Cognition 234 (2023) 105381

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original articles

Mountains of memory in a sea of uncertainty: Sampling the external world despite useful information in visual working memory

Andre Sahakian *, Surya Gayet, Chris L.E. Paffen, Stefan Van der Stigchel

Department of Experimental Psychology & Helmholtz Institute, Utrecht University, Heidelberglaan 1, 3584 CS, Utrecht, The Netherlands

ARTICLE INFO

Dataset link: https://osf.io/pkxdc/, https://osf.io/pkxdc/

Keywords:
Visual working memory
Action threshold
Copying task
Cognitive offloading

ABSTRACT

A large part of research on visual working memory (VWM) has traditionally focused on estimating its maximum capacity. Yet, humans rarely need to load up their VWM maximally during natural behavior, since visual information often remains accessible in the external world. Recent work, using paradigms that take into account the accessibility of information in the outside world, has indeed shown that observers utilize only one or two items in VWM before sampling from the external world again. One straightforward interpretation of this finding is that, in daily behavior, much fewer items are memorized than the typically reported capacity limits. Here, we first investigate whether this lower reliance on VWM when information is externally accessible might instead reflect resampling before VWM is actually depleted. To this aim we devised an online task, in which participants copied a model (six items in a 4x4 grid; always accessible) in an adjacent empty 4x4 grid. A key aspect of our paradigm is that we (unpredictably) interrupted participants just before inspection of the model with a 2-alternative-forced-choice (2-AFC) question, probing their VWM content. Critically, we observed above-chance performance on probes appearing just *before* model inspection. This finding shows that the external world was resampled, despite VWM still containing relevant information. We then asked whether increasing the cost of sampling causes participants to load up more information in VWM or, alternatively, to squeeze out more information from VWM (at the cost of making more errors). To manipulate the cost of resampling, we made it more difficult (specifically, more time-consuming) to access the model. We show that with increased cost of accessing the model (which lead to fewer, but longer model inspections), participants could place more items correctly immediately after sampling, and they kept attempting to place items for longer after their first error. These findings demonstrate that participants both encoded more information in VWM *and* made attempts to squeeze out more information from VWM when sampling became more costly. We argue that human observers constantly evaluate how certain they are of their VWM contents, and only use that VWM content of which their certainty exceeds a context-dependent "action threshold". This threshold, in turn, depends on the trade-off between the cost of resampling and the benefits of making an action. We argue that considering the interplay between the available VWM contents and a context-dependent action threshold, is key for reconciling the traditional VWM literature with VWM use in our day-to-day behavior.

1. Introduction

In order to perform many everyday actions, it is helpful to maintain an internal visual representation of task-relevant objects (e.g., knowing where and in what orientation you left the knife on the counter while cooking). Maintaining a visual representation of an object is straightforward when it is within view, but oftentimes objects are (temporarily) out of sight. This occurs, for example, when switching gaze to another object (you turn to an overboiling pan), or when something occludes the view (someone covers the knife with a dishcloth). In such cases, it might be necessary to internally maintain an accurate visual representation of the object. This maintenance is subserved by visual working memory (VWM) (Baddeley & Hitch, 1974). VWM has mainly been studied as an isolated memory storage, where the focus lay on finding the maximum capacity of this storage. To this end, most studies were set up in such way that the task required observers to tax their VWM maximally. Subsequently probing their memory (in one way or another) provided a measure of how much information is contained in VWM (Ma et al., 2014). How VWM is usually employed in our everyday life, however, is quite different from how it is employed in such studies. Contrary to typical VWM tasks (where objects are removed and asked about after some delay), objects in the world usually remain accessible: they can therefore be inspected and reinspected, simply by reorienting

* Corresponding author.
E-mail address: a.sahakian@uu.nl (A. Sahakian).

https://doi.org/10.1016/j.cognition.2023.105381
Received 19 August 2022; Received in revised form 14 December 2022; Accepted 17 January 2023
Available online 30 January 2023

towards the object in the external world. Leaving information externally and retrieving it when desired reduces the need to maximally load up VWM (Ballard et al., 1995; Van der Stigchel, 2020).

A handful of studies have examined how VWM is employed in such naturalistic settings (Ballard et al., 1995; Draschkow et al., 2021; Somai et al., 2020). These studies investigated VWM from a goal-oriented perspective: the objective was not to force maximal loading of VWM, but rather to observe how VWM was employed during goal-directed behavior. These studies mostly make use of so-called copying tasks. In such tasks, an arrangement of items (commonly referred to as the "Model") has to be recreated at another location (see Fig. 1A and Video S1 - osf.io/g64vz for our implementation). Importantly, during execution of the task, the *Model* always remains accessible. Observers are instructed to pick up items from a separate pool of items (referred to as the "Resources") and create a replica of the arrangement of their own in the "Workspace". Researchers applying this paradigm typically found that observers inspected the *Model* relatively often (e.g., they inspected the *Model* to identify which item to pick up, and then inspected the same item in the *Model* again to determine where to place it in the *Workspace*). The frequent (re)inspecting of the *Model* was typically interpreted as observers relying only little on their VWM. More specifically, with a maximum number of items in VWM being 1 or 2, utilization of VWM was deemed minimal compared to typical estimates of VWM capacity, which hover around 4 four items (Luck & Vogel, 1997). This discrepancy between potential and actual use of VWM capacity is indeed striking at face value.

Interpreting exactly how much information is held in VWM based on the frequency of accessing external information is potentially problematic, however. The underlying assumption in this interpretation is that observers put all their VWM content to use and only sample external information when no applicable information is left in VWM. Here, we question this assumption that observers resample the external world only after their VWM is totally depleted. Instead, we argue that there are other factors at play that determine when (and how often) observers decide to sample external information. One such factor could relate to the certainty of the VWM content, that is: how (un)certain is someone of their memory content? Low certainty might prompt observers to inspect the *Model* one more time instead of putting the available information to use. Another factor pertains to the cost of resampling; when accessing external information is undesirable or unfeasible, acting based on imperfect information (rather than strengthening the imperfect information) might be a more favorable option than reinspecting the *Model*. Indeed, previous studies have shown that increasing the cost of accessing external information resulted in less frequent sampling (Ballard et al., 1995; Draschkow et al., 2021; Somai et al., 2020). Together, these considerations lead us to question whether the frequency of resampling provides an accurate reflection of the information stored in VWM.

We hypothesize that in copying tasks (such as discussed above) there is more VWM content present than is measured by copying behavior alone. To test this hypothesis, we designed a copying task similar to those used in the studies described above (Ballard et al., 1995; Draschkow et al., 2021; Somai et al., 2020), but with one crucial addition to the paradigm: while creating a replica of the *Model*, observers were unpredictably interrupted, and the content of their VWM was probed. Importantly, the interruptions were timed to occur when observers made an attempt to resample the external information from the *Model* (but just before they actually did). Above chance performance on these probes would indicate that VWM is not depleted before sampling, and that more information is present in VWM than what is inferred from copying behavior alone. This finding would directly challenge the assumption that observers refer to external information only when VWM is depleted.

If it is true that the frequency of resampling cannot be assumed to reflect VWM content, then this also casts doubt on another important characteristic of naturalistic VWM use. Previous studies have shown that increasing the cost of sampling resulted in less frequent sampling (Ballard et al., 1995; Draschkow et al., 2021; Somai et al., 2020). The conclusion in those studies is that increasing the cost of sampling causes observers to load up more information in VWM. We hypothesize instead, that the reduced number of reinspections when the cost of sampling is higher do not necessarily imply a higher VWM load, but could also reflect that a larger portion of the (same amount of) VWM load is put to use, before observers choose to reinspect the *Model* (i.e., they have a stronger tendency to "give it a try"). To distinguish between these two possibilities we also manipulated the sampling cost in our experiment. A higher sampling cost (in our study) meant it was more time-consuming to reach the *Model*. In order to assess whether observers (1) load up more items in memory, or (2) squeeze out more items from memory when sampling costs increase, we also manipulated the moment when VWM content was probed: *just before* or *just after* sampling.

To summarize, in the current study we questioned (1) whether the number of items that is put to use before resampling the external world actually reflects the entire content of VWM, and (2) whether increasing the cost of external sampling leads to either loading up more content or using a larger portion of available VWM content (or both). To preface our results: we found that there was a substantial amount of useful information left in VWM when VWM was probed just before an attempt to access external information. This indicates that factors other than merely VWM content play a role in the decision to sample external information. Furthermore, we found that when the cost of sampling was increased, observers invested more time to sample external information (i.e., loading up more), *and* made more attempts to apply information before sampling again (i.e., squeezing out more of the available VWM content). These results show that changing task constraints might not (only) change the content of VWM, but (also) how it is put to use. Accordingly, we argue that in understanding and interpreting behavior in naturalistic VWM tasks, more factors must be considered than merely the capacity of VWM. To facilitate such interpretations, we conclude this article by proposing a model to describe behavior in naturalistic tasks, which require VWM use. This relatively simple model captures the continuous nature of VWM representations, the certainty of one's VWM content, and the cost–benefit analysis underlying the decision to either act on the available VWM content, or to gather more information from the external world.

2. Methods

2.1. Participants

Participants were recruited via the online platform Prolific (www.prolific.co). By applying Prolific's built-in screening tools we only included participants who (1) indicated to have normal or corrected-to-normal vision, (2) indicated to be fluent in English, (3) had an approval rate higher than 95% and (4) had not taken part in earlier pilot versions of this experiment. We set out to collect data from 80 participants (40 per between-observer condition). As there were no prior comparable studies run online to base our sample-size on, we based the current sample-size on prior experience and practical constrains (task duration, financial compensation for online participants). We included only participants for whom the data of more than 95% (i.e., 46 or more) of the 48 experimental trials was recorded. We included 88 participants to be able to run certain analyses with at least 40 participants per between-observer condition (some participants did not encounter a single probe question in some conditions, providing no relevant data for probe question analyses). All 88 included participants had completely viable data for analyses regarding copying strategies. Therefore, we included all of them for the copying strategy analyses, as we had no justifiable reason to exclude their data. The experiment complied with all ethical guidelines set out in the Declaration of Helsinki, and was approved by the Ethics Committee of the Faculty of Social and Behavioral Sciences of Utrecht University. The approval is filed under number 21-0297. The monetary reward for successful completion of the task was (the equivalent of) 6.25 GBP.

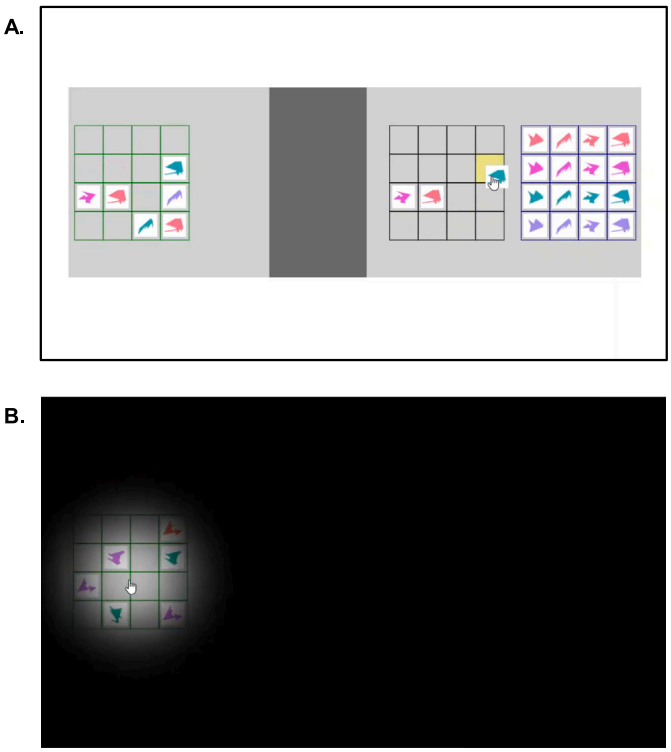

Fig. 1. Panel **A.** illustrates the layout of the copying task. The grid on the left, the *Model*, had to be recreated on the middle grid, the *Workspace*, by dragging the items from the rightmost grid, the *Resources*, and dropping them in the correct cell in the *Workspace* (also see Video S1 - osf.io/g64vz). Panel **B.** shows a still from an experimental trial, as seen by the participants. An opaque black overlay covered the display, and only a circular area was made transparent (fully transparent in the center, but less towards the edges). In the baseline aperture-speed condition this aperture was continuously centered on the current cursor position (see Video S2 - osf.io/w7zag). In the slow aperture-speed condition, the aperture moved with a reduced speed across the dark gray area towards the *Model* on the left (see Video S3 - osf.io/3z8xn). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2. Apparatus and stimuli

The experiment was programmed using the JavaScript libraries jsPsych (version 6.3.0) (de Leeuw, 2015) and Fabric.js (version 4.3.1; www.fabricjs.com), and was hosted online on the web service Cognition (www.cognition.run).

As the current experiment was conducted online, it is likely that many types of devices and displays were used. We strongly encouraged participants to conduct the experiment on a laptop or desktop computer with the use of a computer mouse. When asked, all participants reported that they conducted the experiment on a desktop or laptop device, and the majority reported that they controlled the cursor with a computer mouse (some indicated to have used a touchpad).

To account for varying sizes of the displays, we implemented a calibration procedure to end up with equal stimulus sizes: We asked participants to hold up a credit card (or any other standard sized card, commonly 8.56 cm wide) against the screen, and resize a rectangle on the screen to match the card's size. If done correctly this procedure ensured that the experiment was contained in a light gray rectangle of about 25 cm wide, and 8.5 cm high, and that the stimuli were contained in 1 by 1 cm white boxes (see Fig. 1A). We refer to the colored polygons in boxes as items.

The shapes of the stimulus set consisted of 20 polygons adopted from Arnoult (1956). In addition, we used 20 colors, selected from the HSLuv (www.hsluv.org) color space. We selected 20 equidistant hues on the color wheel, with the saturation set to 90% and luminance to 65% (see Fig. 2). Given the various displays (and lighting settings) participants likely used, the same RGB-values will have inevitably resulted in different monitor outputs for each participant in terms of both luminance and hue.

In order to track what part of the experimental display participants were looking at in this online study, we implemented a cursor-directed aperture in our task (Anwyl-Irvine et al., 2021). Effectively, this meant that a black overlay covered the view of the display, and only a circular area surrounding the location of the cursor was made transparent (see Fig. 1B and Video S2 - osf.io/w7zag). The transparency followed a Gaussian function: it was maximally transparent at the center and less transparent at the edges. The size of the aperture was set to be just large enough to make the whole *Model* visible at once. Specifically, the radius of the aperture was set to 9% of the width of the light gray experiment rectangle, and the standard deviation of the Gaussian

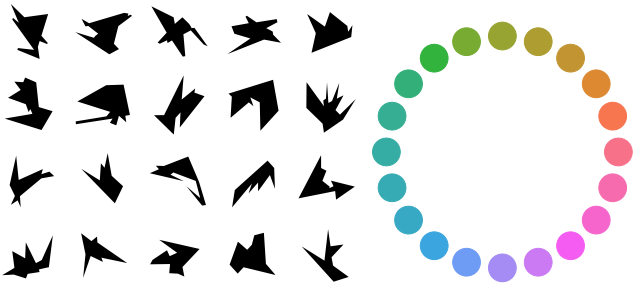

Fig. 2. The twenty shapes and twenty colors that were combined to create the stimuli in the experiment. Given 20 shapes and 20 colors, we could create 400 unique stimuli. For each trial, a random selection (*without* replacement) of four shapes and four colors was used to create 16 unique stimuli. From this pool of 16 stimuli, 6 were randomly selected (*with* replacement), and randomly positioned in the *Model* grid for each trial. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

function (for the transparency) was set to half the size of the aperture's radius. The experience of the participant is similar to looking around in a darkened environment using a focal flashlight.

To modulate the cost of sampling we manipulated the speed of the aperture to make it more time-consuming to reach the *Model*. In the slow aperture-speed condition, the aperture moved with a reduced speed across the dark gray band towards the *Model* (but not slowly back towards the *Workspace*; see Video S3 - osf.io/3z8xn). In the slow aperture-speed condition the aperture crossed the dark gray border in approximately 1.67 s, while in the baseline aperture-speed condition the aperture moved synchronously with the cursor (i.e., there was no additional delay).

2.3. Procedure

Prior to starting the experimental trials, we presented a video explaining the task, asked demographic questions, and asked participants to complete two practice trials: one without and one with the overlay and aperture. To complete an experimental trial, participants needed to recreate a model grid (*Model* for short) by dragging and dropping items from the resources grid (the *Resources*) onto the workspace grid (the *Workspace*) using the cursor. While an item was hovered above the *Workspace*, the nearest grid cell was highlighted in yellow. When an item was released on the correct cell, it snapped to the center of the cell; when released on an incorrect location the item shot back to its original location in the *Resources*.

Occasionally, participants were interrupted during a trial, and had to answer a two-alternative forced choice (2-AFC) question (see Fig. 3). We refer to these questions as "probe questions" (or "probes" for short). These probe questions showed one highlighted grid cell in an empty *Model* grid, and required participants to indicate which of two items was located in this cell of the *Model* of the current trial. Of the two given alternatives, one was the correct item, and the other item was chosen from the *Resources* of the current trial. The probed item (or more specifically: the item's location) was randomly chosen to be one of the *Model* items which had not yet been copied in the *Workspace*. Participants answered the probe question by placing their chosen item in the highlighted cell, and by indicating how confident they were about their choice. They could choose from the following options to indicate their confidence: "completely confident", "fairly confident", "somewhat confident", "slightly confident", "not at all confident". Importantly, probe questions could be initiated at two different time points: For half of participants probe questions could occur when they tried to inspect (i.e., move the aperture towards) the *Model*, but before actually seeing the *Model* (specifically, when the aperture's center crossed the right border of the dark gray band; see Video S2 - osf.io/w7zag). For the other half of participants, probe questions could occur after *Model* inspection, when participants moved the aperture towards the *Workspace*, but before actually reaching the *Workspace* (specifically, when the aperture's center crossed the left border of the dark gray band; see Video S3 - osf.io/3z8xn). To ensure that the probe questions always occurred unpredictably, there was a 1/6th chance of one occurring every time the aperture crossed the specified border (i.e., this was either immediately before a *Model* inspection, or immediately after a *Model* inspection). One exception to this rule was that a probe question could only occur *after* the *Model* was inspected at least once in every new trial, in order to prevent a question about a *Model* that was never seen. Note that the initiation of the probe questions was randomly determined (and *not predetermined*), because it is unknown—a priori—how often a participant will inspect the model. We therefore did not control the number of probe question per trial, per condition, or per participant. Participants were instructed to prioritize performing the copying task and not to focus on these probe questions too much; they were advised to simply answer the questions to the best of their knowledge whenever they occurred.

2.4. Experimental design

The main factor of interest in the current experiment was the *cost* to reach the *Model* for inspection. We manipulated this factor within-observer by adding a condition where it was more time-consuming to reach the *Model*. Participant completed 48 trials in two blocks of 24 trials. One block comprised only baseline aperture-speed trials, the other only slow aperture-speed trials. The block order was counterbalanced across participants. We opted to block aperture-speed conditions such that participants could settle on a consistent strategy for each aperture-speed condition.

Furthermore, we ensured that the manipulation of *when* probe questions could occur was balanced across participants: for half of the participants probe questions could only occur just before inspecting the *Model*, and for the other half of the participants probe questions could only occur just after inspecting the *Model*.

We were interested in several outcome measures and, in particular, how these were modulated by an increase in the cost to reach the *Model*. First and foremost, we were concerned with the performance on the probe questions *just before Model* inspection. More specifically, whether the accuracy for these questions was above chance. If so, that would show that an attempt to (re)sample the *Model* does not imply that visual working memory is depleted of useful information. Furthermore, we were interested in measures which informed us about the copying-strategies employed by participants. These measures were (1) the number of *Model* inspections per trial, (2) the duration of a single *Model* inspection, and (3) the number of errors (i.e., incorrect placements) per trial. These measures were informative about the strategies that participants chose to perform the task, and how their strategies were affected by the increased cost of sampling. Additionally, these measures allowed us to check whether the cost manipulation of our online task yielded (at least qualitatively) comparable results as previous lab-based studies (Draschkow et al., 2021; Somai et al., 2020). Finally, we were interested in how the cost manipulation affected the performance on probe question both before and after sampling.

2.5. Analysis

We conducted all our analyses using Bayesian statistics with the JASP software using the default priors wherever relevant for conducting Bayesian statistics, and always setting the seed value to 1 for reproducibility (JASP Team, 2022). We opted to use the labels suggested by Kass and Raftery (1995) for the interpretation of Bayes factors.

Regarding the probe questions *before* sampling, we determined for each participant the proportion of correct responses and, using a Bayesian one-sample *t*-test, tested whether the average proportion correct responses was higher than chance (0.5 for the current task). Furthermore, we tested whether the proportion of correct responses differed between the two probe timing conditions (*before* versus *after* sampling) using a Bayesian independent samples *t*-test. For these analyses we used data of the 80 participants who encountered at least one probe question in each aperture-speed condition.

Next, we analyzed and compared the three outcome measures of the copying task described above (i.e., number of *Model* inspections, *Model* inspection durations, and number of erroneous item placements) in the two aperture-speed conditions for each participant. For each measure we performed a Bayesian repeated measures analysis of variance (RM ANOVA), to test whether there was a difference between the two means in the baseline aperture-speed condition and the slow aperture-speed condition (within-observers). The factor probe timing was included to account for possible systematic differences in strategy that could emerge as a result of the different probe timings (before or after *Model* inspection, manipulated between-observers). For these analyses of copying strategies we used data of all 88 included participants.

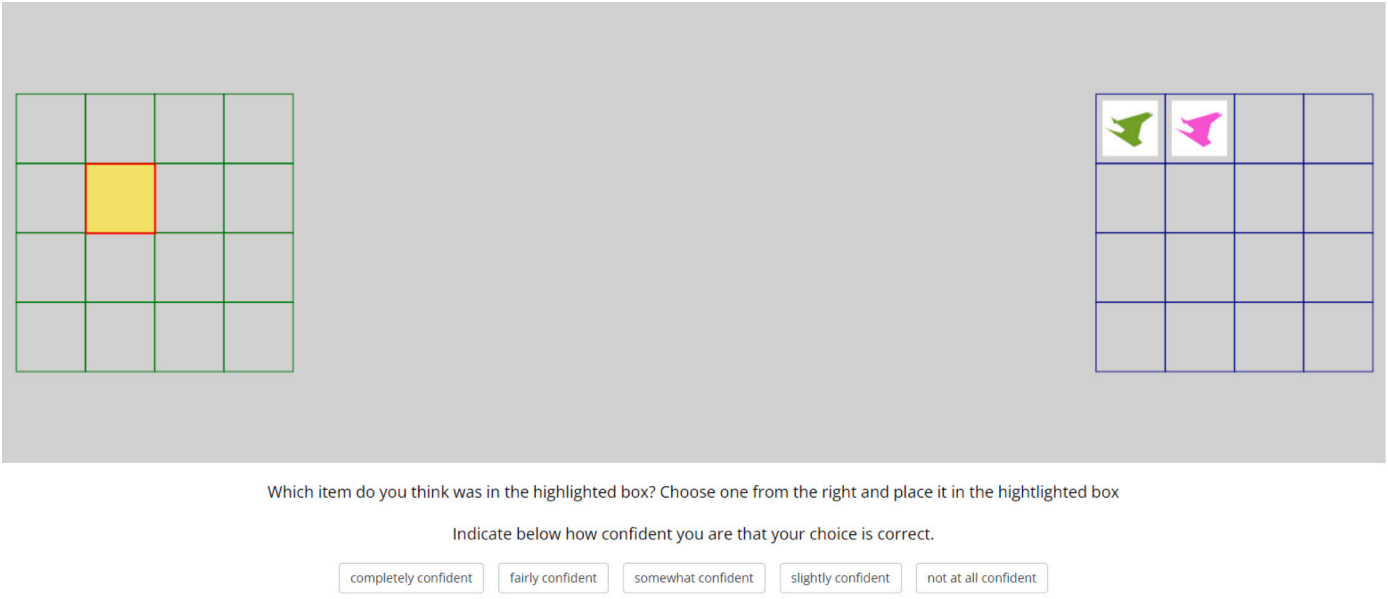

Fig. 3. Example of a probe question. One grid cell was highlighted, and the participant was asked to indicate which of the two items belonged to that cell in the *Model* of the current trial, and to indicate how confident they were that their choice was correct. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Finally, to test for the interaction between sampling cost (high and low) and probe timing (just before or just after sampling) on probe question performance, we performed a Bayesian RM ANOVA. Here we used again the data of those 80 participants, who encountered one or more probe question in each aperture-speed condition.

For all Bayesian RM ANOVAs we performed an analysis of effects across matched models (Mathôt, 2017). This analysis compares models that contain the effect in question to equivalent models stripped of the effect. Here, the "Inclusion Bayes Factor" (BF_{incl}) reflects the amount of evidence specifically for the (main or interaction) effect in question. We have uploaded the complete statistical results (e.g., model comparisons for RM ANOVAs, etc.) on the OSF-platform, which can be accessed via the link: osf.io/pkxdc/.

3. Results

3.1. VWM content before and after sampling

As a consequence of our design choices (and participants' sampling behavior), probe questions were not equally distributed across participants and experimental conditions. In the baseline aperture-speed condition participants encountered on average 15.91 (SD = 9.13) probe questions, and in the slow aperture-speed condition the average was 8.58 (SD = 5.33). Regarding the accuracy on the probe questions just *before* sampling (in the baseline aperture-speed condition) we found that participants answered, on average, 73% (SD = 17) of the probe questions correctly (see Fig. 5). We found decisive evidence ($BF_{+0} = 4.51 \times 10^{7}$) that this accuracy was higher than chance (which was 50%). A higher than chance accuracy when probed before *Model* inspection suggests that there was *at least some* information in VWM when the decision was made to retrieve information by inspecting the *Model*. The accuracy on probe questions (again, in the baseline aperture-speed condition), but just *after* sampling, was on average, 82% (SD = 11). There was substantial evidence ($BF_{10} = 9.35$) that the accuracies between the probe timing conditions (i.e., before versus after sampling) differed. This difference suggests that the VWM load was higher after sampling the model than before sampling the model (which is unsurprising, but serves as a sanity check for the usefulness of the probe questions in measuring differences in VWM load).

3.2. Copying strategies

3.2.1. Sampling frequency

We found that, on average, participants made 4.83 (SD = 2.08) *Model* inspections per trial in the baseline aperture-speed condition, while they made 2.66 (SD = 1.12) *Model* inspections per trial in the slow aperture-speed condition (see Fig. 4). As there were always six items to be copied, participants correctly placed ($\frac{6}{4.83}$ =) 1.24 items per *Model* inspection in the baseline aperture-speed condition, and ($\frac{6}{2.66}$ =) 2.26 items per *Model* inspection in the slow aperture-speed conditions. The analysis of effects (across matched models) showed decisive evidence ($BF_{incl} = 1.12 \times 10^{16}$) that aperture-speed condition modulated the number of *Model* inspections. Qualitatively and quantitatively, these results are in line with previous findings of object copying tasks: Roughly one item is copied per *Model* inspection under "normal" viewing conditions, and this number almost doubles when accessing the *Model* is made more costly (Somai et al., 2020). Furthermore, we found no evidence for a main effect of probe timing ($BF_{incl} = 1.16$), and substantial evidence *against* an interaction effect of probe timing and aperture-speed ($BF_{incl} = 0.27$) on the number of *Model* inspections. The absence of these effects of probe timing merely show copying strategy was not affected by the timing of the probe questions (before or after a *Model* inspection).

3.2.2. Sampling duration

Regarding the durations of *Model* inspections, we found that in the baseline aperture-speed condition, participants on average inspected the *Model* for 3.09 s (SD = 2.57) before returning to the *Workspace*. In the slow aperture-speed condition, they took 6.20 s (SD = 3.34) per *Model* inspection (see Fig. 4). Again, the analysis of effects (across matched models) showed decisive evidence ($BF_{incl} = 8.77 \times 10^{20}$) that the mean duration of each *Model* inspection was modulated by aperture-speed conditions. Clearly, participants took more time to inspect the *Model* when accessing it was made more difficult. As was the case for the number of model inspections, there was no evidence that probe timing affected *Model* inspection durations ($BF_{incl} = 0.37$), and there was substantial evidence against an interaction effect of probe timing and aperture speed ($BF_{incl} = 0.28$) on the *Model* inspection durations. This indicates that the timing of the probe questions did not influence the sampling durations in any way.

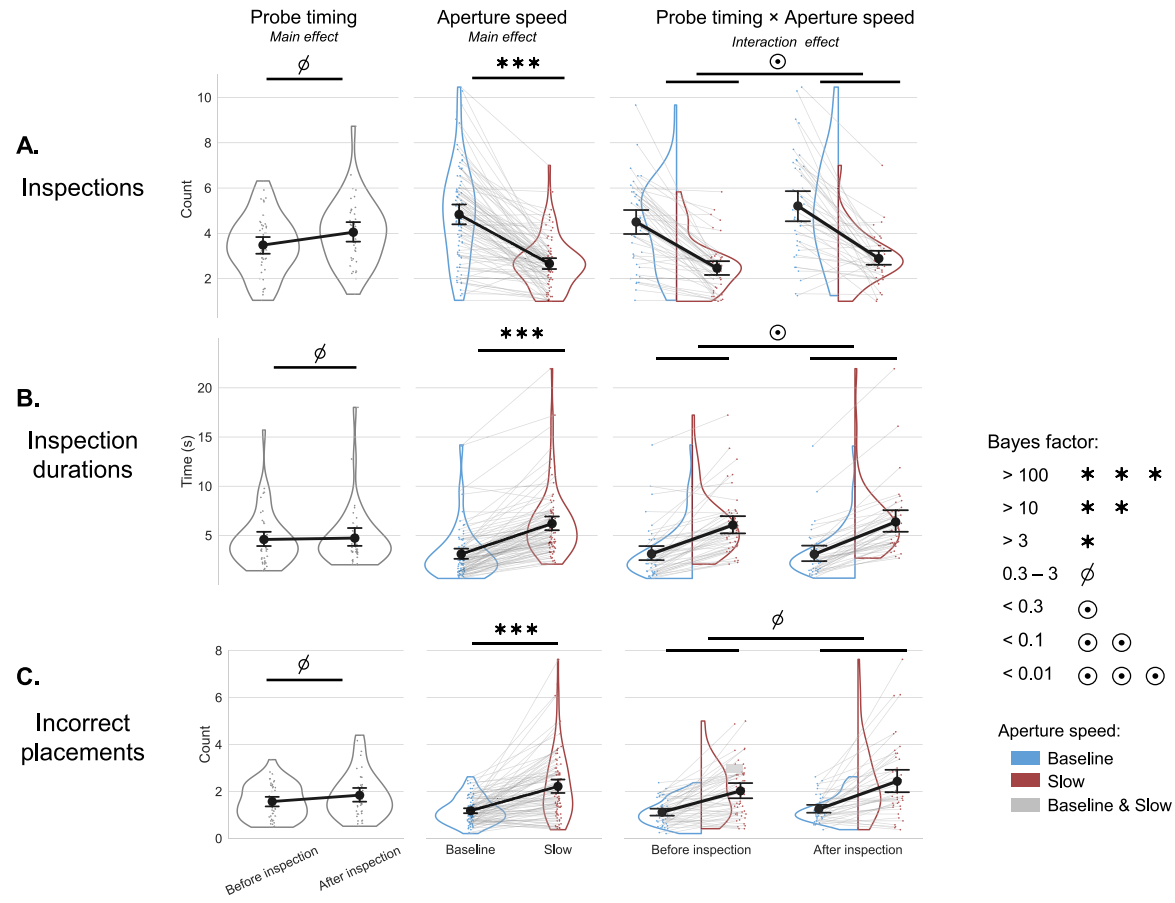

Fig. 4. The results for the three copying strategy measures. **Panels A., B, and C.** respectively show the participant means of the number of *Model* inspection inspections per trial, the duration of single inspections, and the number of incorrect placements per trial. These measures are split across the two probe timing conditions, the two aperture-speed conditions, and both conditions together in the three columns. The results of Bayesian repeated measures ANOVAs (inclusion Bayes factors (BF_{incl}) for analyses of effects across matched models) are shown with symbols above the plots. BFs larger than 3 are marked with asterisks (∗), signifying substantial (or more) evidence *in favor of* the effect in question; BFs between 0.3 and 3 are marked with the null sign (ø), signifying no conclusive evidence in favor of or against the effect; BFs smaller than 0.3 are marked with circled dots (⊙), signifying substantial (or more) evidence *against* the effect. In all plots, the small (blue, red and gray) dots represent individual participants' means, the large dots represent group means, and the error bars represent (bootstrapped) 95% confidence intervals. The violin plots (i.e., kernel density plots) show the spread of the participants' means. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2.3. Errors

On average, participants made about 1.18 (SD = 0.54) errors per trial in the baseline aperture-speed condition, while they made about 2.22 (SD = 1.40) errors per trial in the slow aperture-speed condition (see Fig. 4). Here too, the analysis of effects (across matched models) showed decisive evidence ($BF_{incl} = 1.71 \times 10^9$) that there was a main effect of aperture-speed condition. More errors in the high sampling cost condition indicated that participants made more attempts to put VWM contents to use. Taking into account the number of *Model* inspections, we find on average that participants made ($\frac{1.18}{4.83}$ =) 0.24 errors per inspection in the baseline aperture-speed condition, while they made ($\frac{2.22}{2.66}$ =) 0.83 errors per inspection in the slow aperture-speed condition. This means that more than 3 times as many errors were made per inspection in the slow aperture-speed condition compared to the baseline condition. Also for the number of errors, there was no evidence for a main effect of probe timing ($BF_{incl} = 0.55$) and no evidence for an effect of the interaction of probe timing and aperture speed condition ($BF_{incl} = 0.35$). This indicates that there was no evidence that the timing of the probes affected the number of errors that participants made during a trial.

3.3. VWM content before and after sampling depending on sampling cost

To reiterate, the probe question performance in the *baseline* aperture-speed condition: the accuracy on probes before *Model* inspection was 73% (SD = 17), while after *Model* inspection the accuracy was 82% (SD = 11). The probe question performance in the *slow* aperture-speed condition was 75% (SD = 17) before sampling, while after sampling the accuracy was 85% (SD = 11). The analysis of effects (across matched models) showed strong evidence ($BF_{incl} = 62.41$) that there was a main effect of probe timing, showing that—unsurprisingly—participants had a higher accuracy on probe trials just after model inspection compared to just prior to model inspection. We found no evidence, however, for a main effect of aperture-speed condition ($BF_{incl} = 0.52$) and substantial evidence against an interaction effect of probe timing and aperture speed condition ($BF_{incl} = 0.23$). These results do not provide conclusive evidence in favor of (or against) the hypothesis that more information is loaded up in VWM, or that VWM is being depleted more, in the high compared to low sampling cost condition. As such, we have no evidence that participants loaded up more information in VWM in the high sampling cost condition, despite inspection durations being nearly twice as long as in the low sampling cost condition. At the same time, we have no evidence either that participants had less VWM content left when they decided to reinspect the *Model* in the high

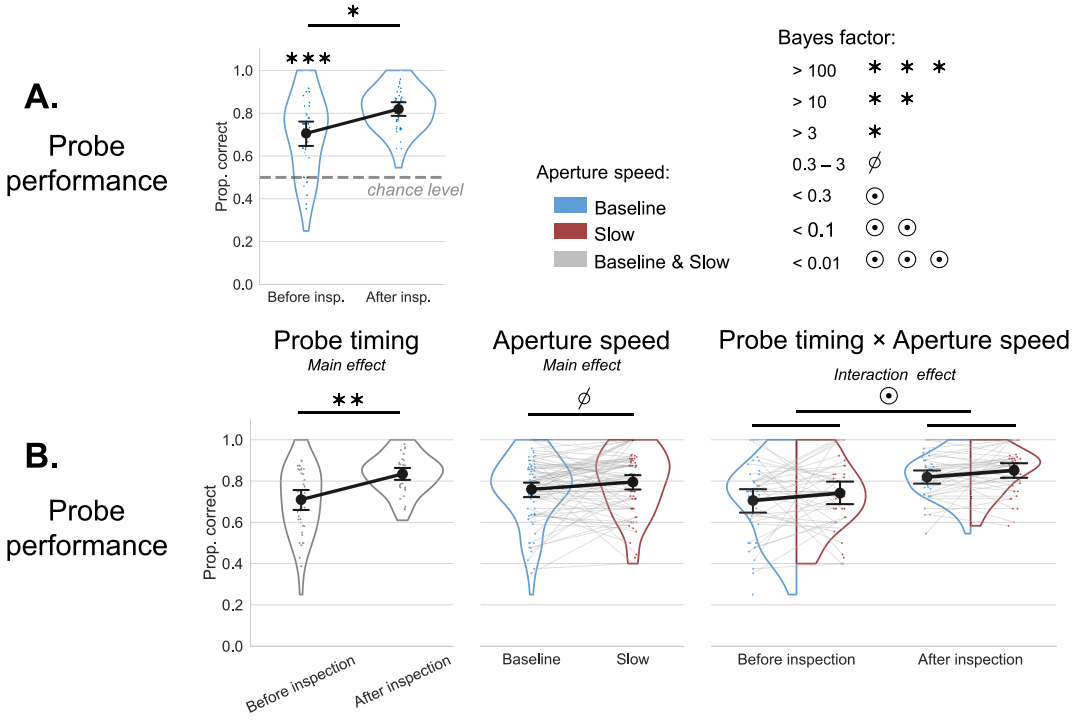

Fig. 5. The performance on probe questions. **Panel A.** shows the proportion of correct responses on probe questions only in the baseline aperture-speed condition split across the two probe timing conditions. The results of a Bayesian one-sample *t*-test (testing whether the proportion correct responses on probes before sampling is higher than the chance level of 0.5) and of a Bayesian two-samples *t*-test (testing whether the performance on the probes is different in the two probe timing conditions) are shown with symbols above the plots. Bayes factors larger than 3 are marked with asterisks (∗), signifying substantial (or more) evidence *in favor of* the effect in question; Bayes factors between 0.3 and 3 are marked with the null sign (ø), signifying no conclusive evidence in favor of or against the effect; Bayes factors smaller than 0.3 are marked with circled dots (⊙), signifying substantial (or more) evidence *against* the effect. **Panel B.** shows the performance on the probe questions, split across the two probe timing conditions, the two aperture-speed conditions, and both conditions together. The results of Bayesian repeated measures ANOVAs (inclusion Bayes factors for analyses of effects across matched models) are shown with symbols above the plots following the convention described above. In all plots, the small (blue, red and gray) dots represent individual participants' means, the large dots represent group means, and the error bars represent (bootstrapped) 95% confidence intervals. The violin plots (i.e., kernel density plots) show the spread of the participants' means. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sampling cost condition, despite participants making almost twice as many errors (i.e., squeezing out more information from VWM) as in the low sampling cost condition.

3.4. Alternative load and squeeze measures

While the probe questions were sensitive enough to answer one important question in this study: "Is there information left in VWM before an inspection?", they were not sensitive enough to conclusively answer the second main question of this study: "Was more information encoded in VWM ('loading') or was more information applied from VWM ('squeezing') when sampling cost were higher?". Therefore, we devised two post hoc measures based on copying behavior that can directly address this question. For the measure of how much information was loaded in VWM after a *Model* inspection, we took the number of successive correct placements immediately after *Model* inspection. To elaborate, we argued that a placement sequence (directly after an inspection) such as 'Correct–Correct–Incorrect–Incorrect' (which would have a load of 2) results from a better representation of the *Model* and a higher VWM load than the placement sequence 'Correct–Incorrect–Incorrect–Correct' (which has a load measure of 1, even though it has the same number of correct and incorrect placements). With this new measure for VWM load after sampling, we found decisive evidence ($BF_{10} = 2.19 \times 10^{8}$) that the load is higher in the slow aperture-speed condition (Mean = 2.00, SD = 0.91) than in the baseline aperture-speed condition (Mean = 1.32, SD = 0.68; see Fig. 6A). This result suggests that generally more information was loaded up in VWM when the cost of sampling was higher (and when inspection durations were longer, as reported earlier). Note that this reasoning is based on the assumption that the items with the best representations are attempted first (one might reason instead that the best items are kept for last). To verify that the best remembered items were placed first, we computed the proportion of correct placements for every N^{th} placement after a *Model* inspection. We indeed found that the first placement has the highest proportion of correct placements (Mean = 0.88 SD = 0.07), and the proportion of correct placements decreases with every successive placement: there was decisive evidence ($BF_{incl} = 3.74 \times 10^{15}$) for a main effect of ordinal placement (see Fig. 6B). This finding confirms that participants attempt to place their best representations first and then work their way down towards the worst representation.

Following this same line of reasoning, we can use the number of placement attempts *after* the aforementioned streak of correct placements as a measure of 'squeezing'. We reasoned that more attempts after an incorrect placement implies a tendency to apply more VWM content (albeit of a lesser quality). To illustrate with the previous two examples, the sequence 'Cor. - Cor. - Incor. - Incor.' would have a squeeze measure of 2, while the sequence 'Cor. - Incor. - Incor. - Cor.' would have a squeeze measure of 3. The results showed decisive evidence ($BF_{10} = 2.69 \times 10^{16}$) that there were more attempts after the initial streak of correct placements in the slow aperture-speed condition (Mean = 1.59; SD = 0.85), than in the baseline aperture-speed condition (Mean = 0.55; SD = 0.52; see Fig. 6C). This result shows that when

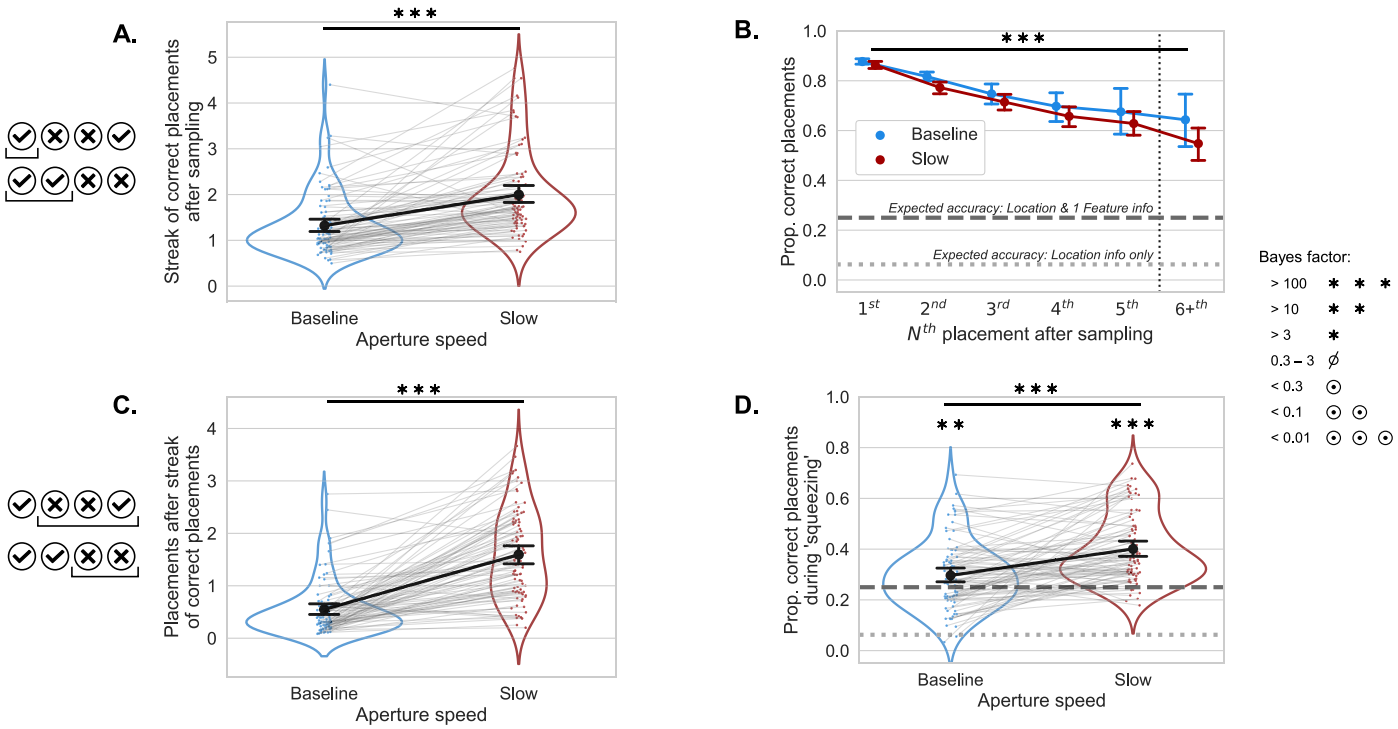

Fig. 6. Data of the post hoc analyses for alternative measures for load and squeeze measures. **Panel A.** shows the number of successive correct placements immediately after sampling, separately for the baseline and slow aperture-speed conditions. **Panel B.** shows the proportion of correct placements for the N^{th} placement after inspection, separately for the two aperture-speed conditions. The dashed dark-gray line represents the expected performance for placements which are based on knowledge of the location and one feature (either color or shape) of the item in question (i.e., 1/4). The dotted line represents the expected performance if only the location of an item was known (i.e., 1/16). **Panel C.** shows the number of placement attempts after the initial streak of correct placements (i.e., the number of placements following those of Panel A.). Note that the streak of incorrect placements is independent of the streak of correct placements, since the total length of the sequence is unbound. **Panel D.** shows the proportion of correct squeeze placements (the placement sequences of Panel C.) for the two aperture-speed conditions. The dashed and dotted lines represent the same as in panel B. In all plots, the small connected (blue and red) dots represent individual participants' means, the large dots represent group means, and the error bars represent (bootstrapped) 95% confidence intervals. The violin plots (i.e., kernel density plots) show the spread of the participants' means. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the cost of sampling was higher, more attempts were made to put the available VWM information to use.

An implicit assumption here was that the 'squeeze' placements were based on information in VWM, rather than being pure guesses. To verify that this was indeed the case we computed the proportion of correct 'squeeze' placements. The data showed that participants on average indeed had a relatively high proportion of correct placements when squeezing: 0.30 (SD = 0.13) in the baseline aperture-speed condition, and 0.40 (SD = 0.14) in the slow aperture-speed condition (see Fig. 6D). This performance is much higher than would be expected if placements were pure guesses (i.e., pure chance level is approximately 0.04). We even found strong evidence ($BF_{+0} = 34$ and $BF_{+0} = 20.02 \times 10^{14}$, for the baseline and slow aperture-speed conditions respectively) that performance is better than what would be expected if a placement was based on knowledge of an item's location and one feature (i.e., the color or shape), which was 0.25. As such, even the attempts after an initial streak of correct placements, are based on quite strong useful information in VWM. Taken together, when sampling costs increase, participants tend to both load up more information in memory (i.e., encode more items), and squeeze out more information from memory (applying less certain items) before resampling the model.

4. Discussion

In the current study, we investigated the use of visual working memory (VWM) in a naturalistic setting in which relevant visual information remained externally available for reinspection. Our primary goal was to examine whether the amount of information put to use from VWM is an accurate index of the full amount of information stored in VWM. We therefore implemented a copying task in which the content of VWM was infrequently probed with a 2-alternative forced choice (2-AFC) question when an attempt was made to sample information from the external world. Our results show that, under these circumstances, VWM actually contains more information than is put to use in a goal-oriented task requiring VWM usage (such as the copying task). This finding implies that less VWM content is utilized than is initially encoded, and therefore also implies that the amount of information that is utilized is a poor index of VWM load. One important consideration about probe question performance is that any information sufficient to perform above-chance on a probe question (e.g., knowing only the color), is also useful information for the copying task (i.e., knowing an item's color provides a fourfold increase in the chance of selecting the correct item). Put differently, to perform above chance on the probe question, a participant needs *some* information about the model. In turn, any information about the model is useful knowledge when deciding, for example, which item to pick up or where to place it. Indeed, our data shows that participants did act on partial or imperfect information, since participants regularly made incorrect placements. A second aim was to test a prediction stemming from the following observation in naturalistic VWM use: when the cost of sampling external information increases, external information is sampled less often (Draschkow et al., 2021; Somai et al., 2020). We reasoned that less frequent sampling could be due to (1) encoding more information with each sample, or (2) applying a larger portion of what is encoded. Note that specifically the latter hypothesis was warranted by our primary finding (i.e., not all VWM content is applied before resampling). The data show that when

the sampling cost was higher, participants spent more time viewing the *Model* during each inspection (suggesting they attempted to encode more information in VWM), in line with observations from earlier studies (Draschkow et al., 2021; Somai et al., 2020). Interestingly, our data also show that participants made more errors when the cost of sampling was higher. More errors suggest that participants indeed tried harder to put more of the VWM content to use, and that VWM should have been depleted more. Despite the clear difference in strategy (e.g., longer inspections and more errors), we found no conclusive evidence that the probe question performance was modulated by the cost of sampling. Both when probes occurred before sampling, and when probes occurred after sampling, we did not find (conclusive) evidence in favor of a different performance in the low sampling cost condition versus the high sampling cost condition. This was surprising, as a similar probe performance *after* sampling would suggest that VWM contained a similar amount of information in the high versus low sampling cost condition. It would be equally surprising if there was a similar probe performance *before* sampling, because this would suggest that information in VWM was depleted to a similar extent in the high versus low sampling cost condition. In contrast, two post hoc analyses of the copying behavior did provide definitive conclusions where the probe question analyses did not. When we took the number of successive correct item placements immediately after sampling as a measure for VWM load, the data showed that more information was encoded when sampling cost was higher. Also, when we took the number of (either correct or incorrect) placements following this first streak of correct placements as a measure for how much information was squeezed out of VWM, we found that, in the high sampling cost condition, more VWM content was depleted before resampling.

Generally, our findings relating to copying strategies and strategy changes (when sampling becomes more costly) are in line with previous studies employing copying tasks in particular (Ballard et al., 1995; Draschkow et al., 2021; Somai et al., 2020), and with the literature on cognitive offloading in general (Hu et al., 2019; Risko & Gilbert, 2016). As the name already suggests, cognitive offloading refers to the act of reducing the cognitive demand of a task or activity: using a spoken navigation tool when driving a car, or going through a (readily accessible) shopping list one item at a time when doing groceries. In tasks where putting less tax on (visual) working memory is a viable option, observers prefer to do so (Risko & Dunn, 2015). In the current study frequent sampling can very well be interpreted as a cognitive offloading strategy: copying items one at a time (thus sampling often) puts less cognitive demand on VWM. Note that the act of sampling comes at a cost, since sampling takes time and effort (e.g., redirecting attention, making an eye or body movement, etc.). When deciding what to do next, observers therefore need to weigh the benefit of cognitive offloading against the cost of sampling the external world. Our results show that participants indeed engaged in cognitive offloading, by sampling approximately one time for each item that they copied. Our results also show a marked decrease in cognitive offloading when sampling was made more costly (i.e., copying multiple items at a time when sampling costs were higher). This implies that the task context determines how worthwhile observers consider cognitive offloading to be.

With this study we add two important new findings to our understanding of goal-oriented VWM use. Using a naturalistic task setting, in which relevant information remains externally available, we first show that observers often reinspect the external world despite still having relevant information available in VWM. In other words, the frequency of sampling from the outside world does not necessarily reflect the amount of information that is stored in VWM. This finding has important implications for future studies employing copying tasks and other similar paradigms. When interpreting copying behavior, or—more generally—any other (goal-oriented) actions based on VWM, researchers should consider how much information might actually be in VWM and which part of that information is put to use. Second, we show three distinct changes in strategy when retrieving task-relevant information becomes harder: observers (1) sample less frequently from the external world, but (2) spend more timing doing so, and (3) make more attempts to put VWM content to use (thus leading to more errors). This third change of strategy is particularly interesting, as it has not received much interest in VWM research, yet has a potentially large effect on estimates of VWM capacity. In line with the theory of resource models of VWM (Bays & Husain, 2008; Ma et al., 2014; Wilken & Ma, 2004), we suggest that errors made in VWM tasks are not the result of random guessing. Rather, as we showed in a post hoc analysis, when the task becomes harder, the number of errors increase as actions are based on more noisy or uncertain information.

To capture VWM use during goal-directed behavior (such as in the copy-task, employed here), we propose a theoretical model that describes how we act based on task context and certainty of VWM content during day-to-day behavior. We model certainty of VWM representations as a continuous function across stimulus space (i.e., feature dimensions like location, color, shape, etc.). A similar type of modeling of VWM content is also applied by the target confusability competition (TCC) model of visual memory (Schurgin et al., 2020), and other continuous resource models of VWM (Bays & Husain, 2008; Ma et al., 2014; Wilken & Ma, 2004). The novelty of our proposed model lies in the addition of a flexible, context-dependent decision threshold, which we call an "action threshold" (see Fig. 7). This action threshold allows to model what information in VWM is acted on and what information is not. The threshold reflects what information is deemed worthwhile to act on, given the current situation. If the certainty of a VWM representation surpasses the action threshold, that piece of information will be put to use. If the certainty is below the action threshold, it will not be put to use. The action threshold we describe is reminiscent of the *criterion* of signal detection theory (Banks, 1970; Macmillan & Creelman, 2004). Much like the criterion, the action threshold determines how liberally (or conservatively) a decision is made to utilize information. From this model it becomes evident that there can be information which can be reported upon (forced) request, but might not be put to use when unfavorable in a certain task context. This model also explains how in some situations, given the same VWM content, one can act differently. For example, if one's action threshold is high, only very certain information will be put to use. This might be the case when sampling new information from the world is relatively easy: an effective strategy then (which does not tax VWM too much) would be to sample very often and only act on very certain information, lowering the chance of errors. Conversely, when their action threshold is low (e.g., because sampling information from the world is costly), observers might adopt a more liberal tendency to put VWM content to use, even when information is very uncertain. Acting on less certain information would consequently lead to more errors. This is precisely what we found experimentally: frequent sampling, with few errors when cost of sampling was low; but less sampling and more errors when sampling costs were high. In situations where a response is forced (where there is no possibility to improve VWM before responding), the action threshold is lowered completely, and essentially any information (no matter how weak) will be applied. In such cases, the model becomes equivalent to the TCC model (Schurgin et al., 2020). Our proposed model also generalizes to all other situations where actions or decisions are based on (working) memory. In all such cases, the expected cost and benefits are weighed to produce an appropriate threshold for the required certainty of information for it to be worthwhile to act on. To promote a more intuitive understanding of our model, we metaphorically describe our model as an inundated landscape (take another look at Fig. 7). We believe that this metaphor facilitates reasoning and discourse about VWM-guided behavior, by visually capturing the key aspects of our model (i.e., certainty of VWM content and the action threshold).

Our proposed model can also elegantly reconcile the discrepancy between typical estimates of maximum VWM capacity (i.e., about 4 items), and the much lower VWM capacity utilized (about 1–2 features)

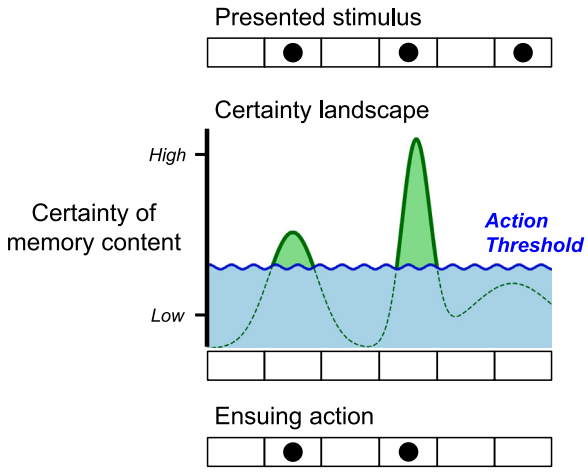

Fig. 7. The proposed model describing visual working memory (VWM) usage, based on certainty of information and an "action threshold". This action threshold describes the minimal level of certainty of information that is deemed sufficient to warrant action. The action threshold depends on a balance between costs (of sampling or action) that depend on the current task context and constraints (but is independent of memory strength). To illustrate this framework, a basic case is provided in the figure above: After inspecting the positions of three items in an external stimulus, the memory of their locations can be visualized as a continuous function, where the certainty of items is plotted across the locations (the curve in green; metaphorically referred to as a mountainous landscape). If the certainty surpasses the action threshold (in blue; metaphorically referred to as the sea level), that piece of information will be put to use. Conversely, if the certainty remains below the action threshold, no action will follow (and observers might opt to resample the external world instead). In summary, the peaks of the certainty landscape that rise above sea level (i.e., the islands) are acted upon, whereas submerged hills are not. Depending on task settings the sea level might rise or fall, thereby submerging islands or exposing submerged hills respectively.

as reported in goal-oriented paradigms. Paradigms designed to measure VWM capacity usually encourage observers to load up VWM maximally, and—more importantly—probe the VWM contents with forced-choice responses. From the perspective of our model, the forced nature of these paradigms causes the action threshold to be lowered completely, so that any information that is available in VWM—no matter how uncertain—is acted on. This results in a high estimate for VWM load or capacity. In contrast, in goal-oriented (or naturalistic) paradigms, the amount of information that is loaded up in VWM, and whether or not it is utilized, depends on the observer's cost–benefit analysis. In the latter paradigms some content will be acted upon, while some content will remain below threshold and will therefore not be reflected in self-initiated behavior. This in turn can result in lower estimates of VWM *use*, while VWM load is potentially just as high as in forced-choice paradigms. Considering an action threshold can thus resolve the diverging estimates of VWM load in different paradigms. In terms of the inundated landscape metaphor: The goal-oriented paradigms (Ballard et al., 1995; Draschkow et al., 2021; Somai et al., 2020) only count the number of islands rising above sea level, disregarding any landscape hidden below sea level. Typical forced-choice paradigms that measure maximum capacity however, disregard the presence of a sea level altogether, and treat the entire mountain landscape equally, irrespective of whether hills rise above sea level or not. At the risk of repeating ourselves, we stress the relevance of simultaneously considering both the landscape and the sea level (i.e., the VWM content and the action threshold) in understanding working memory use during day-to-day behavior.

Reflecting on the probe questions, we remark that it is curious that the performance was not affected by the cost of sampling. We observed that when the cost of sampling increased, participants made half the number of inspections to copy one *Model*, which effectively means they copied twice as many items per inspection. Being able to apply more information per inspection can either be explained by loading up more in VWM beforehand (soaking up more information), or applying more from VWM (squeezing out more information), or both. Since the probe question performance is a direct measure of VWM content, we expected that an increased sampling cost would produce better performance on probe question *after sampling* (if more was soaked up), or worse performance *before sampling* (if more was squeezed out). Surprisingly, we found neither. We think the most parsimonious explanation is that the measure of VWM content (i.e., probe question performance) was not sensitive enough to detect an effect of the cost of sampling. Although more probe questions might have yielded more conclusive results, the number of probe question was constrained by several necessary choices in the experimental design: First, initiation of probe question, either immediately before or immediately after inspection, was purposefully set to be determined at random on an inspection-by-inspection basis (because it was unknown—a priori—how often a participant would inspect the model). Therefore, it was impossible to attain a fixed number of probes per participant or per condition. Second, including too many probe questions would have interfered with the main task performance. Finally, limitations imposed by reasonable task durations and financing online participants restricted us from collecting much more probe question data. Yet, it should be noted that the probe question accuracy was sensitive enough to measure a difference in VWM content before sampling compared to after sampling (reflecting that—unsurprisingly—participants put VWM content to use while they performed the task). As such, the probe questions were sensitive enough to pick up on relatively large changes in VWM content, but might be not sensitive enough to pick up on more subtle changes in VWM content caused by a change in sampling cost. To rule out that small numbers of probe questions per participant might have impacted our results, we redid the probe question analyses including only those participants who encountered at least 5 or at least 10 probe questions per condition. The results of these analyses (which can be found at: osf.io/pkxdc) yielded the same conclusions as the results reported above, which reassured us that our results for this analysis were not affected by the choice for a particular inclusion threshold.

Another potential issue regarding the probe questions is that participants might have adapted their copying behavior, because they anticipated these questions and wanted to perform well on them. This issue is indeed inherent to any paradigm with such probes. We tried to minimize this undesired effect by (1) explicitly instructing participants that only the main task was relevant and needed to be prioritized, (2) keeping the number of probe questions low, and (3) ensuring the probe questions appeared completely unpredictably. When asked about the probe questions afterwards the majority of participants indicated to have put either no or little effort in changing their strategy just to perform well on the probe questions. Nevertheless, we suggest that future studies might want to consider more sensitive and less disruptive methods of probing VWM content during naturalistic VWM tasks.

5. Conclusion

Our findings showed that there was more information in VWM than was put to use when studying VWM in a naturalistic task in which visual information always remained available for inspection. Furthermore, we showed that when sampling was made more costly, the copying strategy changed in three distinct ways: (1) information was sampled less frequently, (2) information was inspected longer, and (3) more attempts were made to use whatever information was in VWM before sampling again. Using measures for VWM load, derived from copying actions, we found that when sampling became more costly, more information was loaded up in VWM and more information is squeeze out of VWM. We argue that actions in such naturalistic VWM tasks are based on the certainty of information currently in memory, and on a cost–benefit analysis of "acting on what is currently known" versus "improving the internal information". Our findings also highlight that, to better understand how VWM is employed in real life, human behavior needs to be investigated in paradigms that both

track the actions *and* reveal the contents of VWM. In order to advance our understanding of VWM-guided behavior, we propose a model that parsimoniously captures the continuous nature of VWM, as well as the influence of task context—which determines what part of VWM content is put to use, and what part is not.

CRediT authorship contribution statement

Andre Sahakian: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Surya Gayet:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition. **Chris L.E. Paffen:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Stefan Van der Stigchel:** Conceptualization, Methodology, writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data, analyses, and experimental code are uploaded on the OSF platform, and accessible via the link: https://osf.io/pkxdc/. We believe this study adds a rich dataset to the field. This dataset enables researchers to test various predictions and investigate many open questions which are beyond the scope of the current study. For example, investigating scan paths of the aperture on the *Model* could yield valuable information about encoding strategies and visual foraging for information (Manohar & Husain, 2013; Wolfe, 2013). As another example, the confidence judgments we collected with the answers to probe questions (but did not further discuss in the current study) can provide insights in observers' metacognitive abilities and how confidence in one's memory relates to their behavior (see the https://osf.io/pkxdc/ OSF repository for descriptive and statistical analyses of this measure). These examples highlight the rich and diverse nature of the dataset and the potential forthcoming research.

Acknowledgments

All authors have read and approved the final manuscript.

Funding

This project was supported by an ERC Consolidator Grant [grant number ERC-CoG-863732] to Stefan Van der Stigchel, and a Veni grant from Netherlands Organisation for Scientific Research (NWO) [grant number: Vl.Veni.191G.085] to Surya Gayet.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105381. Screen recordings showing one practice trial, one trial in the baseline, and one in the slow aperture speed condition.

References

Anwyl-Irvine, A. L., Armstrong, T., & Dalmaijer, E. S. (2021). MouseView.js: Reliable and valid attention tracking in web-based experiments using a cursor-directed aperture. *Behavior Research Methods*, 1–25.

Arnoult, M. D. (1956). Familiarity and recognition of nonsense shapes. *Journal of Experimental Psychology*, *51*(4), 269–276.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation - Advances in Research and Theory*, *8*(C), 47–89.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*, 66–80.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *74*, 81–99.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*, 851–854.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

Draschkow, D., Kallmayer, M., & Nobre, A. C. (2021). When natural behavior engages working memory. *Current Biology*, *31*(4), 869–874.e5.

Hu, X., Luo, L., & Fleming, S. M. (2019). A role for metamemory in cognitive offloading. *Cognition*, *193*, Article 104012.

JASP Team (2022). JASP (version 0.16.1)[computer software].

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281, 1997 390:6657.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*, 347–356.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press.

Manohar, S. G., & Husain, M. (2013). Attention as foraging for information and value. *Frontiers in Human Neuroscience*, 711.

Mathôt, S. (2017). Bayes like a baws: Interpreting Bayesian repeated measures in JASP. https://www.cogsci.nl/blog/interpreting-bayesian-repeated-measures-in-jasp Accessed: 2022-07-26.

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.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*, 676–688.

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*, 1156–1172, 2020 4:11.

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*, *122*, 108–114.

Van der Stigchel, S. (2020). An embodied account of visual working memory. *Visual Cognition*, *28*, 414–419.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*, 11.

Wolfe, J. M. (2013). When is it time to move to the next raspberry bush? Foraging rules in human visual search. *Journal of Vision*, *13*, 10.