	**−0.276**	0.248
Job Simulator Time-Based Acc.	3	0.187	0.119		***0.848***	**−0.352**	0.074	−0.163	0.234	−0.184	−0.150	−0.023	−0.134	**−0.344**	−0.013
Job Simulator Total Acc.	4	0.097	***0.673***	***0.814***		−0.263	−0.075	−0.065	**0.348**	−0.224	−0.048	−0.011	−0.177	***−0.372***	0.139
Breakfast Task Time-Based Acc.	5	−0.016	−0.129	**−0.307**	**−0.306**		−0.071	0.199	−0.003	0.002	0.124	−0.148	−0.144	***0.355***	0.159
Belongings Task Event-Based Acc.	6	−0.195	−0.241	−0.195	**−0.290**	0.061		−0.040	−0.026	**−0.315**	0.115	0.206	−0.068	0.102	**−0.299**
Call-back Task Time-Based Acc.	7	−0.060	**−0.326**	−0.175	**−0.323**	0.190	−0.077		−0.087	0.242	−0.006	−0.113	−0.107	0.205	−0.080
OSPAN	8	−0.224	**0.275**	0.136	**0.264**	−0.086	−0.034	***−0.362***		−0.215	0.070	0.003	0.034	0.066	0.101
Psychomotor Vigilance	9	−0.132	−0.236	−0.164	**−0.262**	0.186	0.110	***0.542***	−0.182		−0.037	−0.186	0.084	0.170	−0.102
Open-mindedness	10	−0.010	−0.003	−0.157	−0.121	0.165	−0.115	**0.269**	−0.073	**0.271**		0.186	0.107	−0.064	0.138
Conscientiousness	11	−0.015	0.215	0.020	0.145	0.075	−0.076	−0.034	0.040	−0.095	0.013		0.183	−0.022	−0.106
Extraversion	12	0.175	−0.017	0.017	0.004	0.105	0.045	0.185	−0.142	0.141	0.255	***0.380***		0.168	**−0.283**
Agreeableness	13	−0.189	0.152	−0.181	−0.047	0.171	0.139	0.016	0.008	0.114	0.137	***0.474***	***0.438***		***−0.375***
Neuroticism	14	0.190	−0.199	0.124	−0.024	−0.190	0.048	***−0.340***	0.199	**−0.272**	−0.028	−0.133	***−0.388***	***−0.447***	

relatively familiar demands embedded within them (e.g., meal preparations, convenience store transactions), we found large age-related deficits in PM performance. Of note, these age-related deficits could not be explained by age- and individual-differences in time-monitoring behaviors or planning ability. In contrast, we found a lack of age differences in remembering to retrieve and return personal belongings after the in-lab session (i.e., real-world event-based PM), and an age *benefit* in remembering to call the lab back after relatively short intervals while participants were at home going about their daily lives (i.e., real-world time-based PM). That is, despite our efforts to assess similar naturalistic PM performance in the same young and older adults across in lab and real-world settings, the paradoxical pattern of age differences across settings was replicated.

By including time-interval PM intentions that were as similar as logistically possible across the VR, Breakfast, and real-world tasks, we were able to test our novel hypothesis by creating a "PM-paradox-index" that assayed individual differences in young and older adults' PM performance across the lab and real-world contexts. Using hierarchical regression modeling, we found that our independent measures of WM, vigilance, and personality traits were able to predict young and older adults' PM performance. Critically, however, the specific predictive variables from these measures were different between age groups and their predictive power differed between task contexts, despite similar PM demands. In the lab, PM performance was generally predicted by WM in both groups and by agreeableness within the older adult group (Table 3). In contrast, real-world PM performance on the Call-back task was positively associated with WM, vigilance and neuroticism, primarily within the young adult group (Table 3). Across task contexts, individual differences in WM, vigilance, agreeableness, and neuroticism predicted patterns of PM performance, even after controlling for age and individual differences in other personality traits. The roles of vigilance and neuroticism were replicated across both the VR and Breakfast task indices, attesting to the reliability of this novel finding. Below, we discuss these findings, the importance of ecological validity for PM assessments, the validation of our novel VR-based task, and the theoretical implications for resolving the age-PM paradox.

Validating measures of PM during immersive virtual reality gameplay

Previous attempts to address the role of ecological validity in the age-PM paradox have employed the use of VR-based tasks to recreate PM scenarios as they exist outside of the lab. We used an immersive VR method that preserves real-world perceptual and motor interactions and tried to validate this method by showing convergent findings with an objective measure of naturalistic PM. This approach builds on previous work that has established convergence between immersive VR and a subjective measure of everyday memory (Ouellet et al., 2018). We hypothesized and observed that PM performance during the *Job Simulator* game significantly predicted naturalistic PM performance measured with the Breakfast task.

Our regression analyses showed that performance on *Job Simulator* was able to predict performance on the Breakfast task beyond the variance due to age, particularly for time-based PM performance. This supports both the ecological validity of the novel VR task and the predictions of the multi-process model of PM (McDaniel & Einstein, 2000). Specifically, while young adults outperformed older adults across all PM task types in *Job Simulator*, performance differences were substantially larger for tasks that had time-based cues than event-based cues. This pattern suggests that tasks requiring conscious monitoring of external cues, rather than spontaneous retrieval, show greater age deficits.[6] The results also showed that individual differences in WM played a greater predictive role in time-based PM task accuracy than in event-based tasks. This supports a tenet of models of PM that propose that the degree of controlled, strategic processing involved in a given PM task modulates the demand on WM (Brandimonte et al., 2001; d'Ydewalle et al., 2001, Cona et al., 2014; Einstein et al., 1995; Park et al., 1997; N. S. Rose et al., 2015).

That said, the greater amount of time monitoring (i.e., checking the clock) performed by young adults compared to older adults in both *Job Simulator* and the Breakfast task was unable to fully account for the sizable age differences in performance. This suggests that the older adults who updated their WM by checking the clock as often as young adults were still unable to execute PM intentions as effectively as the young adults did. It may be the case that age differences in PM performance rely less on the proactive use of strategic monitoring, and more so on the reactive use of spontaneous retrieval processes triggered by the presentation of PM cue-related information in the environment. This proposal applies a prominent theory of cognitive control to PM and aging and is consistent with some age differences in proactive versus reactive control mechanisms in other domains (Lamichhane et al., 2018; McDaniel et al., 2013; Rummel & Kvavilashvili, 2023). It should also be noted that additional factors likely contribute to the pattern of findings as the proportion of unexplained variance is considerable. Prior experience with video games, especially VR, or computerized technologies may play a role in the ease or speed with which the participants could learn to perform the tasks (or even occupational/retirement status, especially given that many PM tasks assess occupation-related behaviors). Future studies should assess and control for such factors including age differences in experience with (or potential aversions to) such digital/gaming technologies.[7] For example, age-related stereotype threat or older adults' self-perceptions about their proficiency with new technologies or memory functioning could have impacted the results by causing them to believe that they would not be able to remember all of the tasks or by causing them to engage in extra mnemonic strategies or strategic monitoring processes than the young adults who, conversely, were likely *over*confident in their ability to remember to get their cellphone back, return the experimenter's activity tracker, and call the lab back at the instructed times.

On the benefits of experience to naturalistic PM performance

Differences in age-related patterns of performance on the Breakfast task also provide insight into how real-world experience affects naturalistic task performance. Older adults showed comparable global task performance relative to young adults in 1) their preparatory planning, 2) their ability to finish all of the food items together, and 3) remembering to turn off cooking appliances. In contrast, older adults showed deficits in local task performance relative to young adults in 1) meeting the time constraints for each individual food item, 2) serving the food according to the instructed arrangement, and 3) finishing all items within 13 minutes. Compared to a computerized version of the Breakfast Task, which also involved setting the table and preparing foods, this pattern

Table 4. Results from Regression Models predicting PM performance from age and the cognitive and personality predictors. For associations within each group, see Table 3. Bolded values indicate $p < .05$.

Model	Variable		β	*F* or *t*	*p*	Model	Variable		β	*F* or *t*	*p*
Job Simulator (Total Accuracy)						**Belongings task (Event-Based Accuracy)**					
Model 1		**0.497**		**108.574**	**< .001**	Model 1		−0.010		0.065	0.800
	Age Group		**−0.708**		**< .001**		Age Group		0.026	0.254	0.800
Model 2		**0.564**		**3.383**	**0.003**	Model 2		−0.054		0.414	0.891
	Age Group		**−0.507**	**−5.745**	**< .001**		Age Group		0.097	0.675	0.501
	Working Memory		**0.288**	**3.349**	**0.001**		Working Memory		0.078	0.556	0.580
	Vigilance		−0.097	−1.424	0.158		Vigilance		0.115	1.037	0.302
	Open-mindedness		−0.030	−0.457	0.649		Open-mindedness		0.001	0.013	0.989
	Conscientiousness		0.097	1.307	0.194		Conscientiousness		−0.017	−0.137	0.891
	Extraversion		−0.031	−0.431	0.667		Extraversion		0.047	0.395	0.694
	Agreeableness		**−0.193**	**−2.529**	**0.013**		Agreeableness		−0.126	−1.012	0.314
	Neuroticism		−0.091	−1.223	0.224		Neuroticism		0.047	0.387	0.700
Breakfast task (Time-Based Accuracy)						**Call-Back task** (Time-Based Accuracy)					
Model 1		**0.449**		**89.721**	**< .001**	Model 1		**0.119**		**17.366**	**< .001**
	Age Group		**−0.674**	**−9.472**	**< .001**		Age Group		**−0.375**	**−4.167**	**< .001**
Model 2		**0.494**		**2.364**	**0.028**	Model 2		**0.329**		**4.179**	**< .001**
	Age Group		**−0.692**	**−7.280**	**< .001**		Age Group		**−0.579**	**−5.236**	**< .001**
	Working Memory		−0.044	−0.476	0.635		Working Memory		**−0.258**	**−2.399**	**0.018**
	Vigilance		0.025	0.339	0.735		Vigilance		**0.335**	**3.941**	**< .001**
	Open-mindedness		−0.134	−1.899	0.060		Open-mindedness		0.127	1.550	0.124
	Conscientiousness		0.111	1.382	0.170		Conscientiousness		−0.030	−0.323	0.747
	Extraversion		0.039	0.495	0.622		Extraversion		−0.015	−0.161	0.873
	Agreeableness		**−0.278**	**−3.381**	**0.001**		Agreeableness		−0.037	−0.391	0.696
	Neuroticism		−0.058	−0.716	0.476		Neuroticism		−0.182	−1.947	0.054

Table 5. Results from Regression Models predicting JS-CB and BF-CB PM paradox indices from age and the cognitive and personality predictors. Bolded values indicate $p < .05$.

Model	Variable		β	*F* or *t*	*p*	Model	Variable		β	*F* or *t*	*p*
JS-CB PM Index						**BF-CB PM Index**					
Model 1		**0.498**		**107.162**	**< .001**	Model 1		**0.458**		**91.565**	**< .001**
	Age Group		**−0.710**	**−10.352**	**< .001**		Age Group		**−0.680**	**−9.569**	**< .001**
Model 2		**0.534**		**14.633**	**< .001**	Model 2		**0.548**		**15.441**	**< .001**
	Age Group		−0.760	−1.560	0.122		Age Group		−0.570	−1.198	0.234
	Working Memory		−0.080	−0.825	0.411		Working Memory		**−0.210**	**−2.215**	**0.029**
	Vigilance		**0.149**	**2.097**	**0.039**		Vigilance		**0.205**	**2.924**	**0.004**
	Open-mindedness		0.005	0.073	0.942		Open-mindedness		−0.010	−0.162	0.871
	Conscientiousness		0.003	0.044	0.965		Conscientiousness		0.051	0.658	0.512
	Extraversion		0.008	0.104	0.917		Extraversion		0.021	0.282	0.779
	Agreeableness		−0.150	−1.926	0.057		Agreeableness		**−0.200**	**−2.601**	**0.011**
	Neuroticism		**−0.160**	**−2.056**	**0.042**		Neuroticism		**−0.160**	**−2.096**	**0.039**

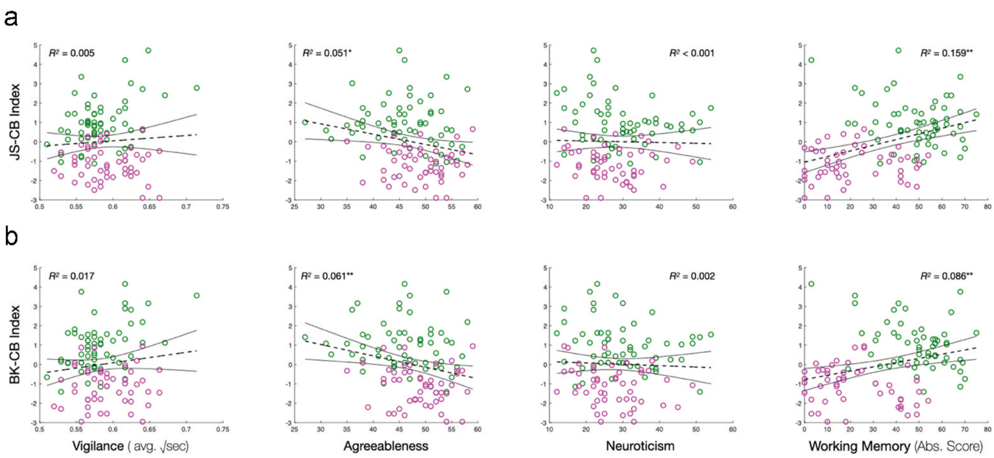

Figure 5. Predictors of each PM-paradox index. The black dashed and solid lines indicate the best-fitting regression models and 95% confidence intervals, respectively. R^2 values indicate the proportion of variance explained across individuals in both age groups. For associations within each group, see Table 3. *p < 0.05; **p < 0.01.

of both intact and deficient performance in global and local task demands, respectively, is both partially consistent with and contrary to findings in Craik and Bialystok (2006). In their computerized version of the Breakfast Task, there were significant age-related deficits in both global and local task performance. It may be the case that in our version of the Breakfast Task that required using realistic items and appliances and physically switching between cooking and table settings allowed the participants to better engage the action systems, which may have helped translate their real-world experience to the controlled research settings. Indeed, older adults performed comparably to young adults on global task goals of preparing all food items so that they were all ready at the same time, which is a common goal of cooking typical meals. In contrast, older adults showed deficits on local task goals that contained relatively strict, somewhat arbitrary time deadlines, such as flipping the eggs at exactly 3 minutes. It may be the case that the local task time constraints and serving arrangement rules contained a greater sense of arbitrariness in their requirements, while the tasks showing comparable performance across age groups were more universal and allowed older adults to better apply their real-world cooking experience (Altgassen et al., 2010; Hering et al., 2014; Kliegel et al., 2008). This is consistent with findings of age differences in PM being larger for experimenter vs. self-generated tasks and Schnitzspahn et al. (2020) proposal that differences in the use of such tasks between lab and real-world studies likely contributes to the age-PM paradox.

Experimenter-assigned tasks like those of the computerized version of the Breakfast Task (as well as the lack of immersion and engagement of action systems) may require greater executive control and demands on WM, even though they also require the same simulated cooking and table setting tasks. Indeed, factor analysis of the performance measures on the computerized Breakfast Task has shown that the task largely taps individual differences in executive functions (assessed with measures of working memory, processing speed, inhibition, reasoning) and, to a lesser degree, PM (Rose et al., 2010). Moreover, the computerized Breakfast Task does not include a planning phase or similar measures of global task performance. Future research should elaborate on these findings by investigating why some real-world experience (e.g., finishing cooking foods together, turning appliances on and off) might be easier to apply and translate into successful performance in lab-based simulations of everyday behavior relative to (non-immersive) computer game simulations.

Predicting PM performance across task contexts

In addition to establishing convergent findings between *Job Simulator* and the Breakfast task, we also explored the role of WM, attentional vigilance, and personality traits in predicting performance on PM tasks across varying contexts. In order to accomplish this, we created a PM index of difference scores between time-based PM performance on *Job Simulator* and the Call-back task, and between the Breakfast task and the Call-back task. Regression analyses and within-group associations indicated that the paradoxical age differences in performance between the lab and real-world measures of PM were predicted by WM, vigilance, agreeableness and neuroticism.

Given that the immersive VR gameplay in *Job Simulator* required switching between the ongoing game narrative and the experimenter-instructed PM tasks, participants likely had substantial difficulty maintaining all goal-relevant representations in focal attention throughout the game. Thus, age-related deficits and associations with WM capacity were likely due to the demands placed by switching between actively maintaining one's current task goal in focal attention and retrieving subsequent task goals from memory. This account is consistent with the dual-component model of WM (Unsworth & Engle, 2007) and findings showing that a substantial proportion of variance in age-differences in WM across the adult lifespan are attributable to the component associated with retrieval from memory (Hale et al., 2011). By controlling for age group, we were able to verify that variable in PM performance across the contexts were predicted by individual differences in WM. Importantly, these findings were observed for PM task performance despite the fact that all participants could accurately recall these tasks on the subsequent retrospective memory test. Thus, the contribution of switching between active maintenance and retrieval from memory cannot be attributed to young or older adults' failures to encode or retrieve the tasks in retrospective memory.

In contrast to the age differences in *Job Simulator* and the Breakfast task, findings in the Belongings task and Call-back task replicated patterns consistent with the age-PM paradox (Rendell & Thomson, 1999), with older adults performing comparably and even outperforming young adults, respectively. As stated earlier, individual differences in WM, vigilance and neuroticism within the young adult group were shown to drive age-group differences in real-world PM performance. Unlike *Job Simulator* and the Breakfast task, on the Call-back task, participants may have been able to arrange their daily life schedule during the test period to enable them to maintain PM tasks in focal attention while monitoring for relevant time cues. This suggests that the "ongoing" situational factors of everyday life did not tax controlled attention and WM retrieval processes to the extent required for switching between the ongoing and PM tasks. In particular, although the nature of the time-based PM tasks across the *Job Simulator*, Breakfast Task, and Call-back task shared many similarities, especially with regards to their durations and cues (which contrasts with many previous comparisons of time-based PM performance across lab and real-world contexts, Haines et al., 2020), the situations did differ in terms of their load/cognitive demands. It may be that older adults possess lifestyles that permit individuals with lower WM and vigilance abilities greater freedom to keep intentions in focal awareness compared to young adults, particularly those scoring higher on the neuroticism scale.

However, tasks that can be performed with processes associated with vigilance alone are not typically deemed to be true measures of PM. For example, Brandimonte et al. (2001) asserted that tasks modulated exclusively by attentional vigilance do not meet the demand criteria of a PM measure, which suggests that, at least in the way that older adults performed the Call-back task, the task measured vigilance more than PM processes. Alternatively, there is some evidence using experience-sampling methods to probe the frequency of thoughts (monitoring) about PM tasks in young and older adults' daily lives which suggests that older adults report thinking about PM tasks they have to do about twice as much as young adults (Gardner & Ascoli, 2015). This may help explain older adults' superior PM in real world settings. Because we could not directly assess participants' lifestyle demands, the frequency of monitoring, or the specific cues that participants used to complete the task during the at-home testing period, our ability to interpret their role in these findings is limited. Future research is required to observe potential differences in the ways in which young and older adults structure their environments, monitor, and cue themselves to perform PM tasks during at-home testing periods during situations that match the load/demands of lab-based comparison tasks (Bailey et al., 2010).

With regards to the Belongings task, young adults did not outperform older adults, yet both groups were better at remembering to retrieve their cell phone than remembering to return the tracker to the researcher. This suggests that both young and older adults were more motivated or incentivized to remember the personally-relevant item than the item relevant to the researcher (Walter & Meier, 2014). This is especially notable given that young adults presumably hold a greater attachment and habit of always having their cell phone with them than most older adults (Forgays et al., 2014). Aberle et al. (2010) have shown that using incentives to increase motivation in naturalistic PM tasks eliminates the paradox and young adults perform comparably to older adults. Although it was hypothesized that age- and individual differences in personality traits may have affected participants' ability to retrieve and return the items on this Belongings task after the experiment, we found no evidence for such associations. Our recent study showed that performance on this Belongings task was unaffected whether or not participants were explicitly instructed to remember the tasks, which suggests that there was sufficient intrinsic motivation to remember the tasks (N. Rose et al., 2023).

Associations between personality traits and our PM paradox indices revealed interesting patterns, yet questions about the roles of personality traits in the age-PM-paradox remain. Previous research suggested that developmental increases in conscientiousness are associated with improved performance on lab-based (Cuttler & Graf, 2007; Einstein & McDaniel, 1996) and self-reported real-world measures of PM (Rummel et al., 2022). For example, Rummel et al. (2022) recently reported a positive association between young adults' level of conscientiousness and self-reported PM performance in the real-world using experience sampling methods, as well as a negative association with openness. We found associations between agreeableness and time-based performance in *Job Simulator* and Breakfast task performance, especially in older adults. While a small, but statistically significant relationship between lab-based measures of PM and agreeableness has been previously reported (Salthouse et al., 2004), that study did not report increases in agreeableness as a function of age group in their sample, which contrasts previous work that has shown such a developmental pattern (Allemand et al., 2008, Lodi-Smith et al., 2009; McCrae et al., 1999). In contrast, the association between neuroticism and real-world PM performance seems to be driven by a developmental pattern, with young adults showing higher neuroticism than older adults on average.

To summarize the associations with personality, while it seemed plausible for developmental increases in conscientiousness and agreeableness to be associated with "paradoxical" age-benefits in real-world PM, the pattern of results here, like the broader literature on associations between personality traits and PM, is mixed. Personality traits seem like they should relate to individuals' habits and preferences to either structure their life according to planned activities or postpone planned activities in favor of more pressing, immediately-gratifying activities (procrastination), both of which impact PM performance (Zuber & Kliegel, 2020). Additionally, age differences in beliefs about the social importance of PM tasks impact age differences in PM (Altgassen et al., 2010). One issue is that, although there may be developmental differences in personality traits between young and older adults on average, it is important to acknowledge that the rank-order of individual personality traits may be stable over adulthood. Another limitation is that we only considered the "big five" dimensions of personality. Future research should attempt to differentiate the roles of age-group/cohort differences and individual differences in multiple dimensions of personality and their association with PM and cognitive aging. For example, according to Kuhl and colleagues' Personality Systems Interactions theory, the critical dimension of personality that impacts PM (and age differences in PM) is the extent to which an individual has a volitional disposition of being action-oriented versus "state-oriented" (Kaschel et al., 2017). Specifically, state-oriented individuals have reduced WM and try to maintain PM intentions via WM-demanding monitoring processes, which results in age deficits in high WM-load conditions. Future research with sufficient methodological rigor will be needed to capture the age- and individual-differences in the personality and cognitive predictors that likely underlie the age-PM-paradox and elucidate their sources.

Limitations and future directions

The naturalistic PM tasks that were assessed were mostly instructed by the experimenters – they were not self-generated by participants. The only exception was the real-world task where participants were to retrieve their own phone before leaving the lab (N. Rose et al., 2023). Schnitzspahn et al. (2020) recently showed that allowing participants to self-generate their own PM tasks to complete over the assessment period in their daily lives resulted in no age differences. The self-referential processes associated with generating PM demands for oneself may partially explain the lack of age differences found here in our Belongings task. However, these processes are unlikely to have contributed to the age benefits that were observed on the Call-back task, where times were still instructed by the experimenter. It may be the case that the naturalistic nature of the Call-back task disproportionately increased the intrinsic motivation of older adults compared to young adults, despite containing demands that were generated by the researcher (Aberle et al., 2010; Peter & Kliegel, 2018). Future work should seek to disentangle the role of self-generated PM tasks from intrinsic motivation in studies of PM and aging, especially those that measure performance across multiple settings, where the effects of these variables may differ.

Although the sample sizes were sufficient for the assessment of age-differences in PM, which was justified by our power analysis, stronger and possibly more reliable conclusions could be drawn with larger numbers of participants contributing data for all measures. Having all participants perform all measures is most appropriate for addressing the age-PM paradox, and the within-subjects comparisons are necessary for the calculation of our novel index of the age-PM paradox; however, doing so is logistically challenging due to the time and costs that are required. Limitations of the current study include the cross-sectional age comparison with assessments of only single measures of a few cognitive and motivational factors associated with personality traits. Future studies should assess larger samples, with age continuously varying across a wide range, and collect more assessments of cognitive and motivational predictors of PM performance longitudinally over multiple timepoints to further assess age- and individual-differences in predictors that may underlie the age-PM paradox.

Conclusion

While the PM tasks presented here were designed to capture similar real-world PM processes, the correlational and regression analyses of age- and individual-differences in performance across contexts suggest that the real-world tasks may be measuring different processes than the naturalistic PM tasks performed under controlled laboratory conditions. This study demonstrates the value of incorporating multiple different assessments of PM and represents a fruitful step toward finally resolving the age-PM paradox. Lab-based PM tasks that are used to evaluate age-related differences should be developed and validated with real-world PM tasks to match the cognitive demands and personality traits that are associated with performance across task settings. Future work should systematically manipulate task variables and contexts to investigate age and individual differences in PM, as well as their predictors. Longitudinal assessments could track changes in cognitive abilities and personality traits over time and determine how PM performance changes due to development of the underlying processes.

In conclusion, this study offers one of the first objective validations of immersive VR gameplay for measuring PM functioning in both young and older adults. In doing so, we found that naturalistic measures of PM relate to age- and individual differences in cognitive abilities and personality traits differently across lab and real-world contexts, thus helping to resolve the age-PM paradox.

Notes

1. Some suggest that older adults care more about the social contract of doing the task for the experimenter than young adults due to generational or cohort differences. However, this has been observed for 4 decades, so a simple cohort-effect seems implausible.

2. Analyses of PM performance were repeated using an ANCOVA to control for differences in retrospective memory performance. This confirmed that none of the relationships between our variables of interest were significantly altered by differences in retrospective memory performance.
3. Mean square-root transformed deviation scores were used as the dependent measure of interest because it was assumed to demonstrate the best psychometric properties given that there were only 4 observations. Different scoring procedures were formally assessed to ensure that the conclusions did not depend on the scoring procedure that was selected, including a transformation of the continuous data to accuracy scores (0, .25, .5, .75, or 1.0). The age difference was similar for every version of scoring. For example, the mean accuracy for the young and older adults were 60% and 78%, $p < .001$.Note that the index that was calculated to assess the age-PM-paradox pattern for each participant that is described below used time deviation scores to facilitate comparison between the *Job Simulator*, Breakfast, and Call-back tasks.
4. Due to time constraints during the first session, 48 participants (40 old, 8 young) completed the convenience store clerk simulation at the start of the second session; 22 (16 old, 6 young) completed BFI-II and PVT at the start of the second session. We re-ran our subsequent PM analysis including session-of-completion as a factor and found no significant between-subject effects for session, $F(1,107) = 0.70$, $p = 0.41$ nor group-by-session interactions $F(1,107) = 1.17$, $p = 0.28$.
5. This range was appropriate because the clock that participants used to monitor time was displayed in hours and minutes, not seconds. Event-based PM accuracy thresholds were set to ten seconds surrounding the ideal time at which the event cue occurred. The time-accuracy threshold for one event-based task was set to within 60 seconds surrounding event-cue onset (i.e., turning the sink on and off every 3 orders) because the time between completing an order and beginning the next one was approximately 60 seconds, and accurate performance could reasonably fall outside of the ten second window established for the other event-based tasks and up to 60 seconds, at which point the next order occurred. Performance was also coded for different response windows ("a little late," "very late," or those that were completely forgotten, i.e., not performed at all). We re-ran our mixed-design ANOVA on the proportion of correct PM tasks to include responses that were binned as "a little late." Using this more lenient scoring criterion did not change any of the age group relationships. There was a main effect of regularity due to performance being better on irregular tasks than regular tasks. This was qualified by a cue type by regularity interaction due to this pattern being more pronounced in event-based measures than time-based measures. This suggests that the benefit of task regularity was not observed when a more lenient criterion was used.
6. Note that overall performance was higher for time-based PM than event-based PM, which is atypical in PM research, and was likely due to the stricter response window for event- (10 sec) than time-based (60 sec) tasks which we enforced due to hardware limitations that only allowed us to display time in minutes rather than seconds in *Job Simulator*. When we included responses that fell into the "little late" category, results regarding the pattern of age-differences in PM accuracy were preserved.
7. Note that a recent study showed that executive/PM performance on a VR task was not associated with gaming experience (Seesjärvi et al., 2022).

Acknowledgments

The authors thank Dr. Mike Villano, Co-Director of the ND-VR Lab, for his help with setting up and implementing the VR task protocol, and Delaney Schrenk, Anna Rohrer, Annie Birmingham, Mary Glass, and Lucie Moore for their help with data collection.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by funds from the University of Notre Dame's Digital Learning Initiative and by the William P. and Hazel B. Collegiate Chair endowment awarded to N.S.R. Preparation of this manuscript was supported by the National Science Foundation CAREER Grant [1848440] awarded to No. S.R.

ORCID

Nathan S. Rose http://orcid.org/0000-0002-7619-4441

References

Aberle, I., Rendell, P. G., Rose, N. S., McDaniel, M. A., & Kliegel, M. (2010). The age prospective memory paradox: Young adults may not give their best outside of the lab. *Developmental Psychology*, *46*(6), 1444. https://doi.org/10.1037/a0020718

Allemand, M., Zimprich, D., & Hendriks, A. A. (2008). Age differences in five personality domains across the life span. *Developmental Psychology*, *44*(3), 758. https://doi.org/10.1037/0012-1649.44.3.758

Altgassen, M., Kliegel, M., Brandimonte, M., & Filippello, P. (2010). Are older adults more social than younger adults? Social importance increases older adults' prospective memory performance. *Aging, Neuropsychology & Cognition*, *17*(3), 312–328. https://doi.org/10.1080/13825580903281308

Altgassen, M., Rendell, P. G., Bernhard, A., Henry, J. D., Bailey, P. E., Phillips, L. H., & Kliegel, M. (2015). Future thinking improves prospective memory performance and plan enactment in older adults. *Quarterly Journal of Experimental Psychology*, *68*(1), 192–204. https://doi.org/10.1080/17470218.2014.956127

Anderson, F. T., & McDaniel, M. A. (2019). Hey buddy, why don't we take it outside: An experience sampling study of prospective memory. *Memory & Cognition*, *47*(1), 47–62. https://doi.org/10.3758/s13421-018-0849-x

Bailey, P. E., Henry, J. D., Rendell, P. G., Phillips, L. H., & Kliegel, M. (2010). Dismantling the "age–prospective memory paradox": The classic laboratory paradigm simulated in a naturalistic setting. *Quarterly Journal of Experimental Psychology*, *63*(4), 646–652. https://doi.org/10.1080/17470210903521797

Banville, F., Nolin, P., Lalonde, S., Henry, M., Dery, M. P., & Villemure, R. (2010). Multitasking and prospective memory: Can virtual reality be useful for diagnosis? *Behavioural Neurology*, *23*(4), 209–211. https://doi.org/10.1155/2010/320707

Bopp, K. L., & Verhaeghen, P. (2005). Aging and verbal memory span: A meta-analysis. *Journal of Financial Stability*, *60*(5), 223–P233. https://doi.org/10.1093/geronb/60.5.P223

Brandimonte, M. A., Ferrante, D., Feresin, C., & Delbello, R. (2001). Dissociating prospective memory from vigilance processes. *Psicológica*, *22*(1), 97–113.

Brandt, J., Spencer, M., & Folstein, M. (1988). The telephone interview for cognitive status. *Neuropsychiatry, Neuropsychology, & Behavioral Neurology*, *1*(2), 111–117.

Chen, G., King, J. A., Burgess, N., & O'Keefe, J. (2013). How vision and movement combine in the hippocampal place code. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *110*(1), 378–383. https://doi.org/10.1073/pnas.1215834110

Cona, G., Bisiacchi, P. S., & Moscovitch, M. (2014). The effects of focal and nonfocal cues on the neural correlates of prospective memory: Insights from ERPs. *Cerebral Cortex*, *24*(10), 2630–2646. https://doi.org/10.1093/cercor/bht116

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W.(2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*, 769–786.

Craik, F. I. M., & Bialystok, E. (2006). Planning and task management in older adults: Cooking breakfast. *Memory & Cognition*, *34*(6), 1236–1249. https://doi.org/10.3758/BF03193268

Crovitz, H. F., & Daniel, W. F. (1984). Measurements of everyday memory: Toward the prevention of forgetting. *Bulletin of the Psychonomic Society*, *22*(5), 413–414. https://doi.org/10.3758/BF03333861

Cuttler, C., & Graf, P. (2007). Personality predicts prospective memory task performance: An adult lifespan study. *Scandinavian Journal of Psychology*, *48*(3), 215–231. https://doi.org/10.1111/j.1467-9450.2007.00570.x

Cuttler, C., & Graf, P. (2008). Sub-clinical checking compulsions are related to impaired prospective memory independently of depression, anxiety and distractibility. *Journal of Anxiety Disorders*, *22*(4), 642–654. https://doi.org/10.1016/j.janxdis.2007.06.003

Debarnot, U., Crépon, B., Orriols, E., Abram, M., Charron, S., Lion, S., Roca, P., Oppenheim, C., Gueguen, B., Ergis, A.-M., Baron, J.-C., & Piolino, P. (2015). Intermittent theta burst stimulation over left BA10 enhances virtual reality-based prospective memory in healthy aged subjects. *Neurobiology of Aging*, *36*(8), 2360–2369. https://doi.org/10.1016/j.neurobiolaging.2015.05.001

Dong, D., Wong, L. K. F., & Luo, Z. (2018). Assess BA10 activity in slide-based and immersive virtual reality prospective memory task using functional near-infrared spectroscopy (fNIRS). *Applied Neuropsychology: Adult*, *26*(5), 465–471. https://doi.org/10.1080/23279095.2018.1443104

d'Ydewalle, G., Bouckaert, D., & Brunfaut, E. (2001). Age-related differences and complexity of ongoing activities in time- and event-based prospective memory. *The American Journal of Psychology*, *114*(3), 411–423. https://doi.org/10.2307/1423688

Einstein, G. O., Holland, L. J., McDaniel, M. A., & Guynn, M. J. (1992). Age-related deficits in prospective memory: The influence of task complexity. *Psychology and Aging*, *7*(3), 471–478. https://doi.org/10.1037/0882-7974.7.3.471

Einstein, G. O., & McDaniel, M. A. (1990). Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 717–726. https://doi.org/10.1037//0278-7393.16.4.717

Einstein, G. O., & McDaniel, M. A. (1996). Retrieval processes in prospective memory: Theoretical approaches and some new empirical findings. In M. Brandimonte, G. O. Einstein, & M. A. McDaniel (Eds.), *Prospective memory: Theory and applications* (pp. 115–141). Lawrence Erlbaum Associates Publishers.

Einstein, G. O., & McDaniel, M. A. (2008). Prospective memory and metamemory: The skilled use of basic attentional and memory processes. In A. S. Benjamin & B. H. Ross (Eds.), *Skill and strategy in memory use* (pp. 145–173). Elsevier Academic Press.

Einstein, G. O., McDaniel, M. A., Richardson, S. L., Guynn, M. J., & Cunfer, A. R. (1995). Aging and prospective memory: Examining the influences of self-initiated retrieval processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 996–1007. https://doi.org/10.1037/0278-7393.21.4.996

Ellis, J., & Kvavilashvili, L. (2000). Prospective memory in 2000: Past, present, and future directions. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory & Cognition*, *14*(7), S1–S9. https://doi.org/10.1002/acp.767

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309–331. https://doi.org/10.1037/0096-3445.128.3.309

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Feinkohl, I., Cress, U., & Kimmerle, J. (2016). Reheating breakfast: Age and multitasking on a computer-based and a non-computer-based task. *Computers in Human Behavior*, *55*, 432–438. https://doi.org/10.1016/j.chb.2015.09.040

Festini, S. B., McDonough, I. M., & Park, D. C.(2016). The busier the better: Greater busyness is associated with better cognition. *Frontiers in aging neuroscience*, *8*, Article 98.

Forgays, D. K., Hyman, I., & Schreiber, J. (2014). Texting everywhere for everything: Gender and age differences in cell phone etiquette and use. *Computers in Human Behavior*, *31*, 314–321. https://doi.org/10.1016/j.chb.2013.10.053

Gardner, R. S., & Ascoli, G. A. (2015). The natural frequency of human prospective memory increases with age. *Psychology and Aging*, *30*(2), 209. https://doi.org/10.1037/a0038876

Gondo, Y., Renge, N., Ishioka, Y., Kurokawa, I., Ueno, D., & Rendell, P. (2010). Reliability and validity of the prospective and retrospective memory questionnaire (PRMQ) in young and old people: A Japanese study. *Japanese Psychological Research*, *52*(3), 175–185. https://doi.org/10.1111/j.1468-5884.2010.00433.x

Gonneaud, J., Piolino, P., Lecouvey, G., Madeleine, S., Orriols, E., Fleury, P., Eustache, F., & Desgranges, B. (2014). Assessing prospective memory in young healthy adults using virtual reality. In P. M. Sharkey & J. Merrick (Eds.), *Virtual reality: People with special needs* (pp. 107–120). Nova Science Publishers.

Graf, P., & Uttl, B. (2001). Prospective memory: A new focus for research. *Consciousness and Cognition: An International Journal*, *10*(4), 437–450. https://doi.org/10.1006/ccog.2001.0504

Haines, S. J., Randall, S. E., Terrett, G., Busija, L., Tatangelo, G., McLennan, S. N., Rose, N. S., Kliegel, M., Henry, J. D., & Rendell, P. G. (2020). Differences in time-based task characteristics help to explain the age-prospective memory paradox. *Cognition*, *202*, Article 104305. https://doi.org/10.1016/j.cognition.2020.104305

Hale, S., Rose, N. S., Myerson, J., Strube, M. J., Sommers, M., Tye-Murray, N., & Spehar, B. (2011). The structure of working memory abilities across the adult life span. *Psychology and Aging*, *26*(1), 92–110. https://doi.org/10.1037/a0021483

Helson, R., Kwan, V. S., John, O. P., & Jones, C. (2002). The growing evidence for personality change in adulthood: Findings from research with personality inventories. *Journal of Research in Personality*, *36*(4), 287–306. https://doi.org/10.1016/S0092-6566(02)00010-7

Henry, J. D., MacLeod, M. S., Phillips, L. H., & Crawford, J. R. (2004). A meta-analytic review of prospective memory and aging. *Psychology and Aging*, *19*(1), 27–39. https://doi.org/10.1037/0882-7974.19.1.27

Hering, A., Cortez, S. A., Kliegel, M., & Altgassen, M. (2014). Revisiting the age-prospective memory-paradox: The role of planning and task experience. *European Journal of Ageing*, *11*(1), 99–106. https://doi.org/10.1007/s10433-013-0284-6

Hering, A., Kliegel, M., Rendell, P. G., Craik, F. I. M., & Rose, N. S. (2018). Prospective memory is a key predictor of functional independence in older adults. *Journal of the International Neuropsychological Society*, *24*(6), 640–645. https://doi.org/10.1017/S1355617718000152

Ihle, A., Schnitzspahn, K., Rendell, P. G., Luong, C., & Kliegel, M. (2012). Age benefits in everyday prospective memory: The influence of personal task importance, use of reminders and everyday stress. *Aging, Neuropsychology & Cognition*, *19*(1–2), 84–101. https://doi.org/10.1080/13825585.2011.629288

Jylkkä, J., Ritakallio, L., Merzon, L., Kangas, S., Kliegel, M., Zuber, S., Hering, A., Salmi, J., & Laine, M. (2023). Assessment of goal-directed behavior with the 3D videogame EPELI: Psychometric features in a web-based adult sample. *Public Library of Science One*, *18*(3), e0280717. https://doi.org/10.1371/journal.pone.0280717

Jylkkä, J., Ritakallio, L., Merzon, L., Kangas, S., Kliegel, M., Zuber, S., & Laine, M.(2023). Assessment of goal-directed behavior with the 3D videogame EPELI: Psychometric features in a web-based adult sample. *Public Library of Science ONE*, *18*(3), e0280717.

Kalpouzos, G., Eriksson, J., Sjölie, D., Molin, J., Nyberg, L., & Sporns, O. (2010). Neurocognitive systems related to real-world prospective memory. *Public Library of Science One*, *5*(10), e13304. https://doi.org/10.1371/journal.pone.0013304

Kane, M. J., Conway, A. R. A., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 21–46). Oxford University Press.

Kaschel, R., Kazén, M., & Kuhl, J.(2017). State orientation and memory load impair prospective memory performance in older compared to younger persons. *Aging, Neuropsychology & Cognition*, *24*(4), 453–469. https://doi.org/10.1080/13825585.2016.1225664

Kliegel, M., Ballhausen, N., Hering, A., Ihle, A., Schnitzspahn, K. M., & Zuber, S. (2016). Prospective memory in older adults: Where we are now and what is next. *Gerontology*, *62*(4), 459–466. https://doi.org/10.1159/000443698

Kliegel, M., & Martin, M. (2003). Prospective memory research: Why is it relevant?[Editorial]. *International Journal of Psychology*, *38*(4), 193–194. https://doi.org/10.1080/00207590344000114

Kliegel, M., McDaniel, M. A., & Einstein, G. O. (Eds.). (2008). *Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives*. Taylor & Francis Group/Lawrence Erlbaum Associates.

Lamichhane, B., McDaniel, M. A., Waldum, E. R., & Braver, T. S. (2018). Age-related changes in neural mechanisms of prospective memory. *Cognitive, Affective, & Behavioral Neuroscience*, *18*(5), 982–999. https://doi.org/10.3758/s13415-018-0617-1

Lodi-Smith, J., Geise, A. C., Roberts, B. W., & Robins, R. W.(2009). Narrating personality change. *Journal of Personality & Social Psychology*, *96*(3), 679–689. https://doi.org/10.1037/a0014611

Martin, M., & Park, D. C. (2003). The Martin and park environmental demands (MPED) questionnaire: Psychometric properties of a brief instrument to measure self-reported environmental demands. *Aging Clinical and Experimental Research*, *15*(1), 77–82. https://doi.org/10.1007/BF03324483

Maylor, E. A. (1990). Age and prospective memory. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *42*(3–A), 471–493. https://doi.org/10.1080/14640749008401233

McCrae, R. R., Costa, P. T., de Lima, M. P., Simões, A., Ostendorf, F., Angleitner, A., Marušić, I., Bratko, D., Caprara, G. V., Barbaranelli, C., Chae, J.-H., & Piedmont, R. L. (1999). Age differences in personality across the adult life span: Parallels in five cultures. *Developmental Psychology*, *35*(2), 466–477. https://doi.org/10.1037/0012-1649.35.2.466

McDaniel, M. A., & Einstein, G. O. (2000). Strategic and automatic processes in prospective memory retrieval: A multiprocess framework. *Applied Cognitive Psychology*, *14*(SpecIssue), S127–S144. https://doi.org/10.1002/acp.775

McDaniel, M. A., LaMontagne, P., Beck, S. M., Scullin, M. K., & Braver, T. S. (2013). Dissociable neural routes to successful prospective memory. *Psychological Science*, *24*(9), 1791–1800. https://doi.org/10.1177/0956797613481233

Moscovitch, M. (1982). A neuropsychological approach to memory and perception in normal and pathological aging. In F. I. M. Craik & S. Trehub (Eds.), *Aging and cognitive processes* (pp. 55–78). Plenum Press.

Niedźwieńska, A., & Barzykowski, K. (2012). The age prospective memory paradox within the same sample in time-based and event-based tasks. *Aging, Neuropsychology & Cognition*, *19*(1–2), 58–83. https://doi.org/10.1080/13825585.2011.628374

Okahashi, S., Seki, K., Nagano, A., Luo, Z., Kojima, M., & Futaki, T. (2013). A virtual shopping test for realistic assessment of cognitive function. *Journal of NeuroEngineering and Rehabilitation*, *10*(1), 1–13. https://doi.org/10.1186/1743-0003-10-59

O'Rear, A. E., & Radvansky, G. A. (2019). Location-based prospective memory. *Quarterly Journal of Experimental Psychology*, *72*(3), 491–507. https://doi.org/10.1177/1747021818758608

Ouellet, E., Boller, B., Corriveau-Lecavalier, N., Cloutier, S., & Belleville, S. (2018). The virtual shop: A new immersive virtual reality environment and scenario for the assessment of everyday memory. *Journal of Neuroscience Methods*, *303*, 126–135. https://doi.org/10.1016/j.jneumeth.2018.03.010

Park, D. C. (2000). The basic mechanisms accounting for age-related decline in cognitive function. In D. C. Park & N. Schwarz (Eds.), *Cognitive aging: A primer* (pp. 3–21). Psychology Press.

Park, D. C., Hertzog, C., Kidder, D. P., Morrell, R. W., & Mayhorn, C. B. (1997). Effect of age on event-based and time-based prospective memory. *Psychology and Aging*, *12*(2), 314–327. https://doi.org/10.1037/0882-7974.12.2.314

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging*, *17*(2), 299–320. https://doi.org/10.1037/0882-7974.17.2.299

Parsons, T. D., Barnett, M., & Robillard, J. (2017). Validity of a newly developed measure of memory: Feasibility study of the virtual environment grocery store. *Journal of Alzheimer's Disease*, *59*(4), 1227–1235. https://doi.org/10.3233/JAD-170295

Patton, G. W., & Meit, M. (1993). Effect of aging on prospective and incidental memory. *Experimental Aging Research*, *19*(2), 165–176. https://doi.org/10.1080/03610739308253929

Peter, J., & Kliegel, M. (2018). The age-prospective memory paradox: Is it about motivation? *Clinical and Translational Neuroscience*, *2*(2), 35. https://doi.org/10.1177/2514183x18807103

Phillips, L. H., Henry, J. D., & Martin, M.(2008). Adult aging and prospective memory: The importance of ecological validity. In M. Kliegel, M. A. McDaniel, & G. O. Einstein (Eds.), *Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives* (pp. 161–185). Taylor & Francis Group/Lawrence Erlbaum Associates.

Rendell, P. G., & Craik, F. I. M. (2000). Virtual week and actual week: Age-related differences in prospective memory. *Applied Cognitive Psychology*, *14*(7), S43–S62. https://doi.org/10.1002/acp.770

Rendell, P. G., & Thomson, D. M. (1999). Aging and prospective memory: Differences between naturalistic and laboratory tasks. *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences*, *54*(4), 256–P269. https://doi.org/10.1093/geronb/54B.4.P256

Rose, N., Doolen, A. C., & O'Rear, A. E. (2023). They forgot their "baby"?!: Factors that lead students to forget their cell phone. *Journal of Applied Research on Memory and Cognition*. https://doi.org/10.1037/mac0000110

Rose, N. S., Myerson, J., Sommers, M. S., & Hale, S. (2009). Are there age differences in the executive component of working memory? Evidence from domain-general interference effects. *Aging, Neuropsychology & Cognition*, *16*(6), 633–653. https://doi.org/10.1080/13825580902825238

Rose, N. S., Rendell, P. G., Hering, A., Kliegel, M., Bidelman, G. M., & Craik, F. I. M. (2015). Cognitive and neural plasticity in older adults' prospective memory following training with the virtual week computer game. *Frontiers in Human Neuroscience*, *9*, Article 592. https://doi.org/10.3389/fnhum.2015.00592

Rose, N. S., Rendell, P. G., McDaniel, M. A., Aberle, I., & Kliegel, M. (2010). Age and individual differences in prospective memory during a "virtual week": The roles of working memory, vigilance, task regularity, and cue focality. *Psychology and Aging*, *25*(3), 595–605. https://doi.org/10.1037/a0019771

Rummel, J., & Kvavilashvili, L. (2023). Current theories of prospective memory and new directions for theory development. *Nature Reviews Psychology*, *2*(1), 40–54. https://doi.org/10.1038/s44159-022-00121-4

Rummel, J., Snijder, J. P., & Kvavilashvili, L. (2022). Prospective memories in the wild: Predicting memory for intentions in natural environments. *Memory & Cognition*, *51*(5), 1–15. https://doi.org/10.3758/s13421-022-01379-y

Rypma, B., & D'Esposito, M. (2000). Isolating the neural mechanisms of age-related changes in human working memory. *Nature Neuroscience*, *3*(5), 509–515. https://doi.org/10.1038/74889

Sakai, H., Nagano, A., Seki, K., Okahashi, S., Kojima, M., & Luo, Z. (2018). Development of a cognitive function test using virtual reality technology: Examination in healthy participants. *Aging, Neuropsychology & Cognition*, *25*(4), 561–575. https://doi.org/10.1080/13825585.2017.1351916

Salthouse, T. A. (1994). The aging of working memory. *Neuropsychology*, *8*(4), 535–543. https://doi.org/10.1037/0894-4105.8.4.535

Salthouse, T. A., Berish, D. E., & Siedlech, K. L. (2004). Construct validity and age sensitivity of prospective memory. *Memory & Cognition*, *32*(7), 1133–1148. https://doi.org/10.3758/BF03196887

Sarter, M., & Bruno, J. P. (2004). Developmental origins of the age-related decline in cortical cholinergic function and associated cognitive abilities. *Neurobiology of Aging*, *25*(9), 1127–1139. https://doi.org/10.1016/j.neurobiolaging.2003.11.011

Schaffer, G., & Poon, L. W. (1982). Individual variability in memory training with the elderly. *Educational Gerontology*, *8*(3), 217–229. https://doi.org/10.1080/0360127820080301

Schnitzspahn, K. M., Ihle, A., Henry, J. D., Rendell, P. G., & Kliegel, M. (2011). The age-prospective memory-paradox: An exploration of possible mechanisms. *International Psychogeriatrics*, *23*(4), 583–592. https://doi.org/10.1017/S1041610210001651

Schnitzspahn, K. M., Kvavilashvili, L., & Altgassen, M. (2020). Redefining the pattern of age-prospective memory-paradox: New insights on age effects in lab-based, naturalistic, and self-assigned tasks. *Psychological Research*, *84*(5), 1370–1386. https://doi.org/10.1007/s00426-018-1140-2

Seesjärvi, E., Puhakka, J., Aronen, E. T., Hering, A., Zuber, S., Merzon, L., Kliegel, M., Laine, M., & Salmi, J. (2022). EPELI: A novel virtual reality task for the assessment of goal-directed behavior in real-life contexts. *Psychological Research*, *87*(6), 1899–1916. https://doi.org/10.1007/s00426-022-01770-z

Smith, R. E., Persyn, D., & Butler, P. (2011). Prospective memory, personality, and working memory: A formal modeling approach*. *Zeitschrift für Psychologie/Journal of Psychology*, *219*(2), 108–116. https://doi.org/10.1027/2151-2604/a000055

Soto, C. J., & John, O. P. (2017). The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Taube, J. S., Valerio, S., & Yoder, R. M. (2013). Is navigation in virtual reality with FMRI really navigation? *Journal of Cognitive Neuroscience*, *25*(7), 1008–1019. https://doi.org/10.1162/jocn_a_00386

Terry, W. S. (1988). Everyday forgetting: Data from a diary study. *Psychological Reports*, *62*(1), 299–303. https://doi.org/10.2466/pr0.1988.62.1.299

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

Unsworth, N., & Engle, R. W.(2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132. https://doi.org/10.1037/0033-295X.114.1.104

Unsworth, N., McMillan, B. D., Brewer, G. A., & Spillers, G. J. (2012). Everyday attention failures: An individual differences investigation. *Journal of Experimental Psychology*, *38*(6), 1765–1772. https://doi.org/10.1037/a0028075

Uttl, B., & Greene, E. (2008). Transparent meta-analysis of prospective memory and aging. *Public Library of Science ONE*, *3*(2), e1568. https://doi.org/10.1371/journal.pone.0001568

Uttl, B., & Kibreab, M. (2011). Self-report measures of prospective memory are reliable but not valid. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *65*(1), 57. https://doi.org/10.1037/a0022843

Uttl, B., White, C., Wong Gonzalez, D., McDouall, J., & Leonard, C. (2013). Prospective memory, personality, and individual differences. *Frontiers in Psychology*, *4*, 130. https://doi.org/10.3389/fpsyg.2013.00130

Verhaeghen, P., Martin, M., & Sędek, G. (2012). Reconnecting cognition in the lab and cognition in real life: The role of compensatory social and motivational factors in explaining how cognition ages in the wild. *Aging, Neuropsychology & Cognition*, *19*(1–2), 1–12. https://doi.org/10.1080/13825585.2011.645009

Walter, S., & Meier, B. (2014). How important is importance for prospective memory? A review. *Frontiers in Psychology*, *5*, Article 657. https://doi.org/10.3389/fpsyg.2014.00657

Watson, D. C. (2001). Procrastination and the five-factor model: A facet level analysis. *Personality and Individual Differences*, *30*(1), 149–158. https://doi.org/10.1016/S0191-8869(00)00019-2

Watson, D., & Clark, L. A. (1992). On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model. *Journal of Personality*, *60*(2), 441–476. https://doi.org/10.1111/j.1467-6494.1992.tb00980.x

West, R. L. (1988). Prospective memory and aging. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (Vol. 2, pp. 119–125). Wiley.

Wilson, B., Cockburn, J., & Baddeley, A. D. (1985). *The rivermead behavioural memory test*. Bury St Edmunds.

Woods, S. P., Weinborn, M., Li, Y. R., Hodgson, E., Ng, A. R. J., & Bucks, R. S. (2015). Does prospective memory influence quality of life in community-dwelling older adults? *Aging, Neuropsychology & Cognition*, *22*(6), 679–692. https://doi.org/10.1080/13825585.2015.1027651

Zuber, S., & Kliegel, M. (2020). Prospective memory development across the lifespan. *European Psychologist*, *25*(3), 162–173. https://doi.org/10.1027/1016-9040/a000380