# The Geneva Space Cruiser: A Fully Self-Administered Online Tool to Assess Prospective Memory across the Adult Lifespan

**Article** *in* Memory · October 2021

DOI: 10.1080/09658211.2021.1995435



**7 authors**, including:

Sascha Zuber
University of Victoria


Maximilian Haas
University of Geneva


David Framorando
University of Geneva


Nicola Ballhausen
Tilburg University

# The Geneva Space Cruiser: A Fully Self-Administered Online Tool to Assess Prospective Memory across the Adult Lifespan

Zuber, S.[1,2,3*], Haas, M.[1,3], Framorando, D.[4], Ballhausen, N.[1,5], Gillioz, E.[3], Künzi, M.[1,2,3], & Kliegel, M.[1,2,3]

[1] Centre for the Interdisciplinary Study of Gerontology and Vulnerability, University of Geneva, Geneva, Switzerland
[2] Swiss National Centre of Competences in Research LIVES–Overcoming vulnerability: life course perspectives, Lausanne and Geneva, Switzerland
[3] Department of Psychology, University of Geneva, Geneva, Switzerland
[4] School of Psychology, The University of Queensland, Brisbane, Australia
[5] Department of Developmental Psychology, Tilburg University, Tilburg, The Netherlands

**Correspondence:*
Sascha Zuber
University of Geneva
Center for the Interdisciplinary Study of Gerontology and Vulnerabilities (CIGEV)
Boulevard du Pont d'Arve 28
CH- 1205 Genève
Office 103

**Contact:**
E-Mail: Sascha.Zuber@unige.ch
Phone: +41 22 379 83 81

This is an Accepted Manuscript (AM) of an article accepted for publication by Taylor & Francis in *Memory* on October 14, 2021. Available online soon.

Reference:

Zuber, S., Haas, M., Framorando, D., Ballhausen, N., Gillioz, E., Künzi, M., & Kliegel, M. (in press).  The Geneva Space Cruiser: A Fully Self-Administered Online Tool to Assess Prospective Memory across the Adult Lifespan. *Memory*.

# The Geneva Space Cruiser: A Fully Self-Administered Online Tool to Assess Prospective Memory across the Adult Lifespan

The current study aimed to examine whether the Geneva Space Cruiser – a new online adaptation of the Cruiser – represents a valid, reliable and useful tool to assess prospective memory (PM) across the adult lifespan via fully self-administered online testing. Therefore, an adult lifespan sample of 252 adults (19-86 years old) performed the Geneva Space Cruiser in the laboratory and online, at home, and also performed a more traditional laboratory PM task. A second sample of 224 young adults (19-35 years old) participated in a test-retest online assessment of the Geneva Space Cruiser. Bayesian analyses showed that the Geneva Space Cruiser yielded similar results when administered in the laboratory versus online, both in terms of data distribution as well as of key outcome measures (i.e., PM performance and monitoring). Results further showed very good test-retest reliability and acceptable construct validity. Finally, the online tool was sensitive for detecting age-differences similar to those typically observed in laboratory studies. Together, our findings suggest that the Geneva Space Cruiser represents a rather valid, moderately to highly reliable, and generally useful tool to assess PM in online testing across wide ranges of the adult lifespan, with certain limitations for the oldest participants and for women.

**Key words:** prospective memory; online assessment; serious game; validation; lifespan; monitoring; time-based; older adults



## Introduction

Prospective Memory (PM) encompasses our ability to remember to perform planned intentions at the appropriate moment in the future (Einstein & McDaniel, 1990). Everyday examples of PM tasks are remembering to attach a file before sending off an email, remembering to text someone on their birthday, or remembering to take one’s medication on time. PM contributes to various aspects of our everyday lives, such as our professional achievement, our social relations, and our health and well-being (e.g., Hering et al., 2018;

Loft et al., 2019; Woods et al., 2015). To assess PM, researchers have so far mostly relied on face-to-face assessment conducted in laboratory settings (Rummel & Kvavilashvili, 2019). Although other assessment approaches exist (such as using memory diaries or giving participants more naturalistic tasks that they have to perform outside the laboratory, see e.g., Haas et al., 2020; Kvavilashvili & Rummel, 2020; Schnitzspahn et al., 2020), they typically also include some part of laboratory assessment (e.g., participants first coming to the laboratory to receive instructions, to fill out questionnaires, and to complete other cognitive tasks).

Face-to-face laboratory sessions have multiple advantages, such as allowing direct contact with the participant or providing a more controlled assessment environment, and are therefore the gold standard of cognitive assessment. However, they also are costly in terms of resources (e.g., office space, technical material), personnel (e.g., paying experimenters and other staff), planning (e.g., finding participants that are willing to meet at the laboratory; scheduling meetings that are convenient for participant and experimenter), and additional time (e.g., commuting to laboratory; arriving in advance). Further, as recent restrictions related to the COVID-19 pandemic have illustrated, data collection in the laboratory can become complicated or even be made impossible due to factors that the researchers cannot control. Hence, more and more research areas have started to explore online assessment as potential alternative (or helpful addition) to laboratory assessment.

Self-administered online assessment, which takes place remotely at participants' home, importantly differs from face-to-face assessment in the laboratory (for an in-depth comparison of the two testing approaches, see Feenstra et al., 2017; also see Backx et al., 2020). Besides many other differences, one of the key distinctions is the physical presence of an experimenter in laboratory assessments, which can increase the rigor of the assessment (e.g., answering participants' questions, verifying that instructions are properly understood, providing a motivating and encouraging setting), but can also bias participants' performance (e.g., inadvertently inducing expectations of participants' performance, stereotypes, social desirability, stress, and anxiety). A second key difference represents the degree of control that the researchers have over the assessment environment (e.g., distractions, noise, interruptions, time of assessment, use of help, cheating), which is very high in the laboratory but almost absent when participants perform tasks at home. Thus, laboratory assessment allows for more controlled and standardized observations, whereas home assessment has the advantage of providing a more familiar environment, potentially allowing for more natural and spontaneous behaviors to occur. A third key difference represents the researchers' control over the assessment material (e.g., computer hardware, screen size, performance of processor, audio equipment, internet connection), which again can affect participants' performance but also how tasks are presented and how data is recorded, which is particularly important when measuring reaction times (Backx et al., 2020). Related to this, a fourth key difference between the two assessment approaches is that in the laboratory the technical equipment is provided by the researchers, whereas online assessments rely on participants' personal equipment. Research shows that access to computer technology and internet are affected by multiple socioeconomic factors, such as income, education, and age, but also by the region-specific state of digitalization (e.g., Elena-Bucea et al., 2020; Haight et al., 2014; Kontos et al., 2007; Kronenfeld & Penedo, 2020). Thus, online studies can be subject to certain biases (e.g., self-selection of sample, observations being influenced by quality or availability of hardware) and may lead to results that are less representative of the general population.

Even though they are not aimed to replace traditional face-to-face instruments and present certain challenges of their own, online tools can have a series of considerable advantages over laboratory testing. Most importantly, they have the advantage of significantly reducing aforementioned costs related to personnel, equipment, and time. In this regard, previous work shows that online assessment has the potential to reach a larger number of participants in shorter periods of time (Reips, 2002). It can facilitate and accelerate data collection and thereby make results available much faster (also see Finley & Penningroth, 2015). Online testing can also be beneficial when assessing populations that are vulnerable or otherwise difficult to reach (Bohner et al., 2002; Schmidt, 1997). For example, studies show that individuals with high social anxiety are much more willing to participate in online rather than in laboratory studies (e.g., Doorley et al., 2020). In the context of reaching out to various populations, online assessment may be particularly helpful for lifespan research. For example, previous research shows that mobility decreases across the lifespan (Ferrucci et al., 2016), making it more difficult to come to the laboratory as individuals get older (and thus complicating the recruitment of lifespan samples). Related to this, studies also demonstrate that older adults find the laboratory environment more stressful than younger adults, which can negatively affect their performance (Sindi et al., 2013) and make it more difficult to assess and compare actual abilities. Further, in lifespan studies, working-age adults are typically the most difficult group to recruit; they may find it easier to participate online, at a location and time that suits their (busy) schedule rather than making an appointment to come to the laboratory.

Since the initial launch of the first online experiments, multiple studies have suggested that computerized online assessment can provide results that are reliable and comparable to laboratory assessment across different cognitive domains (such as working memory, executive functioning, episodic memory, verbal fluency, attention, vigilance, or complex psycholinguistic abilities), and even for full-scale neuropsychological test batteries such as the Cambridge Neuropsychological Test Automated Battery or the Cognitive Function Test (Backx et al., 2020; Feenstra et al., 2017; Kim et al., 2019; Luna et al., 2020; Trustram Eve & de Jager, 2014). However, certain studies also highlighted that performance significantly varies between the two settings, that participants are more distracted during self-assessment, and that currently most online batteries lack normative data (Al Baghal, 2017; Gates & Kochan, 2015; Madero et al., 2021). This highlights the need for more research on the utility, comparability, and reliability of online assessments.

In fact, despite the illustrated benefits of online assessment, today, there still is surprisingly little solid empirical evidence on how to appropriately assess PM online. One of the first behavioral studies in this area was conducted by Logie and Maylor (2009) using a single-item online task to assess PM in the context of a BBC TV program (i.e., participants had to remember to click on a smiley-face at the end of a questionnaire). Only a handful of studies have since relied on online testing to assess PM in a similar, single-item paradigm (Horn & Freund, 2020) or in a more fine-grained manner using paradigms with multiple prospective items (Gilbert, 2015a, 2015b; Gladwin et al., 2020; Juncu et al., 2020; Scarampi & Gilbert, 2020). Although these studies suggest (or at least assume) that PM can be assessed reliably in an online setting, it is important to highlight that they did not comprehensively examine the psychometric properties of their task as they, for example, did not include any laboratory measures of PM that would allow for comparisons with online performance. So far, only Finley and Penningroth (2015) directly compared laboratory and online assessment of the same PM task. However, their main focus was to compare data quality and data loss in laboratory versus online testing in order to examine the feasibility of online PM studies. Finley and Penningroth (2015) found that feasibility was lower and that more data was lost online – which partially resulted from participants failing to follow instructions (i.e., they did not perform the PM task). Importantly, the authors did not examine participants' PM performances per se (i.e., whether participants have comparable performances in the two environments). Thus, as of today, the appropriate way of remotely administering PM via fully self-administered online tools remains an open question and it was the purpose of the present paper to comprehensively examine the psychometric properties, feasibility and usefulness of one possible tool that takes a serious game approach to assess PM.

This approach has two major advantages: First, it advances the available literature by going beyond the predominant single-item approach so far mostly used in online studies. Those single-item paradigms reduce variance in participants' performance, making it difficult to compare reliability of laboratory versus online assessment and are less powerful when planning for individual difference analyses. Second, simply adopting more complex PM paradigms that are typically used in laboratory studies may be less engaging and more demanding for the participants in an online context, where no experimenter is present to check on the participant and answer potential questions. Indeed, using a complex, more traditional PM paradigm, Finley and Penningroth (2015) found that participants paid less attention and had more difficulties in following the instructions in the online compared to the laboratory assessment. Using a serious game-like tool could improve such issues, because it should be more engaging than a complex laboratory paradigm, thereby helping participants to stay focused for longer periods of time. Further, serious game-like tasks also make it easier to include brief practice tutorials, which allow verifying that participants correctly understand instructions and are able to execute them correctly (e.g., by asking participants to repeat the practice tutorial until performance reaches a specific threshold or by splitting a task into smaller components and sequentially verifying that participants correctly perform each of them before being able to advance to the next one).

A final issue of the previous literature on online assessment concerns age-effects and lifespan development of PM. In this context, the literature generally suggests that PM performance decreases with age when it is assessed via complex, computerized tasks in the laboratory, but that such age-effects can be absent when PM is assessed with more naturalistic tasks outside of the laboratory (for reviews of PM lifespan development, see

Ballhausen et al., 2019; Zuber & Kliegel, 2020; for a recent study on laboratory versus naturalistic PM performance, also see Schnitzspahn et al., 2020). It is important to highlight that previous studies have administered different paradigms to assess PM in versus outside the laboratory (e.g., Schnitzspahn et al., 2020). Thus, today it is unclear whether administering the same computerized laboratory PM task outside the laboratory in an online assessment would yield age-differences similar to what is observed in laboratory assessment, or whether PM performance would be spared when assessed in a naturalistic setting (i.e., online, at participants' home) despite the computerized, non-naturalistic features of the PM task itself. Similarly, it is unclear whether such differences between laboratory and online performance mainly result from age-related differences in computer literacy, or whether they would reflect actual differences in PM abilities across the adult lifespan.

***The current study***

The current study set out to assess the validity, reliability and utility of the Geneva Space Cruiser for the assessment of PM across the adult lifespan. The Geneva Space Cruiser was adapted from a previously established serious game-like task that has been successfully used in several variants in the laboratory to assess PM in children (Kerns, 2000; Kliegel et al., 2013; Kretschmer et al., 2014; Lewis et al., 2016; Mahy et al., 2015; Rendell et al., 2009; Souchay et al., 2018; Talbot & Kerns, 2014; Voigt et al., 2011; Voigt et al., 2014; Zuber et al., 2019). Specifically, to investigate whether the Geneva Space Cruiser represents a valid, reliable and useful tool to assess PM in a fully self-administered online assessment across the adult lifespan, we aimed to examine the following questions: *(1) Data distribution and sensitivity:* Is there a sufficiently wide distribution in the data (i.e., variance in performance between participants), while avoiding floor or ceiling effects (meaning that the task is neither too easy nor too difficult across the adult lifespan)? Are data distributions comparable between laboratory and online assessment? *(2) Laboratory-*

*online reliability of the Geneva Space Cruiser:* Do outcome measures (i.e., ongoing task performance, PM performance, PM monitoring count) of laboratory and online assessment correlate? *(3) Laboratory-online comparability of the Geneva Space Cruiser:* Are outcome measures in the laboratory and in the online assessment comparable (meaning that there are no significant differences in difficulty levels)? *(4) Test-retest reliability of the online assessment:* Do the outcome measures of the online assessment correlate between test and retest? *(5) Construct validity of the online assessment:* Are the outcome measures of the online version of Geneva Space Cruiser predicted by the outcome measures of a more traditional laboratory PM task? *(6) Age effects and utility for adult lifespan studies:* To assess whether the Geneva Space Cruiser represents a comparable alternative to more traditional laboratory tasks for the assessment of PM in adult lifespan studies (i.e., whether it finds similar age-effects as most laboratory studies), we aimed to examine whether PM performance would be negatively affected by age, and if so, whether negative age effects persisted beyond age-related differences in computer literacy.

## Methods

### *Participants*

*Main sample: validation of laboratory versus online assessment*

Two hundred and sixty-two participants were recruited to participate in the main study. They were recruited through direct contact with the experimenters (i.e., experimenters' networks), word-of-mouth advertising of former participants, and flyers distributed in public places (e.g., at the University, sports clubs, or associations for older adults). Participants were also recruited from a Bachelor course in Psychology. In exchange for participation, students received eligibility to take the course exam (but were not awarded any course credits). All participants were volunteers and did not receive any monetary reward. Inclusion criteria were that they had a computer at home or had access to a computer in a quiet environment in which they could perform the online task. Further, participants had to be native speakers or fluent in French. Because our goal was to examine the feasibility of our online tool in a rather broad population that closely reflects potential populations of future online studies, we primarily aimed to characterize the sample and to exclude participants with rather unusual patterns of behavior that might affect results in unwanted ways. Specifically, we aimed to exclude participants with extreme, outlying values on crystallized intelligence, fluid intelligence, and processing speed as key indicators of cognitive functioning. These variables were not intended for a neuropsychological assessment, as they would not necessarily be sensitive to neurocognitive decline nor to detecting individuals potentially at risk of developing any cognitive impairments, such as mild cognitive impairment or Alzheimer's disease. Thus, we applied a cutoff of 2.5 *SD*, which was more inclusive than typical cutoffs for neuropsychological assessments (e.g., Goldman et al., 2013; Jak et al., 2016). Consequently, six participants were excluded because their crystallized intelligence (i.e., verbal comprehension assessed via the Mill Hill Vocabulary Scale; Deltour, 1993), their fluid intelligence or their processing speed (i.e., perceptual reasoning assessed via the Code subtest and processing speed via the Matrices subtest of the WAIS-IV respectively; Wechsler, 2008) was below 2.5 *SD* of the mean of the entire sample (note that WAIS-IV scores were age-standardized). Three participants were excluded because their monitoring count on the Geneva Space Cruiser was over 2.5 *SD* above the group mean (i.e., they pushed the monitoring button between 123 and 352 times, which equals to a button push every 1 to 3 seconds). One participant was excluded because data for the Geneva Space Cruiser was not properly recorded (i.e., zero-values on all variables). The final sample consisted of 252 adults (67.9 % women) between 19 and 86 years old (*M* = 47.07 years, *SD* = 17.46 years).

*Secondary sample: participants for test-retest reliability assessment*

In the context of a different study, 224 younger adults (between 19 and 35 years old) were recruited and performed the online assessment of the Geneva Space Cruiser a first time (i.e., test) and a second time after one week (i.e., retest). They were recruited from a Bachelor course in Psychology. In exchange for participation, students received eligibility to take the course exam (but were not awarded any course credits). All participants were volunteers and did not receive any monetary reward. Inclusion criteria were the same as for the main sample (see above). Eleven participants were excluded because their crystallized intelligence (i.e., verbal comprehension assessed via the subtest of the NV5-R test battery; Thiébaut & Richoux, 2005) or because their processing speed (assessed via an adapted version of the letter string comparison; Posner & Mitchell, 1967) was below 2.5 *SD* of the mean of the entire sample. Ten participants were excluded because their monitoring count on the Geneva Space Cruiser was over 2.5 *SD* above the group mean. The final sample consisted of 203 younger adults between 19 and 34 years old (age *M* = 22.55, *SD* = 3.01; 82.1% women) who performed the online task two times, with an interval of approximately one week between the two sessions.

***Measures***

*Computer Literacy Scale*

Participants' computer literacy was assessed via the Computer Literacy Scale (CLS; Sengpiel & Dittberner, 2008), which was translated into French. The CLS is a brief two-part questionnaire, which targets the basic understanding of symbols and terms commonly used in interactive computer technology. The first part evaluates diversity and frequency of computer use: participants have to indicate how frequently they use the computer for eleven specific tasks (e.g., searching information, e-banking) on a 4-point Likert scale (1 = never; 2 = seldom; 3 = sometimes; 4 = often). *Diversity and frequency of computer use* is calculated by summing participants' answers and dividing it by 44. The second part evaluates objective knowledge of computer technology: Participants see a series of 26 computer related icons and words (e.g., a triangular play-button (►), the word "hyperlink") and a list of possible answers. Participants have to attribute the correct answer to each computer item (e.g., attributing the answer "play" to the play-button). *Objective computer knowledge* is calculating by summing the number of correct answers divided by 26. Finally, *computer literacy* is calculated as the mean of the two sub-scores.

*Traditional laboratory PM task*

Traditional laboratory PM performance was assessed via an N-back PM paradigm, as N-back working memory paradigms have been used repeatedly in previous PM studies (e.g., Altgassen et al., 2010; Ballhausen et al., 2017; Chen et al., 2017; Cona et al., 2015; Einstein et al., 1992; Einstein & McDaniel, 1990; Jäger & Kliegel, 2008; Oksanen et al., 2014). As ongoing task (=OT), participants performed a 2-back updating task in which they had to decide for each trial whether the currently presented picture was the same as the picture presented two trials before. If the picture was the same, they had to push the "green button" (i.e., right arrow on the keyboard), else, they had to push the "red key" (i.e., left arrow on the keyboard). Each picture was presented during 1000ms preceded by a 1500ms fixation cross. After successfully completing a practice block (i.e., succeeding at least 10 out of 14 practice trials), participants completed an OT only block of 48 trials. Then, the PM task was presented: Participants were instructed that, in addition to the OT, they had to remember to push the "Enter" button at every full minute (i.e., at 60 seconds, 120 seconds, etc.). Participants could check the progression of time by pushing the spacebar, which displayed a timer at the bottom of the screen for 3 seconds (format: mm:ss). After short delay phase, during which they completed the Mill Hill Vocabulary Scale, participants performed the ongoing and PM task for six minutes (i.e., 144 OT trials, six PM target times). Finally, participants were asked to describe the task instructions to ensure they remembered the PM task retrospectively. All participants of the final sample were able to recall the PM instructions. *Outcome measures:* OT performance was indexed as “hits minus false alarms” (“proportion of correct responses on hit trials” – “proportion of false alarms on non-hit trials”; for similar scoring, see e.g., Zuber et al., 2019). PM performance was indexed as PM accuracy (“number of PM responses” / “number of possible PM responses”; note that, as in previous studies, responses given ± 5 seconds to the target time were counted as correct PM responses).

*Geneva Space Cruiser*

Laboratory and online PM performance was assessed via the Geneva Space Cruiser, a new, adapted version of the previously established “Cruiser” or “CyberCruiser” (Kerns, 2000; Kliegel et al., 2013; Kretschmer et al., 2014; Lewis et al., 2016; Mahy et al., 2015; Rendell et al., 2009; Souchay et al., 2018; Talbot & Kerns, 2014; Voigt et al., 2011; Voigt et al., 2014; Zuber et al., 2019). Specifically, to allow collecting data online, the current version of the Geneva Space Cruiser was programmed by David Framorando using the JavaScript programming language for the online display of the task, and PHP and MySQL for data collection and storage. Participants were instructed that they would be navigating a spaceship in a cosmic-like environment in which series of stars and asteroids cross the screen from right to left (see Figure 1 for an illustration). They could navigate their spaceship horizontally and vertically (using the four arrow keys) and were instructed that their goal was to gain as many points as possible (ongoing task, OT). Collecting stars (by maneuvering the spaceship into them) would gain 30 points, whereas colliding with asteroids would lose 100 points. Participants were further informed that by pushing the F-key the spaceship could launch missiles and that destroying asteroids with missiles would also gain 30 points. The current score was displayed in the top left corner of the screen. After receiving these general instructions, participants completed a practice tutorial. In a first step, participants had to maneuver the spaceship to the right, left, top and bottom border of the screen. They were only allowed to advance to the next action (e.g., moving to left border of the screen) after having successfully completed the previous action of the tutorial (e.g., moving to the right border of the screen). This allowed familiarizing participants with the navigation and assuring that they correctly followed instructions. In a second step, participants were familiarized with firing missiles: they were required to fire two missiles before being allowed to advance in the practice. Then, the PM task was presented: Participants were told that their spaceship also needed fuel. They could display the fuel gauge by pushing the C-key, which displayed the fuel gauge for 3 seconds at the top right corner of the screen. The fuel gauge turned from green to yellow to red, and participants were told that they could refuel whenever the gauge was in the red section by pushing the E-key. The fuel gauge turned red every 50 seconds and was fully emptied every 60 seconds (note that these time aspects were not explicitly communicated to the participants). Refueling in the red time window would gain 300 points, whereas not refueling would not gain or lose any points and the spaceship would refuel automatically. After receiving these instructions, in a third tutorial, participants had to display the fuel gauge twice before being able to advance. They were then shown a summary of the point system and of the different tasks. Next, they performed a final practice during which they could perform all actions explained previously (including refueling). This practice lasted for 70 seconds and participants were reminded in-game that they should display the fuel gauge and that they refuel (if they did not do so spontaneously). Before starting the actual test phase, participants then completed the subtest Codes of the WAIS-IV in the laboratory session and the Computer Literacy Scale in the online session, which both took approximately 5 minutes to complete. After this delay phase, participants worked on the Geneva Space Cruiser task for six minutes and ten seconds (i.e., six PM target times). Finally, in the laboratory session, participants were asked to describe the task instructions to ensure they remembered the PM task retrospectively. Note that the Geneva Space Cruiser was programmed so that it pre-loads each task block in its entirety (i.e., practice block; actual game with PM task) before it starts. Similarly, during the game, data on participants' performance are temporarily stored in the computers' data buffer and only transferred to the online database at the end of each block. Thus, temporarily instable or slower internet connections would only affect the time to pre-load the task block, but should not affect the smoothness of the game running nor data collection (including performance and response time recording). *Outcome measures:* OT performance was calculated as the total points received for the OT only (i.e., without points that were gained for refueling). PM performance was indexed as PM accuracy ("number of refuels" / "number of possible refuels"). The monitoring count represents the total number of times participants pushed the button to display the fuel gauge.

***Procedure***

*Procedure for main sample*

All illustrated in Figure 2, the current study consisted of three assessment sessions. For all participants of the main sample, the first session was conducted in the laboratory, during which participants gave their informed consent, filled out a socio-demographic questionnaire, completed the traditional laboratory PM task, the Mill Hill Vocabulary Scale (Deltour, 1993) as delay phase of the PM task, and finally the Matrices subtest of the WAIS-IV (Wechsler, 2008). Participants then assisted in a second laboratory session and in an online session, during both of which they performed the Geneva Space Cruiser (note that in both assessment environments, the Geneva Space Cruiser was the exact same task, with the only exception that in the laboratory assessment, the experimenter was present during the instruction phase to answer potential questions but left the room as soon as the actual task started). These two sessions were administered in counter-balanced order, with half of the participants first doing the second laboratory session and then the online session, and vice versa. As delay phase of the Geneva Space Cruiser, participants completed the subtest Codes of the WAIS-IV (Wechsler, 2008) in the laboratory session and the Computer Literacy Scale in the online session. There was approximately one week between each of the three sessions. If the final session was conducted in the laboratory, participants were debriefed at the end of the session (i.e., explanation of the main study goals and answering participants' questions). If the final session was online, participants were called on the telephone for the debriefing.

*Procedure for secondary sample*

In the context of a different study, the secondary sample participated in two online sessions. During the first online session, participants gave their informed consent, filled out a socio-demographic questionnaire, and completed other questionnaires and cognitive tasks that are not of principal interest for the current analyses (e.g., stress, sleep, affect, verbal comprehension task, processing speed). Then they performed the Geneva Space Cruiser, with the State-Trait Anxiety Inventory questionnaire delay phase. Finally, they completed a task evaluating overall cognitive functioning. The second online session was administered approximately one week later. Participants again completed multiple questionnaires (e.g., on persevering thoughts, resilience). Then, they performed the Geneva Space Cruiser, and filled out the different questionnaires (i.e., on personality, stress, and anxiety) as delay phase. Finally, participants again completed a task evaluating overall cognitive functioning.

***Statistical analyses***

All analyses and results are based on the final samples after exclusions. Missing data was treated with pairwise deletion for respective analyses. Bayesian analyses were conducted using version 1.2 of the free statistics sofware jamovi (The jamovi project, 2020). All other analyses were conducted using version 25 of the SPSS Statistics software package. Figures were created in Python, version 3.8.3, using the libraries 'seaborn' and 'matplotlib'. According to guidelines provided by Dancey and Reidy (2017) and Everitt and Skrondal (2010), the following interpretation of correlation strengths and Cronbach's alphas was applied: Correlations: .10-.40 = weak, .40-.70 = moderate, .70-1 = strong;

Cronbach's alphas: < 0.60 unacceptable, 0.60-0.65 undesirable, 0.65-0.70 acceptable, 0.70–0.80 respectable, 0.80-0.90 very good.

## Results

### *Descriptive statistics of main sample per age group*

Table 1 shows number of participants, age, percentage of women, education, and descriptive statistics of control variables per (~10-year) age group. We conducted a series of one-way ANOVAs with age group as factor on education and on our control variables. There was a significant effect of age group on education, $F(5,245) = 2.28$, $p = .048$. However, all post-hoc Bonferroni-corrected comparisons were non-significant (*p*s > .104). There was a significant effect of age group on the Mill Hill score, $F(5,245) = 5.05$, $p < .001$. Post-hoc Bonferroni-corrected comparisons showed that the 19 – 29 group had significantly lower scores compared to all three groups that were 50 years or older (*p*s ≤ .033). There was no significant effect of age group neither on the Matrices score, $F(5,244) = 1.51$, $p = .186$, nor on the Code score, $F(5,238) = 0.81$, $p = .547$. There was a significant effect of age group on the Computer Literacy Scale score, $F(5,228) = 20.58$, $p < .001$. Post-hoc Bonferroni-corrected comparisons showed that the 70+ group had significantly lower computer literacy than all other groups (*p*s < .001), that the 60 – 69 group had lower literacy compared to 19 – 29 and to the 30 – 39 group (*p*s < .001), and that the 50 – 59 group had lower literacy only compared to 19 – 29 group ($p = .011$).

### *Missing data*

To examine missing data, we conducted a post-hoc analysis of missing values using Little's test of missing completely at random (MCAR). Based on expectation-maximization as estimation method, this revealed that data of the three main outcome measures were missing completely at random in both assessment environments, $\chi^2 = 2.56$, $p = .861$, meaning that missingness should not have systematically affected results.

### *Validity, reliability, and utility of the Geneva Space Cruiser*

(1) *Data distribution and sensitivity:* Histograms in Figure 3 depict distributions of participants' PM performance on the laboratory and on the online assessment, suggesting similar distributions of PM performance in both environments and no floor nor ceiling effects. A two-sample Kolmogorov-Smirnov Z test confirmed that the data followed similar distributions in both assessment environments, $Z = 0.04$, $p = .989$.

(2) *Laboratory-online reliability of the Geneva Space Cruiser:* Table 2 displays descriptive statistics and correlations of outcome measures on laboratory and online version of the Geneva Space Cruiser. Correlations show moderate links between laboratory and online assessment (*r*s between .56 and .68).

(3) *Laboratory-online comparability of the Geneva Space Cruiser:* Table 2 also displays Student and Bayesian paired-sample *t*-tests between the laboratory and the online outcome measures of the Geneva Space Cruiser. Bayes factors indicate moderate evidence for differences in OT performance, but strong evidence that there were no differences in PM performance nor in monitoring counts (for an interpretation of Bayes factors, see Schönbrodt & Wagenmakers, 2018).

(4) *Test-retest reliability of the online assessment:* Table 3 shows descriptive statistics and test-retest reliability of the Geneva Space Cruiser in the secondary sample. With Cronbach's alphas ≥ .80 and correlations ≥ .66, results suggest very good test-retest reliability with moderate to strong correlations between the two assessment times.

(5) *Construct validity of the online assessment:* To examine construct validity of the online version of the Geneva Space Cruiser, we regressed its three main outcome measures (i.e., OT and PM performance, monitoring count) on the same measures assessed with the more traditional laboratory PM task in three separate regressions. Results show that OT performance on the traditional laboratory PM task significantly predicted the OT performance on the online assessment, $\beta = .33$, $t(207) = 5.08$, $p < .001$, $R^2 = .11$. Similarly, PM performance on the traditional laboratory PM task significantly predicted the PM

performance on the online assessment, β = .29, $t(183) = 4.17$, $p < .001$, $R^2 = .09$. Finally, monitoring count on the classical laboratory PM task significantly predicted the monitoring count on the online assessment, β = .39, $t(174) = 5.52$, $p < .001$, $R^2 = .15$.

(6) *Age effects and utility for adult lifespan studies:* To examine whether the Geneva Space Cruiser represents a comparable alternative to laboratory tasks for the assessment of PM in adult lifespan studies (i.e., whether it finds similar age-decreases as traditional laboratory studies beyond differences in computer literacy), we regressed laboratory and online PM performance on age (step 1) and on age and computer literacy (step 2) in hierarchical regressions (see Table 4). Results show that age and computer literacy significantly predicted PM performance in both assessment environments, and that age had a significant and larger effect over and above computer literacy. Figure 4 shows regression plots with PM performance regressed on age. Similarly, Figure 5 shows bar plots of PM performance per age group. Together, hierarchical regression and Figures 4 and 5 show age-related decreases in PM performance. Importantly, both figures also show large inter-individual differences and high variability in performance across the adult lifespan (no ceiling nor floor effects for any of the age groups). Finally, to further examine the utility of the Geneva Space Cruiser for the assessment of PM across the lifespan, we compared PM performance in the laboratory versus online per age group. Results of frequentist and Bayesian *t*-tests showed anecdotal to moderate evidence that there were no differences in PM performance between the two assessment environments for any of the age groups (all *p*s > .05; $BFs_{10}$ ranging from 0.44 to 0.17).

***Gender effects on validity and reliability***

Following up on suggestions from an anonymous reviewer, we examined gender effects on validity and reliability of the Geneva Space Cruiser as post-hoc analyses. First, we assessed gender effects on laboratory-online reliability by separately examining correlations for women and men of the main sample. In women, correlations between laboratory and online assessments of OT performance, PM performance, and monitoring count were $r = .50$, $r = .61$, and $r = .62$, respectively (all *p*s < .001), whereas in men they were $r = .59$, $r = .59$, $r = .79$, respectively (all *p*s < .001). Second, we examined the comparability of online and laboratory assessment per gender in the main sample. In women, there was anecdotal evidence for lower OT performance online compared to the laboratory ($t(135) = 2.60$, $p = .01$, $BF_{10} = 2.42$), but moderate evidence for neither differences in PM performance nor in monitoring counts (*p*s > .05; $BF_{10\ PM\ performance} = 0.36$, $BF_{10\ monitoring} = 0.10$). In men, there was anecdotal to moderate evidence for no differences on all three outcomes (all *t*-tests were non-significant; $BF_{10\ OT\ performance} = 0.27$, $BF_{10\ PM\ performance} = 0.53$, $BF_{10\ monitoring} = 0.19$). Third, we examined test-retest reliability of the online assessment per gender in the secondary sample. In women, correlations between laboratory and online assessments of OT performance, PM performance, and monitoring count were $r = .72$, $r = .64$, and $r = .77$, respectively (all *p*s < .001), whereas Cronbach's alphas were $\alpha = .84$, $\alpha = .78$, and $\alpha = .87$. In men, test-retest correlations were $r = .88$, $r = .69$, and $r = .71$ respectively, (all *p*s < .001), whereas Cronbach's alphas were $\alpha = .93$, $\alpha = .82 = \alpha = .87$. Fourth, we directly compared performance on the different outcome measures between genders in the main sample. There was strong to very strong evidence that women performed significantly worse on the OT compared to men, both in the laboratory ($t(229) = - 3.04$, $p = .003$, $BF_{10} = 11.04$) and online ($t(219) = - 3.74$, $p < .001$, $BF_{10} = 93.49$). In contrast, all other comparisons were non-significant (all *p*s > .05) and provided moderate evidence for no gender differences in PM performance ($BF_{10\ laboratory} = 0.17$, $BF_{10\ online} = 0.35$), and anecdotal evidence for no gender differences in monitoring ($BF_{10\ laboratory} = 0.51$, $BF_{10\ online} = 0.87$). Finally, we examined gender effects on our control variables, which showed moderate evidence that there were no gender differences on computer literacy, education, fluid or crystallized intelligence (all *p*s > .05; all $BFs_{10}$ <

.33). However, women had significantly faster processing speed than men ($t(242) = 3.80$, $p < .001$, $BF_{10} = 116.15$).

**Discussion**

The current study set out to examine the validity, reliability and utility of the Geneva Space Cruiser for the assessment of PM across the adult lifespan via online testing. First, regarding *data distribution and sensitivity* of the tool, present findings show that the Geneva Space Cruiser yields similar data distributions in the online and in the laboratory setting. In both environments, we observe large variance in performance between participants while avoiding floor or ceiling effects. Together, this indicates good sensitivity of the tool, which allows detecting inter-individual differences in PM performance across the adult lifespan using the same task for all age groups. Second, regarding *reliability*, the tools' main outcome measures (i.e., OT performance, PM performance and PM monitoring) correlate moderately between laboratory and online assessment and there are no significant differences neither in PM performance nor in PM monitoring between the two settings. Further, performances on the online task are stable over multiple assessments (i.e., very good test-retest reliability). Together, these findings suggest moderate online-laboratory reliability (across the lifespan) and very good overall retest reliability (in younger adults) of the Geneva Spacer Cruiser to assess PM online. Third, regarding *construct validity*, results show that all three outcome measures on the Geneva Space Cruiser are significantly predicted by outcomes on a more traditional laboratory PM task, suggesting acceptable construct validity of the new online assessment tool. However, it is important to note that the explained variances on the three outcome measures of the Geneva Space Cruiser were rather small (ranging from 9 to 15%), which suggests that other factors, such as participants' motivation, planning skills, prior task experience, or their personal strategies, may also have affected their performance (for factors contributing to PM performance, see e.g., Zuber & Kliegel, 2020).

Fourth, regarding *age effects and utility for adult lifespan studies*, results of hierarchical regressions demonstrate that the Geneva Space Cruiser is a sensitive tool for detecting negative age effects on PM (i.e., age-related declines in PM performance) over and above inter-individual differences in computer literacy. Inspection of associations between PM performance and age further shows that the Geneva Space Cruiser yields inter-individual differences and variance in performance across the adult lifespan (i.e., no ceiling nor floor effects for any of the age groups). Together, this suggests that the Geneva Space Cruiser represents a rather useful tool for adult lifespan studies, finding similar age effects as studies using more traditional laboratory paradigms (see Zuber & Kliegel, 2020 for an overview). However, although inter-individual variability was large in all age groups, we highlight that for participants of age 70 and older, mean PM performance was rather low. This suggests that both the online as well as the laboratory assessment may have been challenging for the oldest participants.

Finally, regarding *gender effects*, our findings suggest that for PM performance as well as for monitoring validity and reliability slightly varied but generally were relatively comparable between men and women. Both genders also had comparable education, computer literacy, fluid and crystallized intelligence (but women had significantly faster processing speed than men). However, when assessed online, women performed worse on the OT compared to men and compared to when women were assessed in the laboratory. Similarly, women's OT performance was less reliable compared to men. Together, this suggests that both genders showed relatively similar patterns of results for PM performance and for monitoring, but that the online assessment led to gender effects on OT performance, which could not be explained by differences in any of the control variables.

In sum, findings of the present study indicate that the Geneva Spacer Cruiser represents a rather valid, moderately to highly reliable, and generally useful tool for the assessment of PM across wide ranges of the adult lifespan, forming a fairly comparable alternative to more traditional PM tasks used in laboratory assessments. Although previous studies assumed that PM could be assessed online, this is the first study to validate a fully self-administered online tool for the assessment of PM and to demonstrate that PM performances can be relatively similar online as when assessed in the laboratory. Validating an online tool as alternative to laboratory assessment of PM has multiple advantages, such as requiring less resources, less personnel, and less planning. A further advantage of online tools is that they can create a less stressful testing environment (e.g., Sindi et al., 2013) and thereby provide easier access to populations that are vulnerable or otherwise difficult to reach (e.g., Bohner et al., 2002; Schmidt, 1997). Further, as recent events related to COVID-19 have illustrated, online tools are particularly handy to maintain research activities in times where face-to-face contact with participants may be restricted. In this context, online tools are particularly useful because they allow collecting data in larger samples and in shorter periods of time (Reips, 2002), providing faster results to potentially pressing research questions (also see Finley & Penningroth, 2015). PM represents one of the key indicators of cognitive functioning in everyday life, which highlights the relevance and the need for a new tool that can be used to assess PM in large-scale online studies targeting everyday relevant outcome measures. To facilitate study preparation and data collection, it is therefore particularly handy to have a task available that can be used across wide ranges of the adult lifespan without necessarily having to adapt its characteristics to specific age groups (although certain adaptations for the oldest participants may be useful).

***Limitations and outlook***

Despite the various advantages mentioned above, in subsequent sections we aim to highlight challenges that come with online testing in general, limitations of the Geneva Space Cruiser in particular, and to provide an outlook for future research. Previous studies targeting feasibility and utility of online assessment tools suggest that downsides of online testing are higher dropout rates and lower percentages of the initial data that can be used for the final analyses (Birnbaum, 2004; Finley & Penningroth, 2015). Indeed, in our own study, for certain analyses the sample was up to 30% smaller than the full sample tested. However, examining missing data revealed that data of the three main outcome measures were missing completely at random in both assessment environments. This indicates that missing data were independent of any observed and unobserved variables and not due to any systematic mechanisms of missingness (Li, 2013). Indeed, missing data were mostly due to accidental technical issues with recording the data and to certain participants not completing all parts of the assessment. Thus, for studies that are fully online this should not be very problematic per se, as dropouts and missing data can be remedied easily and at little cost by recruiting larger samples. However, it seems important to highlight this fact, as it implies that, when planning target sample sizes, researchers should take a rather conservative and cautious approach and aim for larger participant numbers, if possible.

Previous studies also suggest that a further challenge of online testing consists in the lack of control over participants, their compliance, and the testing environment itself (e.g., Goodman et al., 2013; Skitka & G., 2006). For example, Finley and Penningroth (2015) find that their participants were less attentive and understood PM instructions less well when assessed online. Although in our study we find strong evidence that PM performance and PM monitoring were similar in the laboratory and in the online assessment, we also observe moderate evidence that OT performance was lower in the online assessment. One strength of the Geneva Space Cruiser is that it includes a detailed practice tutorial, allowing researchers verifying that participants correctly understood and were (theoretically) able to perform ongoing and PM task correctly. Therefore, differences in OT performance were unlikely to be the result of difficulties understanding and executing the instructions. However, it is possible that they resulted from participants being less attentive to the OT when tested at home. Although we again highlight that this did not affect PM performance per se, future studies will have to further investigate how this affects other aspects of performing a PM task (e.g., whether participants commit more false PM responses [i.e., responding outside of the PM target window] or whether OT performance could be improved with additional experimental manipulations). Further, although previous versions of the Cruiser paradigm have mainly been used to assess PM in children, the current study focused on adults only and future studies will have to confirm that the Geneva Space Cruiser is suitable to assess children similarly to its predecessors. In this context, assessing minors represents an additional challenge in online studies, as it also requires the consent of children's parents or their legal guardians and thus presents additional technical, legal, or ethical hurdles.

Further, although we found rather similar results between women and men on PM performance and monitoring, it is important to highlight that women generally displayed worse performance and lower reliability on the OT, particularly when assessed online. Although this does not seem to impact the assessment of PM (i.e., performing the PM task), a possible explanation of gender differences on the OT (i.e., playing the video game) could be that men are generally more experienced in video gaming and in using the internet (e.g., Helsper, 2010; Li & Kirkup, 2007; Lucas & Sherry, 2004; Phan et al., 2012). Although we did not find any gender differences in computer literacy per se, future studies will have to investigate the role of previous gaming experience and internet use in more detail. As in-depth examinations of online tools are still relatively scarce, today there is little discussion of gender differences in this context and more research is required, addressing whether gender differences consistently exist in online assessment, and if so, whether those vary across the lifespan and between cohorts.

In terms of age-related cognitive impairments, we highlight that we intentionally set rather broad in- and exclusion criteria for the present study, because we aimed to examine the feasibility and utility of the Geneva Space Cruiser for future online studies in naturally sampled and therefore rather diverse populations. Thus, we aimed to include participants that had access to the necessary technical equipment and were able to perform the online tasks autonomously at a distance, but we did not aim to exclude individuals potentially at risk of developing any cognitive impairments, such as mild cognitive impairment or Alzheimer's disease. Our results are encouraging in that they suggest the Geneva Space Cruiser to be feasible for a large portion of the general population. However, a shortcoming of the present findings is that they do not allow to draw specific conclusions about feasibility and utility for different at-risk populations.

In this regard, it is also important to highlight that the present study included a relatively small number of participants above the age 70. Further, although they were highly educated and had high scores on control variables (and may thus not be completely representative of the general population of that age), their PM performance was rather low. In addition, test-retest reliability was not assessed in this group. Thus, future studies will have to examine in more detail how the present findings apply to the oldest individuals in particular, for example by matching education and cognitive performance with the general population and by considering age-related changes in performance after age 70.

Related to this, a final limitation of online studies is that they require participants to have access to a computer (or other computer-based technology such as tablets). This criterion may pre-select or at least favor a particular profile of individuals (e.g., younger; wealthier or more educated; cognitively more performing or more at ease with the use of technologies; see Elena-Bucea et al., 2020; Haight et al., 2014; Kontos et al., 2007; Kronenfeld & Penedo, 2020). Our recruitment strategies may have further contributed to such (self-)selection biases, as they likely favored participation of individuals that were more motivated, had higher education and higher intellectual abilities, and had more experience with and easier access to computer technologies. In the context of adult lifespan studies, this could be particularly problematic, because middle-aged and older adults that meet this criterion may be highly selective sample that is little representative of the general population. Although research shows that computer use and computer access become increasingly common across the lifespan (Federal Statistical Office, 2020a, 2020b; Hunsaker & Hargittai, 2018; Mitzner et al., 2010; Pew Research Center, 2019), future studies will have to further examine which participants are selected via this criterion and to which degree results generalize to the general population, and maybe older adults in particular.

Regarding future research, we also highlight that the present study examined a time-based PM version of the Geneva Space Cruiser, which are taxing PM tasks in terms of cognitive demands and allocation of attentional resources (see Zuber & Kliegel, 2020, for a review). Consequently, we argue that if a time-based version yielded rather valid and reliable results when administered online, less resource-demanding versions should be similarly feasible. However, this will have to be corroborated by future studies. For this purpose, the Geneva Space Cruiser – as its predecessors – could be easily adapted to assess different types of PM (e.g., by using asteroids or less relevant elements of the game as event-based PM cues).

***Conclusion***

In conclusion, in times when conducting research in face-to-face laboratory settings may be restricted and access to participants limited, it seems particularly timely and valuable to have alternative assessment tools available that allow conducting research remotely. Within this scope and its aforementioned limitations, the current study demonstrates that the Geneva Space Cruiser represents a rather valid, moderately to highly reliable, and therefore generally useful alternative to classical laboratory testing for the assessment of PM across wide ranges of the adult lifespan. However, the present findings also suggest that online assessment may be somewhat more challenging for women and for the oldest participants.

**Acknowledgements**

We thank Hadir Elhanafi, Vanessa Garcia, Annick Hottelier, Laurène Kuhn, Clémentine Laboret, Clara Mourin, Laura Pede, and Malory Sugnaux for their assistance in study preparation and data collection. This work was supported by the Swiss National Science Foundation (SNSF) under Grant 100019_165572 and by the Swiss National Center of Competences in Research LIVES – Overcoming vulnerability: life course perspectives, which is financed by the Swiss National Science Foundation (grant number: 51NF40-160590).

**Disclosure of interest statement**

The authors report no conflict of interest.

**Data availability statement**

The datasets generated and analyzed for the current study are not publicly available because participant consent forms did not include authorization for public data sharing. However, data are available from the corresponding author [SZ] on reasonable request.

**References**

Al Baghal, T. (2017). The Effect of Online and Mixed-Mode Measurement of Cognitive Ability. *Social Science Computer Review*, *37*(1), 89-103. https://doi.org/10.1177/0894439317746328

Altgassen, M., Phillips, L. H., Henry, J. D., Rendell, P. G., & Kliegel, M. (2010). Emotional target cues eliminate age differences in prospective memory. *Quarterly Journal of Experimental Psychology*, *63*(6), 1057-1064. https://doi.org/10.1080/17470211003770920

Backx, R., Skirrow, C., Dente, P., Barnett, J. H., & Cormack, F. K. (2020). Comparing Web-Based and Lab-Based Cognitive Assessment Using the Cambridge Neuropsychological Test Automated Battery: A Within-Subjects Counterbalanced Study [Original Paper]. *Journal of Medical Internet Research*, *22*(8), e16792. https://doi.org/10.2196/16792

Ballhausen, N., Hering, A., Rendell, P. G., & Kliegel, M. (2019). Prospective memory across the lifespan. In J. Rummel & M. A. McDaniel (Eds.), *Prospective Memory*. Routledge. https://doi.org/https://doi.org/10.4324/9781351000154

Ballhausen, N., Schnitzspahn, K. M., Horn, S. S., & Kliegel, M. (2017). The interplay of intention maintenance and cue monitoring in younger and older adults' prospective memory. *Memory & Cognition*, *45*(7), 1113-1125. https://doi.org/10.3758/s13421-017-0720-5

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, *55*(1), 803-832. https://doi.org/10.1146/annurev.psych.55.090902.141601

Bohner, G., Danner, U. N., Siebler, F., & Samson, G. B. (2002). Rape myth acceptance and judgments of vulnerability to sexual assault: an Internet experiment.

*Experimental Psychology*, *49*(4), 257-269. https://doi.org/10.1026/1618-3169.49.4.257

Chen, Y. Z., Lian, R., Yang, L. X., Liu, J. R., & Meng, Y. F. (2017). Working Memory Load and Reminder Effect on Event-Based Prospective Memory of High- and Low-Achieving Students in Math. *Journal of Learning Disabilities*, *50*(5), 602-608. https://doi.org/10.1177/0022219416668322

Cona, G., Kliegel, M., & Bisiacchi, P. S. (2015). Differential effects of emotional cues on components of prospective memory: an ERP study. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00010

Dancey, C. P., & Reidy, J. (2017). *Statistics Without Maths for Psychology* (7th ed.). Pearson Education.

Deltour, J. J. (1993). *Echelle de vocabulaire de Mill Hill de JC Raven: Adaptation française et normes européennes du Mill Hill et du Standard Progressive Matrices de Raven (PM38)*. Editions L'Application des Techniques Modernes.

Doorley, J. D., Volgenau, K. M., Kelso, K. C., Kashdan, T. B., & Shackman, A. J. (2020). Do people with elevated social anxiety respond differently to digital and face-to-face communications? Two daily diary studies with null effects. *Journal of Affective Disorders*, *276*, 859-865. https://doi.org/https://doi.org/10.1016/j.jad.2020.07.069

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

Elena-Bucea, A., Cruz-Jesus, F., Oliveira, T., & Coelho, P. S. (2020). Assessing the Role of Age, Education, Gender and Income on the Digital Divide: Evidence for the European Union. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-020-10012-9

Everitt, B. S., & Skrondal, A. (2010). *The Cambridge Dictionary of Statistics* (4th ed.). Cambridge University Press.

Federal Statistical Office. (2020a). *Digitale Kompetenzen* Retrieved 04.01.2021 from https://www.bfs.admin.ch/bfs/de/home/statistiken/kultur-medien-informationsgesellschaft-sport/informationsgesellschaft/gesamtindikatoren/haushalte-bevoelkerung/digitalekompetenzen.html

Federal Statistical Office. (2020b). *Internetnutzung*. Retrieved 04.01.2021 from https://www.bfs.admin.ch/bfs/de/home/statistiken/kultur-medien-informationsgesellschaft-sport/informationsgesellschaft/gesamtindikatoren/haushalte-bevoelkerung/internetnutzung.html

Feenstra, H. E. M., Vermeulen, I. E., Murre, J. M. J., & Schagen, S. B. (2017). Online cognition: factors facilitating reliable online neuropsychological test results. *The Clinical Neuropsychologist*, *31*(1), 59-84. https://doi.org/10.1080/13854046.2016.1190405

Ferrucci, L., Cooper, R., Shardell, M., Simonsick, E. M., Schrack, J. A., & Kuh, D. (2016). Age-related change in mobility: perspectives from life course epidemiology and geroscience. *The Journals of Gerontology: Series A*, *71*(9), 1184-1194. https://doi.org/10.1093/gerona/glw043

Finley, A. J., & Penningroth, S. L. (2015). Online versus in-lab: pros and cons of an online prospective memory experiment. In A. M. Columbus (Ed.), *Advances in Psychology Research* (Vol. 113, pp. 135-162). Nova Science Publishers, Inc.

Gates, N. J., & Kochan, N. A. (2015). Computerized and on-line neuropsychological testing for late-life cognition and neurocognitive disorders: are we there yet? *Current Opinion in Psychiatry*, *28*(2), 165-172. https://doi.org/10.1097/yco.0000000000000141

Gilbert, S. J. (2015a). Strategic offloading of delayed intentions into the external environment. *Quarterly Journal of Experimental Psychology*, *68*(5), 971-992. https://doi.org/10.1080/17470218.2014.972963

Gilbert, S. J. (2015b). Strategic use of reminders: Influence of both domain-general and task-specific metacognitive confidence, independent of objective memory ability. *Consciousness and Cognition*, *33*, 245-260. https://doi.org/10.1016/j.concog.2015.01.006

Gladwin, T. E., Jewiss, M., Banic, M., & Pereira, A. (2020). Associations between performance-based and self-reported prospective memory, impulsivity and encoding support. *Acta Psychologica*, *206*, 103066. https://doi.org/https://doi.org/10.1016/j.actpsy.2020.103066

Goldman, J. G., Holden, S., Bernard, B., Ouyang, B., Goetz, C. G., & Stebbins, G. T. (2013). Defining optimal cutoff scores for cognitive impairment using Movement Disorder Society Task Force criteria for mild cognitive impairment in Parkinson's disease. *Movement Disorders*, *28*(14), 1972-1979. https://doi.org/10.1002/mds.25655

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, *26*(3), 213-224. https://doi.org/10.1002/bdm.1753

Haas, M., Zuber, S., Kliegel, M., & Ballhausen, N. (2020). Prospective memory errors in everyday life: does instruction matter? *Memory*, *28*(2), 196-203. https://doi.org/10.1080/09658211.2019.1707227

Haight, M., Quan-Haase, A., & Corbett, B. A. (2014). Revisiting the digital divide in Canada: the impact of demographic factors on access to the internet, level of online activity, and social networking site usage. *Information, Communication & Society*, *17*(4), 503-519. https://doi.org/10.1080/1369118X.2014.891633

Helsper, E. J. (2010). Gendered Internet Use Across Generations and Life Stages. *Communication Research*, *37*(3), 352-374. https://doi.org/10.1177/0093650209356439

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

Horn, S. S., & Freund, A. M. (2020). How Do Gain and Loss Incentives Affect Memory for Intentions Across Adulthood? *The Journals of Gerontology: Series B*. https://doi.org/10.1093/geronb/gbaa140

Hunsaker, A., & Hargittai, E. (2018). A review of Internet use among older adults. *New Media & Society*, *20*(10), 3937-3954. https://doi.org/10.1177/1461444818787348

Jäger, T., & Kliegel, M. (2008). Time-based and event-based prospective memory across adulthood: Underlying mechanisms and differential costs on the ongoing task. *The Journal of General Psychology*, *135*(1), 4-22.

Jak, A. J., Preis, S. R., Beiser, A. S., Seshadri, S., Wolf, P. A., Bondi, M. W., & Au, R. (2016). Neuropsychological Criteria for Mild Cognitive Impairment and Dementia Risk in the Framingham Heart Study. *Journal of the International*

*Neuropsychological Society*, *22*(9), 937-943. https://doi.org/10.1017/S1355617716000199

Juncu, S., Blank, H., Fitzgerald, R. J., & Hope, L. (2020). Do image variability and names in missing person appeals improve prospective person memory? *Journal of Applied Research in Memory and Cognition*, *9*(3), 410-418. https://doi.org/https://doi.org/10.1016/j.jarmac.2020.07.001

Kerns, K. A. (2000). The CyberCruiser: An investigation of development of prospective memory in children. *Journal of the International Neuropsychological Society*, *6*(1), 62-70.

Kim, J., Gabriel, U., & Gygax, P. (2019). Testing the effectiveness of the Internet-based instrument PsyToolkit: A comparison between web-based (PsyToolkit) and lab-based (E-Prime 3.0) measurements of response choice and response time in a complex psycholinguistic task. *Plos One*, *14*(9), e0221802. https://doi.org/10.1371/journal.pone.0221802

Kliegel, M., Mahy, C. E. V., Voigt, B., Henry, J. D., Rendell, P. G., & Aberle, I. (2013). The development of prospective memory in young schoolchildren: The impact of ongoing task absorption, cue salience, and cue centrality. *Journal of Experimental Child Psychology*, *116*(4), 792-810. https://doi.org/10.1016/j.jecp.2013.07.012

Kontos, E. Z., Bennett, G. G., & Viswanath, K. (2007). Barriers and Facilitators to Home Computer and Internet Use Among Urban Novice Computer Users of Low Socioeconomic Position [Original Paper]. *J Med Internet Res*, *9*(4), e31. https://doi.org/10.2196/jmir.9.4.e31

Kretschmer, A., Voigt, B., Friedrich, S., Pfeiffer, K., & Kliegel, M. (2014). Time-based prospective memory in young children-Exploring executive functions as a developmental mechanism. *Child Neuropsychology*, *20*(6), 662-676. https://doi.org/10.1080/09297049.2013.841881

Kronenfeld, J. P., & Penedo, F. J. (2020). Novel Coronavirus (COVID-19): telemedicine and remote care delivery in a time of medical crisis, implementation, and challenges. *Translational Behavioral Medicine*, *11*(2), 659-663. https://doi.org/10.1093/tbm/ibaa105

Kvavilashvili, L., & Rummel, J. (2020). On the nature of everyday prospection: a review and theoretical integration of research on mind-wandering, future thinking, and prospective memory. *Review of General Psychology*, *24*(3), 210-237. https://doi.org/10.1177/1089268020918843

Lewis, C. E., Thomas, K. G., Molteno, C. D., Kliegel, M., Meintjes, E. M., Jacobson, J. L., & Jacobson, S. W. (2016). Prospective Memory Impairment in Children with Prenatal Alcohol Exposure. *Alcoholism: Clinical and Experimental Research*, *40*(5), 969-978. https://doi.org/10.1111/acer.13045

Li, C. (2013). Little's test of missing completely at random. *Stata Journal*, *13*(4), 795-809. https://EconPapers.repec.org/RePEc:tsj:stataj:v:13:y:2013:i:4:p:795-809

Li, N., & Kirkup, G. (2007). Gender and cultural differences in Internet use: A study of China and the UK. *Computers & Education*, *48*(2), 301-317. https://doi.org/https://doi.org/10.1016/j.compedu.2005.01.007

Loft, S., Dismukes, K., & Grundgeiger, T. (2019). Prospective memory in safety-critical work contexts. In J. Rummel & M. A. McDaniel (Eds.), *Prospective Memory*. Routledge.

Logie, R. H., & Maylor, E. A. (2009). An Internet study of prospective memory across adulthood. *Psychology and Aging*, *24*(3), 767-774. https://doi.org/10.1037/a0015479

Lucas, K., & Sherry, J. L. (2004). Sex Differences in Video Game Play:: A Communication-Based Explanation. *Communication Research*, *31*(5), 499-523. https://doi.org/10.1177/0093650204267930

Luna, F. G., Roca, J., Martín-Arévalo, E., & Lupiáñez, J. (2020). Measuring attention and vigilance in the laboratory vs. online: The split-half reliability of the ANTI-Vea. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01483-4

Madero, E. N., Anderson, J., Bott, N. T., Hall, A., Newton, D., Fuseya, N., Harrison, J. E., Myers, J. R., & Glenn, J. M. (2021). Environmental Distractions during Unsupervised Remote Digital Cognitive Assessment. *The Journal of Prevention of Alzheimer's Disease*, *8*(3), 263-266. https://doi.org/10.14283/jpad.2021.9

Mahy, C. E. V., Voigt, B., Ballhausen, N., Schnitzspahn, K., Ellis, J., & Kliegel, M. (2015). The impact of cognitive control on children's goal monitoring in a time-based prospective memory task. *Child Neuropsychology*, *21*(6), 823-839. https://doi.org/10.1080/09297049.2014.967202

Mitzner, T. L., Boron, J. B., Fausset, C. B., Adams, A. E., Charness, N., Czaja, S. J., Dijkstra, K., Fisk, A. D., Rogers, W. A., & Sharit, J. (2010). Older adults talk technology: technology usage and attitudes. *Computers in Human Behavior*, *26*(6), 1710-1721. https://doi.org/10.1016/j.chb.2010.06.020

Oksanen, K. M., Waldum, E. R., McDaniel, M. A., & Braver, T. S. (2014). Neural mechanisms of time-based prospective memory: evidence for transient monitoring. *Plos One*, *9*(3), e92123. https://doi.org/10.1371/journal.pone.0092123

Pew Research Center. (2019). *Internet/Broadband Fact Sheet*. Retrieved 04.01.2021 from https://www.pewresearch.org/internet/fact-sheet/internet-broadband/

Phan, M., Jardina, J., Hoyle, S., & Chaparro, B. (2012). Examining the Role of Gender in Video Game Usage, Preference, and Behavior. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*, 1496 - 1500.

Posner, M. I., & Mitchell, R. F. (1967). Chronometric analysis of classification. *Psychological Review*, *74*(5), 392–409. https://doi.org/10.1037/h0024913

Reips, U.-D. (2002). Standards for Internet-Based Experimenting. *Experimental Psychology*, *49*(4), 243-256. https://doi.org/10.1027//1618-3169.49.4.243

Rendell, P. G., Vella, M. J., Kliegel, M., & Terrett, G. (2009). Effect of delay on children's delay-execute prospective memory performance. *Cognitive Development*, *24*(2), 156-168. https://doi.org/10.1016/J.Cogdev.2008.12.002

Rummel, J., & Kvavilashvili, L. (2019). Take the field! Investigating prospective memory in naturalistic and real-life settings. In J. Rummel & M. A. McDaniel (Eds.), *Prospective Memory* (pp. 157-169). Routledge.

Scarampi, C., & Gilbert, S. J. (2020). The effect of recent reminder setting on subsequent strategy and performance in a prospective memory task. *Memory*, *28*(5), 677-691. https://doi.org/10.1080/09658211.2020.1764974

Schmidt, W. C. (1997). World-Wide Web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments, & Computers*, *29*(2), 274-279. https://doi.org/10.3758/BF03204826

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

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128-142. https://doi.org/10.3758/s13423-017-1230-y

Sengpiel, M., & Dittberner, D. (2008). The computer literacy scale (CLS) for older adults - development and validation. In M. Herczeg & M. C. Kindsmüller (Eds.), *Mensch & Computer 2008: Viel Mehr Interaktion* (pp. 7-17). Oldenbourg Verlag.

Sindi, S., Fiocco, A. J., Juster, R.-P., Pruessner, J., & Lupien, S. J. (2013). When we test, do we stress? Impact of the testing environment on cortisol secretion and memory performance in older adults. *Psychoneuroendocrinology*, *38*(8), 1388-1396. https://doi.org/https://doi.org/10.1016/j.psyneuen.2012.12.004

Skitka, L. J., & G., S. E. (2006). The Internet as psychological laboratory. *Annual Review of Psychology*, *57*(1), 529-555. https://doi.org/10.1146/annurev.psych.57.102904.190048

Souchay, C., Dubourg, L., Ballhausen, N., Schneider, M., Cerf, C., Schnitzspahn, K., Faivre, L., Kliegel, M., & Eliez, S. (2018). Time-based prospective memory in children and adolescents with 22q11.2 deletion syndrome. *Clinical Neuropsychologist*, *32*(5), 981-992. https://doi.org/10.1080/13854046.2017.1403652

Talbot, K. D. S., & Kerns, K. A. (2014). Event- and time-triggered remembering: The impact of attention deficit hyperactivity disorder on prospective memory performance in children. *Journal of Experimental Child Psychology*, *127*, 126-143. https://doi.org/10.1016/j.jecp.2014.02.011

The jamovi project. (2020). *jamovi (Version 1.2) [Computer Software]*. In Retrieved from https://www.jamovi.org

Thiébaut, E., & Richoux, V. (2005). Element of predictive validity to the NV5-R battery of tests scores. *Pratiques psychologiques*, *11*(4), 403-416. https://doi.org/10.1016/j.prps.2005.09.010

Trustram Eve, C., & de Jager, C. A. (2014). Piloting and validation of a novel self-administered online cognitive screening tool in normal older persons: the Cognitive Function Test. *Int J Geriatr Psychiatry*, *29*(2), 198-206. https://doi.org/10.1002/gps.3993

Voigt, B., Aberle, I., Schonfeld, J., & Kliegel, M. (2011). Time-based prospective memory in schoolchildren: The role of self-initiation and strategic time monitoring.

*Zeitschrift Fur Psychologie-Journal of Psychology*, *219*(2), 92-99. https://doi.org/10.1027/2151-2604/A000053

Voigt, B., Mahy, C. E. V., Ellis, J., Schnitzspahn, K., Krause, I., Altgassen, M., & Kliegel, M. (2014). The development of time-based prospective memory in childhood: The role of working memory updating. *Developmental Psychology*, *50*(10), 2393-2404. https://doi.org/10.1037/a0037491

Wechsler, D. (2008). *Wechsler adult intelligence scale - fourth edition*. Pearson Assessment.

Woods, S. P., Weinborn, M., Li, Y. Q. 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 and Cognition*, *22*(6), 679-692. https://doi.org/10.1080/13825585.2015.1027651

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

Zuber, S., Mahy, C. E. V., & Kliegel, M. (2019). How executive functions are associated with event-based and time-based prospective memory during childhood. *Cognitive Development*, *50*, 66-79. https://doi.org/https://doi.org/10.1016/j.cogdev.2019.03.001

**Table 1**

*Number of participants, age, percentage of women, education, and descriptive statistics of control variables per (~10-year) age group (for main sample).* [back to text]

| Age group | | Age | | Gender | Education | | Mill Hill | | Matrices | | Code | | CLS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *N* | *Mean* | *SD* | *women* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| 19 – 29 | 56 | 23.64 | 2.29 | 79% | 16.23 | 2.16 | 22.38 | 3.89 | 10.15 | 2.91 | 11.35 | 2.74 | .83 | .06 |
| 30 – 39 | 39 | 34.01 | 2.98 | 54% | 16.80 | 3.14 | 23.87 | 4.16 | 11.61 | 2.51 | 11.50 | 2.66 | .82 | .07 |
| 40 – 49 | 37 | 45.67 | 3.00 | 68% | 15.32 | 3.82 | 23.89 | 4.49 | 10.68 | 2.22 | 11.49 | 3.30 | .78 | .08 |
| 50 – 59 | 46 | 54.60 | 3.04 | 72% | 15.00 | 3.52 | 25.22 | 5.28 | 10.83 | 2.31 | 10.98 | 2.30 | .76 | .10 |
| 60 – 69 | 50 | 63.47 | 2.55 | 67% | 15.84 | 4.18 | 26.36 | 4.50 | 10.54 | 2.74 | 11.57 | 2.32 | .72 | .11 |
| 70+ | 24 | 76.56 | 4.55 | 58% | 14.38 | 3.64 | 25.79 | 5.02 | 10.92 | 2.70 | 12.25 | 1.80 | .62 | .15 |
| ANOVA | | | | | $p = .048$* | | $p < .001$*** | | $p = .186$ | | $p = .547$ | | $p < .001$*** | |

*Note.* Education = education in years; Mill Hill = raw score; Matrices = age-norm standardized scores; Code = age-norm standardized scores; CLS = Score on Computer Literacy Scale; * $p < .05$; *** $p < .001$.

**Table 2**

*Descriptive statistics, correlations, as well as Student and Bayesian paired-sample t-test comparisons of outcome measures on laboratory and online version of the Geneva Space Cruiser*. [back to text]

| Variable | Laboratory assessment | | | | Online assessment | | | | Correlations | Comparisons | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *min* | *max* | *Mean* | *SD* | *min* | *max* | *r* | *t* | $BF_{10}$ |
| OT performance | 8876.45 | 3449.41 | 280 | 18870 | 8170.00 | 3602.25 | 1140 | 18850 | .56*** | 2.80** | 3.50 |
| PM performance | 0.58 | 0.37 | 0 | 1 | 0.56 | 0.38 | 0 | 1 | .60*** | 0.27 | 0.08 |
| Monitoring count | 17.81 | 13.34 | 0 | 70 | 18.13 | 14.93 | 0 | 84 | .68*** | -0.54 | 0.09 |

*Note.* OT performance = score without points gained for refueling; PM performance = PM accuracy (number of refuels / number of possible refuels); Monitoring count = number of times participants displayed fuel gauge. *r* = Pearson correlations. *t* = Student t. $BF_{10}$ = Bayes factor (with $H_1$ assuming that a difference between laboratory and online testing). ** $p < .01$; *** $p < .001$.

**Table 3**

*Descriptive statistics and test-retest reliability of the Geneva Space Cruiser in a secondary sample of younger adults.* [back to text]

| Variable | Test | | Retest | | Reliability | |
|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | α | *r* |
| OT performance | 10503.35 | 2731.58 | 11768.94 | 2788.00 | .88 | .78*** |
| PM performance | .61 | .34 | .67 | .34 | .80 | .66*** |
| Monitoring count | 17.37 | 11.23 | 18.46 | 11.92 | .87 | .77*** |

*Note*. OT performance = score without points gained for refueling; PM performance = PM Accuracy (number of refuels / number of possible refuels); Monitoring count = number of times participants displayed fuel gauge. α = Cronbach's alpha reliability coefficients; *r* = Pearson correlations. *** $p < .001$

**Table 4**

*Hierarchical regression with age (step 1), and age and computer literacy (step 2) predicting PM performance in laboratory and online version of the Geneva Space Cruiser.* [back to text]

| Variable | Laboratory PM performance | | | | | Online PM performance | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | β | $t$ | $p$ | $R^2$ | $\Delta R^2$ | β | $t$ | $p$ | $R^2$ | $\Delta R^2$ |
| Age | -.54 | -9.43 | .000*** | .29 | .29*** | -.56 | -9.87 | .000*** | .31 | .31*** |
| Age | -.45 | -6.68 | .000*** | .31 | .02* | -.47 | -7.02 | .000*** | .32 | .01*** |
| CLS | .17 | 2.44 | .015* | | | .15 | 2.15 | .033* | | |

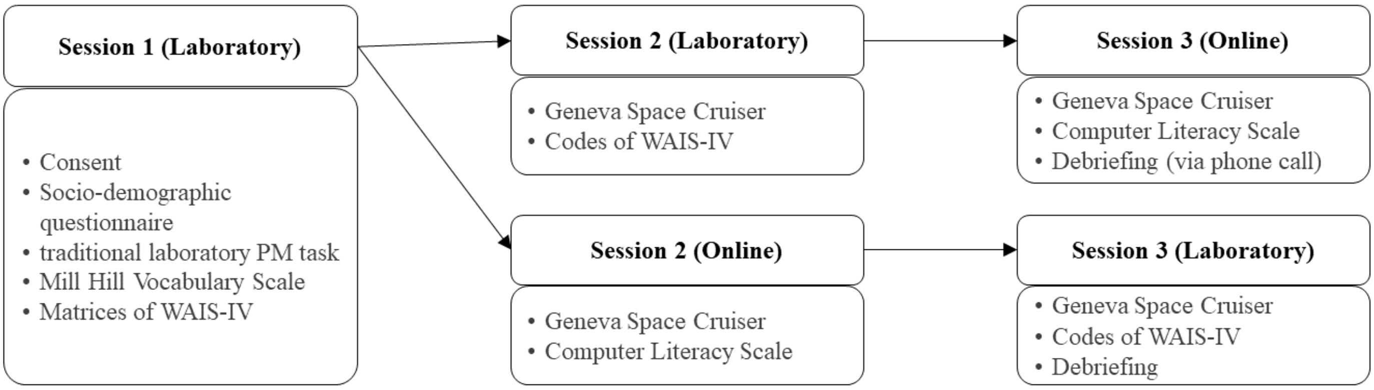
*Note.* CLS = Score on Computer Literacy Scale. * $p < .05$; *** $p < .001$.

**Figure 1**

In-game illustration of the Geneva Space Cruiser. The fuel gauge is displayed for three seconds after participants push the C-key. [back to text]

**Figure 2**

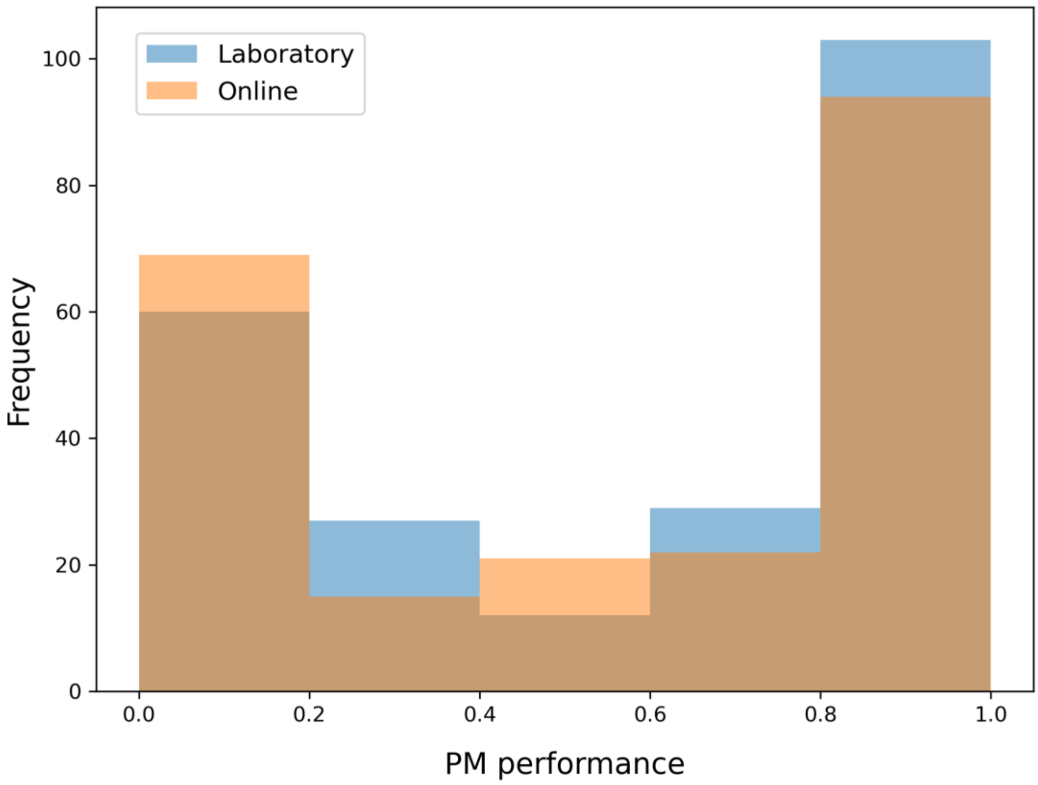
Overview of the assessment procedure. [back to text]




*Note.* All participants of the main sample first participated in a laboratory session (session 1). In counter-balanced order, half of the participants then participated in another laboratory (session 2) and finally in an online session (session 3), whereas the other half then participated in the online (session 2) and finally in another laboratory session (session 3).

**Figure 3**

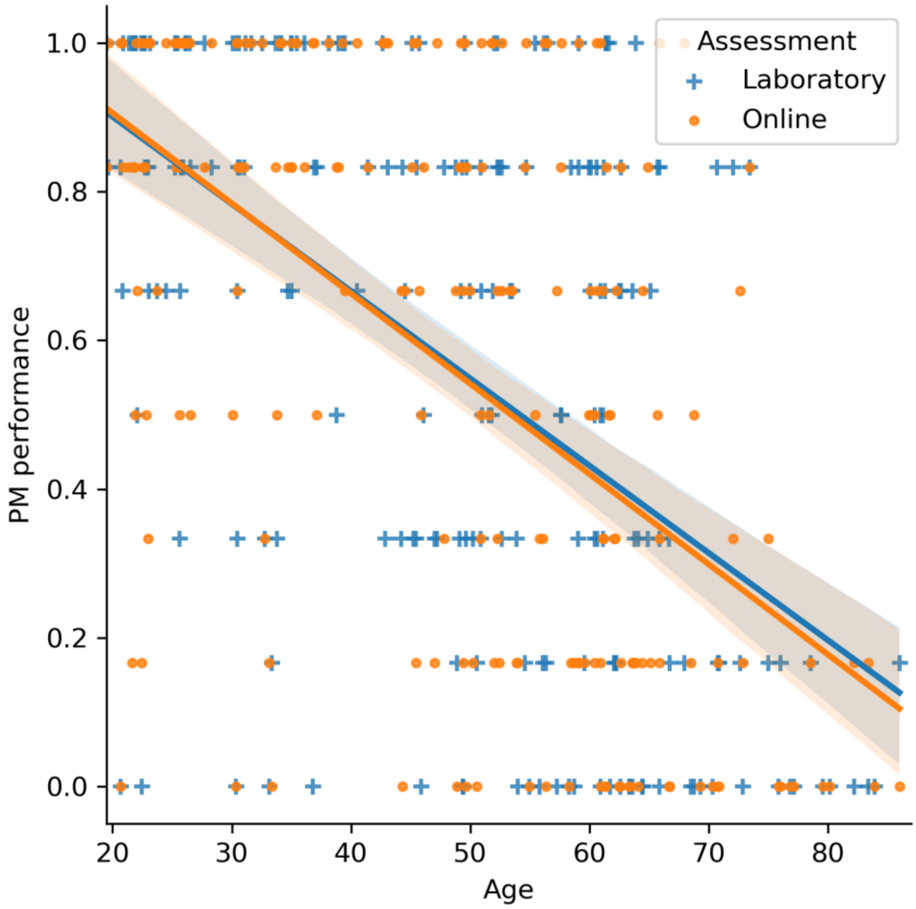
Overlapping histograms of frequencies of participants' PM performance on the laboratory and on the online assessment. [back to text]




*Note*. PM performance = PM Accuracy (number of refuels / number of possible refuels).

**Figure 4**

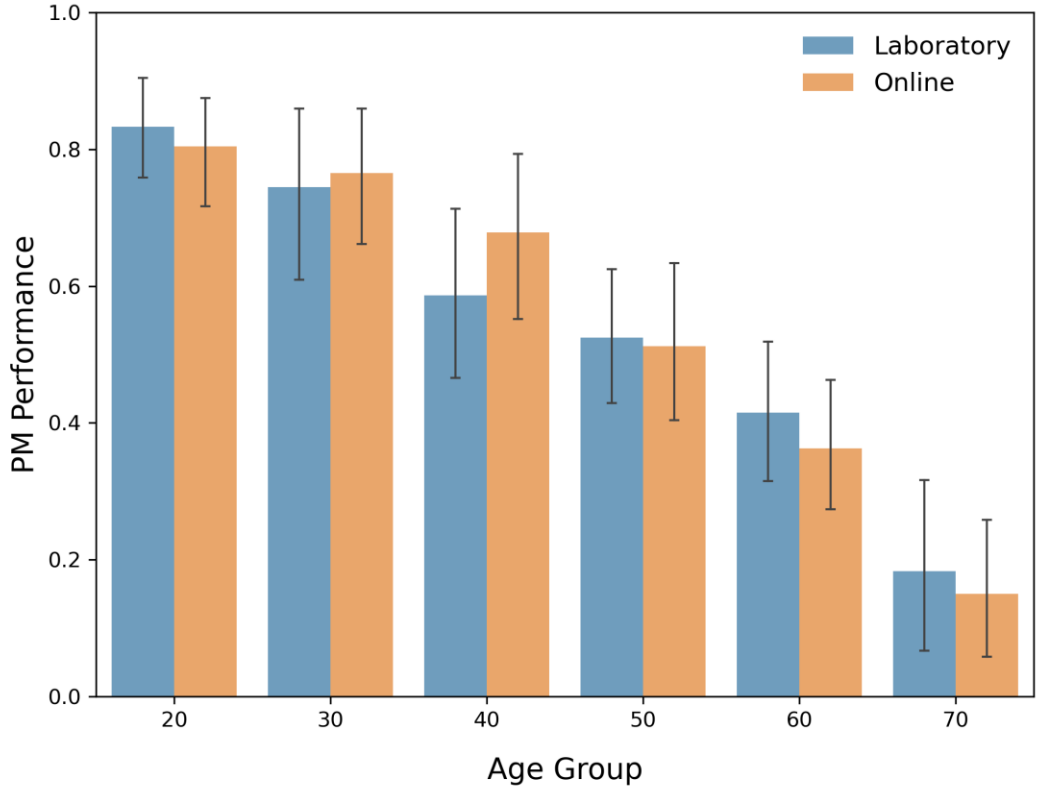
Overlapping regression plots with PM performance regressed on age. [back to text]




*Note*. PM performance = PM Accuracy (number of refuels / number of possible refuels).

**Figure 5**

Bar plots with PM accuracy and standard errors across the adult lifespan, separated per age group and assessment environment. [back to text]




*Note*. PM performance = PM Accuracy (number of refuels / number of possible refuels). Note that frequentist and Bayesian *t*-tests showed anecdotal to moderate evidence that there were no differences in PM performance between the two assessment environments for any of the age groups.