Journal of Clinical and Experimental Neuropsychology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ncen20

Inspecting the external world: Memory capacity, but not memory self-efficacy, predicts offloading in working memory

Sanne Böing, Antonia F. Ten Brink, Carla Ruis, Zoë A. Schielen, Esther Van den Berg, J. Matthijs Biesbroek, Tanja C. W. Nijboer & Stefan Van der Stigchel

To cite this article: Sanne Böing, Antonia F. Ten Brink, Carla Ruis, Zoë A. Schielen, Esther Van den Berg, J. Matthijs Biesbroek, Tanja C. W. Nijboer & Stefan Van der Stigchel (2024) Inspecting the external world: Memory capacity, but not memory self-efficacy, predicts offloading in working memory, Journal of Clinical and Experimental Neuropsychology, 46:10, 943-965, DOI: 10.1080/13803395.2024.2447263

To link to this article: https://doi.org/10.1080/13803395.2024.2447263

JOURNAL OF CLINICAL AND EXPERIMENTAL NEUROPSYCHOLOGY
2024, VOL. 46, NO. 10, 943–965
https://doi.org/10.1080/13803395.2024.2447263

Inspecting the external world: Memory capacity, but not memory self-efficacy, predicts offloading in working memory

Sanne Böing[a], Antonia F. Ten Brink[a], Carla Ruis[a,b], Zoë A. Schielen[a,c], Esther Van den Berg[d], J. Matthijs Biesbroek[b,e], Tanja C. W. Nijboer[a] and Stefan Van der Stigchel[a]

[a]Experimental Psychology, Helmholtz Institute, Utrecht University, Utrecht, The Netherlands; [b]Department of Neurology and Neurosurgery, University Medical Center Utrecht, Utrecht, The Netherlands; [c]Department of Geriatrics, University Medical Center Utrecht, Utrecht, The Netherlands; [d]Department of Neurology and Alzheimer Center Erasmus MC, Erasmus MC University Medical Center, Rotterdam, The Netherlands; [e]Department of Neurology, Diakonessenhuis Hospital, Utrecht, The Netherlands

ABSTRACT
Individuals with memory impairments may need to rely often on the external world (i.e. offloading). By memorizing only a fraction of the items at hand, and repeatedly looking back to the remainder of items (i.e. inspecting), they can avoid frailty or effortful memory use. However, individuals with subjective concerns may also prefer to rely on the external world even though their capacity is intact. Crucially, capacity assessment fails to recognize offloading strategies, while inspection assessment may reveal how people choose to deploy memory in everyday life. To disentangle the relative contributions of memory capacity and memory self-efficacy to offloading behavior, we recruited 29 individuals who were referred to a memory clinic and 38 age-matched individuals. We assessed memory capacity using neuropsychological measures, and memory self-efficacy using questionnaires. Inspection behavior was assessed in a copy task that allowed participants to store information to their preferred load or to rely on the external world. Referred individuals had lower capacity scores and lower memory self-efficacy. They inspected as often as controls, but used longer inspections and performed worse. Across all subjects, memory capacity – but not memory self-efficacy – explained inspection frequency and duration, with higher capacity associated with fewer and shorter inspections. Capacity measures thus translate to how people choose to deploy their memory in tasks that do not force full capacity use. However, people generally avoided remembering more than two items per inspection, and thus avoided using their full capacity. Inspection behavior was not further explained by memory self-efficacy, suggesting that inspections are not a sensitive measure of constraints experienced in everyday life. Although we provide support for the predictive value of capacity tasks in tasks with more degrees of freedom, capacity tasks overlook offloading behavior that individuals may employ to avoid using their full memory capacity in everyday life.

ARTICLE HISTORY
Received 9 July 2024
Accepted 21 December 2024

KEYWORDS
Offloading; sampling; working memory; Metamemory; neuropsychological assessment

1. Introduction

Memory complaints are common in the general aging population (Ponds et al., 1997). Although some degree of memory loss is inherent to getting older, subjectively experienced memory problems are a precursor to cognitive impairment and may be indicative of underlying pathology (Drag & Bieliauskas, 2010; Jessen et al., 2010; Saykin et al., 2006; Steinberg et al., 2013). Concerns about memory functioning may therefore warrant a referral to a memory clinic. To discriminate between intact, below average, and impaired memory capacity, the referred individual is asked to encode, maintain and report as much information as possible. The resulting score is used to construct a cognitive profile and subsequently to guide diagnosis. However, memory capacity scores that are obtained in a clinical setting could fall within the normal range even though the individual may report subjective memory complaints in daily life (Beaudoin & Desrichard, 2011). This discrepancy may be due in part to the fact that traditional tasks force people to use a particular strategy (i.e., to memorize as much as possible). Yet, when people can choose whether or not to load memory to maximum capacity, they are likely to minimize the internal cognitive effort involved in performing a task and rely on information from the external world (Burnett & Richmond, 2023; Gilbert, 2015; Meyerhoff et al., 2021; Risko & Dunn, 2015). In other words, they choose to use an offloading strategy (e.g., writing things down, creating cues as reminder (Ballard et al., 1995; Böing et al., 2023, subm.; Draschkow et al., 2021; Gray et al., 2006; Meyerhoff

CONTACT Sanne Böing s.boing@uu.nl Experimental Psychology, Helmholtz Institute, Utrecht University, Heidelberglaan 1, Utrecht 3584CS, The Netherlands

Supplemental data for this article can be accessed online at https://doi.org/10.1080/13803395.2024.2447263

et al., 2021; Morrison & Richmond, 2020; Risko & Dunn, 2015; Sahakian et al., 2023; Somai et al., 2020).; Offloading may not only minimize effort, but also support accurate task completion in healthy individuals (Burnett & Richmond, 2023; Gilbert, 2015; Gilbert et al., 2020, 2023). The use of external memory strategies is frequently reported among older adults (although there are mixed findings for clinical samples) and its usage is even found to increase with increasing age (Pizzonia & Suhr, 2022). Traditional capacity tasks disregard this element of choice in the employment of memory strategies, and, consequently, do not necessarily capture the actual use of memory in everyday life. This means that subtle deviations (e.g., increased reliance on external strategies) leading to the subjective experience of memory failure might go unnoticed in memory assessment. Therefore, rather than thinking of memory as a fixed capacity entity that is always fully utilized, we should consider *how* one uses their memory. In this study, we approach working memory as a system that people use differently depending on the accessibility of information, their maximum memory capacity, *and* their expectations of how their memory will function.

The act of memorizing information is highly dependent on the accessibility of information. This becomes apparent in (visual) working memory paradigms that allow individuals to choose how much information they internalize in working memory and how often they fall back onto external information. *Sampling* behavior – the act of (re)orienting to and (re)inspecting information-to-be-used from the environment once it becomes relevant – is used as an indicator of such reliance on the external world; sampling is shown to occur often when information is relatively easily accessible, and to decrease when it is more effortful to access external information (Ballard et al., 1995; Böing et al., 2023; Draschkow et al., 2021; Droll & Hayhoe, 2008; Melnik et al., 2018; Sahakian et al., 2023; Somai et al., 2020). Given that the visual environment is generally stable, this implies a strong preference for external sampling in activities of daily living (e.g., looking back and forth at a grocery list rather than learning it by heart). Moreover, this reliance on sampling from the external world is even stronger when it is difficult to memorize information; people with impaired memory adhere to sampling rather than using working memory, *even* when sampling becomes costly (Böing et al., 2023). The extent to which people rely on external sampling versus internal working memory storage thus appears to depend on the interplay between information accessibility and working memory capacity.

Although lower levels of working memory functioning are to some extent associated with increased reliance on the external world (Meyerhoff et al., 2021; Morrison & Richmond, 2020; Risko & Dunn, 2015), there is no robust linear relationship between memory capacity and sampling behavior (Böing et al., 2023). Even healthy individuals who are able to remember multiple items (i.e., capacity of two or more items) show frequent inspecting. They only memorize up to two items at a time when information remains accessible in the external world. In fact, *re*inspecting (more than 1 inspection per item) is often observed. This reinspecting behavior has recently been interpreted as an expression of strengthening memory traces before acting on them (e.g., reaching an action threshold; Sahakian et al., 2023). In other words, people may have some residual information in working memory, but are not confident enough to use it, and therefore decide to inspect again. Along this line, we hypothesize that (re)inspecting is a proxy for an individual's *belief* about their own memory functioning. Speculating reinspecting to be an act of reassurance about the accuracy of the representation of the information to be used (i.e., checking oneself), we expect that individuals with negative beliefs about their memory functioning or self-reportedmemory failures may engage in reinspecting behavior more often than individuals with more positive expectations about their memory functioning. Such beliefs can be captured by measures of memory self-efficacy,where low levels of memory self-efficacyindicate uncertainty or negative beliefs about memory functioning. Negative beliefs about memory functioning may be co-occurringwith impaired memory capacity, but may also exist in the absence of impaired memory capacity (Ponds & Jolles, 1996a). Further, the construct of memory self-efficacyis related, but not synonymous, to subjective cognitive decline. Subjective cognitive decline regards the perceived decline in memory function within a person over time, whereas memory self-efficacy refers to the subjective judgment of one's memory functioning at a certain point in time. Even though a person may perceive cognitive decline in their memory function over time, they may still consider their memory functioning at the later timepoint to be adequate (thus, having a sufficient level of memory self-efficacy). On the other hand, perceived cognitive decline may be experienced by the individual to such an extent that it lowers the level of memory self-efficacy.Moreover, the level of memory self-efficacywithin an individual may be low but stable over time, thus without subjective cognitive *decline*. In sum, experiencing perceived decline can, but does not necessarily, lower the level of memory self-efficacy,and memory self-efficacyis not necessarily congruent with objective functioning.

Low levels of memory self-efficacy may lead to greater reliance on the external world, even when this is not necessary given the objectively intact memory capacity. This over-reliance on the external world could, in turn, be experienced as a memory failure by the individual, strengthening the drive to obtain a referral to a memory clinic. Crucially, capacity assessment does not capture memory concerns, while subtle deviations in inspection behavior may be a result of both capacity limitations and underlying memory uncertainty. Assessment of inspection behavior may, therefore, bridge the gap between clinically objectifiable deficits and the subjective experience of memory decline or failure for which the individual is referred to the hospital, and can help to integrate the co-occurring effects of both objective and subjective aspects of memory functioning. Further, inspection behavior may serve as an objective measure of external memory strategies, called for by Pizzonia and Suhr (2022).

In an attempt to disentangle the relative contributions of information accessibility, memory capacity, and memory self-efficacy to reliance on the external world, we assessed inspection behavior of individuals with different levels of memory capacity (as determined by objective metrics in the verbal and visuospatial domain for short- and longer term maintenance) and different levels of self-reported memory complaints, on a copy task that either facilitated inspecting or encouraged memorizing by varying the availability of external information. To this aim, we recruited individuals who had been referred to a memory clinic, as well as age-matched individuals who had not been referred to a memory clinic. This allowed us to include individuals with a wide range of subjective memory complaints and objective memory capacity impairments, resulting in memory profiles with different combinations of subjective and objective performance. As age and the level of education are known to be associated with performance on memory tasks (Brockmole & Logie, 2013; Park et al., 2002), the non-referred group was matched to the referred group on these characteristics. We compared memory use in the two groups across two conditions that differed in the cost (low or high) of accessing information from the external world. Both referred and non-referred individuals were expected to reduce inspecting behavior when information was less readily available (Böing et al., 2023). In addition, the referred individuals were expected to rely more on the external world (due to higher effort to store information or higher levels of memory complaints; Hurt et al., 2012) than non-referred individuals, even when information would not be readily available (Böing et al., 2023). For both referred and non-referred individuals, lower levels of memory capacity and higher levels of subjective memory complaints were expected to predict increased inspection frequency. As depression has been found to be associated with decreased memory performance and subjective memory complaints (Johansson et al., 1997; Schmand et al., 1996; Turvey et al., 2000), we also explored this attribute as a potential (confounding) factor influencing inspection behavior. Assessing inspection behavior and its underlying attributes may be an elegant and much needed way to approximate memory use in daily life, and may serve as a starting point to increase our understanding of patients' objective, subjective and interactive memory functioning.

2. Materials and methods

2.1. Participants

Individuals referred for memory assessment were recruited via the outpatient memory clinics of the University Medical Centre Utrecht (UMCU), the Erasmus MC University Medical Center Rotterdam, and Diakonessenhuis Hospital. These clinics have different specializations, and the types of referrals vary accordingly. The memory outpatient clinic of the neurology department of the UMCU sees a heterogeneous group of adults of all ages who experience memory problems due to, for example, neurodegenerative diseases, traumatic brain injury, an as yet unknown cause, or as a result of psychological factors. The memory clinic of the gerontology department of the UMCU specifically focuses on older adults (>65 years). MCI and dementia are regularly diagnosed. The route of referral is similar for both clinics of the UMCU: individuals may have initiated a referral themselves or are referred by their general practitioner or by other clinicians within the hospital (e.g., endocrinology, nephrology) who suspect cognitive decline. The outpatient memory clinic at the Erasmus MC specializes in Alzheimer's disease but also diagnoses other types of dementia. At the memory clinic of the neurology department of the Diakonessenhuis Hospital, a heterogeneous group of adults of all ages are seen; patients are mainly referred by general practitioners and either a brief cognitive screening tool or extensive neuropsychological testing is used, depending on the differential diagnosis and complexity of the case. Note that despite a referral to any of the clinics, a medical diagnosis may not be made after assessment.

The eligibility of referred individuals was based on the judgment of a neuropsychologist and/or a multidisciplinary team within the outpatient memory clinic. To be eligible for participation, referred individuals had to either self-report memory complaints, have objective memory impairment based on neuropsychological assessment, or have memory impairment observed by a clinician. Referred individuals had to be between 18 and 85 years old, speak Dutch fluently, and be able to give consent. They were excluded if there was evidence of visuospatial neglect, deficits in visual perception, aphasia, or if motor impairments prevented the use of a computer mouse.

Partners or family members accompanying the referred person were actively approached to act as matched controls. In addition, age- and education-matched controls were recruited via various public and university platforms (e.g., social media, family members, university intranet, community centers).

We recruited two groups (i.e., referred to the memory clinic and matched controls) with the aim of having at least 25 participants in each group. These numbers were determined by considering previous studies that have tested sampling behavior, and a power analysis. The original trade-off effect on sampling versus storing has been observed in a group of only 7 participants (no mention of effect size; Ballard et al., 1995), which was replicated by Somai et al. (2020) in a group of 12 participants (only unstandardized β coefficients for linear mixed-effect models mentioned). As we expected greater variability in our target groups due to the heterogeneity of referral reasons and a wider age range, we aimed to recruit at least twice as many participants in each group. A previous study from our research group showed that this number was sufficient to detect differences in eye movement behavior between patients with Korsakoff syndrome and controls (detected effect sizes β in the range of 0.05–0.38; Böing et al., 2023).

All participants gave written informed consent prior to the start of the experiment. Participants were compensated for their participation with 7EU per hour paid in increments of 30 minutes, and received compensation for travel costs.

We included 29 referred individuals (see Supplementary Figure S1 for a flowchart) and 38 non-referred controls. With the current sample size, for a one-tailed non-parametric Wilcoxon-Mann-Whitney t-test ($\alpha = .05$) with a power of .8, we should be able to reliably detect effects of Cohen's $d = 0.63$ (Faul et al., 2009). Effects commonly reported in comparable paradigms are similarly large (Draschkow et al., 2021; Sahakian et al., 2023). Furthermore, the linear mixed-effects models we used have higher power than t-tests. Therefore, we were confident that our study would have a sufficient power.

The project was approved by the Faculty Ethics Review Board of the Faculty of Social and Behavioural Sciences at Utrecht University (protocol numbers 21–0076 and 21–0269). The protocol was conducted in accordance with the Declaration of Helsinki.

2.2. Procedure

Parts of the methods section are similar to those described in a previous study (Böing et al., 2023).

After participants agreed to participate, they received an online link to fill out questionnaires (see "Questionnaires" for a description) at home in the period 14 to 1 day(s) before their test session. Individuals that already completed the Hospital Anxiety and Depression Scale in the outpatient memory clinic were exempt from filling in this questionnaire online. Questionnaires were administered to characterize the referred and non-referred group, and to rule out depression as a potential confound.

At the university testing facility, the rest of the test protocol (see "Experimental computer tasks" and "Neuropsychological tasks" for a description) was administered in a single visit. The first and second session of the experiment were separated by a break of 10 to 20 minutes, and the total test duration was a maximum of 3 hours. All tasks that were administered were memory tasks. These memory tasks were included to get an idea of the memory capacity across the groups, but also with the aim of integrating them into a memory compound score (see Analysis) that takes into account memory capacity in both the verbal and visual domains for both short-term and longer term delays. Task administration in session 1 comprised (in this order): Location Learning Task – direct recall, Copy Task – first session, Location Learning Task – delayed recall, Digit Span WAIS IV, and if time allowed: a Fixation and Free viewing task (not taken into account in the current analysis). Task administration in session 2 comprised (in this order): Rey Auditory-Verbal Learning Task – direct recall, Copy Task – second session, Rey Auditory-Verbal Learning Task – delayed recall, Corsi Block Tapping Task, and if time allowed: Change Detection Task.

At the end of the test protocol, the Metamemory In Adulthood questionnaire was administered. This was the case only for a subset of participants as the questionnaire was added later to the test protocol. This questionnaire was added to get an extra measure on beliefs about one's memory function, and was used in the calculation of the subjective memory compound score (see Analysis). See Supplementary Table S1 for a schematic overview of the test procedure.

Before their visit, we checked whether individuals that were referred to the hospital had already performed some of the neuropsychological tasks as part of standard care. If this were the case, they were exempt from that task; previously reported scores on those tasks were used in order to prevent unnecessary workload and avoid potential practice effects. Practice effects can occur after short time intervals between testing sessions, and can last up to 7 years (Calamia et al., 2012). Therefore, we should be wary with "overtesting" people. As a rule of thumb, task administration in the hospital had to be within a period of six months before their visit to the study site to remain valid. The six month rule was based upon clinical practice where six months is believed to be long enough to have general task effects to wear off. The majority of referred individuals was scheduled within three months after their visit to the clinician. The period between assessment in the clinic and assessment in the research facility was sometimes less than three months, but never more than six months. It is improbable that substantial cognitive changes have occurred within this time frame. Almost all referred individuals were exempt from the Digit Span Test and Rey Auditory-Verbal Learning Test (see below) as these tasks are commonly used in both screening or extensive neuropsychological testing.

2.3. Measurements

2.3.1. Experimental computer tasks

2.3.1.1. Apparatus. Experimental tasks were run on a Windows 10 Enterprise computer with an Intel Core i7–4790 CPU and 16GB RAM, and displayed on a 27 inch LCD monitor at a resolution of 2560x1440 pixels at 100 hz. An EyeLink 1000 eye tracker (SR Research Ltd., Canada) was placed at the desktop to track the eyes at a sample rate of 1 kHz. Participants were seated with their heads in a chin-rest at ~67.5 cm from the monitor, and the lights were dimmed during administration of the experimental tasks. Eye-tracker calibration and validation were performed manually with a 9-point grid attempting to achieve a calibration error of less than 2 degrees of visual angle (dva).

2.3.1.2. Copy Task. Identical to our previous study (Böing et al., 2023), we adapted a Copy Task that was originally used in our research group (Somai et al., 2020) to better fit our participant population. The task aimed to provoke a strategy switch in relying on internal visual working memory versus inspecting information from the outside world. The experiment was programmed in Python 3.7 using the PyQt5 library (Riverbank Computing Limited, 2019) for visual presentation and mouse and keyboard interaction. PyGaze (Dalmaijer et al., 2014) was used to interact with the eye tracker.

A model puzzle consisting of 6 items in a 3 × 3 example grid was shown at the left-hand side of the screen (see Figure 1). At the right-hand side of the screen, a 3 × 3 empty grid was presented, with a 2 × 3 resource grid presented below. The resource grid only contained items that were needed to copy the model; no distractors were present. Items were adopted from Arnoult (Arnoult, 1956; Figure 1(a) and consisted of black geometrical shapes that could not easily be named to measure reliance on VWM instead of verbalization strategies (Somai et al., 2020).

The task consisted of two experimental conditions. In the baseline or "low-cost" condition, the example grid was visible throughout the trial (Figure 1(b)). In this way, the "cost" to gather information from the outside world was low. In the experimental "high-cost" condition, we raised the cost to inspect information from the external world by introducing a gaze-contingent waiting time: the example appeared after fixating the left side of the screen for a total of 2000 ms. During the waiting time an hourglass was presented (Figure 1(c)). If participants looked back to the right *during* the waiting interval, the delay-clock would pause, and would restart as soon as the eyes were redirected to the hourglass again, so that gaze-contingent waiting always was 2000 ms, and never more. Once the example became visible, it remained on screen until the participant would move their eyes toward the right side of the screen after which it would disappear.

Participants were instructed to rebuild the model puzzle as quickly and accurately as possible by dragging items from the resource grid to the empty grid using a computer mouse. Participants received direct feedback: if an item was placed incorrectly, the item disappeared and the background of the cell turned red for 700 ms, after which subjects could make another attempt. If the item was placed correctly, the background of the cell turned green for 700 ms and the item remained fixed. A trial ended after correct placement of six items, or when the time-limit of 42 seconds had passed. The time-limit of 42 seconds was based on the study of Somai et al. (2020) in which high-cost conditions with 200, 1500 and 3000 ms delays were used. The authors observed maximum completion times of 30 seconds for placing six items in either of

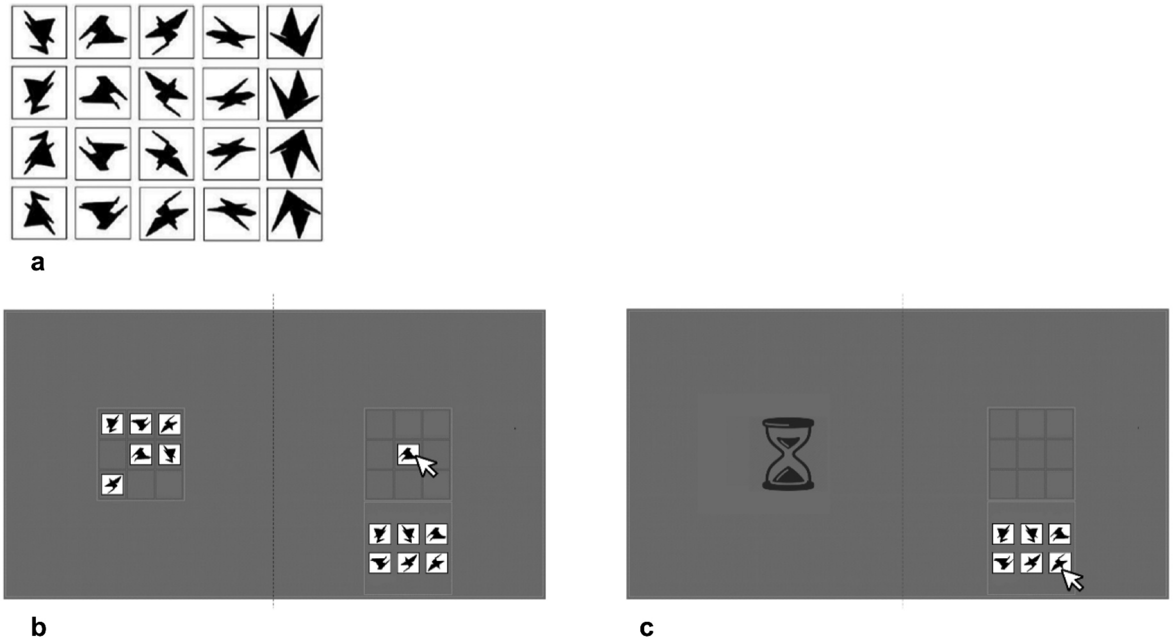

Figure 1. a) All possible stimuli in the Copy Task. Adopted from Arnoult (1956). An example trial is depicted for the low-cost condition (b) and high-cost condition (c) of the Copy Task. At the left-hand side of the screen, the example grid is either visible or replaced by an hourglass for 2000 ms (i.e., gaze-contingent occlusion). At the right-hand side of the screen, the empty grid to place the items (top) and the resource grid (bottom) are presented. A trial ended after 42 seconds. Note: the dotted midline is depicted for illustrative purposes and not visible in the experiment. The Copy Task layout is adopted and adjusted from Somai et al. (2020), and Böing et al. (2023).

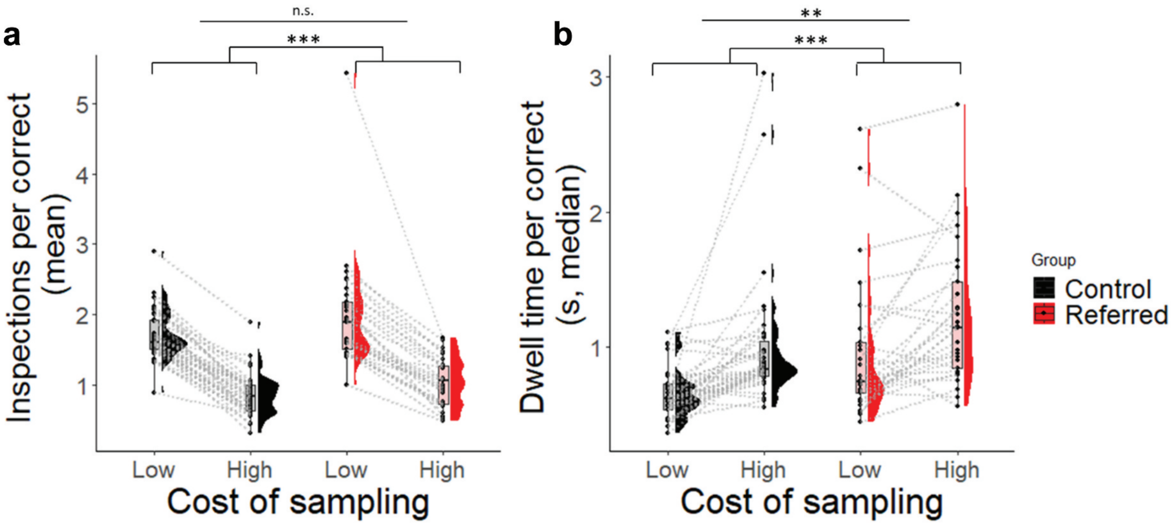

Figure 2. Eye-movement measures as indicator for inspection behaviour. (a) Mean number of inspections needed to make one correct placement, (b) Median dwell time in seconds per correct placement for non-referred controls (black) and referred individuals (red) across conditions (low-cost, high-cost). Black dots and grey lines represent outcomes of individual participants. Asterisks indicate significant effects. n.S = non-significant, *$p \le .05$, **$p \le .01$,*** $p \le .001$,.

the three variations. As we tested older adults and patients with potential cognitive decline, we anticipated our subjects to need more time. We therefore complemented the maximum observed completion time of Somai and colleagues by adding the gaze-contingent delay of 2000 ms for each item that had to be placed in the high-cost condition. In case someone would inspect once per item (which seems plausible from Somai et al., 2020), this would result in an additional 12 seconds. The choice to impose a time-limit at all was made because we wanted to have some control over the maximum task administration time, as we were bound to a larger protocol with limited testing time. After successful completion of a trial, positive feedback was shown (a thumbs up symbol). If subjects failed to correctly place all items within the time-limit, they were shown feedback that they ran out of time. By introducing the time-limit, we encouraged subjects to adopt a time-efficient strategy (Melnik et al., 2018). There was no specific incentive to increase the importance of accuracy as compared to speed or vice versa. Faster trial completion would yield faster task completion, serving as an incentive to increase working pace, but trying to be accurate also serves faster task completion, as making mistakes may also lead to increased completion times. This speed-accuracy trade-off is taken into account by the analysis of task performance (see "Performance measures" below).

We administered two sessions of the Copy Task, each session consisting of two blocks. First, three practice trials were performed in the low-cost condition to get familiar with the task. Calibration and validation of the eye-tracker were performed after the practice trials. Both sessions started with a low-cost block of 15 trials, followed by a high-cost block of 15 trials, resulting in a total of 30 trials per condition when combining data of the two sessions. This block design could have led to carry-over effects (Patrick et al., 2015), but we have deliberately chosen for this non-counterbalanced design a priori. We opted for this to make sure that our participants (especially older adults and/or cognitively impaired individuals) understood the basics of the task before being introduced to the more complex gaze-contingent high-cost condition.

A drift check (max. 2 dva) was performed before each trial, and recalibration was performed when deemed necessary. After each block, participants answered questions on their experience of commitment to and difficulty of the task (not considered in the current analysis). Each session of the Copy Task took 25 to 45 minutes, dependent on the calibration time, the participants' work pace, and the number and length of breaks.

Performance measures. We defined and calculated several outcome measures to describe between-group performance on the Copy Task (see Supplementary Materials). For between-group analysis, we calculated the *linear integrated speed-accuracy score* (*LISAS;* Vandierendonck, 2017) per individual per condition (low-cost, high-cost) as:

$$LISAS = RTij + PEijx\frac{SRT}{SPE}$$

where *RTij* (reaction time) denotes the trial *i net copying time* (completion time minus hourglass waiting time) divided by the number of correct placements for individual *j*. The reaction time data was log transformed to account for skewness associated with time measures. *PEij* refers to the proportion of errors on trial *i* and equals 1 minus the number of correct placements divided by the total attempts in that trial. *SRT*denotes the individual *j*'s overall net copying time standard deviation, and *SPE*is the individual *j*'s overall *PE* standard deviation. Standard deviations were calculated for individual *j* by collapsing all trials without split on condition (Vandierendonck, 2017). The *LISAS* was chosen as it combines two outcomes of performance (accuracy and speed) and weighs their importance equally. Lower LISAS reflects better (i.e., more accurate and faster) performance.

Eye-movement measures. We defined and calculated several outcome measures to describe between-group inspection behavior on the Copy Task (see Supplementary Materials). For the between-group analysis, the *number of inspections per correct placement* was chosen as it reflects eye movement inspection behavior regardless of overall performance (i.e., "per correct placement"). *Dwell time per correct* was analyzed as well.

2.3.1.3. *Change Detection Task.*

(see Supplementary Materials for an extensive description).

To assess visual working memory capacity in a traditional lab paradigm, we used a simplified version of the Change Detection Task from Luck and Vogel (Luck & Vogel, 1997; Oudman et al., 2020). Participants completed 80 trials in which they verbally reported whether or not they detected a change in the orientation of one bar amongst 2, 3, 4, or 6 bars before and after a white noise mask was presented. *D' (*dprime) was calculated as capacity outcome measure. *D'* is stated to yield a robust outcome for visual working memory performance that is less prone to biases in response tendency than, for example, Kmax (Williams et al., 2022).

$$d' = \mathrm{z}[\mathrm{p}(\mathrm{hits})] - \mathrm{z}[\mathrm{p}(\mathrm{falsealarms})]$$

2.3.2. Neuropsychological tasks

(see Supplementary Materials for extensive descriptions).

The neuropsychological tasks that are described below all have a similar task instruction: to memorize and report back as much as possible. Therefore, they are all grafted on obtaining a maximum capacity score. We transform (part of) these capacity scores into one memory compound score that takes into account both verbal and visual (working) memory performance (see 2.4.3. Objective memory capacity).

2.3.2.1. Location Learning Task.

To assess visuospatial immediate and long-term recall the standard stimulus set B of the modified Location Learning Task was used (Kessels et al., 2006, 2014). From this task displacement errors (sum of errors over five trials) can be calculated. A higher number of placement errors indicates worse memory performance (Kessels et al., 2014). Further, a learning index can be derived, and a delayed recall score can be obtained after prompting the individual to place as many items as possible after ~30 minutes. Only the displacement errors are used in the memory compound score, because these reflect short-term encoding success. Higher displacement error scores indicate worse performance. This score is reversed in pre-processing of the data to ensure that higher numbers reflect better performance. The delayed recall scores reflect longer-term retrieval processes, which are only of secondary interest in the current study.

2.3.2.2. Rey Auditory Verbal Learning Task.

To assess verbal immediate and long-term recall, the Rey Auditory Verbal Learning Task (15 items, Dutch version; Bouma et al., 2012; Saan & Deelman, 1986) was administered. The outcome measure used here is the total number of correctly recalled words over the course of five trials (range: 0–75). Higher scores reflect better memory function. A delayed recall score is obtained after prompting the individual to recall as many words as possible after ~25 minutes. Only the direct recall score is used in the memory compound score, because it reflects short-term encoding success. The delayed recall scores reflect longer-term retrieval processes, which are only of secondary interest in the current study.

2.3.2.3. Digit Span (WAIS-IV).

The Digit Span subtask forward and backward from the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV; Wechsler, 2012) were administered to assess short-term auditory memory and verbal working memory. The longest sequence that was correctly repeated was used as an outcome measure for maximum capacity (span range 2–8 or 2–9, for forward and backward respectively). As such, higher scores indicate better performance. Deviating from our protocol, some patients completed the Digit Span task from the WAIS-III as these were part of the standard administration in the hospital. WAIS-III has a different item score system than WAIS-IV, and therefore yields a different classification of scale scores. However, this has no implications for raw span scores, and therefore, the span scores obtained from the hospital could be used without conversion problems.

2.3.2.4. Corsi Block Tapping Task.

A digitized version (2D) of the Corsi Block Tapping Task was used to assess visuospatial working memory (Brunetti et al., 2014; Claessen et al., 2015; Corsi, 1972; Kessels et al., 2000). The forward subtask assesses short-term visuospatial attention; the backward subtask assesses VWM. To quantify maximum capacity, the span of the longest sequence that was correctly repeated was used (forward range 2–9, backward range 2–8). Higher scores indicate better performance.

2.3.3. Questionnaires

2.3.3.1. Memory complaints.

Participants were asked whether they experienced memory problems (yes/no). This answer was used to categorize participants with and without subjective memory problems. As this question is inclusive but fairly unspecific (Abdulrab & Heun, 2008), we included the Cognitive Failures Questionnaire and the Metamemory In Adulthood questionnaire to obtain a better idea about subjective memory experience.

2.3.3.2. Cognitive Failures Questionnaire.

As a measure of subjective cognitive functioning in the broader term, the Dutch Cognitive Failure Questionnaire is a 25-item questionnaire inquiring about the frequency with which participants experience small mistakes in daily life, on a 5-point scale, globally targeting attention and memory (Broadbent et al., 1982; Ponds et al., 2006), for example: “Do you find to forget whether you’ve turned off a light or a fire or locked the door?.” Items 2, 6, 16, 17, 18, 23, and 24 together make up for a subscale “absentmindedness” (Ponds et al., 2006) containing items about memory. We used this subscale as an outcome of self-reported memory *failure* occurrences; it was used in the memory self-efficacy compound.

2.3.3.3. Fatigue.

We used the 4-statement Dutch Verkorte Vermoeidheidsvragenlijst to assess experienced fatigue over the previous two weeks (Alberts et al., 1997; Bleijenberg et al., 2009). One of the statements is: “I feel tired.” On a 7-point scale, participants were asked to indicate to what extent the statement held true, where higher scores indicate more fatigue. One statement was rephrased ("I feel fit"), so that lower scores indicated more fatigue, and needed to be reversed in scoring. Total scores range from 4 to 28, and a score ≥ 18 indicates severe fatigue. These are reported as a group descriptive.

2.3.3.4. Hospital Anxiety and Depression Scale. The Dutch Hospital Anxiety and Depression Scale is a 14-item self-report questionnaire that is often administered in clinical care as a screener to assess complaints of anxiety (7 items) and depression (7 items), without focusing on physical complaints (Spinhoven et al., 1997). Scores can be interpreted per subdomain. Scores within the range of 0–7 to indicate no anxiety or depression, 8–10 to indicate possible anxiety or depression, and scores of 11–21 to indicate probable anxiety or depression (Jungen et al., 2019). Note that these results alone are not used to make a clinical diagnosis, but rather serve as an indicator of the presence of distress (Spinhoven et al., 1997). The depression scale is taken into consideration to account for the potential influence of depression on task performance.

2.3.3.5. Metamemory in Adulthood. The abridged version of the Dutch Metamemory in Adulthood questionnaire was adapted from Ponds and Jolles (1996b). It consists of 58 items that inquire about memory and attention, and an additional 16 items that ask about strategies people apply to support memory in daily life. Participants indicated the extent to which they agree with the statement on a 5-point scale. Several scale scores can be computed: Task, Capacity, Change, Anxiety, Achievement, Locus, External Strategies, and Internal Strategies. A memory self-efficacy score – the outcome of self-reported memory *functioning* – was derived from the Capacity, Change, and Anxiety subscale together. This score was used in the memory self-efficacy compound.

2.4. Pre-processing

2.4.1. Referral

We created a binary variable "Referral" to indicate whether or not the individual was referred, independently of the outcome of their assessment at the outpatient clinic.

2.4.2. Inspection behaviour

Saccades, fixations, and timestamps were extracted using the EyeLink 1000 parser (default EyeLink saccade detection algorithm, SR Research Ltd., Canada). Data pre-processing was implemented using Python 3.10. Every first trial in each block was removed from analysis: this trial served to check whether the instructions had been retained (additional instructions were given when needed) and to habituate the participant to the new situation (e.g., from low-cost to high-cost). If additional instructions (on mouse use, task, posture) were provided, the trial number was logged and the invalid trial was excluded from analysis. This was the case for 43 of the trials in the group of referred individuals and for 14 of the trials in the group of non-referred controls (see 3.2.1. Data Loss). Variables were calculated as described in "Measurements." Data analyses were conducted using R 4.1.2 (R Core Team, 2017).

2.4.3. Objective memory capacity

To get an estimate of objective memory capacity across groups, raw capacity scores (x) were transformed to z-scores for separate tasks. To bundle these into a single memory domain compound score, we averaged the z-scores of the separate capacity tasks to get a single value for general memory performance. Z was denoted as:

$$z = \frac{x - \mu}{\sigma}$$

where x denotes the raw score of the individual, μ refers to the mean for the complete sample, and σ refers to the standard deviation within the complete group. The "general memory" z-compound is then calculated by summing all available memory z-scores for the individual and dividing it by the number of tasks administered. The scores used in calculation of the objective memory compound score are: Rey Auditory Verbal Learning Task – direct recall (over five trials), Location Learning Task – displacement errors (over five trials), Digit Span forward span, Digit Span backward span, Corsi Block Tapping Task forward span, Corsi Block Tapping Task backward span, and dprime. Note that the delayed recall scores are not taken into account in the compound score. We decided not to do this, as we could not assure that the delay period was equally long for all the participants; the Copy Task often took too long, and the delayed recall may have only taken place after 45 minutes, which is almost twice the time window that is used in clinical care and valid interpretation of the score. We therefore consider the delayed recall scores for the Rey Auditory Verbal Learning Task and Location Learning Task with a grain of salt, but descriptively report them nonetheless.

2.4.4. Memory self-efficacy

The subscale “absentmindedness” of the Cognitive Failure Questionnaire was taken as an outcome of subjective memory failure. Further, a memory self-efficacy score can be derived from the Capacity, Change and Anxiety subscale of the Metamemory In Adulthood questionnaire. This memory self-efficacy score and the subscale Absentmindedness were transformed into a memory self-efficacy compound *z*-score. The Metamemory In Adulthood questionnaire was added to the protocol later (as a result of advancing insights), so we only have this data for a smaller part of the participants ($n = 18$ for referred individuals, $n = 15$ for non-referred controls).

2.5. Data analyses

2.5.1. Group characteristics

To assure similarity between groups in terms of age and education, Mann-Whitney U tests were performed. A chi-squared test was performed to compare sex distributions between groups. Scores on neuropsychological tasks and questionnaires were reported to characterize groups, and chi-squared tests and proportion z-tests were performed to test group differences.

2.5.2. Inspection strategies and performance across groups

For descriptive purposes, we reported inspection behavior and performance across the referred and non-referred group. For each individual, we aggregated outcome measures by the mean over trials per condition (low-cost, high-cost), except for time-based outcome measures which were aggregated by the median. Group scores (i.e., medians) were then calculated from these individual values.

To assess group differences in inspection behavior and performance, we ran linear mixed-effects models (LMM; Singmann & Kellen, 2019) on either of the outcomes across both conditions. Factors included in the LMM were Referral, Condition, Referral*Condition, and random intercept and slope to control for individual differences. After fitting the model, the significance of factors was judged using an alpha of 0.05. The normality of the residuals was visually examined and confirmed for every linear mixed-effects model. Effect sizes were reported as standardized beta-coefficients (β) with a 95% confidence interval. LMM were chosen over mere ANOVAs because of their robustness against deviations from normality of the outcome variables, and because they control for missing data and individual differences (Schielzeth et al., 2020).

Initially, datasets of all participants were analyzed without the removal of outliers. To rule out the possibility that our findings were solely driven by outliers, we removed participants whose aggregated scores were ≥ 1.5 times the interquartile range apart from the total group median for that specific outcome measure in that specific condition (low-cost or high-cost). When identified as outlier in either of the conditions, data of this participant were removed from both conditions. After outlier exclusion, the analyses were run again. Information on the effect of outliers is mentioned in the section of the respective analyses.

2.5.3. Inspecting behaviour based on memory capacity and memory self-efficacy

To investigate the effects of objective and subjective memory functioning on inspection behavior and related performance, we included the memory capacity compound z-score and memory self-efficacy compound z-score in regression models to predict the *number of inspections per correct placement* and *dwell time per correct placement* (as measures of inspection behavior), and *LISAS* (as measure of performance). Condition (low-cost, high-cost) was also included as explanatory factor in the model. Age, level of education, and depression score were included as covariates in all models. The significance of factors was judged using an alpha of 0.05.

2.5.4. Predictive value of memory capacity subtasks

To explore the predictive value of memory capacity in verbal or visual working and/or long term memory on inspection behavior and related performance, we ran (non-parametric) regression models to predict *LISAS*, and *number of crossings per correct placement* and *dwell time per correct placement* in both conditions as a function of *z*-transformed memory capacity task scores, with covariates age and education. Each of the capacity tasks was included in a separate regression model. The significance of factors was judged using an alpha of 0.05. Results are reported in the Supplementary Material.

2.6. Code and software

Experiment code, raw and pre-processed eye movement data, raw scores on neuropsychological assessment, and analysis scripts are publicly available and can be found at Open Science Framework: https://osf.io/ys67b/

3. Results

3.1. Group characteristics

We approached 66 referred individuals through the outpatient memory clinics. Thirty-seven were interested in participation and were invited to the testing facility. Six of these canceled their appointment without wanting to reschedule, and two test sessions were prematurely ended because the participant was not able to complete the copy task. Eventually, we were able to obtain a valid dataset (with copy task completion being the lead criterium) of 29 referred individuals (see Table 1 for demographic characteristics; see Supplementary Figure S3 for a patient flow chart; see Supplementary Figure S4 for information on suspected neurological etiology). All individuals were without known visual field defects and had normal or corrected-to-normal visual acuity.

Forty-eight non-referred individuals were recruited as control group. Four dropped out, four were not tested on the copy task due to technical problems, and one participant did not meet our inclusion criteria. For one of the participants, we were unable to track the eyes. Eventually, we obtained a valid dataset (with copy task completion being the lead criterium) of 38 non-referred individuals (see Table 1 for demographic characteristics and see Supplementary Figure S5 for a control flow chart).

Group characteristics, scores on neuropsychological assessment and questionnaires, and statistical comparisons between groups are displayed in Table 1. Note that the level of education is characterized according to the classification of Verhage (1964, 1965), that is commonly used in Dutch clinical care, and classifies the level of education (ranging from 1 to 7) based on the number of education years.

3.2. Inspection strategies and performance across groups

3.2.1. Data loss

Across 29 participants in the referred group, 1740 trials were planned to be collected. All first trials of each block were removed to assure task comprehension (116 trials). Fourteen trials were lost due to technical issues. Any reason that could possibly interfere with performance (excessive movement of the participant, forgetting the task instructions, problems controlling the mouse) was logged, and the corresponding trials (43 trials) were removed from further analysis. These included trials in which the eye-tracker lost signal. Despite the implementation of a drift check, some trials were started with a drift check above the 2 degrees visual angle threshold. When exceeding 5 degrees visual angle, trials were excluded (14 trials). Finally, 1553 trials were left for analysis.

Across 38 participants in the non-referred group, 2280 trials were planned to be collected. Again, all first trials of each block were removed (152 trials). In the non-referred group, one participant did not complete the second session of the Copy Task (minus 28 trials). Trials that were invalid due to signal loss, excessive movement of the participant, forgetting the task instructions, or problems controlling the mouse were removed (14 trials). For one participant, we were urged to exclude the entire second session because the majority of trials exceeded the drift check threshold. In total, 45 trials needed to be excluded because of exceeding the drift check threshold. In this group, 2041 trials were left for analysis.

3.2.2. Descriptive values

Group scores for inspection behavior and performance across conditions (low-cost and high-cost) were calculated and reported in Table 2. We confirmed that there was no differential effect (interaction) of session number across groups on our outcome measures of interest (in bold, Table 2) to ensure that pooling the conditions across sessions was a valid practice and outcome measures would not be confounded by differences in session effects between groups.

3.2.3. Inspection behaviour analysis

A linear mixed-effect model was fit to predict the *number of inspections per correct placement* by referral (referred, non-referred) and condition (low-cost, high-cost). A main effect of condition was found ($t = -11.178$, $p < .001$, $\beta = -0.87$ [−1.02, −0.72]), with more model inspections in the low-cost condition as compared to the high-cost condition. No effect of referral was found ($p = .069$), nor was there an interaction effect ($p = .5$): referred individuals inspected the model just as often as non-referred controls to place one item correctly.

The same factors were included in a model with *dwell time per correct placements* as dependent variable. Again, a main effect of condition was found ($t = 4.15$, $p < .001$, $\beta = 0.38$ [0.20, 0.56]), showing that inspection durations of the model increased in the high-cost condition. Here, a main effect of referral was found ($t = 2.87$, $p = .005$, $\beta = 0.35$ [0.11, 0.59]), showing that referred individuals took more time to inspect the model for one correct placement compared to non-referred controls. No interaction effect between group and condition was found ($p = .8$). Figure 2 visualizes findings for the inspection behaviour analysis.

Table 1. Demographic characteristics, scores on memory capacity tasks, and questionnaires, split per group (i.e., individuals referred or not referred to a memory clinic), medians (IQR) or frequencies are depicted.

	Individuals referred to memory clinic			Non-referred matched controls			
	n	n (%)/ Mdn (IQR)	Range	*n*	n (%)/ Mdn (IQR)	Range	Test Statistic ^{a}
Demographics							
Sex, male	29	17 (58.6%)		38	15 (39.5%)		χ^2 =1.710, *p*=.191, *d*=0.324
Age in years	29	67 (10)	37–80	38	60 (11.8)	40–81	*U*=404, *p*=.063, *r*=-0.27
Level of education	29	6 (1)	4–7	38	6 (1.75)	4–7	*U*=589, *p*=.621, *r*=0.07
Suspected neurological etiology		17 (58.6%)					
yes		11 (37.9%)					
no		1 (3.4%)					
ambiguous diagnosis							
Do you experience memory problems?		26 (89.7%)			9 (23.7%)		χ^2= 26.108, *p* <.001**
yes							
Fatigue, % severe fatigue		11 (37.9%)			6 (15.8%)		χ^2=3.169, *p*=.076
Anxiety		19 (65.5%)			31 (81.6%)		
Not present (score 0–7)		5 (17.2%)			7 (18.4%)		
Potential (score 8–10)		5 (17.2%)			0 (0%)		
Likely (score ≥11)							
Depression		24 (82.8%)			36 (94.7%)		
Not present (score 0–7)		4 (13.8%)			2 (5.3%)		
Potential (score 8–10)		1 (3.4%)			0 (0%)		
Likely (score ≥11)							
HADS Total score		10 (9)	2–22		5 (7.75)	0–19	*U* =280, *p* <.001**, *r*=-0.49
Neuropsychological task scores							
Location learning task	29			38			
Total displacement score^b		49 (35)	5–150		27.5 (26)	0–75	*U*=262, *p* <.001***, *r*=-0.53
Learning index (0–1)		0.287 (0.209)	0.054–1		0.523 (0.419)	0.101–1	*U*=772, *p*=.005**, *r*=0.40
Delayed recall: Placement errors	28	3.5 (10.5)	0–39	37	1 (4)	0–19	*U*=312, *p*=.005**, *r*=-0.40
Rey auditory–verbal learning task	28			38			
Immediate recall: Total correct (0–75)^b		36.5 (15.2)	13–51		42 (18)	30–67	*U*=778, *p* <.001***, *r*=0.52
Delayed recall: Total correct (0–15)		4.5 (5)	0–13		8 (6)	3–14	*U*=796, *p* <.001***, *r*=0.54
Digit span (WAIS-III/IV)	29			38			
Forward span (2–9)^b		5 (1)	4–7		6 (2)	4–9	*U*=777, *p*=.002**, *r*= 0.41
Backward span (2–8)^b		4 (1)	2–7		5 (2)	2–8	*U*=708, *p*=.041*, *r*=0.28
Corsi block-tapping task	29						
Forward span (2–9)^b		5 (2)	2–7	38	5 (1)	3–8	*U*=642, *p*=.229, *r*=0.16
Backward span (2–8)^b		5 (2)	2–7	37	6 (1)	2–7	*U*=626, *p*=.234, *r*=0.17
Change detection paradigm	23			36			
d′^{b}		1.79 (1.17)	0.817–3.36		2.23 (0.985)	0.246–3.8	*U*=498, *p*=.197, *r*=0.20
Impairment within memory domain^c							
impaired		5 (17.25%)			2 (5.3%)		
below average		15 (51.7%)			5 (13.2%)		
within normal range		9 (31.05%)			31 (81.5%)		
Memory capacity compound, *z*	29	−0.358 (1.07)	−1.53–0.69	38	0.242 (0.789)	−0.88–1.64	*U*=846, *p* <.001***, *r*=0.54
Memory questionnaires							
Cognitive Failure Questionnaire	29			38			
Total score (max. 125)		63 (21)	38–93		54.5 (13.5)	25–109	*U*=344, *p*=.009*, *r*=-0.38
Scale absentmindedness (max. 35) ^{e}		17 (6)	10–25		14 (4.75)	7–30	*U*=308, *p*=.002**, *r*=-0.44
Metamemory In Adulthood	18			15			
Scale Anxiety (lower is better) ^{d}		3.33 (0.58)	2.17–3.83		2.67 (1.04)	1.25–4	*U*=69.5, *p*=.019*, *r*=-0.49
Scale Capacity (higher is better)		2.54 (0.7)	1.75–3.67		3.25 (0.58)	2.58–4.5	*U*=226, *p*=.001***, *r*=0.67
Scale Change (higher is better)		2.35 (0.65)	1.5–3.2		3.2 (0.65)	2.11–4.3	*U*=230, *p* <.001***, *r*=0.70
Sum score Memory Self-Efficacy ^{e}		2.62 (0.7)	2.11–3.32		3.27 (0.57)	2.26–4.42	*U*=228, *p* <.001***, *r*=0.69
Memory self-efficacy compound, *z*	29	−0.281 (1.11)	−1.55–1.17	38	0.253 (0.891)	−2.98–1.79	*U*=804, *p*=.001***, *r*=0.46

n = sample size, Mdn = median, IQR = interquartile range, range (min.–max.). Sample size may differ per outcome variable. ^{a} Non-parametric test statistics indicating group differences and effect sizes: chi-squared, p-value, and d for binomial variables, or Mann – Whitney – Wilcoxon U, p-value, and rank-biserial correlation r for continuous data. ^{b} Capacity scores used in memory capacity compound z-score; Location Learning Task displacement errors are reversed so that higher scores indicate better performance on all capacity tasks. ^{c} Impaired: a score < 2nd percentile on ≥ 2 sub tasks^b (without d′); Below average: a score < 2nd percentile on 1 sub task^b (without d′) and/or a score between 2nd – 9th percentile on ≥ 2 sub tasks^b (without d′); Within normal range: does not fit criteria for impairment or below average. ^{d} Anxiety scale is reversed in calculation of the Memory Self-Efficacy sum score, so that higher scores indicate better subjective memory experience. ^{e} Scores used in calculation of memory self-efficacy compound z-score; higher scores indicate better subjective memory experience. * $p \leq 0.05$, ** $p \leq 0.005$, *** $p \leq 0.001$.

3.2.3.1. Sensitivity analyses. After outlier removal, the effect of condition was still present for both the number of inspections as well as the dwell time per correct placement (both $p < .001$). The effect of referral on the number of inspections remained insignificant for the number of inspections (p = .06), and vanished for dwell time per correctly placed item (p = .09).

Table 2. Group scores (referred, non-referred) for outcomes of performance and inspection behavior across conditions (low-cost and high-cost). Variables in bold are used in subsequent analyses.

	Individuals referred to memory clinic			Non-referred matched controls		
Copy Task Scores	*n*	Mdn (IQR)	Range	*n*	Mdn (IQR)	Range
Completion time, s	29			38		
Low-cost[a]		21.5 (11)	13.3–42		18.5 (5.95)	12.1–33.4
High-cost		38.2 (6.42)	30.8–42		33.8 (10)	24–42
Net copying time, s	29			38		
Low-cost [a]		21.5 (11)	13.3–42		18.5 (5.95)	12.1–33.4
High-cost		28 (5.45)	22.4–32.33		23.8 (5)	18.5–36
Correct placements (0–6)	29			38		
Low-cost		6 (0.154)	4.39–6		6 (0.027)	5.57–6
High-cost		5.46 (1.52)	3.07–6		5.83 (0.567)	2.23–6
Success rate (0–1)	27			36		
Low-cost		0.973 (0.031)	0.779–1		0.97 (0.036)	0.844–1
High-cost		0.874 (0.058)	0.738–0.981		0.91 (.0115)	0.497–0.98
Speed score, s	29			38		
Low-cost		3.7 (2.18)	2.35–11.3		3.25 (1.18)	2.09–6.73
High-cost		5.24 (3.16)	3.75–11.3		4.16 (1.23)	3.28–15.2
LISAS				36		
Low-cost	26	1.38 (0.63)	0.94–2.52		1.22 (0.43)	0.77–2.14
High-cost	27	2.03 (0.54)	1.49–2.75		1.67 (0.42)	1.26–3.71
Number of crossings	29			38		
Low-cost		9.89 (3.29)	5.96–18.6		9.62 (2.52)	5.29–14.1
High-cost		4.73 (1.16)	2.93–6.18		4.15 (1.87)	1.82–6.96
Dwell time per crossing, s	29			38		
Low-cost		0.465 (0.137)	0.27–1.058		0.39 (0.114)	0.262–0.564
High-cost		1.172 (0.661)	0.557–3.729		1.13 (0.79)	0.562–5.616
Number of inspections per correct placement	29			38		
Low-cost		1.88 (0.673)	0.994–5.44		1.61 (0.42)	0.881–2.9
High-cost		1.05 (0.546)	0.488–1.66		0.84 (0.374)	0.304–1.9
Dwell time per correct placement, s	29			38		
Low-cost		0.742 (0.376)	0.444–2.61		0.62 (0.184)	0.36–1.11
High-cost		1.14 (0.638)	0.558–2.8		0.84 (0.255)	0.55–3.03

Valid datasets *n*, median Mdn, interquartile range (IQR), and range (min.–max.). [a] Completion time and net copying time in the low-cost condition are the same. In the high-cost condition, the net copying time is the completion time minus the hourglass waiting time.

3.2.4. *Performance analysis*

A linear mixed-effect model was fit to analyze the influence of referral (referred, non-referred) and condition (low-cost, high-cost) on speed-accuracy performance (*LISAS*). The model yielded a main effect of condition ($t = 9.19$, $p < .001$, $\beta = 0.53$ [0.41, 0.64]) with performance decreasing with high-cost inspecting as compared to low-cost inspecting. A main effect of referral ($t = 2.66$, $p = .01$, $\beta = 0.26$ [0.07, 0.46]) was present, indicating that referred individuals performed worse than non-referred individuals. No interaction between group and condition was found ($p = .7$). Figure 3 visualizes findings for the performance analysis.

3.2.4.1. Sensitivity analysis. When running the same models after outlier removal, the effects of condition and referral persisted ($p < .001$, $p < .005$, respectively), and again, no interaction was found ($p = .35$).

3.3. *Inspection behaviour based on memory capacity and memory self-efficacy*

Referred and non-referred individuals did not differ in the number of inspections per correct placement, and the difference between groups for the dwell time needed to place one item correctly was mostly driven by outliers. This indicates that the distinction between groups is not so clear-cut, which can be attributed to the fact there is large overlap between groups (see Supplementary Figure S6). Some referred individuals showed no objective memory impairments, while some non-referred individuals did show objective memory impairments. Although the groups statistically differed on both objective memory capacity and memory self-efficacy (see Table 1), referral as a sole factor appears not to be sensitive enough to explain inspection behavior. To investigate the effects of objective memory functioning and memory self-efficacy on inspection behavior and related performance, we tested the predictive value of the memory capacity compound and the memory self-efficacy compound. As there were no strong indicators of severe depression in our sample (see Table 1), we decided to exclude this covariate from the model to reach higher power. The level of education and age were included as covariate.

The number of inspections per correct placement was predicted by condition ($t = -12.8$, $p < .001$, $\beta = -0.9$), and the memory capacity compound score ($t = -4.56$, $p < .001$, $\beta = -0.36$), but no effect of memory self-

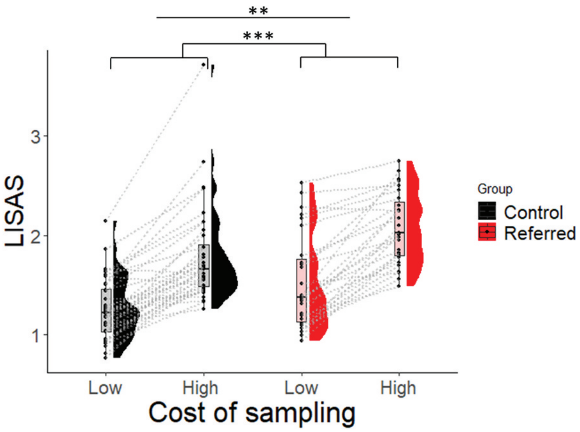

Figure 3. Performance outcome expressed in LISAS for non-referred controls (black) and referred individuals (red) across conditions (low-cost, high-cost). Black dots and grey lines represent outcomes of individual participants. Lower LISAS indicates better performance. Asterisks indicate significant effects. *$p \leq .05$, **$p \leq .01$,*** $p \leq .001$.

efficacy was found ($p = .28$). Interaction effects were absent (all $p > .3$). Dwell time per correct placement was influenced by condition ($t = 4.34$, $p < .001$, $\beta = 0.32$), and memory capacity ($t = -3.83$, $p < .001$, $\beta = -0.32$). Again, no effect of the subjective component was found ($p = .74$), and no interaction effects were apparent (all $p > .05$). Finally, LISAS was significantly predicted by condition ($t = 8.6$, $p < .001$, $\beta = 0.54$), memory capacity ($t = -5.12$, $p < .001$, $\beta = -0.36$), but not by memory self-efficacy ($p = .67$). The effects of condition and memory capacity held under nonparametric tests. Figures 4, 5 and 6 visualize the observed effects per outcome variable. Covariates are not taken into account in these figures.

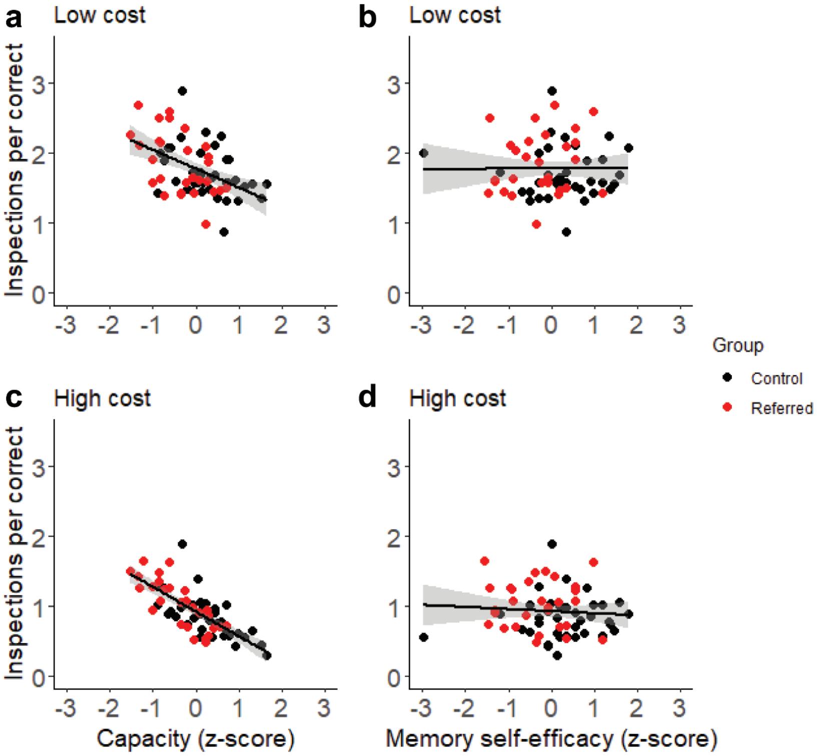

Figure 4. The relation between the number of inspections per correct placement and the capacity compound z-score, and memory self-efficacy compound z-score. Non-referred controls are depicted in black, individuals referred to the outpatient memory clinic are depicted in red. A smoothed linear coefficient is added in black with confidence intervals in grey.

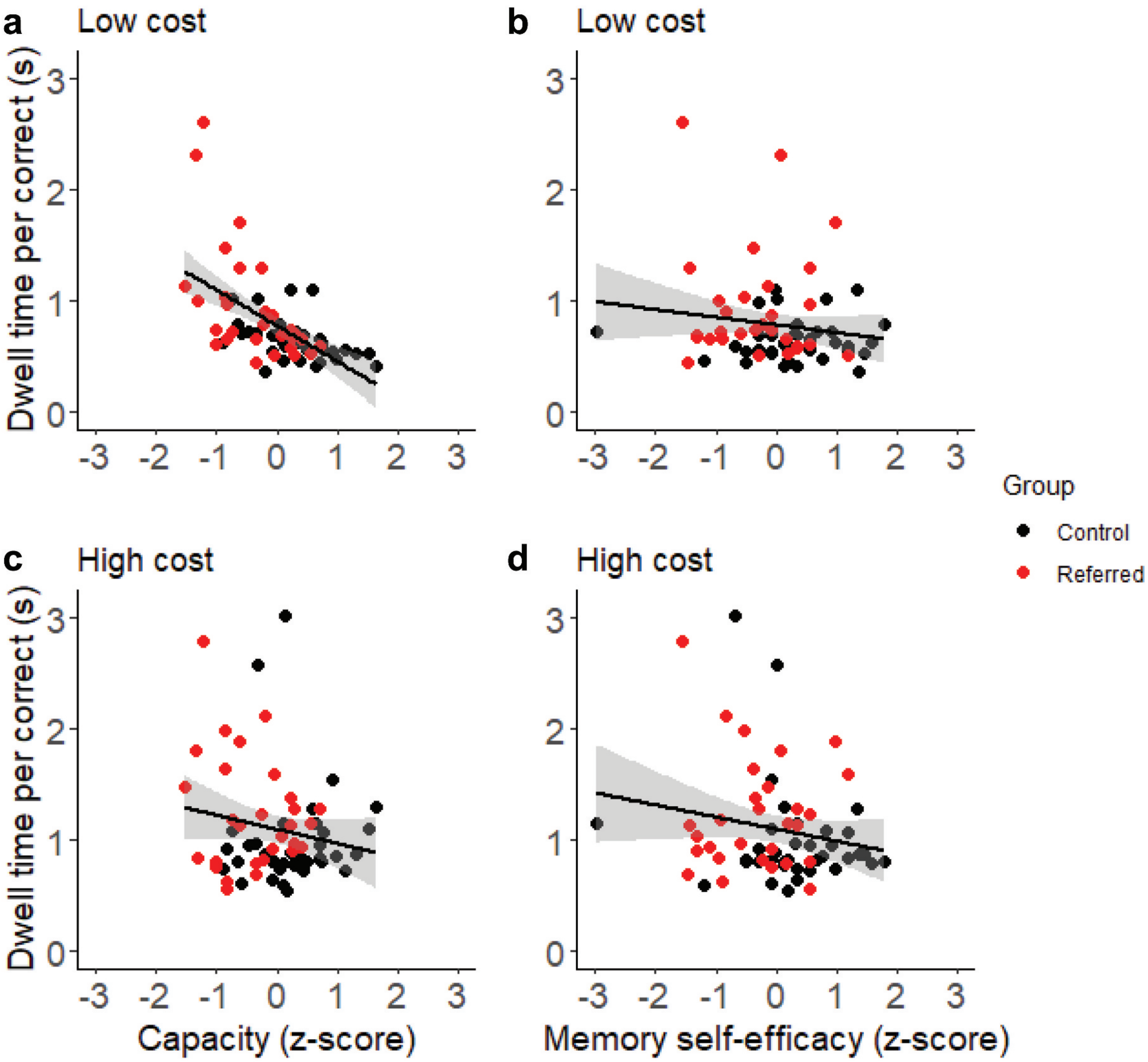

Figure 5. The relation between dwell time per correct placement and the capacity compound z-score, and memory self-efficacy compound z-score. Non-referred controls are depicted in black, individuals referred to the outpatient memory clinic are depicted in red. A smoothed linear coefficient is added in black with confidence intervals in grey.

In summary, higher capacity was associated with fewer and shorter inspections per correctly placed item, and better performance, regardless of whether or not information was readily available. Memory self-efficacy was associated with neither of these outcomes. The same conclusions were drawn when running the analyses with only the CFQ as a measure of memory self-efficacy (to make sure that the hiatus in number of valid data sets of people completing the MIA would not bias the results for this component; see Supplementary Materials for elaboration).

3.4. Predictive value of memory capacity subtasks and level of memory functioning

The results of the non-parametric regression models to explore the *number of crossings per correct placement* and *dwell time per correct placement* in both conditions as a function of raw memory capacity subtask scores, with covariates age and level of education, are reported in the Supplementary Material. We have analyzed all subtasks (also the delayed recall scores) to explore any relation between memory subprocesses (e.g., short-term encoding versus long-term retrieval) and inspection behavior. After correcting for multiple tests, we found that verbal attentional span (Digit Span forward), visual working memory capacity (d'), and verbal encoding (Rey Auditory-Verbal Learning Task total score) were related to inspection frequency and duration when information was freely available. When inspecting information became more costly, the other subtasks also started to exert their influence on inspection frequency; we found that all but two (Location Learning

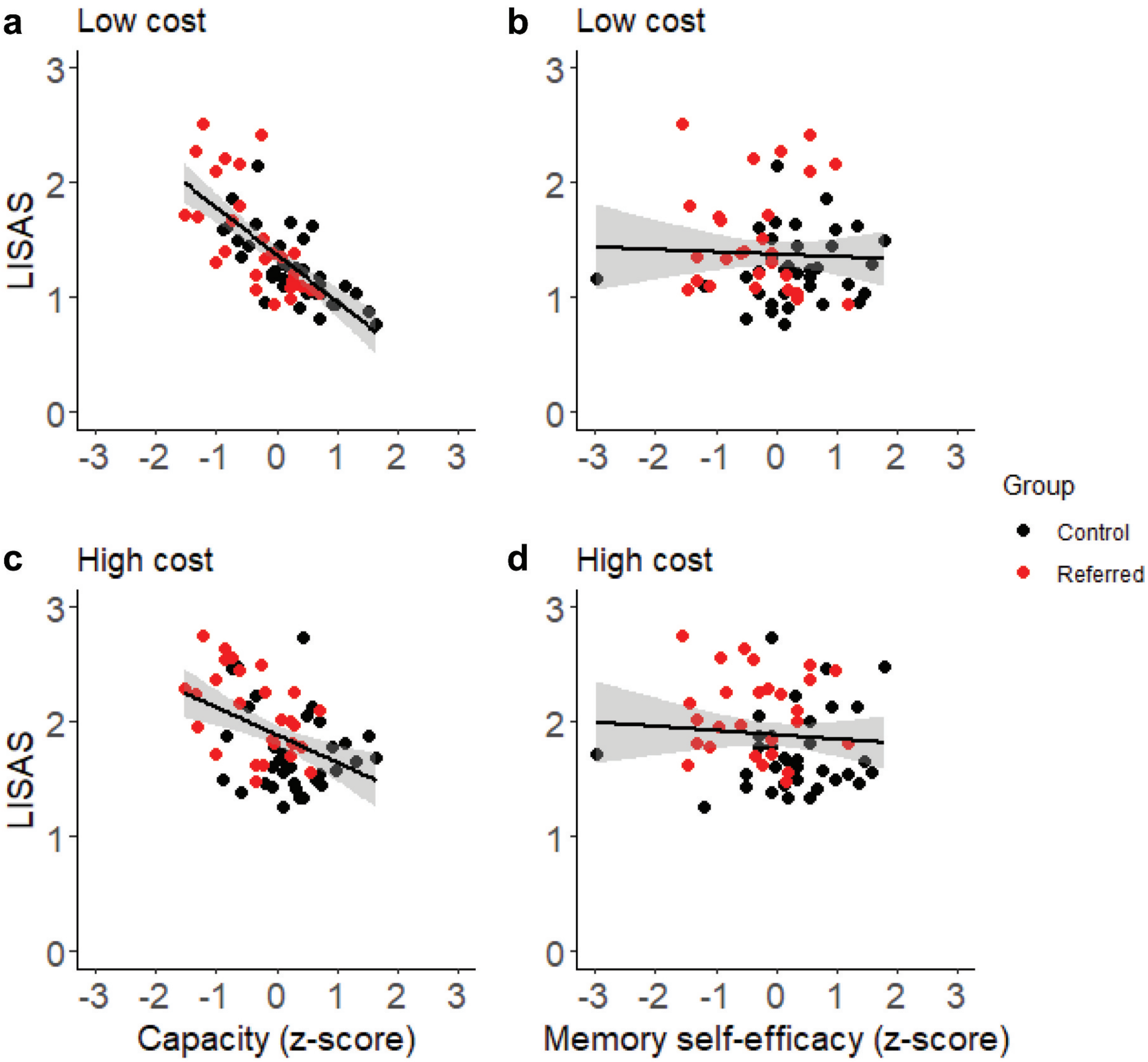

Figure 6. The relation between LISAS and the capacity compound z-score, and memory self-efficacy compound z-score. Non-referred controls are depicted in black, individuals referred to the outpatient memory clinic are depicted in red. A smoothed linear coefficient is added in black with confidence intervals in grey.

Task displacement errors and Location Learning Task delayed recall) subtests were predictive of inspection frequency in the high-cost condition, implying that higher memory capacity on each of these subtasks resulted in fewer inspections needed to place one item correctly. Interestingly, none were related to dwell time in the high-cost condition. Although interpreting these results is premature due to the relatively small sample size, we cautiously infer that there may be a benefit of both verbal and visual attentional and working memory span, resulting in fewer inspections and shorter inspection duration in stable visual environments, and that one may benefit further from higher capacity in situations where information is less readily available and memorization is prompted. Individuals with higher capacity rely less on the outside world.

To get a gist about clinical value, we further visualized inspections per correct as a function of level of memory performance category (intact, below average, impaired; see Supplementary Material). The impaired group (performance <2nd percentile on two or more subtasks (American Psychiatric Association, 2013; Hendriks et al., 2014, 2020), inspected significantly more than the intact and below average group (performance <2nd percentile on one subtask (and/)or below the 9th percentile on two or more subtasks), but the intact and below average group did not differ from each other.

4. Discussion

Relying on the external world by (re)inspecting (i.e., “sampling” or “offloading”) visual information alleviates the need to load internal memory to its full capacity. Since visual information remains available in the external world, it is unlikely that the full capacity of visual memory will be used in everyday life if given the option not to. However, to clinically objectify memory complaints that have warranted a referral for cognitive assessment, it is precisely this capacity characteristic that is examined, and it is generally not taken into account that the individual can exploit their environment as a support system. Crucially, this means that current memory assessment fails to incorporate the possibility that people may choose to memorize information at the preferred rather than the maximum load, and instead (re)inspect information from the external world. To complement the clinical approximation of memory use in everyday life, we assessed memory capacity, subjective experience of memory functioning (i.e., memory self-efficacy), *and* inspection behavior in individuals who were either referred to an outpatient clinic for cognitive assessment or not.

As expected, compared to the non-referred group, the referred group had a lower memory capacity as measured with standard neuropsychological memory tests, and lower levels of memory self-efficacy as measured with memory self-efficacy questionnaires. When participants were asked to copy an example puzzle to an empty grid, there was no difference in inspection frequency: both groups inspected the example puzzle equally often to place one item correctly. Only the degree of availability of information showed to influence inspection frequency: both groups made fewer (re) inspections when they had to wait every time they wanted to inspect the example puzzle (high cost condition) as compared to when the information was continuously available (low cost condition), replicating results from previous studies (Draschkow et al., 2021; Sahakian et al., 2023; Somai et al., 2020). Interestingly, the referred group did not inspect more often than the non-referred group in either of the conditions, while an effect of memory impairment was previously shown in a previous study where a group of severely memory-impaired individuals with Korsakoff syndrome inspected more often as compared to age-matched healthy controls (Böing et al., 2023). The absence of a group difference in the current sample can partly be explained by the heterogeneity of both the referred and non-referred group: some individuals were referred to a memory clinic but performed only slightly below average, others were not referred but showed impaired memory performance. As we aimed for a group that varied in both objective memory impairments and subjective memory complaints, it was no surprise that groups overlapped regarding memory functioning, and that there was no clear-cut impaired versus non-impaired difference. Our main aim was to include all individuals within one model and investigate the independent effects of objective memory capacity and subjective memory functioning, rather than referral. We found that people with a lower memory capacity inspected more frequently as compared to those with a higher memory capacity. As a subsequent exploration, we interpreted the raw capacity span scores against appropriate norm scores (controlled for age and education), to check the effect of clinical memory impairment. Those who would be classified as clinically impaired, inspected more frequently as compared to those whose performance was below average or intact. These results align with our previous study showing distinctly different behavior for memory-impaired individuals than healthy controls (Böing et al., 2023). This effect thus only arises with more profound memory deficits and shows that inspection frequency is not a sensitive measure to map subtle memory deficiencies; measuring inspection frequency only distinguishes two subgroups (impaired vs. non-impaired) rather than three (impaired vs. below average vs. intact, tested with standard neuropsychological capacity assessments). Future studies with a larger sample size would allow to dissociate subgroups based on inspection behavior not only in terms of the level of memory function, but also in terms of clinical status (e.g., mild cognitive impairment, dementia, Parkinson’s, or presence of psychological factors).

Although there was no difference between the referred and non-referred groups regarding how often they inspected information, there was a difference in inspection duration. The referred group showed increased dwell times, indicating a potential necessity for longer encoding times or a slower evidence (here, confidence) accumulation to reach an action threshold (Lee et al., 2023; Sahakian et al., 2023), although it should be noted that this effect disappeared when outliers were removed. The referred group also showed worse performance than the non-referred group in terms of task speed-accuracy, which may be a reflection of the longer information uptake, but which could also arise because more errors were made, or slower information processing speed in general. In summary, the referred group did not rely *more often* but potentially *longer* on the outside world, and showed weakened performance as compared to the non-referred group.

As memory capacity varies on a continuous rather than dichotomous scale, we were particularly interested to see whether and how the objective memory capacity span and memory self-efficacy would influence inspection frequency, regardless of referral or clinical status. Surprisingly, we found that memory capacity, but not memory self-efficacy, was related to both inspection frequency and dwell time. Higher memory capacity related to fewer and shorter (re)inspections to place one item correctly. These results show that the standard neuropsychological memory capacity tasks used in clinical care generalize to behavior in a more free setting where participants can choose whether or not to load memory. In other words, the amount of information one *is able to* memorize translates to how often someone falls back onto the external world, and consequently, how much information someone *will* memorize in day-to-day memory use. This relation is clinically implied but not so often directly tested, although there have been some earlier attempts in which capacity scores assessed with specific neuropsychological measures (e.g., Corsi Forward Span) were related to different types of cognitive offloading (Meyerhoff et al., 2021). Cognitive offloading is a broad concept and refers to any type of behavior that is executed to decrease the effort associated with cognitive processes (e.g., using a calendar to support prospective memory, but also tilting one's head to avoid mental rotation). We consider inspection behavior as a subordinate of cognitive offloading behavior. Our results add to the offloading literature by taking several clinically relevant memory capacity subtasks into account and testing them as an integrated memory concept as well as displaying their individual predictive value on inspection behavior.

One remark on estimating memory capacity is that the standard neuropsychological task outcomes are based on a single measurement in time, making them susceptible to measurement errors and attentional lapses. Composing a capacity score that combines performance on multiple memory subtasks (such as in the present study, and in, e.g., Morrison & Richmond, 2020) partly accounts for such momentary influences, but cannot fully eliminate them. Surely it would be more robust to extract an average capacity span from repeated trials for each subtask (as Meyerhoff et al. (2021) did for the Corsi Block Tapping Task Forward Span), but this is time-costly and burdensome for patients in a clinical setting. Other momentary influences relate to motivational and personal aspects of task completion, such as the desire to be accurate or certain rather than quick, or the drive to perform at a personally set maximum or a "just" sufficient level. Consider also that undergoing assessment as part of a scientific study comes with a different incentive, mind-set and setting than completing these for the sake of diagnosis in a clinical setting.

Apart from using an integrated memory concept rather than single capacity measures, the current results also add another clinical perspective to the offloading literature: while most offloading studies are based upon healthy student populations, we extend our findings to adults with memory impairments (including Böing et al., 2023). Importantly, the various lines of research on strategic use of the external world emphasize the wide variety of memory support strategies, ranging from either the trade-off in memorizing versus sampling in working memory (Böing et al., 2023; Draschkow et al., 2021; Hoogerbrugge et al., 2024; Sahakian et al., 2023; Somai et al., 2020; Van der Stigchel, 2020) to more conscious and deliberate cognitive offloading (e.g., writing things down to aid memory, placing a cue for oneself to remember intentions; see Gilbert et al. (2023) and Risko and Gilbert (2016) for a review). With regards to the relation between memory capacity and any type of offloading, findings are mixed, showing that they are subject to specific characteristics of the memory task being used. For example, prospective memory (i.e., remembering an intention to act out in the future) requires a different allocation of subsystems and operates on a different timescale than making sure to remember and dial the correct number from an appointment note, and memory capacity may differentially affect these processes. It is therefore difficult to generalize findings of the current study on visual working memory capacity, but also those of the aforementioned studies, to a general concept of real-world memory usage. Nevertheless, our findings emphasize the need to take memory strategies into consideration when trying to approximate freedom-of-choice memory use alongside memory capacity. The relevance of such considerations is subserved by our current observation that, although higher memory capacity relates to fewer inspections, people still *avoid maximum capacity usage*, and prefer to memorize one up to three items maximum in working memory. Even when information is not always readily available (high cost condition), people often take some degree of time or effort expenditure for granted (in waiting, annoyance, physical exertion) to avoid full memory capacity use. What's more, increasing the amount of information to be remembered increases the likelihood of offloading (Risko & Gilbert, 2016). This adaptive behavior is missed in the regular assessment of memory capacity.

It is intriguing that subjective and objective capacity measures are frequently discrepant; people may experience memory failure in the absence of impaired memory capacity (Beaudoin & Desrichard, 2011; Mattos et al., 2003; Ponds & Jolles, 1996a). In our results, similarly, the lack of a relation between subjective memory performance on the one hand, and capacity and inspection behavior on the other, illustrates that memory self-(in)efficacy does not translate to maximum capacity nor actual visual working memory usage. Whereas we expected people with lower confidence to check more often, memory self-efficacy was not associated with the frequency with which one relied on the external world. This contrasts with Gilbert (2015), who posed that both task-specific as well as domain-general metacognitive confidence (i.e., like our quantification of memory self-efficacy) explain offloading. The incongruency between their and our findings likely arises due to different operationalizations of offloading, with the study of Gilbert (2015) focusing on intention offloading (hinting at prospective memory) and ours on working memory. In a task that engages working memory in a way that is more similar to ours, Grinschgl et al. (2021) manipulated domain-general confidence by facilitating fake feedback, and found, like us, no effect of confidence on offloading behavior. Following the same line as Hertzog and colleagues (1987, as cited in Beaudoin & Desrichard, 2011), the authors propose a distinction between metacognitive *beliefs*, i.e., a more generic gist about one's memory, and metacognitive *experiences*, reflecting confidence about performance on a specific task (or trial) that had just been completed. We captured metacognitive beliefs, but not metacognitive experiences across the test procedure (e.g., self-efficacy after a capacity task, versus after an (un)successful copy task trial), while the latter might have exerted an effect on inspection strategy in the way that Gilbert (2015) described. This line of reasoning suggests that memory confidence fluctuates depending on prior experience, task-specific characteristics and the moment in time one is asked to judge confidence. Further diving into the waters of memory uncertainty, Sahakian et al. (2023) showed that, even *within trials*, people were not keen on acting upon content that they were not confident enough about, although they had some residual information in working memory. This aligns with the observation that people may use offloading even though it does not necessarily benefit performance but mostly serves to safeguard a feeling of security (Risko & Dunn, 2015). *Re*inspecting could then be seen as an act to accumulate confidence, and the threshold of certainty that needs to be reached would then be described as an action threshold (Sahakian et al., 2023). We expected our measure of memory self-efficacy to express the individual's general action threshold: some people would be more confident than others, and would therefore need less reassurance, resulting in fewer inspections. Yet, individual differences in memory self-efficacy could not account for inspection frequency (nor duration). If one wants to make claims about reinspecting – or offloading in the broader sense – as an expression of reassurance behavior, it might be valuable to test people with specific tendencies as observed in, for example, individuals with obsessive-compulsive disorder (Karadag et al., 2005; Tolin et al., 2001), or high levels of performance failure anxiety. A further peril of judging one's own memory (that seeps through to the memory self-efficacy construct and also into the reason for referral) is that the general public often falsely attributes cognitive failures to memory dysfunction, while they are often the result of deficits in other domains (e.g., executive function, attention, motivation; Hendriks et al., 2014). We cannot rule out the influence of these factors within our sample.

The heterogeneity in tasks, groups, cognitive profiles and personal characteristics adds complexity to disentangling factors that influence memory usage, but at the same time reflects the complexity with which the clinician is faced when assessing cognitive functioning in the clinic. We therefore embrace this complexity when trying to approximate memory usage in daily life and underline that memory assessment is multifaceted. Different tasks may evoke different self-perceived memory ability, capacity estimates may vary across the verbal and visual modalities, the ease with which external stimuli are internalized may vary because of stimulus characteristics, semantic labeling, and familiarity (e.g., shopping items versus abstract geometric shapes), and *how* one chooses to use either the internal memory load or the external world is differentially balanced based on both internal and external demands. Further, the ability to learn over the course of trials and draw from long-term memory (e.g., stronger memories for repeated stimuli) may differentially influence individuals. When adding degrees of freedom to a task, as in everyday life, other cognitive domains may further start to interact with how memory is engaged. For example, our task not only required working memory, but also aspects of executive functioning (attention, planning, monitoring). As we have tested memory functioning and not executive functioning, we can only make an attempt to attribute our findings to working memory. Yet, one should be wary that this does not nearly explain the additional cognitive processes that may play a role when interacting with the environment and engaging working memory in everyday life. Taken together, approaching memory functioning by *only* examining memory capacity does not do justice to the many layers of memory usage: individuals – with and without memory capacity constraints – may employ a variety of compensatory fallbacks dependent on the task at hand, explaining why we do see differences between referred and non-referred individuals on tasks that *force* maximum capacity use, but lesser so on a task where people have more degrees of freedom and can *choose* how many items they memorize at once.

To put this in a clinical perspective: memory complaints, as repeatedly shown, do not necessarily translate to memory capacity measures, and memory capacity does only to a certain extent translate to actual memory use when given the choice. Although forced capacity tasks appear to be more sensitive to categorize individuals' performance as clinically impaired, below average or intact as compared to our freedom-of-choice copy task, they fail to capture the workaround that people use to prevent maximum capacity usage. When given some wiggle room, those with lower memory capacity can compensate by increasing their reliance on the external world. These individuals might need to inspect information somewhat more often or longer to use information correctly, but this may be the relatively "cheap" price one pays to work around capacity limits.

Acknowledgments

The authors thank T. Hamers and B. de Zwart for helping with data collection. We thank A.J. Hoogerbrugge for programming the Copy Task, and supporting with data pre-processing.

Disclosure statement

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

Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme [grant agreement 863732].

Author contributions

Conceptualization: S.B., A.F.T.B., C.R., E.V.d.B., Z.A.S., T. C.W.N., and S.V.d.S; data curation: S.B.; formal analysis: S.B.; funding acquisition: S.V.d.S.; investigation: S.B.; methodology: S.B., A.F.T.B., C.R., E.V.d.B., Z.A.S., T.C. W.N., and S.V.d.S.; project administration: S.B.; resources: S.B., A.F.T.B., C.R., E.V.d.B., Z.A.S., and M.J.B.; software: S.B., and A.F.T.B.; supervision: A.F.T.B., T.C.W.N., and S. V.d.S.; visualization: S.B.; writing – original draft: S.B.; writing – review and editing: S.B., A.F.T.B., C.R., E.V.d. B., Z.A.S., M.J.B., T.C.W.N., and S.V.d.S. All authors have read and agreed to the published version of the manuscript.

Institutional review board statement

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of the Faculty Ethics Review Board of the Faculty of Social and Behavioural Sciences at Utrecht University (protocol numbers protocol numbers 21–0076 and 21–0269).

Informed consent statement

Informed consent was obtained from all subjects involved in the study. Written informed consent was obtained from the patient(s) to use their data in this paper.

Data availability statement

The data presented in this study are openly available in Open Science Framework at https://osf.io/ys67b/.

References

Abdulrab, K., & Heun, R. (2008). Subjective memory impairment. A review of its definitions indicates the need for a comprehensive set of standardised and validated criteria. *European Psychiatry*, *23*(5), 321–330. https://doi.org/10.1016/j.eurpsy.2008.02.004

Alberts, M., Smets, E. M. A., Vercoulen, J. H. M. M., Garssen, B., & Bleijenberg, G. (1997). 'Verkorte vermoeidheidsvragenlijst': Een praktisch hulpmiddel bij het scoren van vermoeidheid. *Nederlands tijdschrift voor geneeskunde*, *141*(32), 1526–1530.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Arnoult, M. (1956). Familiarity and recognition of nonsense shapes. *Journal of Experimental Psychology*, *51*(4), 269–276. https://doi.org/10.1037/h0047772

Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*(1), 66–80. https://doi.org/10.1162/jocn.1995.7.1.66

Beaudoin, M., & Desrichard, O. (2011). Are memory self-efficacy and memory performance related? A meta-analysis. *Psychological Bulletin*, *137*(2), 211–241. https://doi.org/10.1037/a0022106

Bleijenberg, G., Knoop, H., & Gielissen, M. (2009). De Verkorte VermoeidheidsVragenlijst voor het vaststellen van de ernst van chronische vermoeidheid. *Bijblijven*, *25*(1), 19–21. https://doi.org/10.1007/BF03087615

Böing, S., de Zwart, B., Ten Brink, A. F., Nijboer, T. C. W., & Van Der Stigchel, S., (SUBM.). *Relying on the external world: Individuals variably use low- and high-loading, but rarely full-loading, strategies when engaging visual working memory*.

Böing, S., Ten Brink, A. F., Hoogerbrugge, A. J., Oudman, E., Postma, A., Nijboer, T. C. W., & Van der Stigchel, S. (2023). Eye movements as proxy for visual working memory usage: Increased reliance on the external world in Korsakoff Syndrome. *Journal of Clinical Medicine*, *12*(11), Article 11. https://doi.org/10.3390/jcm12113630

Bouma, A., Mulder, J., Lindeboom, J., Schmand, B., & Brein en Cognitie (Psychologie, FMG). (2012). *Handboek neuropsychologische diagnostiek*. Pearson. https://dare.uva.nl/personal/pure/en/publications/handboek-neuropsychologische-diagnostiek(633b0632-36fd-4ab5-92cc-3d868ff18266).html

Broadbent, D. E., Cooper, P. F., FitzGerald, P., & Parkes, K. R. (1982). The Cognitive Failures Questionnaire (CFQ) and its correlates. *British Journal of Clinical Psychology*, *21*(1), 1–16. https://doi.org/10.1111/j.2044-8260.1982.tb01421.x

Brockmole, J., & Logie, R. (2013). Age-related change in visual working memory: A Study of 55, 753 participants aged 8-75. *Frontiers in Psychology*, *4*. https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00012

Brunetti, R., Del Gatto, C., & Delogu, F. (2014). eCorsi: Implementation and testing of the Corsi block-tapping task for digital tablets. *Frontiers in Psychology*, *5*. https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00939

Burnett, L. K., & Richmond, L. L. (2023). Just write it down: Similarity in the benefit from cognitive offloading in young and older adults. *Memory & Cognition*, *51*(7), 1580–1592. https://doi.org/10.3758/s13421-023-01413-7

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(4), 543–570. https://doi.org/10.1080/13854046.2012.680913

Claessen, M. H. G., van der Ham, I. J. M., & van Zandvoort, M. J. E. (2015). Computerization of the Standard Corsi Block-Tapping Task affects its underlying cognitive concepts: A pilot study. *Applied Neuropsychology Adult*, *22*(3), 180–188. https://doi.org/10.1080/23279095.2014.892488

Corsi, P. M. (1972). *Human memory and the medial temporal region of the brain* [PhD thesis, McGill University]. Montreal, 85.

Dalmaijer, E. S., Mathôt, S., & Van der Stigchel, S. (2014). PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior Research Methods*, *46*(4), 913–921. https://doi.org/10.3758/s13428-013-0422-2

Drag, L. L., & Bieliauskas, L. A. (2010). Contemporary review 2009: Cognitive aging. *Journal of Geriatric Psychiatry and Neurology*, *23*(2), 75–93. https://doi.org/10.1177/0891988709358590

Draschkow, D., Kallmayer, M., & Nobre, A. C. (2021). When natural behavior engages working memory. *Current Biology*, *31*(4), 869–874.e5. https://doi.org/10.1016/j.cub.2020.11.013

Droll, J., & Hayhoe, M. (2008). Trade-offs between gaze and working memory use. *Journal of Experimental Psychology Human Perception and Performance*, *33*(6), 1352–1365. https://doi.org/10.1037/0096-1523.33.6.1352

Faul, F., Erdfelder, E., Lang, L.-G., & Buchner, A. (2009). *G*power, 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences*. https://link.springer.com/article/10.3758/BF03193146

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

Gilbert, S. J., Bird, A., Carpenter, J. M., Fleming, S. M., Sachdeva, C., & Tsai, P.-C. (2020). Optimal use of reminders: Metacognition, effort, and cognitive offloading. *Journal of Experimental Psychology General*, *149*(3), 501–517. https://doi.org/10.1037/xge0000652

Gilbert, S. J., Boldt, A., Sachdeva, C., Scarampi, C., & Tsai, P.-C. (2023). Outsourcing memory to external tools: A review of 'intention offloading'. *Psychonomic Bulletin & Review*, *30*(1), 60–76. https://doi.org/10.3758/s13423-022-02139-4

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, *113*(3), 461–482. https://doi.org/10.1037/0033-295X.113.3.461

Grinschgl, S., Meyerhoff, H. S., Schwan, S., & Papenmeier, F. (2021). From metacognitive beliefs to strategy selection: Does fake performance feedback influence cognitive offloading? *Psychological Research*, *85*(7), 2654–2666. https://doi.org/10.1007/s00426-020-01435-9

Hendriks, M. P. H., Kessels, R. P. C., Gorissen, M., Schmand, B., & Duits, A. (2014). *Neuropsychologische Diagnostiek. De klinische praktijk*. Uigeverij Boom.

Hendriks, M. P. H., Mol, B. A. W., & Kessles, R. P. C. (2020). Opinie: Uniformiteit in de kwalitatieve beschrijving van scores op prestatietaken. *Tijdschrift voor Neuropsychologie*, *15*(3). https://www.tvnp.nl/inhoud/tijdschrift_artikel

Hoogerbrugge, A. J., Strauch, C., Böing, S., Nijboer, T. C. W., & Stigchel, S. V. D. (2024). Just-in-time encoding into visual working memory is contingent upon constant availability of external information. *Journal of Cognition*, *7*(1), 39. https://doi.org/10.5334/joc.364

Hurt, C. S., Burns, A., Brown, R. G., & Barrowclough, C. (2012). Why don't older adults with subjective memory complaints seek help? *International Journal of Geriatric Psychiatry*, *27*(4), 394–400. https://doi.org/10.1002/gps.2731

Jessen, F., Wiese, B., Bachmann, C., Eifflaender-Gorfer, S., Haller, F., Kölsch, H., Luck, T., Mösch, E., van den Bussche, H., Wagner, M., Wollny, A., Zimmermann, T., Pentzek, M., Riedel-Heller, S. G., Romberg, H.-P., Weyerer, S., Kaduszkiewicz, H., Maier, W., Bickel, H., & German Study on Aging, C. and D. in P. C. P. S. G. (2010). Prediction of dementia by subjective memory impairment: Effects of severity and temporal association with cognitive impairment. *Archives of General Psychiatry*, *67*(4), 414–422. https://doi.org/10.1001/archgenpsychiatry.2010.30

Johansson, B., Allen-Burge, R., & Zarit, S. H. (1997). Self-reports on memory functioning in a longitudinal study of the oldest old: Relation to current, prospective, and retrospective performance. *Journals of Gerontology, Series B:*

Psychological Sciences & Social Sciences, *52B*(3), P139–P146. https://doi.org/10.1093/geronb/52B.3.P139

Jungen, M., van Engelen, E., & Bokhorst, M. (2019). *Uitgebreide toelichting van het meetinstrument: HADS*. https://meetinstrumentenzorg.nl/wp-content/uploads/instrumenten/HADS-form.pdf

Karadag, F., Oguzhanoglu, N., Ozdel, O., & Atesci, F. C. (2005). Memory function in patients with obsessive compulsive disorder and the problem of confidence in their memories: A clinical study. *Croatian Medical Journal*, *46* (2), 282–287.

Kessels, R. P. C., Bucks, R. S., Willison, J. R., & Byrne, L. M. T. (2014). *Location learning test* (3th ed.). Hogrefe Uitgevers.

Kessels, R. P. C., Nys, G. M. S., Brands, A. M. A., van den Berg, E., & Zandvoort, V. M. J. E. (2006). The modified location learning test: Norms for the assessment of spatial memory function in neuropsychological patients. *Archives of Clinical Neuropsychology*, *21*(8), 841–846. https://doi.org/10.1016/j.acn.2006.06.015

Kessels, R. P. C., van Zandvoort, M. J. E., Postma, A., Kappelle, L. J., & de Haan, E. H. F. (2000). The Corsi Block-Tapping Task: Standardization and normative data. *Applied Neuropsychology*, *7*(4), 252–258. https://doi.org/10.1207/S15324826AN0704_8

Lee, D. G., Daunizeau, J., & Pezzulo, G. (2023). Evidence or confidence: What is really monitored during a decision? *Psychonomic Bulletin & Review*, *30*(4), 1360–1379. https://doi.org/10.3758/s13423-023-02255-9

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), Article 6657. https://doi.org/10.1038/36846

Mattos, P., Lino, V., Rizo, L., Alfano, Â., Araújo, C., & Raggio, R. (2003). Memory complaints and test performance in healthy elderly persons. *Arquivos de Neuro-psiquiatria*, *61*(4), 920–924. https://doi.org/10.1590/S0004-282X2003000600006

Melnik, A., Schüler, F., Rothkopf, C. A., & König, P. (2018). The world as an external memory: The price of saccades in a sensorimotor task. *Frontiers in Behavioral Neuroscience*, *12*, 253. https://doi.org/10.3389/fnbeh.2018.00253

Meyerhoff, H. S., Grinschgl, S., Papenmeier, F., & Gilbert, S. J. (2021). Individual differences in cognitive offloading: A comparison of intention offloading, pattern copy, and short-term memory capacity. *Cognitive Research: Principles and Implications*, *6*(1), 34. https://doi.org/10.1186/s41235-021-00298-x

Morrison, A. B., & Richmond, L. L. (2020). Offloading items from memory: Individual differences in cognitive offloading in a short-term memory task. *Cognitive Research: Principles and Implications*, *5*(1), 1. https://doi.org/10.1186/s41235-019-0201-4

Oudman, E., Schut, M. J., Ten Brink, A. F., Postma, A., & Van der Stigchel, S. (2020). Visual working memory capacity in Korsakoff's amnesia. *Journal of Clinical and Experimental Neuropsychology*, *42*(4), 363–370. https://doi.org/10.1080/13803395.2020.1722800

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

Patrick, J., Morgan, P. L., Smy, V., Tiley, L., Seeby, H., Patrick, T., & Evans, J. (2015). The influence of training and experience on memory strategy. *Memory & Cognition*, *43*(5), 775–787. https://doi.org/10.3758/s13421-014-0501-3

Pizzonia, K. L., & Suhr, J. A. (2022). Systematic review of correlates of internal and external memory strategy use in older adults. *Journal of Applied Gerontology*, *41*(5), 1491–1499. https://doi.org/10.1177/07334648211065427

Ponds, R., Commissaris, K., & Jolles, J. (1997). Prevalence and covariates of subjective forgetfulness in a normal population in the Netherlands. *International Journal of Aging & Human Development*, *45*(3), 207–221. https://doi.org/10.2190/MVQ1-WB58-875H-Y4X0

Ponds, R. W. H. M., Boxtel, M., & Jolles, J. (2006). De "Cognitive Failure Questionnaire" als maat voor subjectief cognitief functioneren. *Tijdschrift voor Neuropsychologie*, *1* , 37–45. https://www.researchgate.net/profile/Rudolf-Ponds-2/publication/286212206_De_'Cognitive_Failure_Questionnaire'_als_maat_voor_subjectief_cognitief_functioneren/links/566e684008aea0892c52959f/De-Cognitive-Failure-Questionnaire-als-maat-voor-subjectief-cognitief-functioneren.pdf

Ponds, R. W. H. M., & Jolles, J. (1996a). Memory complaints in elderly people: The role of memory abilities, metamemory, depression, and personality. *Educational Gerontology*, *22*(4), 341–357. https://doi.org/10.1080/0360127960220404

Ponds, R. W. H. M., & Jolles, J. (1996b). The abridged Dutch metamemory in adulthood (MIA) questionnaire: Structure and effects of age, sex, and education. *Psychology and Aging*, *11*(2), 324–332. https://doi.org/10.1037/0882-7974.11.2.324

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Risko, E. F., & Dunn, T. L. (2015). Storing information in-the-world: Metacognition and cognitive offloading in a short-term memory task. *Consciousness and Cognition*, *36*, 61–74. https://doi.org/10.1016/j.concog.2015.05.014

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Riverbank Computing Limited. (2019). *PyQt5 library*.

Saan, R. J., & Deelman, B. G. (1986). *De 15-Woordentests A en B. Een voorlopige handleiding (Intern rapport)*. AZG, afd. Neuropsychologie.

Sahakian, A., Gayet, S., Paffen, C. L. E., & Van der Stigchel, S. (2023). Mountains of memory in a sea of uncertainty: Sampling the external world despite useful information in visual working memory. *Cognition*, *234*, 105381. https://doi.org/10.1016/j.cognition.2023.105381

Saykin, A. J., Wishart, H. A., Rabin, L. A., Santulli, R. B., Flashman, L. A., West, J. D., McHugh, T. L., & Mamourian, A. C. (2006). Older adults with cognitive complaints show brain atrophy similar to that of amnestic MCI. *Neurology*, *67*(5), 834–842. https://doi.org/10.1212/01.wnl.0000234032.77541.a2

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., Araya-Ajoy, Y. G., & Sutherland, C. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Schmand, B., Jonker, C., Hooijer, C., & Lindeboom, J. (1996). Subjective memory complaints may announce dementia. *Neurology*, *46*(1), 121–125. https://doi.org/10.1212/WNL.46.1.121

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. H. Spieler & E. Schumacher (Eds.), *New methods in cognitive psychology*. (pp. 4–31). Routledge.

Somai, R. S., Schut, M. J., & Van der Stigchel, S. (2020). Evidence for the world as an external memory: A trade-off between internal and external visual memory storage. *Cortex, A Journal Devoted to the Study of the Nervous System and Behavior*, *122*, 108–114. https://doi.org/10.1016/j.cortex.2018.12.017

Spinhoven, P., Ormel, J., Sloekers, P. P. A., Kempen, G. I. J. M., Speckens, A. E. M., & Hemert, A. M. V. (1997). A validation study of the Hospital Anxiety and Depression Scale (HADS) in different groups of Dutch subjects. *Psychological Medicine*, *27*(2), 363–370. https://doi.org/10.1017/S0033291796004382

Steinberg, S. I., Negash, S., Sammel, M. D., Bogner, H., Harel, B. T., Livney, M. G., McCoubrey, H., Wolk, D. A., Kling, M. A., & Arnold, S. E. (2013). Subjective memory complaints, cognitive performance, and psychological factors in healthy older adults. *American Journal of Alzheimer's Disease & Other Dementias®*, *28*(8), 776–783. https://doi.org/10.1177/1533317513504817

Tolin, D. F., Abramowitz, J. S., Brigidi, B. D., Amir, N., Street, G. P., & Foa, E. B. (2001). Memory and memory confidence in obsessive-compulsive disorder. *Behaviour Research and Therapy*, *39*(8), 913–927. https://doi.org/10.1016/S0005-7967(00)00064-4

Turvey, C. L., Schultz, S., Arndt, S., Wallace, R. B., & Herzog, R. (2000). Memory complaint in a community sample aged 70 and older. *Journal of the American Geriatrics Society*, *48*(11), 1435–1441. https://doi.org/10.1111/j.1532-5415.2000.tb02634.x

Van der Stigchel, S. (2020). An embodied account of visual working memory. *Visual Cognition*, *28*(5–8), 414–419. https://doi.org/10.1080/13506285.2020.1742827

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods*, *49*(2), 653–673. https://doi.org/10.3758/s13428-016-0721-5

Verhage, F. (1964). *Intelligentie en leeftijd: Onderzoek bij Nederlanders van twaalf tot zevenenzeventig jaar*.

Verhage, F. (1965). Intelligence and Age in a Dutch Sample. *Human Development*, *8*(4), 238–245. https://doi.org/10.1159/000270308

Wechsler, D. (2012). *WAIS-IV-NL: Wechsler adult intelligence scale - Nederlandstalige bewerking. Technische Handleiding*. (4th ed.). Pearson Assessment and Information BV.

Williams, J. R., Robinson, M. M., Schurgin, M., Wixted, J., & Brady, T. (2022). *You can't "count" how many items people remember in working memory: The importance of signal detection-based measures for understanding change detection performance*. PsyArXiv. https://doi.org/10.31234/osf.io/jnt7r