Visual working memory and eye movements in context

How we make use of the external world

Alex J. Hoogerbrugge

Visual working memory and eye movements in context: How we make use of the external world - Alex J. Hoogerbrugge

Visual working memory and eye movements in context:
How we make use of the external world

Alex Jan Hoogerbrugge

ISBN/EAN: 978-90-835486-6-1
DOI: https://doi.org/10.33540/2960

Design: Alex J. Hoogerbrugge
Layout: Alex J. Hoogerbrugge
Printing: proefschriftenprinten.nl

Visual working memory and eye movements in context:
How we make use of the external world

Visueel werkgeheugen en oogbewegingen in context:
Hoe we gebruikmaken van de externe wereld

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. ir. W. Hazeleger, ingevolge het besluit van het College voor Promoties in het openbaar te verdedigen op

maandag 16 juni 2025 des ochtends te 10.15 uur

door

Alex Jan Hoogerbrugge

geboren op 21 maart 1994
te Capelle aan den IJssel

Promotoren:

Prof. dr. S. van der Stigchel
Dr. T. C. W. Nijboer

Copromotor:

Dr. C. Strauch

Beoordelingscommissie:

Prof. dr. F. P. de Lange
Prof. dr. D. L. Oberski
Prof. dr. S. F. te Pas
Prof. dr. M. J. van der Smagt
Prof. dr. J. Theeuwes

Dit proefschrift werd (mede) mogelijk gemaakt met financiële steun van The European Research Council (ERC), toegekend aan S. van der Stigchel.

Table of Contents

Prologue

General Introduction

It's the Easter weekend. The sun has finally reappeared, the trees are starting to turn green again, and you have an extra free day. Time for a spring renovation. With some trepidation you head to your nearest Swedish furniture megastore, seemingly along with the rest of the country. You scrap and bump your way through the hordes, overwhelmed by all the billboards and colours, looking for a new night stand. Finally, you get to the other side of the checkout queue, put the box in the back of the car, and head back home. Relief.

Unfortunately, the relief doesn't last long. In the few remaining hours of daylight, you try to assemble the night stand as fast as possible and with as little frustration as possible. So you open the box, spread the screws, bolts, et cetera out on the floor (this might look like Figure 1), and consult the first step of the instruction manual...

Up until this point you, the reader, will probably recognise the experience. What comes next, however, is most likely to be the point where our experiences start to diverge. Do you memorise the appearance of a single screw, search for it, place it aside, and then repeat that process for all required screws? Or do you memorise two, or three, or even four screws at once? How much time do you spend memorising each screw? Which screw(s) do you start with? How often do you refresh your memory by looking back at the instruction manual? In other words: *How do we interact with our environment, taking into account the richness of stimuli and the limitations of our memory?*

When, *where*, and *how* we make eye movements to interact with the external world and our memory is the main topic that I investigated in this dissertation – and even the not-so-attentive reader will notice that the instruction manual is a recurring example throughout. In this **General Introduction**, I will start by providing a primer

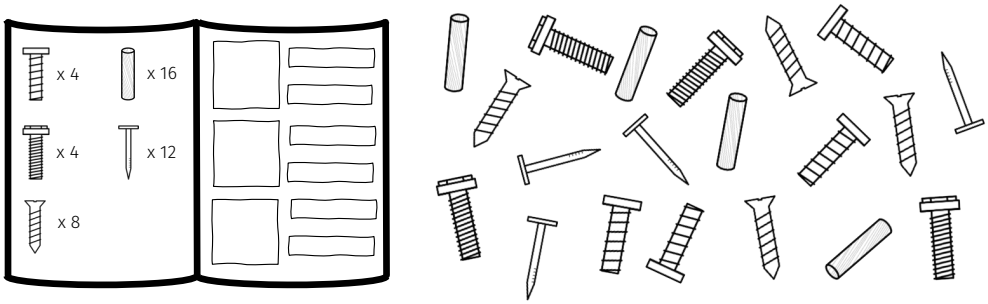


Figure 1: An instruction manual and a pile of hardware.

on eye movements and visual working memory, in which I will outline the background information which readers unfamiliar with the subject at hand might find helpful throughout the rest of this dissertation. Then, I will introduce Chapters **1**, **2**, **3**, and **4** (which make up Part I: Memory in context), and Chapters **5** and **6** (which make up Part II: Individual- and state-dependent influences on eye movements). I will integrate my findings and discuss the theoretical and practical implications of my research in the **General Discussion**.

Primer

Eye movements

We, humans, make on average around two to four eye movements per second (Henderson & Hollingworth, 1998). Spread out over sixteen waking hours, that counts up to more than a hundred thousand eye movements every single day. What's more, we don't just make random eye movements; almost all of them are made purposefully, even if we're not aware of it.

Why do we make eye movements? Let's start at the foundation.

When you or I look at a screw, light from the external world (bouncing off of that screw) comes in through our pupils, and then falls onto the retina at the back of our eyes. From there, signals are sent to the back of the brain, which eventually forms those signals into a coherent picture and allows us to *see* the screw[1]. However, there is actually only a small part of the retina, called the *fovea*, which has the highest acuity (sharpness) of our visual system. Spreading out from the fovea, visual acuity decreases, and details get blurrier. As a result, we arrive at the biggest reason to make eye movements; we need them to survive. Our eyes *must* keep moving around in order to take in the world around us, make out details, detect dangers, explore, and so on (Findlay & Gilchrist, 2003; O'Regan, 1992; Schiller, 1998; van Lieshout et al., 2020).

Having established *why* we move our eyes, we should next discuss *how* we move our eyes. Eye movements can be split into three main types: fixations, saccades and smooth pursuits – and these can be categorised by the speed and distance of the eye's rotation in its socket (Dodge & Cline, 1901; Hessels et al., 2018)[2]. (1) Fixations are periods of relative rest, during which the eye is mostly still, and thus retinal input remains mostly the same[3]. Fixations last at least 50 milliseconds (1/20 of a second), but can last more than a second – and it is during these periods that we usually aim our fovea at the object that we want to see. When the eyes are still, we can process the light that lands on our retina with high detail, and the longer we maintain fixation, the more detail we can gather about the external world (Henderson & Hollingworth, 1998; Theeuwes et al., 1998; Vogel et al., 2006). (2) Fixations are alternated by fast,

[1] For the sake of brevity, this is an enormously simplified version of how vision works. See e.g., Schiller (1998) for a more detailed account.

[2] It can be argued that there are more types of eye movements (e.g., microsaccades, corrective saccades). I have limited the description to the types which are analysed in this dissertation

[3] Our eyes are rarely ever fully still, but during fixations they only move very small distances (Engbert & Kliegl, 2003; Rolfs, 2009).

ballistic movements called saccades. During these saccades, we can gather very little information from the external world (except for some blur; Latour, 1962), and thus we mainly use saccades to transition between different points of fixation (Findlay & Gilchrist, 2003; Melcher & Colby, 2008). (3) Finally, there is one more type of eye movement that we will discuss here. Smooth pursuits are less common, because we can only make them when we are tracking an object that is moving relative to our head. When our eyes follow a moving car, or when we rotate our head while looking at *this* word, our eyes can rotate smoothly, but relatively slowly, in their sockets. Smooth pursuits are therefore a sort of hybrid between fixations and saccades – we can keep foveating and processing an object while moving our eyes (Dodge, 1903; Robinson, 1965).

However, simply fixating an object is not sufficient to gather information. For the light that enters our eye to actually be processed, we also need visual attention (James, 1890). Attention can be described as a sort of flashlight. In this metaphor, the world around us is a dark room, and we can only see wherever we shine our flashlight. This flashlight has a narrow beam, and so – just like our eyes – we need to move the flashlight around in order to process the details of the dark room. Note that attention and eye movements are not necessarily always at the same location; we can overtly (while foveating) as well as covertly (without foveating) attend information (Posner, 1980). Back in the non-metaphorical world, we thus need visual attention for our brain to actually process information from the external world and conclude 'I am looking at a screw!'.

Visual working memory

Once we have fixated and attended a screw, we can use visual working memory (VWM) to temporarily store, manipulate, and use visual representations in order to guide our actions (Baddeley & Hitch, 1974). VWM is used in a myriad of situations, but I will provide an example in the context of visual search. For example, when the instruction manual shows you an image of a screw, you can store the image's appearance in VWM. Then, while actively maintaining the visual representation of the screw in memory, you can use its features (such as shape and colour; Findlay, 1997) as a *template* to find that particular screw in the pile on the floor. To this end, you first create an attentional priority map of your environment; a sort of heatmap which highlights locations where visual input is similar to your internal template (Wolfe, 2021; Zelinsky & Bisley, 2015, see Figure 2 for an illustration). You then move your attention and gaze around the pile, mostly towards items that have a high priority in this map. With every fixation that you make, you compare what you're looking at to the internal VWM template representation of the screw (J. Palmer et al., 2000). If you decide that it's a different item (perhaps a nail), you fixate the next object, and so on, until you find the correct screw. Once you have found a match, you can pick it up and use it to build your furniture.

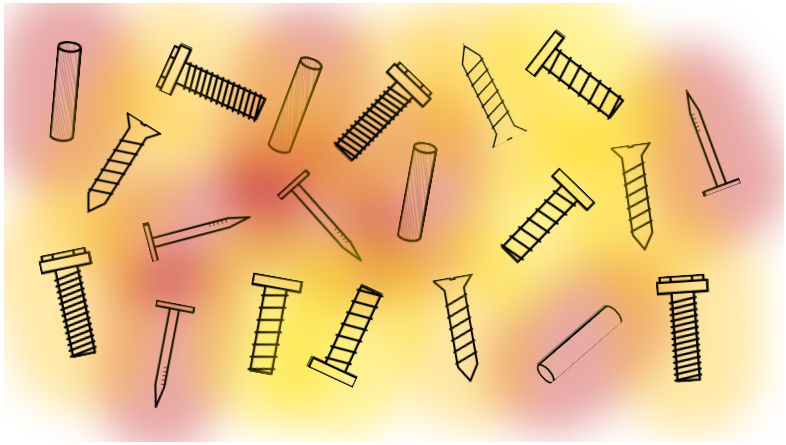


Figure 2: A graphical representation of a possible priority map, ranging from low similarity (red) to high similarity (yellow) to the template in visual working memory.

Part I: Visual working memory in context

In our daily lives – whether we're at a sports match, in a shopping centre, or in the comfort of our own home – we are constantly surrounded by an incredibly rich visual environment. There can be people, plants, screens, chairs, billboards, bikes, and so on, all within view. However, quite often we don't notice whether something in our surroundings has changed after we looked away from it briefly (Simons & Levin, 1997), highlighting that we are actually aware of very little of our surroundings at any given time. In part, this is due to the fact that we need to actively attend visual information in order to process it enough to form a sufficient internal representation (Neisser & Becklen, 1975; Rensink et al., 1997). Attention is a limited resource though, which allows us to only attend certain details of our surroundings at a time (James, 1890; Kahneman, 1973; Simons, 1996). Because of this bottleneck in attentional processing, we spend much of our days attending external information, while constantly having to be selective about what information we attend and what information we ignore. Furthermore, in order to detect changes in scenes, we need to encode and maintain an internal visual representation of what an object looked like in order to be able to compare it to other objects. This requires not only attention, but visual working memory (VWM); a short-term storage system for information which we wish to internally maintain, manipulate or compare – with the goal of using it for subsequent action (Baddeley & Hitch, 1974; Olivers & Roelfsema, 2020). Like attention, VWM is a strongly limited resource in terms of throughput. Because encoding information into VWM requires attention, it is believed that the rate at which information from the external world can be transferred into VWM is restricted (Oberauer, 2019). Moreover, VWM has limited storage capacity of a few items (Brady et al., 2011; Cowan, 2016; Luck & Vogel, 2013; Ma et al., 2014).

Once a representation is encoded in VWM, it needs to be actively maintained if it is to be used (Miller et al., 2018), which requires sustained neural activity and therefore costs physiological energy (Beatty, 1982; Kahneman, 1973). As VWM load is increased,

the required neural activity is assumed to increase monotonically as a result (until capacity is reached; Luria et al., 2016; Vogel & Machizawa, 2004). Moreover, we need to maintain our attention for about 50 milliseconds up to a second on each piece of external information that we wish to encode into VWM (Bays et al., 2011; de Jong et al., 2023; Vogel et al., 2006). It thus becomes clear that encoding and maintaining items in VWM is a costly process, in terms of energy expenditure as well as time. Luckily, most objects in our surroundings are visually stable over brief periods of time. Chairs don't suddenly change shape, trees don't suddenly change colour, and although a basketball may move around, it remains round and orange. It is because of these regularities that, when two people throw a basketball around, we can understand that it is probably the same ball after briefly looking away, even when it is in someone else's hands. These regularities come with the substantial benefit that, during our daily activities, we can forgo memorising the vast majority of objects, thereby limiting VWM load – and thus cost – to a minimum.

Indeed, O'Regan (1992) eloquently described how, in many cases, *the world* can serve as an *external memory* for the veridical representations of items – thereby allowing us to only remember or internally visualise limited details of objects. Shortly thereafter, Ballard et al. (1995) published a seminal study in which they tested this concept experimentally with a blocks-copying task. They found that participants indeed offloaded working memory, keeping VWM load well below assumed capacity. Rather than memorise everything that we attend (which would be extremely effortful and time-consuming), we may instead memorise a single object, and store simple spatial pointers for other possibly relevant objects and locations that we come across, and then carry on with our activity. Once one of those other objects becomes relevant for our current activity, we may activate the pointer, make an eye movement to that location, and then encode its representation into VWM for further use.

Although eye movements are generally considered cognitively 'cheap' (Koevoet, Strauch, Naber, & Van der Stigchel, 2023; Koevoet et al., 2024; Theeuwes, 2012; Theeuwes et al., 1998), they also require some amount of time and energy. To make every single eye movement, attention needs to be shifted, oculomotor musculature needs to be recruited, and neural populations need to be remapped (Bays & Husain, 2007; Duhamel et al., 1992; Jonides, 1983; Melcher, 2007; Rizzolatti et al., 1987). This preparatory phase usually takes around 80 to 250 milliseconds, after which the saccade itself also takes 40 to 400 milliseconds (Bahill et al., 1975; Baloh et al., 1975; Liversedge et al., 2011). As such, *when* and *how much* we use memory is delicately balanced. At its most basic, there exists a trade-off which is a function of the cost of storing content in memory versus the cost of externally sampling information by making eye movements. Studies from our lab and others have consistently indicated that people tend to offload VWM as much as possible. This has most commonly been shown with copying tasks, in which participants are shown an example layout containing multiple items which they need to remember and recreate elsewhere in a workspace. When the example layout can be easily reinspected in these copying tasks, participants often memorise only one item, place it in the workspace, memorise the next item, and so on until the task is finished (cf. Ballard et al., 1995; see Qing et al., 2024 for a meta-analysis). But remember that this behaviour is the result of an internal trade-off which can be shifted by altering the cost of memorizing versus

sampling. As such, simply increasing the distance between the example layout and the workspace – and thus requiring larger saccades – already shifts participants' behaviour towards memorising more information during each inspection (Ballard et al., 1995; Draschkow et al., 2021; Inamdar & Pomplun, 2003). Similarly, briefly delaying access to the example layout each time that participants want to inspect it causes participants to load VWM more (Böing et al., 2023; Gray et al., 2006; Melnik et al., 2018; Sahakian et al., 2023; Somai et al., 2020). Thus, increasing the time- and energy costs of retrieving external information can alter the delicate balance between storing versus sampling, explaining one part of what constitutes 'cost'.

It should be noted, though, that the aforementioned studies manipulated the cost of external sampling in such a way that it was stable and predictable within experimental blocks. As others have mentioned, the storage-sampling trade-off has a strong *just-in-time* component, meaning that we sample external information only if and when it is needed (Droll & Hayhoe, 2007; Droll et al., 2005; Gajewski & Henderson, 2005; Hayhoe et al., 2003; Triesch et al., 2003). When there is a predictable delay of two seconds, this just-in-time component is disrupted, but we may still be able to factor the known cost of sampling (e.g., exactly two seconds) into our internal model. In daily life we frequently encounter situations in which access to external information is unpredictable. Glare from the sun may intermittently occlude your vision while driving, a tall person at a concert may block your view of the stage, or a web page may take a few milliseconds up to several seconds to load. In each of those cases, it is unknown to us when (or if) information will become available to us again, thereby strongly disrupting our ability to sample information just-in-time. To compensate for this disruption, we may build up more elaborate internal representations of external information whenever it becomes available to us. In **Chapter 1**, I used a copying task (Figure 3) to investigate how this unpredictability of access to external information affects the trade-off between internal storage and external sampling. In Experiment 1, we intermittently showed and occluded the example layout, and participants had no control over this. Here, participants attempted to encode more items per inspection than when the layout was constantly available, but this did not consistently result in more correct placements. However, those findings could potentially be explained by inherent differences in how long the example layout could be viewed. In Experiment 2, the example layout only became available after a gaze-contingent delay, which could be constant or variable. Here, the introduction of any delay led participants to increase their VWM load compared to no delay, although the degree of variability in the delay did not affect behaviour. As such, in Chapter 1 I argue that *any* disruption to the continuous availability of external information is the main driver of increased VWM usage, and that predictability of access to external information is less important.

Up until this point, the storage-sampling trade-off had been almost exclusively investigated using aforementioned copying tasks. However, these tasks are quite complex and require several behavioural components from participants (Draschkow et al., 2021). Because of this complexity, inspections of the example layout may occur for several reasons (encoding a shape or its location, tracking which items have been placed, et cetera), and each correct or incorrect placement – or lack of a placement – in the workspace may have several causes (incorrectly remembered shape or location, failed search for the required item, et cetera). Although this paradigm enables us

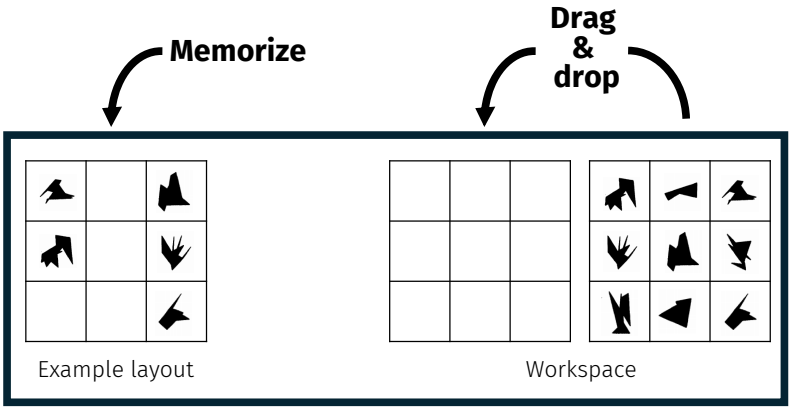


Figure 3: An example of a copying task paradigm. Participants memorise objects from the example layout (left) and drag them to the correct place in the workspace area (right). Stimuli were originally introduced by Arnoult (1956).

to study behaviour in a very naturalistic way, it clearly complicates analyses with respect to lower-level behaviour and cognition. In Chapters 2, 3 and 4, I simplified the paradigm to just a visual search task (Figure 4), in which participants memorised templates on one side of the screen, and searched for those templates amongst distractors in the search array. Essentially, this paradigm retains half of the subprocesses within the copying task (encode and search for items), whilst discarding the subprocesses of dragging and dropping items in their correct locations. Moreover, this novel paradigm comes with the added benefit that visual search has been extensively investigated and modelled (Wolfe, 2021), which allows one to make clearer predictions about expected behaviour. Our key manipulation in **Chapter 2** was that templates could be reinspected throughout the trial in half of conditions, whereas they could only be inspected once before search onset in the other half of conditions. Here, we replicated behaviour which is usually encountered in copying tasks; participants often inspected search templates during the trial when they were able to. In Hoogerbrugge, Sahakian, et al. (2024) I shared three additional unpublished datasets in which we further showed that this visual search paradigm could replicate the findings from copying tasks. Notably, throughout all of our datasets, participants frequently *re*inspected external search templates that they had looked at earlier within a trial. Resampling behaviour in our search paradigm generally scaled with complexity of search, cost of sampling, and was beneficial to task speed, accuracy, and effort; participants could spend less time and fewer cognitive resources to encode templates, and instead memorise or 'refresh' information only when needed (reinforcing our concept that the trade-off contains a strong just-in-time component, as discussed in Chapter 1). In the very first few paragraphs of this General Introduction, I mentioned that our experiences may start to diverge upon commencement of the assembly process. What I found in Chapter 2 is that there are indeed individual differences in how people approach these tasks (some participants relied more on VWM than others), but there does exist a general pattern of behaviour: When people search for screws (Figure 1), many of them encode one screw from the instruction manual as

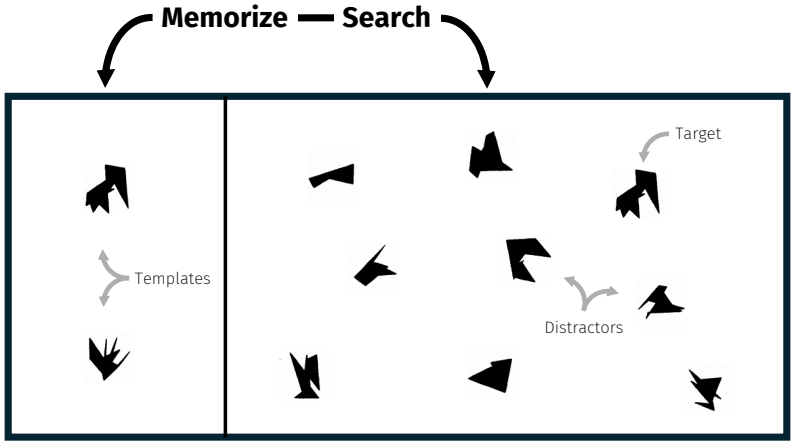


Figure 4: An example of our search task paradigm. Participants memorise templates (left) and search for template-matching targets in the search area (right).

an internal template in VWM, search for it, and then move on to the next screw. On some occasions, people even *re*inspect the image in the instruction manual when they think they may have found the screw.

Having established that people strongly and beneficially rely on the external world, I next asked: How persistent is this behaviour? Let's say you work at the Swedish furniture store, and you need to assemble twenty-five night stands for display. Is it then still so beneficial to constantly reinspect the instruction manual, or is it better to initially take some extra time to elaborately encode representations of the screws into (long-term) memory? Our reasoning was as follows: Repeatedly searching for (or even being exposed to) the same search targets leads to increasingly elaborate internal representations of those targets (likely in interplay with long-term memory; Carlisle et al., 2011; Ebbinghaus, 1885; Hout & Goldinger, 2010; Pashler et al., 2007; Woodman et al., 2001, 2007). Additionally, visual search for items stored in long-term memory is relatively easy and efficient, particularly for many items (Drew & Wolfe, 2014; Drew et al., 2017; Wolfe, 2012; Woodman et al., 2001). Therefore, there should be diminishing speed-, accuracy- and effort benefits of resampling templates as these templates are repeated and become strongly represented in memory. Rather, when you search for the same screw many times consecutively, making saccades towards the instruction manual may become more costly than storing representations in memory – in which case sampling behaviour should decrease or eventually even cease. In **Chapter 3**, I put the persistence of external sampling behaviour to the test by repeating the same template sets for twenty-five consecutive trials. In Experiment 1, search templates remained available throughout all twenty-five consecutive repetitions; only the distractors and target changed between trials. Participants indeed inspected templates less often in the tenth repetition than in the first few repetitions, but behaviour mostly stabilised after that. Strikingly, at the end of all twenty-five repetitions participants still inspected the template area twice per trial when they searched for four templates! Moreover, response times stabilised along with the number of inspections, and accuracy remained stable across all repetitions. In Ex-

periment 2 we tested whether this persistence of resampling behaviour was actually necessary to maintain high accuracy. We made templates unavailable in the last ten repetitions, and found that accuracy remained high even when templates could not be reinspected. Better yet, this seemingly 'excessive' resampling behaviour was beneficial to neither short-term nor longer-term performance. Additional analyses showed that resampling behaviour was at least partially used to boost metacognitive confidence rather than the actual quality of memory representations. Intuitively, this behaviour makes sense: memory representations are prone to errors, interference, and may degrade over time (Baddeley & Hitch, 1974; Desender et al., 2018; Gold et al., 2005; Hardt et al., 2013), thus if the cost of sampling is low enough it may be worth taking some time to verify that our memory is still correct. As such, even when the benefit of offloading memory is eliminated, resampling behaviour persists – although the underlying reason partially shifts from offloading memory representations to boosting confidence in existing representations.

It has become clear that the decision to either use or offload memory in visual search is highly dynamic, dependent on various environmental- or task factors and on individual preferences. Moreover, I have shown that whether people *can* do something does not mean that they *will* do it. Specifically, in Chapter 3 I showed that participants could eventually perform the task from memory, but they still chose to repeatedly reinspect templates. I then realised that this phenomenon may actually be able to reconcile some of the mixed findings in a debated sub-type of visual search – namely *multi-target* search. By some accounts, humans are able to guide search concurrently from multiple items in memory (Beck & Hollingworth, 2017; Beck et al., 2012; Godwin et al., 2015; Grubert et al., 2024; R. S. Williams et al., 2023), whereas others have opposed this finding (Ort et al., 2017, 2019; Van Moorselaar et al., 2014). In **Chapter 4**, I not only asked whether participants *can* perform concurrent multi-target search when instructed (Experiment 1), but also whether they actually *apply* this when given free choice on how to search (Experiments 2a and 2b). Participants were indeed able to search sequentially and concurrently when instructed to do so in Experiment 1. I then used a novel modelling approach to indicate on a trial-by-trial basis whether participants searched sequentially or concurrently in Experiments 2a and 2b. Interestingly, participants used sequential and concurrent search as specific and dissociable modes, and they flexibly adjusted which of the two they used based on task demands as well as individual preferences. Therefore, sequential and concurrent search modes can be considered 'tools in the toolbox' of search strategies; sometimes you need a steel hammer, sometimes you need a wooden mallet – both do similar jobs but are best suited to a specific problem.

Part II: Individual- and state-dependent influences on eye movements

I have discussed that humans generally offload working memory, and that they compensate by frequently sampling external information with the use of saccades. When and where we move our attention (and subsequently our eyes) is therefore

essential in goal-directed tasks – especially those which involve memory (Henderson & Hollingworth, 1998; Le-Hoa Võ & Wolfe, 2015; Mills et al., 2011; Schütz et al., 2011). Better yet, even during non-visual or multimodal tasks, the eyes typically exhibit coupling to, or dominance over, those other modalities (Corneil et al., 2002; Macaluso et al., 2000; Richardson & Spivey, 2000; Stokes & Biggs, 2014). However, in some instances we don't have strong top-down goals. When we stroll through the forest, watch a film, or visit a museum, we often simply want to look at whatever the visual environment has to offer without explicitly searching for things, attempting to remember something, or trying to avoid dangers. In such scenarios, we perform what is called natural viewing or *free-viewing*. When free-viewing, we move our eyes differently than when e.g., searching or memorizing (Borji & Itti, 2014; Buswell, 1935; Henderson & Hollingworth, 1998; Kootstra et al., 2020; Mills et al., 2011; Yarbus, 1967), likely because attention is more strongly driven by bottom-up perceptual input (saliency) than by top-down goals (Itti et al., 1998; van Zoest et al., 2017; but see Awh et al., 2012; Tatler, 2009). Understanding how, when and where people move their eyes during free-viewing can inform us about fundamental attentional mechanisms (Gottlieb et al., 1998; Itti et al., 1998; Koch & Ullman, 1987) but also has more direct commercial applications (e.g., predicting which part of a billboard people will look at first; Bylinskii et al., 2017). It is therefore no surprise that considerable time and resources have been spent on attempting to model gaze behaviour (for example, in Chapter 5 we analysed 21 models which have been cited more than thirty thousand times combined). These models aim to predict where people look based largely on bottom-up visual features, and are therefore commonly referred to as saliency models. Model predictions are tested by comparing them to (human) gaze behaviour, usually over a wide range of images (Kümmerer et al., 2018), but relatively small samples of participants and/or over samples with narrow demographic distributions (e.g., psychology students). Given how often these models are used academically and commercially, it is important that they are validated for many people and in many contexts.

We collected data at the NEMO Science Museum in Amsterdam, where we had a setup in which visitors free-viewed an image while their eye movements were tracked (Figure 5). This provided the opportunity to test how well saliency models can predict where people look within a uniquely large sample and across a wider range of demographics than usual. In **Chapter 5**, I report on 1,600 museum visitors who took part in our experiment, ranging from 6 to 59 years of age. We tested gaze behaviour from our sample against the predictions of 21 popular saliency models, and found that there was significant variability between demographic groups (e.g., 12-year-old children compared to 24-year-old adults) in how well these models predicted where participants would look. Most importantly, our selection of saliency models were best at predicting where 18-29 year-olds would look, but significantly worse at predicting this for e.g., children aged 6-17. As such, I highlight in the discussion of Chapter 5 that it is critical to keep an eye out for potential biases when developing and testing saliency models.

Not only are we likely to look at different locations within images, but our current mental and physiological states also influence how we move our eyes. For example, increased expenditure of mental effort has been linked to increased pupil size (Beatty,

Figure 5: Eye movements made by a participant in the NEMO free-viewing dataset (Chapter 5). White circles indicate fixation locations, circle size represents fixation duration, lines with arrows indicate the direction of saccades.

1982; Kahneman, 1973; Koevoet, Strauch, Van der Stigchel, et al., 2023; Strauch et al., 2022), decreased saccade velocities (Di Stasi et al., 2010, 2013), and a modulation of microsaccade frequency (Pastukhov & Braun, 2010; Siegenthaler et al., 2014). Given that mental effort is closely related to arousal (Hjortskov et al., 2004; Teigen, 1994), and arousal in turn is linked to heart rate (e.g., sympathetic nervous system activation affects heart rate variability; Azarbarzin et al., 2014; Grassi et al., 1998; Mather et al., 2017), it is likely that heart rate and eye movements are physiologically coupled. Indeed, changes in heart rate or heart rate variability have been linked to modulations of eyeblink frequency (Nakano & Kuriyama, 2017) and microsaccade rates (Ohl et al., 2016). Establishing how eye movements are coupled to heart rate additionally provides a new avenue towards remotely and unobtrusively measuring arousal levels. In **Chapter 6**, I describe an integral approach we took to testing whether heart rate and oculomotor metrics were linked during free-viewing of the 1994 Forrest Gump motion picture. We specifically investigated this in a movie-viewing task, such that both spontaneous fluctuations and the movie contents could affect arousal levels while reducing (top-down) goal-directed influences. Using a wide range of oculomotor metrics, we were able to consistently classify above chance level whether participants had high- or low heart rate levels throughout the movie – and which metrics contributed most towards classification. We found that especially blink frequency, as well as metrics related to the velocity and amplitudes of oculomotor movement (rather than frequency or duration), were most informative for the classification of heart rate. Thus, how we move our eyes is coupled to heart rate, with arousal as putative underlying mechanism.

Summary

The external world constantly provides us with incredibly rich visual input, although there is only so much of it that we can actively process and perceive, let alone memorise. When, where, and how we make eye movements to gather information from the external world is therefore an essential aspect of our daily lives. In this dissertation, I outline in the first four chapters that *where* and *when* we make eye movements to sample from the external world is highly intertwined with how we make use of visual working memory, and vice versa. I describe additional mechanisms of the trade-off between internally storing information in VWM versus externally sampling information, and provide an integrated account of how these mechanisms interact. Furthermore, I describe in the last two chapters that *where* we make eye movements differs between people; and *how* we make eye movements is linked to our state of arousal.

Part I

Visual working memory in context

Chapter 1

Just-in-time encoding into visual working memory is contingent upon constant availability of external information

Alex J. Hoogerbrugge
Christoph Strauch
Sanne Böing
Tanja C. W. Nijboer
Stefan Van der Stigchel

Published as:
Hoogerbrugge, A. J., Strauch, C., Böing, S., Nijboer, T. C. W., & Van der Stigchel, S. (2024). Just-in-Time Encoding Into Visual Working Memory Is Contingent Upon Constant Availability of External Information. *Journal of Cognition, 7*(1). doi.org/10.5334/joc.364

Abstract

Humans maintain an intricate balance between storing information in visual working memory (VWM) and just-in-time sampling of the external world, rooted in a trade-off between the cost of maintaining items in VWM versus retrieving information as it is needed. Previous studies have consistently shown that one prerequisite of just-in-time sampling is a high degree of availability of external information, and that introducing a delay before being able to access information led participants to rely less on the external world and more on VWM. However, these studies manipulated availability in such a manner that the cost of sampling was stable and predictable. It is yet unclear whether participants become less reliant on external information when it is more difficult to factor in the cost of sampling that information. In two experiments, participants copied an example layout from the left to the right side of the screen. In Experiment 1, intermittent occlusion of the example layout led participants to attempt to encode more items per inspection than when the layout was constantly available, but this did not consistently result in more correct placements. However, these findings could potentially be explained by inherent differences in how long the example layout could be viewed. Therefore in Experiment 2, the example layout only became available after a gaze-contingent delay, which could be constant or variable. Here, the introduction of any delay led to increased VWM load compared to no delay, although the degree of variability in the delay did not alter behaviour. These results reaffirm that the nature of when we engage VWM is dynamical, and suggest that any disruption to the continuous availability of external information is the main driver of increased VWM usage relative to whether availability is predictable or not.

1.1 Introduction

Imagine laying a jigsaw puzzle. How many puzzle pieces at a time will you encode into memory for subsequent placement? One might initially expect that working memory is loaded to its capacity every time. However, from personal experience you likely recognize that you rarely apply this strategy, but rather memorize only one or two pieces at a time before trying to place them in their right location.

Representations of visual information from our environment – such as puzzle pieces – are stored in visual working memory (VWM); a short-term, limited-capacity system (Baddeley & Herring, 1983; Ma et al., 2014; Salway & Logie, 1995). The limits of VWM capacity have been studied extensively with change detection paradigms, delayed recall, and various other tasks (e.g., Adam et al., 2017; Brady & Tenenbaum, 2013; Cowan, 2016; Luck & Vogel, 2013; Ma et al., 2014; Oberauer et al., 2018). Typically in such studies, displays with to-be-memorized visual information of certain set sizes (e.g., four coloured squares) are briefly and transiently presented, after which this information does not reappear. After a retention interval, participants are required to report on what they remembered; e.g., they are shown a matching- or non-matching display and report whether any stimuli have changed. VWM capacity is then estimated, for example, from task accuracy on each set size. This line of research has been successful, providing insight into how information is represented in VWM and what its limits are (Cowan, 2016; Luck & Vogel, 2013; Ma et al., 2014).

Interestingly, however, VWM is usually not filled to capacity when participants have the option to look back at the to-be-memorized information. When and how much VWM is loaded as a function of task specifics is therefore part of a growing body of literature (T. Kristjánsson et al., 2018; O'Regan, 1992; Van der Stigchel, 2020; Wilson, 2002). For example, in a puzzle with four remaining pieces, one might only memorize one or two pieces at a time, fill them in, and then memorize the remaining two pieces in order to fully complete the puzzle. Therefore, VWM can be regarded as being part of a dynamic system that constantly weighs the costs of maintaining a (high) memory load against the costs of external sampling. Indeed, consistent with the example of a jigsaw puzzle, several studies have found participants to minimally utilize VWM in many circumstances where information could be retrieved *just-in-time* from the environment instead (Ballard et al., 1995; Böing et al., 2023; Draschkow et al., 2021; Droll & Hayhoe, 2007; Gajewski & Henderson, 2005; Gray et al., 2006; Hayhoe et al., 2003; Inamdar & Pomplun, 2003; Melnik et al., 2018; Risko & Dunn, 2015; Risko & Gilbert, 2016; Sahakian et al., 2023, 2024; Somai et al., 2020; Triesch et al., 2003). In this just-in-time approach, external information is only fixated and encoded into memory *if* and *when* it is needed for the task at hand, instead of being processed (and memorized) in advance (Droll & Hayhoe, 2007; Hayhoe et al., 2003). Most notably, previous studies either manipulated the distance between the area where external information could be retrieved and the area where that retrieved information needed to be used, or they delayed access to the required external information. When distance was increased, external sampling would theoretically become more costly, since larger (thus more energy-expensive and time-consuming) eye- or head movements were needed to move the gaze back to the external information. Indeed, the cost of sampling altered the trade-off between storing and just-in-time sampling; shorter distances were linked

to the dominance of external sampling, whereas larger distances were associated with more storing (Ballard et al., 1995; Draschkow et al., 2021; Inamdar & Pomplun, 2003). The second set of studies delayed the access to external information, for example by letting participants wait every time they wanted to sample externally (Böing et al., 2023; Gray et al., 2006; Melnik et al., 2018; Sahakian et al., 2023; Somai et al., 2020). There, participants showed similar patterns of behaviour as in the distance manipulations, providing strong evidence that the cost of access to information in the external world shifts the balance of *when* and *how much* one relies on VWM.

The aforementioned studies all provided external environments which were predictable or stable: External information was removed after an encoding phase (e.g., change detection tasks) or could be revisited throughout the task (e.g., copying tasks). When external information can be revisited, the just-in-time sampling strategy is especially useful if we can make an estimation of how costly (i.e., time and energy) such a revisit will be. Even if we are aware that we will have to wait two seconds before we can resample information, we can factor that delay into our internal model of the cost of just-in-time sampling versus maintaining higher VWM loads. However, we commonly experience situations in which access to external information is not stable. Think, for example, of intermittent glare from the sun in your eyes while driving, which can make relevant external information (e.g., the position of other cars) unavailable at an unpredictable interval. Or think of loading a web page to look up information; depending on your internet speed it may take a few milliseconds up to several seconds to load the page. In these scenarios, the predictability of access to external information is disrupted. Even though information will sometimes be available instantly, the possibility of a delay may disrupt the ability to factor in the cost of just-in-time sampling. Therefore, one may instead rely on building more elaborate internal representations of the external world whenever it is available, rather than sampling information if and when it is needed. It is yet unclear how disruptions to the predictable availability of external information affect the trade-off between storing and sampling. We here hypothesized that participants minimized VWM usage (i.e., primarily sampled externally) when external visual information was readily available, and that (ir)regularly occluding external information would cause a shift towards internal storage. Furthermore, we asked whether this trade-off would change further as information became less and less predictably available.

Across two experiments, participants performed a copying task in which the required external information was constantly available in one condition, and intermittently occluded to varying degrees in other conditions. In Experiment 1, external information was made available and unavailable for different durations across conditions. Although the pacing was mostly predictable, participants had no control over when external information was made available. Experiment 2 followed up on Experiment 1: External information was unavailable by default and only became available after participants fixated an hourglass for a certain delay period. Importantly, this delay period could be constant (and therefore mostly predictable as in Experiment 1) or variable, which made access to external information less predictable.

1.2 Experiment 1

1.2.1 Methods

Participants and procedure

Sample size was determined based on previous, similar, studies (e.g., Draschkow et al., 2021; Melnik et al., 2018; Somai et al., 2020). 26 participants performed the experiment. Gaze data of one participant was corrupted and two participants stopped early. Of the remaining 23 participants (age range 18-29), 10 indicated female and 13 male gender. All had normal or corrected-to-normal sight. Participants were compensated with €7 per hour or course credits. The experiment was approved by the Faculty Ethics Review Board of the Faculty of Social Sciences, Utrecht University (protocol number 21-0297), adhering to the Declaration of Helsinki.

Participants signed an informed consent form, provided their age category and gender, and were then instructed about the task. Each participant first completed five practice trials in a baseline condition. After confirming that they understood the task, the participant started the actual experiment. The experiment took approximately 90-120 minutes to complete.

Apparatus and stimuli

Data and code are available on the Open Science Framework https://osf.io/z2n5x/. The experiment was implemented with Python 3.6 and PyGaze (Dalmaijer et al., 2014). The experiment was displayed on a 27 inch LCD monitor (2560 $\times$ 1440 pixels, 100 Hz). Participants placed their heads in a fixed chin- and forehead rest at 67.5 centimetres from the screen, such that each 100 $\times$ 100 pixel stimulus occupied a visual angle of approximately 2°. The experiment was recorded with an EyeLink 1000 eye tracker (SR Research Ltd., Canada), which measured monocularly at a sampling rate of 1 kHz. We did not standardize whether the left or right eye was tracked. The threshold for eye tracking validation error was 1° (average of 9-point validation) and 1.5° per-point maximum, otherwise the eye tracker was re-calibrated. The stimuli used in this experiment were adapted from Arnoult (1956), and were previously used in Böing et al. (2023), Hoogerbrugge et al. (2023), Sahakian et al. (2023, 2024), and Somai et al. (2020). The stimulus set consisted of five unique shapes, with each shape additionally mirrored horizontally, vertically, and diagonally, creating 20 stimuli in total (Figure 1.1B).

Task

Participants performed a copying task in which they copied a layout of six stimuli within a 3 $\times$ 3 grid on the left side of the screen (*example grid*) to an equally large empty grid on the right side of the screen (*working grid*). The centres of both grids were located at a visual angle of 12° from the centre of the screen, with each of the grids occupying approximately 7.3° $\times$ 8.8° of the visual field. In each trial, six stimuli were randomly selected without replacement and the example grid layout was randomly filled with those six stimuli. On the bottom right of the screen, the same stimuli were presented as in the example grid, but in randomized layout (*resource grid*; Figure 1.1A).

The participants' task was to exactly recreate the layout of the example grid in the working grid, by dragging stimuli from the resource grid to their correct location in the working grid (Figure 1.1A). After correctly placing a stimulus, that working grid location would briefly flash green (750 ms). After incorrectly placing a stimulus, the corresponding grid location would flash red (750 ms) and the stimulus would not lock into place, but instead fly back to the location from which it was dragged. The stimuli in the resource grid remained visible for the whole duration of the trial, even when already placed correctly in the working grid. A trial ended whenever the grid was fully copied or if the task was not completed after 42 seconds. Participants were shown feedback ("Correct"/"Timed out") after each trial.

In the *baseline* condition, the example grid was always visible. In the other three conditions, the example grid was either visible or occluded at specified intervals throughout a trial. Namely, the example grid was repetitiously (1) visible for 4 seconds and subsequently occluded for 2 seconds, such that the availability of external information was *High*; (2) visible for 3 seconds and occluded for 3 seconds, such that the availability of external information was *Medium*; (3) visible for 2 seconds and occluded for 4 seconds, such that the availability of external information was *Low*.

In each trial, the occlusion time was multiplied by a noise factor drawn from a Gaussian distribution (μ =1.0, σ=.1), with the visible time being adjusted accordingly such that the sum of visible time and occlusion time was always 6 seconds. Because the occlusion time was multiplied by a Gaussian noise factor, the possible variation of occlusion times was greater in the *Low* availability condition than in the *High* availability condition – thereby making occlusion durations somewhat less predictable. Whenever the example grid was occluded, a pictogram of an hourglass would appear in its place. The example grid was visible at the start of each trial. Visibility and occlusion of the example grid were repeated until the trial ended. Example videos of trials can be found on OSF (https://osf.io/kz5v6/).

Each of the four conditions was tested in its own block of 35 trials and the block order was randomized between participants. The eye tracker was calibrated and validated before the start of each block. Additionally, a drift check was performed before the start of each trial, by computing the root mean squared error (RMS) between the gaze prediction and a central fixation cross which was shown for two seconds. If the RMS was greater than 1.5° for more than two subsequent trials, the experimenter would recalibrate.

Outcome variables

We computed outcome variables to provide an estimate of how much information was encoded into VWM, and subsequently placed, for each inspection of the example grid. **(A)** The number of *example grid inspections*, which was calculated by counting how many times within a trial the participant made a saccade across the centre of the screen from the right side to the left side. In effect, this variable represents how often participants sampled externally by looking toward the example grid after focusing on the working- and resource area. We did not count crossings in which only the hourglass was fixated, and assumed that short fixations would be unlikely to allow for meaningful encoding (e.g., Bays et al., 2011). Therefore an inspection would only be counted if the example grid was viewed for at least 120ms before the participant

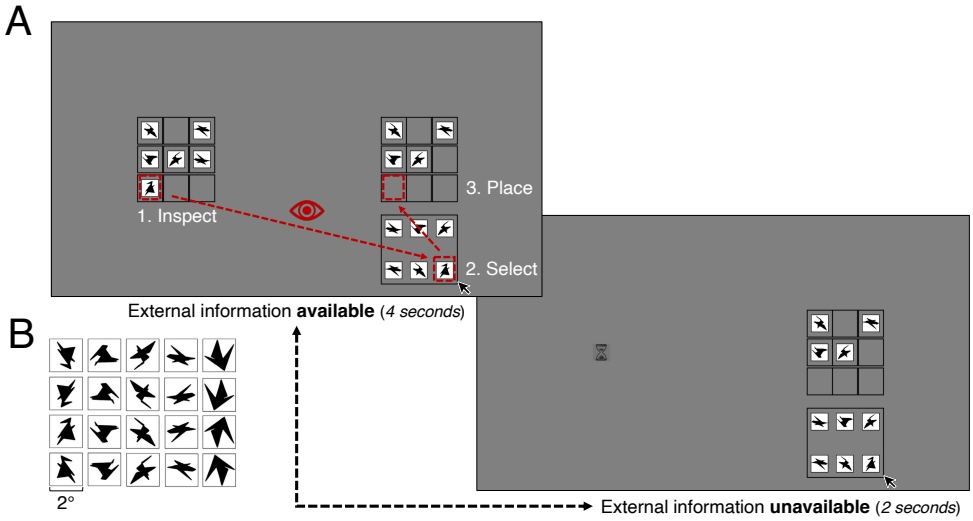


Figure 1.1: **A.** Example of a partially completed trial in the *High* availability condition in Experiment 1. In this example, four items have already been dragged to their correct position. The example grid alternated between available and unavailable throughout a trial, and this was repeated until the trial was completed. **B.** Stimuli as adapted from Arnoult (1956). In each column there is a unique shape, each mirrored horizontally, vertically, and diagonally; 20 stimuli in total. Each stimulus occupied approximately 2° of visual angle.

crossed back towards the working- and resource area. **(B)** The number of *fixations per inspection* was computed by dividing the number of fixations within the boundaries of the example grid by the number of useful inspections. This variable approximates how much information participants attempted to take in each time they placed their overt attention on the example grid. **(C)** The number of correct *items placed per inspection* was computed by dividing the number of correctly placed items per trial by the number of useful inspections made in that trial. It is an estimate of how many items participants (accurately) encoded during each inspection.

We report three additional outcome variables: **(D)** *Completion time (seconds)* was calculated from the start of the trial until all items were placed correctly, or until the 42-second timer was reached. Because the periods during which the example grid was occluded were not useless to participants (i.e., they could still place items during that time), only the time spent gazing at the hourglass in the location of the occluded example grid was subtracted from the completion time. **(E)** The number of *errors per trial*, in which an error constituted the attempted placement of any item in an incorrect slot in the working grid. A greater number of errors may reflect that items were encoded less accurately (Koevoet, Naber, et al., 2023; van den Berg et al., 2012) or that participants had more liberal thresholds for the quality of memory representations that they were willing to act on (Sahakian et al., 2023). **(F)** The *proportion spent waiting* was expressed as the duration that participants spent gazing at the hourglass, divided by the actual duration with which the example grid was occluded during that trial. This measure effectively reflects the proportion of a trial that participants spent unproductively waiting. For example: In a trial in the Low

condition, if the example grid was occluded for 12 seconds in total and a participant spent 600 ms gazing at the hourglass, the proportion spent waiting is 0.05. In the High condition, if the grid was occluded for 6 seconds in total and a participant spent 300 ms gazing at the hourglass, the proportion spent waiting is also 0.05. As such, the proportion that participants spent waiting was standardized between 0 and 1 and could be compared between conditions.

Analyses

Fixations were detected using I2MC (Hessels et al., 2017). All fixation candidates shorter than 60 ms were removed, and fixation candidates which were separated by less than 1° distance were merged. This approach has been shown to remove variation between fixation detection algorithms (Hooge et al., 2022).

Statistical analyses were conducted with JASP 0.18.3 (JASP Team, 2022). The six outcome variables were aggregated per participant, per condition. All were aggregated by the mean, except for *completion time*, which was aggregated by the median. In order to test whether availability of external information affected our outcome variables, we report Repeated Measures ANOVAs. If the assumption of sphericity was violated for an outcome variable, we report corrected ANOVAs (Greenhouse-Geisser if $\epsilon < .75$, otherwise Huynh-Feldt; following Abdi, 2010). Effect sizes of ANOVAs are reported with η^2. Post-hoc, paired samples t-tests are reported, and *p*-values were Bonferroni corrected for six comparisons within each variable. Effect sizes of t-tests are reported with Cohen's d. All statistical test outcomes including Bayes Factors are reported on the Open Science Framework.

1.2.2 Results

Example grid inspections, fixations, and items placed per inspection

In the baseline condition, participants made a median of 7.66 (*median absolute deviation; MAD* = 1.03) example grid inspections per trial, meaning they sampled externally more than once per item (Figure 1.2A). They inspected the example grid less often when availability of visual information was lower, $F(2.7, 59.5) = 6.97$, $p < .001$, $\eta^2 = 0.24$. This main effect was primarily driven by the difference between the baseline and the decreased access conditions, which indicates that the disruption to constant availability was the main driver of increased memory usage. When availability of external information was further reduced, no effect on the number of inspections was found; even in the Low availability condition, participants still inspected the example grid a median of 6.63 (*MAD* = 1.06) times – a small difference compared to the High availability condition (*Mdn* = 6.91, *MAD* = 1.07)

Although the number of example grid inspections was mainly different between the baseline and the decreased access conditions, the number of fixations per example grid inspection steadily increased as availability was reduced ($F(2.1, 46.4) = 21.91$, $p < .001$, $\eta^2 = 0.5$). This indicates that participants at least *attempted* to encode more information per inspection, ranging between *Mdn* = 1.83 (*MAD* = 0.40) fixations in the baseline condition, up to *Mdn* = 2.69 (*MAD* = 0.48) fixations in the Low condition (Figure 1.2B).

This increase in the number of fixations was somewhat reflected in the number of

correct placements per inspection (Figure 1.2C). There was an overall increase across conditions ($F(2.0, 43.8) = 6.60$, $p = .003$, $\eta^2 = 0.23$), although the number of items placed was relatively low overall and post-hoc tests only showed a significant difference ($p < .001$) between the baseline condition and the lowest-availability condition. In the baseline condition, participants correctly placed just less than one item per inspection (*Mdn* = 0.80, *MAD* = 0.16), and interestingly this stayed below one item even in the Low availability condition (*Mdn* = 0.97, *MAD* = 0.21).

The number of example grid inspections and the number of fixations and correct placements per inspection tell us that participants encoded and subsequently placed (slightly less than) one item per crossing, when external information was always available. Removing the ability to always inspect was the main driver of fewer inspections of external information, and an increased number of placements per inspection. When the example grid availability was further decreased, participants did not inspect it less frequently, but they inspected it with more fixations. While participants changed their eye movement strategy, this did not clearly translate into a different number of correct placements per inspection: participants seemed to attempt to *encode* more, but they did not necessarily *place* more items afterwards or make fewer errors. Furthermore, participants still did not regularly place two (or more) items after inspection, even when the example grid was effectively occluded for two-thirds of a trial.

Completion time, errors, and proportion spent waiting

All participants were able to consistently complete the task within the 42-second time limit. However, they seemed not to (be able to) alter their strategy enough to keep completion times consistent across the decreased availability conditions. Participants took longer to finish the task across almost all conditions as availability decreased ($F(3, 66) = 35.97$, $p < .001$, $\eta^2 = 0.62$), except between the baseline and the High availability condition ($p = 0.88$; Figure 1.2D).

Participants made more incorrect placements as availability of external information was reduced ($F(1.8, 39.7) = 15.62$, $p < .001$, $\eta^2 = 0.42$), ranging from a median of 0.09 (*MAD* = 0.09) errors per trial in the baseline condition to 0.23 (*MAD* = 0.26) errors per trial in the Low reliability condition (Figure 1.2E).

Interestingly, participants spent an increasing proportion of trials doing nothing, as the example grid was occluded for longer periods, $F(1.7, 37.0) = 52.69$, $p < .001$, $\eta^2 = 0.71$ (Figure 1.2F). In the Low availability condition, participants spent a median proportion of 0.13 (*MAD* = 0.07) gazing at the hourglass in the location of the occluded information – amounting to around one second of waiting as occlusion lasted 10.17 seconds on median in the Low condition. These findings suggest that, although participants attempted to memorize and place more items as availability was decreased, this adaptation was not necessarily time-efficient; they did not compensate for the decreased availability enough to avoid waiting unproductively.

1.2.3 Interim discussion

We here set out to disrupt the ability to just-in-time sample external information by intermittently occluding the example grid. When the visual information required to

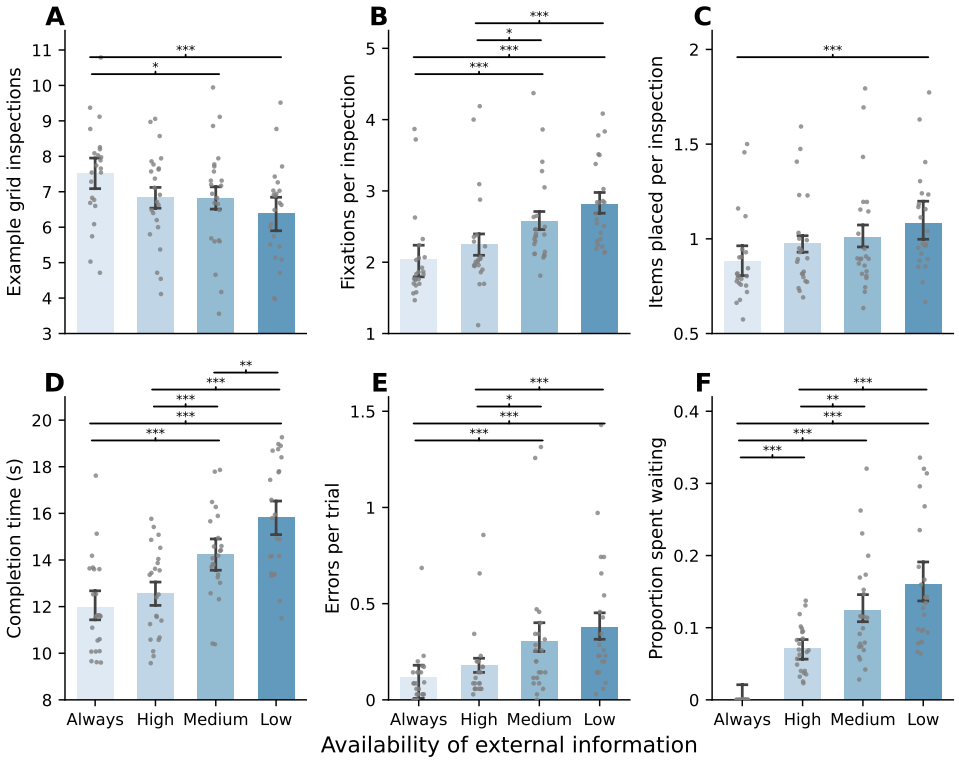


Figure 1.2: Barplots (mean ± 95% within-subjects CI) for each variable, per condition. Individual points represent within-participant aggregates. **A.** The average number of inspections of the example grid per trial. **B.** The average number of fixations made on the example grid per inspection. **C.** The average number of correctly placed items per inspection. **D.** The median completion time (in seconds). Time spent fixating at the example grid while it was occluded was subtracted. **E.** The average number of incorrectly placed items per trial. **F.** The average proportion spent fixating at the example grid location while it was occluded. **Note.** Post-hoc paired samples t-tests (Bonferroni corrected); *** $p < .001$; ** $p < .01$; * $p < .05$.

perform the task was always available, participants memorized and placed just under one item per inspection of external information, consistent with findings from similar paradigms (Sahakian et al., 2023; Somai et al., 2020). When external information was not continuously available throughout a trial, participants adapted their strategy and inspected the example grid less often but with more fixations, which implies that they at least attempted to increase VWM usage. Interestingly, however, the number of placed items did not strongly increase when external availability was decreased: participants placed approximately the same number of items, irrespective of the degree of availability. This provides evidence that the trade-off can be influenced by disrupting participants' ability to sample external information just-in-time, and that it is nonlinear in nature: *Any* removal of self-pacing shifts the trade-off, and this removal seems to influence it more heavily than further decreases in availability of external information.

However, in the current manipulation participants had limited time to view the

example grid, which possibly influenced the amount of external information that could be encoded per inspection (see General Discussion; Koevoet, Naber, et al., 2023). Furthermore, Experiment 1 did not specifically test the removal of self-pacing while keeping other parameters intact. For instance, the increased occlusion duration indirectly introduced a delay of availability, which has been shown to influence the trade-off (Böing et al., 2023; Gray et al., 2006; Melnik et al., 2018; Sahakian et al., 2023; Somai et al., 2020) and may have thus confounded the manipulation. Additionally, the example grid was incidentally available at the right time, regardless of condition, as evidenced by the generally small proportions of trials spent waiting. This means that the just-in-time aspect of availability was partially left intact.

In Experiment 2 we therefore manipulated availability of the example grid without altering the average occlusion durations across conditions. Furthermore, we did not limit how long participants could view the example grid. As in Experiment 1, we expected that participants would predominantly rely on external sampling when external information was readily available. By introducing a delay we expected to observe a shift in the trade-off towards stronger reliance on internal storage in VWM. Critically, we expected that adding variability to the delay period would cause participants to rely even less on external sampling and to encode more items per example grid inspection.

1.3 Experiment 2

1.3.1 Methods

The methods in Experiment 2 were the same as in Experiment 1, unless stated differently.

Participants

16 participants performed the experiment; none of whom had participated in Experiment 1. Gaze data of one participant was corrupted. Of the remaining 15 participants (age range 19-44; M = 25.2), 11 indicated female and 4 indicated male gender.

Task

In Experiment 2, the example grid was occluded by default and showed only if participants gazed at the example grid area for a certain amount of time. Again, the delay was signaled by an hourglass. When the delay period was served, the example grid remained available for as long as participants gazed at it.

In the *baseline* condition, the example grid showed without delay after participants' gaze was detected in that area. In the *constant delay* condition, participants had to gaze at the hourglass for exactly two seconds before the example grid appeared. In the *low variability* condition, the delay period could range between 0 and 4 seconds, drawn from a Gaussian distribution (μ = 2.0s, σ = 0.1s); in the *high variability* condition, the delay period could also range between 0 and 4 seconds, but was drawn from a wider Gaussian distribution (μ = 2.0s, σ = 1.0s). In the variable delay conditions, a new delay duration would be drawn after each time the delay period was fully served – meaning that the delay duration changed multiple times within a

trial (see Supplementary Material Figure 1 for the generated distributions of delay durations). Importantly, on average the delay period was similar across the three non-baseline conditions, thereby ensuring that any behavioural differences between conditions would be caused by uncertainty regarding availability, and not by the inherent difference in delay duration.

Each of the four conditions was tested in its own block of 35 trials and block order was randomized. Participants were instructed before the start of each block whether there was "immediate availability", "a constant delay", "some variance", or "a lot of variance".

Analyses

Instead of the proportion spent waiting, we report **(F)** The *time spent waiting* in seconds. The time spent waiting represents how long participants gazed at the hourglass while the example grid was occluded, and provides an indication whether overall delay durations were similar between conditions in which a delay was present. This outcome variable was aggregated by the median per participant, per condition.

1.3.2 Results

Example grid inspections, fixations, and items placed per inspection

Participants inspected the example grid a median of 5.69 (*MAD* = 1.34) times per trial when there was no delay. Although there was an overall effect of condition on the number of inspections ($F(3, 42) = 16.92$, $p < .001$, $\eta^2 = 0.55$), introducing any delay was the main driver of significantly decreased inspections (all $p < .001$ compared to no delay), but whether the delay was constant or variable did not further affect the number of inspections significantly (Figure 1.3A).

Similar results were observed for the number of fixations per inspection ($F(1.8, 24.5) = 14.27$, $p < .001$, $\eta^2 = 0.51$); participants likely attempted to encode more items when there was a delay compared to no delay (all $p < .001$), but again there was no effect of whether the delay period could vary (Figure 1.3B).

The effect of introducing a delay was also reflected in the number of items placed per inspection ($F(3, 42) = 11.77$, $p < .001$, $\eta^2 = 0.46$). Participants placed slightly more than one item (*Mdn* = 1.35, *MAD* = 0.56) per inspection when appearance of the example grid was not delayed, and placed *Mdn* = 2.41 to *Mdn* = 2.68 items when a delay was introduced (all $p < .01$). However, the number of placements did not differ significantly between any of the delay conditions (Figure 1.3C).

Completion time, errors, and time spent waiting

All participants could consistently complete trials within the 42-second time limit. Participants took *Mdn* = 15.53 (*MAD* = 2.01) seconds to complete the task when there was no delay. When a delay was introduced, median completion time increased to 20.82, 20.37 and 20.58 seconds for the constant delay, low variance, and high variance conditions respectively (all $p < .001$ compared to no delay). Despite an overall effect ($F(3, 42) = 16.45$, $p < .001$, $\eta^2 = 0.54$), there were no significant differences between delay conditions (Figure 1.3D).

Participants made less than one error per trial when there was no delay (*Mdn* = 0.31, *MAD* = 0.20), and this differed significantly from the delay conditions in which they made nearer to one error per trial (*Mdn* = 0.74, 0.86 and 1.00, respectively; all $p < .001$). Again, there was an overall effect ($F(1.6, 22.4) = 8.46$, $p = .003$, $\eta^2 = 0.38$), but no further difference in the number of errors per trial between the delay conditions (Figure 1.3E).

Across the three delay conditions, participants spent an equal amount of time per trial waiting for the example grid to reappear, $F(2, 28) = 3.12$, p = .060, $\eta^2 = 0.18$ (*Mdn* = 3.68, 3.68 and 3.85 seconds, respectively; Figure 1.3F). This indicates that there were no inherent differences in delay duration across those three conditions.

Inspection- and build time

Because the delay itself was excluded from the overall completion time, the current findings show that participants were generally faster at completing the task than

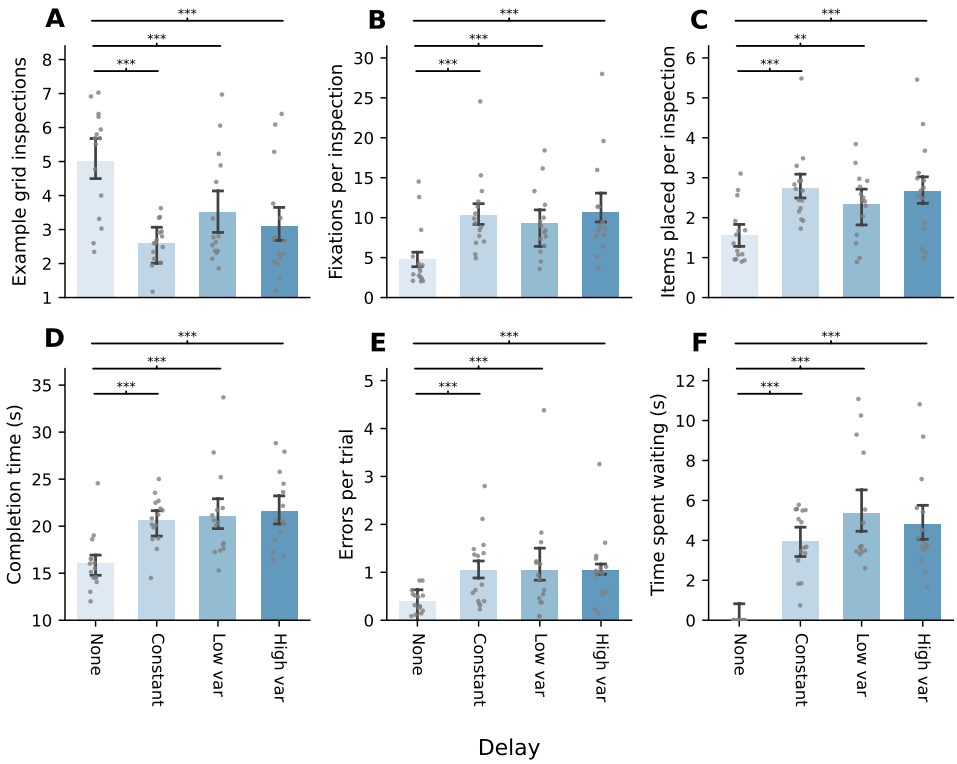


Figure 1.3: Barplots (mean ± 95% within-subjects CI) for each variable, per condition. Individual points represent within-participant aggregates. **A.** The average number of inspections of the example grid per trial. **B.** The average number of fixations made on the example grid per inspection. **C.** The average number of correctly placed items per inspection. **D.** The median completion time (in seconds). Time spent fixating at the example grid while it was occluded was subtracted. **E.** The average number of incorrectly placed items per trial. **F.** The median time per trial spent fixating at the example grid location while serving the delay period. **Note.** Post-hoc paired samples t-tests (Bonferroni corrected); *** $p < .001$; ** $p < .01$; * $p < .05$.

when there was no delay at all. In order to investigate the cause of this increase, we split completion time into its two constituent parts; **(A)** Total inspection time in seconds, computed as the sum fixation duration on the example grid (excluding waiting time); **(B)** Total build time in seconds, computed as the sum fixation duration on the right-hand side of the screen.

In the delay conditions, participants spent an increased amount of time inspecting the example grid compared to the conditions without delay (all $p < .01$), but spent equally long across the delay conditions ($F(2, 28) = 0.20$, $p = .819$, $\eta^2 = 0.01$; Figure 1.4A).

Similarly, participants spent more time building the grid in the low- and high-variance conditions than in the condition without delay ($p = .007$ for both, respectively; Figure 1.4B). As such, increased completion times were caused by increased inspection times as well as increased build times.

Higher memory loads (from increased inspection durations) thus went paired with increased build times, and the proportion of inspection time relative to build time did not change much across conditions; participants spent approximately half as much time encoding information as they spent using that information. Although the ratio between inspecting and building was significantly different between conditions in general ($F(3, 42) = 2.93$, $p = .045$, $\eta^2 = 0.17$), none of the post-hoc t-tests showed significant differences (Figure 1.4C).

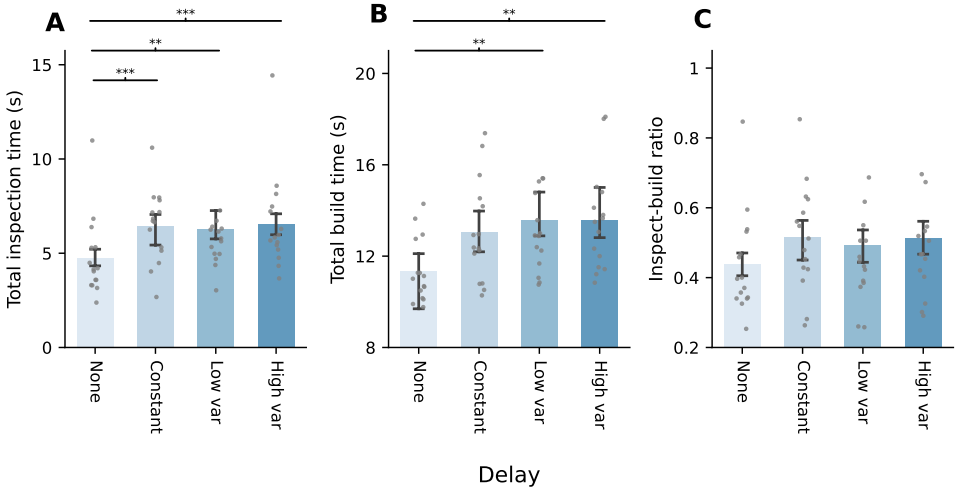


Figure 1.4: Barplots (mean ± 95% within-subjects CI) for each variable, per condition. Individual points represent within-participant aggregates. **A.** The median time spent inspecting the example grid, in seconds. **B.** The median time spent building on the right side of the screen, in seconds. **C.** The ratio between time spent inspecting and time spent building. **Note.** Post-hoc paired samples t-tests (Bonferroni corrected); *** $p < .001$; ** $p < .01$; * $p < .05$.

1.4 General discussion

It has been established that, under full and constant availability of external information, participants predominantly reduce VWM load and sample information from the environment only if and when they need that information for the task at hand. When the cost of sampling from the environment is increased (in previous studies this took the form of a delay or increased distance), participants' strategy shifts towards storing relatively more items in memory compared to when the cost of sampling is low. However, in previous studies the cost of sampling was generally stable and predictable, meaning that the cost of sampling could be factored in to the trade-off between just-in-time external sampling and internal storing in VWM. Here, we investigated whether external sampling remains the dominant strategy when the ability to make accurate estimations of the availability of external information is reduced. To test this, we let participants perform a copying task in which external information became available independent of participants' interaction with it (Experiment 1) and in which access to external information was less predictable (Experiment 2).

In Experiment 1, we intermittently showed and occluded the required external information and participants had no control over this pacing. When the external information was always available, participants memorized and placed just under one item per inspection of external information, indicating a preference to rely on the external world for just-in-time sampling per default, consistent with earlier findings (e.g., Sahakian et al., 2023; Somai et al., 2020). When external information was made less frequently available, participants attempted to encode more information into VWM per inspection of the example grid, although this did not necessarily reflect in more items placed. Notably, participants performed worse as external information was occluded for greater proportions of trials; not only did they take longer to complete the task, they also made more errors and spent more time waiting unproductively.

However, it was unclear whether the observed behaviour was the result of a reluctance to encode more information, or whether participants did not have enough time to do so. In Experiment 2 we therefore introduced a gaze-contingent delay before external information was made available. This delay could be consistently two seconds, or be drawn from a narrow or wide Gaussian distribution centered around two seconds. Participants relied heavily on external sampling when the required information was easily accessible (no delay), and shifted towards using more internal storage when a two-second delay was introduced (conceptually reproducing e.g., Böing et al., 2023; Sahakian et al., 2023; Somai et al., 2020). The trade-off did not shift further with the introduction of variability in the delay. Given that the delay in the high-variance condition ranged between 0 and 4 seconds (see Supplementary Material Figure 1), it is unlikely that the variability of the delay was insufficiently efficacious to reveal meaningful variability-caused effects. Rather, we consider it most likely that predictability of availability does not influence the trade-off between internal storage and external sampling. Additional Bayesian statistics provide moderate evidence for no modulation between the delay conditions on our outcome measures (see Supplementary Material Tables 2 & 3).

Notably, participants relied more on memory in Experiment 2 than in Experiment 1. Due to the forced delay in Experiment 2, participants may have experienced

greater time pressure than in Experiment 1, leading them to encode more information whenever it was available. However, only 53 out of 2,097 trials (2.5%) exceeded the time limit in Experiment 2, which makes it unlikely that time pressure was perceived as very high. More likely, this difference in reliance on memory may be explained by the fact that participants had limited time to view the example grid in Experiment 1, whereas they could inspect the example grid for as long as they wanted in Experiment 2. As a result, participants could encode (and subsequently place) more stimuli per inspection than they could in Experiment 1. This is particularly reflected in the number of fixations per inspection: Participants often made more than six fixations, even though there were only six stimuli to encode. This means that they fixated some stimuli multiple times, indicative of more elaborate encoding (e.g., rehearsing or reinstating; Alfandari et al., 2019; Meghanathan et al., 2019; Zelinsky et al., 2011).

The low number of items placed in Experiment 1 could also be due to the stimuli being complex, making them relatively difficult to encode and maintain in VWM (Bethell-Fox & Shepard, 1988; Eng et al., 2005), especially given that viewing time was limited and somewhat unpredictable (Bays et al., 2011; de Jong et al., 2023). The current stimulus set in combination with limited viewing time may have caused ceiling-effects of participants' ability to encode items. Using simpler stimuli may have led to decreased sampling behaviour and more items placed per inspection (cf. Hoogerbrugge et al., 2023). Avoiding ceiling effects could provide a more sensitive measure of the effect of availability of external information on memory strategies – not only in terms of the number of items memorized, but also on the origin of incorrect placements (Oberauer et al., 2018).

In Experiment 1, participants had no control over availability of the example grid and could not self-initiate the occlusion period, which could have contributed to the relatively low memory usage in Experiment 1 compared to Experiment 2. Namely, participants may have been hampered in their ability to prepare for a shift of attention in the periods just before external information became available (reminiscent of task switching costs; Nieuwenhuis & Monsell, 2002; Rogers & Monsell, 1995; Rubinstein et al., 2001). This lack of preparedness may have affected how much information participants *could* encode. Previous work showed that tonic alerting is linked to how much participants (can) encode on a copying task (Koevoet, Naber, et al., 2023); participants placed more items correctly when their state of alertness was higher before encoding than when it was low. This idea fits with the finding that participants encoded more items in Experiment 2 (in which they could prepare to encode), and strengthens our theory that participants prefer to access external information when they need it and are ready to process it. When exactly these states occur, how the brain monitors for this readiness, and how this depends on one's active interaction with external information will require further investigation.

Additionally, we found several inspections of external information per trial without subsequent placement of any items (note that these were *useful* inspections, during which external information was at least briefly viewed). Qualitatively, these inspections occurred somewhat more frequently at the start of trials, but were otherwise evenly distributed throughout trials. Although the current paradigm might not be sufficiently sensitive to attain a complete understanding of why these crossings were made, we speculate that they were explorative or comparative in nature; participants

briefly inspected the whole environment before starting copying (forming an initial strategy), or later briefly checked which items they had not placed yet. We suggest that future research further investigates why these inspections occur, taking into account the theories that differing aspects of external information may be gathered during inspections (e.g., features, locations, and/or chunks; Ballard et al., 1995; Huang & Awh, 2018), and that memory may not always be completely depleted before a new inspection is made (e.g., Ballard et al., 1995; Sahakian et al., 2023).

Furthermore, completion times were longer in the non-baseline conditions in both experiments, even though actual waiting time was subtracted from this measure. What led to this temporal inefficiency beyond waiting alone? In the delay conditions of Experiment 2, participants spent more time inspecting the example grid as well as more time building the layout than in the baseline condition. Upon closer inspection, the observed completion times in Experiment 2 were positively linked to the number of inspections, the number of fixations per inspection, as well as the number of errors per trial, all of which cost time (see Supplementary Material Table 4). Participants also made longer fixations, and fewer fixations per second in delay conditions (Supplementary Material Figure 2), which indicates more time spent encoding (Bays et al., 2011; Hoogerbrugge et al., 2023), and has been linked to higher VWM load as well as general cognitive load (Meghanathan et al., 2015; Mills et al., 2011; Woodman et al., 2001). These findings indicate that the external availability of information benefits the speed with which we can execute tasks (cf. Hoogerbrugge et al., 2023) relative to when availability is delayed, even when correcting for delay durations.

In natural settings, we rarely fully load VWM to capacity (Van der Stigchel, 2020), and as such the current paradigm is not directly aimed at, nor suited for, making statements about the capacity limits of VWM (Oberauer et al., 2018). Rather, the current paradigm allows one to investigate how VWM is used in more naturalistic and noisy settings, where strategies, preferences, and executive functioning, amongst others, all play essential roles. During the task, participants are required to encode content into VWM, perform a search task in the pool of available stimuli, and perform actions with the items that they find – all while maintaining a mental map of which items have been placed and which have not. As such, working memory must be utilized in multiple formats (i.e., visual representations, spatial locations, etc.) and content must be utilized in multiple modalities (i.e., item recognition, recall, memory updating; Oberauer et al., 2018). When provided with such complexity, introducing any additional working memory load may introduce undesired noise (e.g., Bays, 2014; Oberauer & Lin, 2017; Schurgin et al., 2020), thereby making the task not only more effortful, but also more prone to mistakes. We here focused on manipulating the ability to sample just-in-time, but it is clear that the paradigm provides a rich environment in which to study different aspects of working memory and to place them within a formalized model (Ngiam, 2023).

In sum, we here investigated how the trade-off between storing in visual working memory versus sampling from the external world shifts, as the ability to sample external information just-in-time was manipulated. Generally, any disruption to the continuous availability of external information, such as intermittent occlusion or a delay period, were the main drivers of increased memory usage. There was no consistent evidence that further manipulations of the frequency or predictability

with which information became available affected the storage-sampling trade-off. These findings suggest that the cost of external sampling is primarily driven by time costs rather than predictability of those time costs.

Supplementary Materials and Data Availability
All code and data can be retrieved from the Open Science Framework https://osf.io/z2n5x/.

Acknowledgements
This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 863732). The authors thank Wouter Baars and Noa Hoevers for assistance with data collection.

Author contributions
AJH: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing. **SB**: Conceptualization, Methodology, Writing – Review & Editing. **CS**: Conceptualization, Writing – Original Draft, Writing – Review & Editing, Supervision. **TCWN**: Conceptualization, Writing – Review & Editing, Supervision. **SVdS**: Conceptualization, Writing – Review & Editing, Supervision, Funding Acquisition.

Chapter 2

Don't hide the instruction manual: A dynamic trade-off between using internal and external templates during visual search

Alex J. Hoogerbrugge
Christoph Strauch
Tanja C. W. Nijboer
Stefan Van der Stigchel

Published as:
Hoogerbrugge, A. J., Strauch, C., Nijboer, T. C. W., & Van der Stigchel, S. (2023). Don't hide the instruction manual: A dynamic trade-off between using internal and external templates during visual search. *Journal of Vision, 23*(7), 14. doi.org/10.1167/jov.23.7.14

Abstract

Visual search is typically studied by requiring participants to memorize a template initially, for which they subsequently search in a crowded display. Search in daily life, however, often involves templates that remain accessible externally, and may therefore be (re)attended for just-in-time encoding or to refresh internal template representations. Here, we show that participants indeed use external templates during search when given the chance. This behaviour was observed during both simple and complex search, scaled with task difficulty, and was associated with improved performance. Furthermore, we show that participants used external sampling not only to offload memory, but also as a means of verifying whether the template was remembered correctly at the end of trials. We conclude that the external world may not only provide the challenge (e.g., distractors), but may dynamically ease search. These results argue for extensions of state-of-the-art models of search, as external sampling seems to be used frequently, in at least two ways, and is actually beneficial for task performance. Our findings support a model of visual working memory that emphasizes a resource-efficient trade-off between storing and (re)attending external information.

2.1 Introduction

When we shop for groceries, lay a jigsaw puzzle, or attempt to assemble a piece of Swedish furniture, we must perform visual search. In order to find an object (e.g., a specific screw), we must actively keep a search template in working memory, and then search for it amongst other items. Once we are confident that an attended stimulus matches our template, search is finalized (Olivers & Eimer, 2011; J. Palmer et al., 2000; Wolfe, 2021).

It is no surprise that visual search has been well-investigated for almost a century, given how fundamental this process is for everyday life (e.g., Kingsley, 1932; Titchener, 1924; Wolfe, 2010, 2021). In traditional paradigms, a template has to be maintained in visual working memory (VWM) throughout search, after transient and singular presentation (e.g., Wolfe, 2021). After the offset of the template, participants are presented with a search array and have to indicate whether the target was present. Exhaustive and well-established models for this visual search (e.g., Wolfe, 1994, 2021) explain not only the underlying processes, but also when and why search goes wrong. For instance, the difficulty of visual search scales with stimulus complexity, set size, and many other factors (Anderson, 1996; Cain et al., 2013; Hulleman & Olivers, 2017; Wolfe, 1998, 2021).

Although the conventional experimental set-up has provided many insights into search, many instances of search in daily life differ. Think of your personal experience when it comes to assembling a piece of Swedish furniture, for instance: When searching for two unique screws from a bag full of differing types of screws, we may regularly fail to identify both of our targets in the first attempt. Luckily, we can always choose to memorize and search for one screw first, refer back to the instruction manual, and then search for the other. Similarly, we can look back at the instruction manual in order to refresh our template representations in VWM whenever we feel insufficiently confident that we indeed found the screw that we were looking for.

The external world can thus often help us to refresh the template throughout search, effectively lowering the burden on VWM. Then, the external world may not only *provide the challenge* (e.g., the search display), but also *ease the challenge*, by allowing to resample the template. Indeed, during many tasks, humans look back and forth at instructions in order to help them succeed (Alfandari et al., 2019; Droll & Hayhoe, 2007; Hansen et al., 2018; Hayhoe et al., 2003; Sullivan et al., 2021). This behaviour is in line with earlier findings on mental effort and memory, which indicate that offloading memory is preferred as much as possible over storing internally, by (re)sampling from the environment in a just-in-time manner (Draschkow et al., 2021; Droll et al., 2005; Hayhoe et al., 2003; Melnik et al., 2018; O'Regan, 1992; Risko & Dunn, 2015; Risko & Gilbert, 2016; Somai et al., 2020; Triesch et al., 2003; Van der Stigchel, 2020).

It is currently unknown how observers balance between internal storage of the template and sampling of the external world in search. We therefore asked whether - and to which degree - participants make use of the option to resample not only the search array, but also the template, when given the chance. To answer these questions, we adopted the following reasoning:

- If participants resample templates throughout search when given the chance,

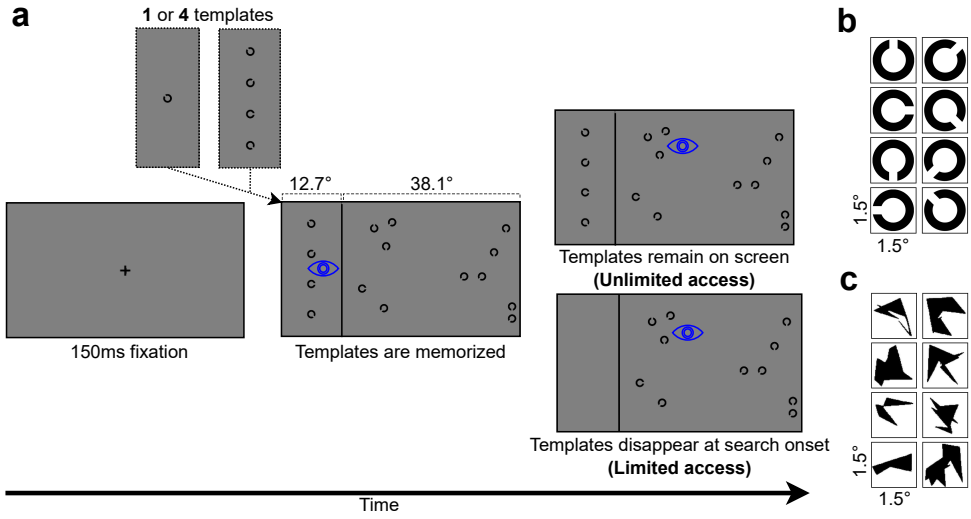


Figure 2.1: **(a)** Sequence of a trial. Trials could contain either one template or four templates. In the unlimited access conditions, templates would remain on the screen throughout a trial. In the limited access conditions, templates would disappear as soon as search started. The vertical line was always present throughout a trial. Stimulus size is not to scale. **(b)** The eight Landolt C's used in Experiment 1. **(c)** The eight original stimuli used in Experiment 2 (Arnoult, 1956). Each stimulus could be shown in one of four rotations, thus creating 32 stimuli. All stimuli occupied approximately 1.5 degrees of visual angle.

this would indicate that they use template availability as a means of **relying on the external world relative to relying on VWM**.

- When templates remain available, the amount of resampling indicates the **degree of reliance on the external world as compared to VWM**. This reliance may also **change as a result of task difficulty**.

- If the amount of resampling is positively associated with better accuracy or completion times, this would indicate a quantifiable **benefit of external sampling** on search.

In order to investigate these questions, participants performed visual search tasks with single- and multi-template search, as well as conditions in which the template(s) remained available throughout a trial, or needed to be encoded up front.

2.2 Experiment 1

2.2.1 Methods

All data together with analysis scripts and supplementary materials may be retrieved via the Open Science Framework https://osf.io/ec7b6/.

Participants and procedure

Nineteen participants performed the experiment, of which two were excluded from analysis due to technical issues and one dropped out during data collection. Thus sixteen participants (8 female, 8 male, age 18-29) were included in the analyses.

Prior to the task, participants read the information letter, signed an informed consent form, and indicated their age and gender. Participants received €7 per hour or course credits, with Experiment 1 taking approximately 60 minutes. The experiment was preceded by four practice trials. The study was approved by the faculty ethics board of Utrecht University, adhering to the declaration of Helsinki.

Apparatus

Monocular gaze location was recorded with an EyeLink 1000+, at a sampling rate of 1kHz. Stimuli were presented on a 27" 2560 × 1440 LCD monitor with a refresh rate of 100Hz. Participants were seated and stabilized with a chin- and forehead rest at 67.5 centimeters from the monitor. The experiment was implemented using PyGaze (Dalmaijer et al., 2014).

All gaze metrics are reported in degrees of visual angle (°). Before the start of the experiment, and between each block, the eye tracker was calibrated and validated with a 9-dot grid, allowing a mean error of 0.5° and a maximum per-dot error of 1.0°. The quality of calibration was automatically evaluated throughout the experiment while each pre-trial fixation cross was presented. If the calibration error exceeded 1.5° over more than two trials, the eye tracker was re-calibrated.

Task and design

Participants performed a visual search task, in which the screen was divided into two sections; a template area and a search area, divided by a vertical line. The template area occupied the leftmost quarter (12.7°) of the screen and contained either one or four templates, dependent on condition. The search area occupied the rightmost three quarters (38.1°) of the screen and contained either one target (matching exactly one of the templates) and ten distractors in target-present trials, or eleven distractors in target-absent trials. Distractors were randomly picked and could therefore be presented multiple times within the search array. Memory loads of one and four templates were chosen such that there were conditions with the minimum required VWM load for any given search task (one template), and conditions which required VWM to be loaded to (or above) capacity if all templates were encoded at once (four templates; Adam et al., 2017; Luck & Vogel, 2013; Vogel & Awh, 2008). 75% of trials were target-present trials. Stimuli were spread out such that participants could not fixate templates and search items simultaneously.

The stimulus set consisted of Landolt C's in eight possible orientations, commonly used in visual search tasks (e.g., Alfandari et al., 2019; Becker, 2011; Carlisle et al., 2011; E. M. Palmer et al., 2019; Smith et al., 2011; Vanyukov et al., 2012). Each stimulus was approximately 1.5° in size (Figure 2.1b).

Before the start of each trial, a central fixation cross was shown, and the trial would only start if a fixation was detected at that location. Participants memorized the template(s) in the template area, and searched for them in the search area; indicating

for each trial whether one of the stimuli in the search area matched a template (by pressing the 'z'-key) or not ('/'-key). After each trial they received feedback, with the screen showing either 'Correct' or 'Incorrect' in blue or red text, respectively (Figure 2.1a). Trials were marked as invalid if the participant indicated that gaze contingent template disappearance did not work as intended.

In the unlimited access condition, templates were visible throughout each entire trial. This allowed participants to gaze back at the templates (resample). The limited access condition followed a classical visual search paradigm by requiring participants to memorize as many templates as possible at once; when participants' gaze crossed the dividing line from the template area towards the search area for the first time, the templates were removed from the screen and could not be sampled again for the remainder of the trial (Figure 2.1a).

The task thus contained four conditions: (1) one template with unlimited access, (2) one template with limited access, (3) four templates with unlimited access, (4) four templates with limited access. These conditions are referred to as 1-Unlimited, 1-Limited, 4-Unlimited, and 4-Limited.

Participants performed 60 trials in each of these four conditions; the sequence of conditions was counterbalanced following a Latin square design.

Analysis

We report three outcome variables. **(1)** Gaze Crossings to Template was extracted by counting the number of saccades which started in the search area and landed in the template area. This variable is representative of the amount of (re)sampling. Since each trial started with a central fixation cross, the minimum number of crossings was always 1, and any value above is indicative of resampling. **(2)** Balanced Accuracy was computed by calculating recall scores (hits divided by number of target-present trials and correct rejections divided by number of target-absent trials, respectively), and taking a weighted average of the two – thereby taking into account the unequal proportion of target-present and -absent trials (implemented using balanced_accuracy_score in scikit-learn; Pedregosa et al., 2011). Balanced Accuracy ranges from 0 to 1, with 0.5 denoting chance-level performance (Brodersen et al., 2010). **(3)** Completion Time was computed as the time in seconds from the first frame in which the trial screen was visible until a keypress was recorded.

For all three outcome variables, trials which were marked as invalid were discarded from the analysis. For Gaze Crossings to Template and Completion Time, only target-present trials and trials with a correct response were considered. Additionally, trials with values beyond the overall 99^{th} percentile were removed. Outcomes of statistical tests with and without these corrections did not substantially differ. In total, 1.5% of trials were marked as invalid, and 1.7% of trials were discarded as outliers.

The median (*Mdn*) and median absolute deviation (*MAD*) are reported for group-level outcomes instead of the mean and standard deviation, in order to better account for non-normally distributed data and group-level comparisons.

Analyses were performed in JASP v0.16.3 (JASP Team, 2022, default priors were used for Bayesian statistics). We report outcomes of Bayesian ANOVAs and *t*-tests, and indicate

whether those tests were performed directionally (BF_{+0}, BF_{-0}) or non-directionally (BF_{10}). Effect sizes (Cohen's d and η_p^2, obtained from classical parametric tests) are reported alongside Bayes Factors. If the assumption of normality was violated, a Bayesian Wilcoxon signed rank test is reported instead, although parametric and non-parametric tests conceptually provided very similar outcomes.

An overview of statistical outcomes is reported in the Supplementary Materials.

2.2.2 Results

Participants would regularly resample if given the chance, even sometimes when only one template needed to be memorized (main effect of template availability BF_{10} = 6706.83, η_p^2 = 0.89; Figure 2.2a). When searching for a single template, participants made slightly more than one gaze crossing per trial from the search area to the template area if templates remained available throughout the trial (1-Unlimited; *Mdn* = 1.07, *MAD* = 0.08; 2.6% of trials contained a second crossing). If the template could only be sampled once, participants did not make additional gaze crossings back to the template area (1-Limited; *Mdn* = 1.0, *MAD* = 0.0; BF_{+0} = 299.7, d = 0.9). This pattern was more accentuated when participants had to memorize four items (main effect number of templates BF_{10} = 497.13, η_p^2 = 0.80; interaction effect BF_{10} = 5.30 $\times 10^{10}$, η_p^2 = 0.79). Here, participants made more crossings when templates remained available (4-Unlimited; *Mdn* = 1.91, *MAD* = 0.46) than when access to the templates was limited, where they made only the initial crossing (4-Limited; *Mdn* = 1.0, *MAD* = 0.0; BF_{+0} = 2.5 $\times 10^5$, d = 2.3).

Overall, there was a main effect of the number of templates on Balanced Accuracy (BF_{10} = 8358.19, η_p^2 = 0.79; Figure 2.2b), but not of template availability (BF_{10} = 2.03, η_p^2 = 0.32). The accuracy was equal between the 1-Unlimited condition (*Mdn* = 0.97, *MAD* = 0.03) and the 1-Limited condition (*Mdn* = 0.97, *MAD* = 0.04; BF_{+0} = 0.4, d = 0.1), suggesting that resampling had no immediate benefit on accuracy in single-template search. With four templates, however, a benefit of unlimited access to the templates was observed, with higher accuracy in the 4-Unlimited condition (*Mdn* = 0.89, *MAD* = 0.05) than in the 4-Limited condition (*Mdn* = 0.84, *MAD* = 0.11; BF_{+0} = 8.5, d = 0.7; interaction effect BF_{10} = 10.26, η_p^2 = 0.33). These findings highlight that the number of templates and template availability dynamically affected accuracy on the task.

A main effect of template availability highlights an overall benefit of the possibility to resample on completion time (BF_{10} = 7.33, η_p^2 = 0.43; Figure 2.2c), although this effect was driven by differences in four-template search and not in single-template search (interaction effect BF_{10} = 91.89, η_p^2 = 0.42). Specifically, template availability did not benefit speed when participants memorized one template. Completion times were similar in the 1-Unlimited condition (*Mdn* = 2.49 s, *MAD* = 0.33) and 1-Limited condition (*Mdn* = 2.60 s, *MAD* = 0.43; BF_{-0} = 0.9, d = -0.3). In the 4-Unlimited condition (*Mdn* = 8.28 s, *MAD* = 1.69), participants were two seconds faster than in the 4-Limited condition (*Mdn* = 10.44 s, *MAD* = 1.11; BF_{-0} = 23.9, d = -0.9). As such, template availability reduced search completion time, but only when searching for multiple templates.

2.2.3 Interim discussion

In Experiment 1, we investigated whether participants preferred to rely on the external world rather than taxing visual working memory (VWM) - and if so, what the extent of this reliance was and whether it changed as a result of task difficulty. Lastly, we investigated whether there was a quantifiable benefit of this reliance on behaviour.

Participants regularly resampled the template area when templates remained available throughout the trial, sometimes even when only one simple template needed to be memorized. Furthermore, this effect was greater in multi-template search compared to single-template search. This indicates that participants often relied on availability of templates when possible, but that the degree of this reliance was dependent on task difficulty.

When memorizing one template, the ability to resample was not linked to an observ-

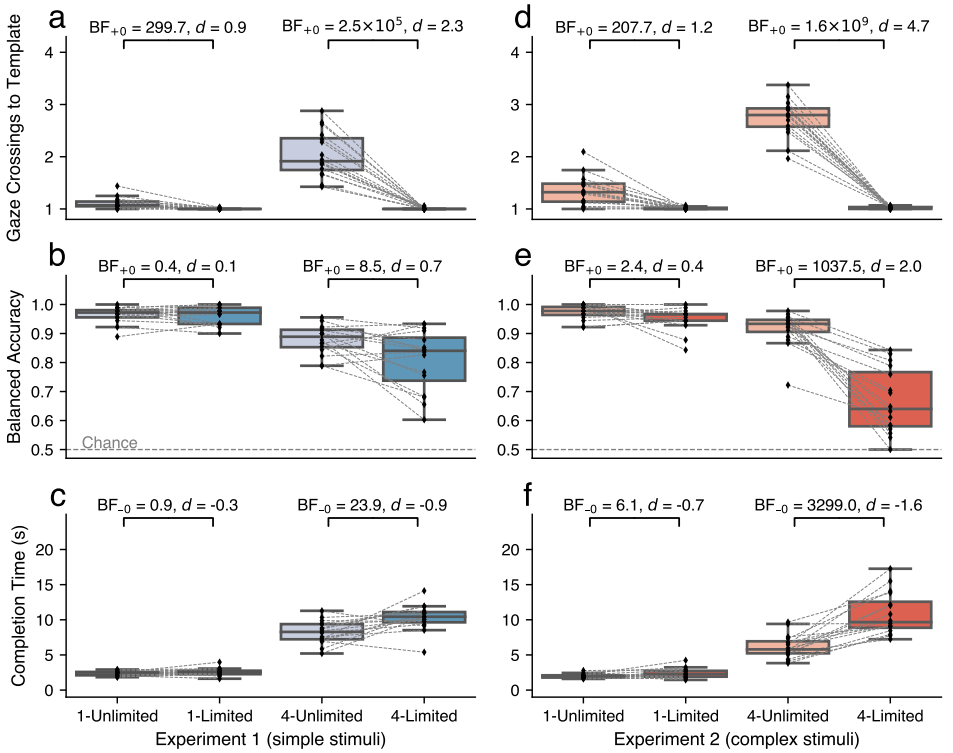


Figure 2.2: Outcome measures of Experiment 1 (left-hand panels) and Experiment 2 (right-hand panels). **(a, d)** The number of times the gaze crossed from the search area to the template area, as a measure of (re)sampling. Since each trial started with a central fixation cross, the minimum number of crossings was always 1, and any value above is indicative of resampling. **(b, e)** Balanced accuracy, which takes into account an unequal proportion of target-present and -absent trials. Chance performance = 0.5. **(c, f)** Trial completion time in seconds, measured from trial start until keypress. Note: All panels except **b** and **e** visualize data of correctly answered and target-present trials only. Diamond markers denote individual participants.

able benefit on classical behavioural outcomes such as accuracy or completion time. However, a benefit of the ability to resample did emerge with four templates instead of one, which indicates that the usefulness of being able to resample becomes more pronounced when the demand on VWM is increased.

Because participants performed very well with one template – which suggests possible floor/ceiling effects – and because the stimuli relied on just one feature (opening direction), we sought to extend the observed phenomena to more complex visual stimuli in Experiment 2.

For Experiment 2 we therefore posited:

- If search difficulty was indeed to affect the degree of reliance on the external world as opposed to VWM, then similar effects as in Experiment 1 should be observed, but **more pronounced in nature with complex stimuli**. This should be observable in single-template search, and be further accentuated in multi-template search.

2.3 Experiment 2

2.3.1 Methods

Experiment 2 followed the same design and procedure as Experiment 1, but with different stimuli.

Participants

Eighteen participants performed the experiment, of which two were excluded due to technical issues. Of the remaining sixteen participants (7 female, 9 male, age 18-29), seven had also participated in Experiment 1. Experiment 2 took approximately 90 minutes to complete.

Stimuli

Stimuli (Figure 2.1c) were a subset of complex shapes introduced by Arnoult (1956), which have been previously employed in VWM research (e.g., Sahakian et al., 2023; Somai et al., 2020).

In order to determine which of the original 30 stimuli were most difficult to verbalize, an online pilot study was run ($N = 48$). Participants indicated which word or name they would assign to each of the stimuli. We then computed the consensus (the percentage of identical or semantically similar responses) and selected the eight stimuli for which consensus was lowest (M consensus of used stimuli = 43%; M of all stimuli = 61%).

Each of the eight selected stimuli could be shown in four configurations (90° rotations), resulting in 32 stimuli. Template and target were considered to be matched only if both the shape and rotation were identical.

2.3.2 Results

Replicating Experiment 1, participants resampled more frequently when templates remained externally available (main effect BF_{10} = 3.69 $\times 10^5$, η_p^2 = 0.92; Figure 2.2d), and this was again stronger in four-template search than in single-template search (interaction effect BF_{10} = 6.68 $\times 10^{21}$, η_p^2 = 0.97). Participants made a greater number of crossings from the search area to the template area in the 1-Unlimited condition (*Mdn* = 1.32, *MAD* = 0.27; 8.7% of trials contained a second crossing) as compared to the 1-Limited condition (*Mdn* = 1.0, *MAD* = 0.0; BF_{+0} = 207.7.6, d = 1.2). They also made a greater number of crossings in the 4-Unlimited condition (*Mdn* = 2.8, *MAD* = 0.26) as compared to the 4-Limited condition (*Mdn* = 1.03, *MAD* = 0.05; BF_{+0} = 1.6 $\times 10^9$, d = 4.7). Beyond replication of Experiment 1, these accentuated effects indicate that the introduction of more complex search templates indeed led to more external sampling.

Overall, unlimited template availability had a positive effect on accuracy (BF_{10} = 1323.97, η_p^2 = 0.81; Figure 2.2e). This effect was not previously present, showing that the introduction of complex stimuli indeed affected task performance. This was again dynamically altered by the number of templates (interaction effect BF_{10} = 2.13 $\times 10^8$, η_p^2 = 0.77). The balanced accuracy was approximately equal between the 1-Unlimited condition (*Mdn* = 0.98, *MAD* = 0.03) and the 1-Limited condition (*Mdn* = 0.97, *MAD* = 0.03; BF_{+0} = 2.4, d = 0.4). Showing a much more pronounced effect than in Experiment 1, however, participants performed the task substantially more accurately in the 4-Unlimited condition (*Mdn* = 0.93, *MAD* = 0.03) than in the 4-Limited condition, where some participants even performed near chance level (*Mdn* = 0.64, *MAD* = 0.11; BF_{+0} = 1037.5, d = 2.0). These findings again highlight that template availability can benefit accuracy on the task, but more substantially so with complex stimuli than with simple stimuli.

Participants were consistently faster when they could resample (main effect BF_{10} = 180.74, η_p^2 = 0.72; Figure 2.2f), but this benefit was greater in four-template search than in single-template search (interaction effect BF_{10} = 2.95 $\times 10^5$, η_p^2 = 0.73). Participants were slightly faster in the 1-Unlimited condition (*Mdn* = 1.96 s, *MAD* = 0.27) than in the 1-Limited condition (*Mdn* = 2.29 s, *MAD* = 0.66; BF_{-0} = 6.1, d = -0.7), which indicates a small benefit of the ability to resample on task completion time. The benefit of template availability on completion times was also observed in the 4-Unlimited condition (*Mdn* = 5.79 s, *MAD* = 01.45), showing almost a halving of the completion time as compared to the 4-Limited condition (*Mdn* = 9.67 s, *MAD* = 2.71; BF_{-0} = 3299.0, d = -1.6). Overall, these findings show again that the benefit of template availability on completion time was more pronounced in four- versus single-template search, and in complex- versus simple templates.

2.3.3 Interim discussion

In both Experiments, participants made use of the possibility to resample templates, primarily when four items needed to be memorized. The possibility to resample templates was associated with shorter completion times and higher accuracy.

But why did template availability benefit completion times, given that this would

require more large saccades back and forth between the template- and search areas? And to what end did participants resample? Was it to encode subsets of templates after each gaze crossing or was it to refresh (double-check) existing representations in VWM? Lastly, one may ask whether double-checking was actually beneficial for search accuracy. We address these questions in the following section.

2.4 How was template availability used?

2.4.1 Just-in-time sampling was linked to shorter completion times than fully loading VWM

Analysis

We report two outcome variables aimed at uncovering why participants were slower at the task when they could not resample. These variables inform us how much time was spent encoding templates (Henderson & Ferreira, 2013; Koevoet, Naber, et al., 2023). **(1)** Total Sampling Duration in seconds provides the overall dwell time in the template area, and was computed as the summed duration of all fixations in the template area within each trial. **(2)** Template Fixation Duration in milliseconds arguably indicates how elaborately participants encoded templates, and was extracted by computing the median duration of all fixations in the template area within each trial.

The outcome variables were aggregated by the median per participant, per condition.

Results

Participants spent more time fixating the template area when they could not resample (Figure 2.3a,c). There were main effects of template availability in both experiments (Experiment 1 BF_{10} = 362.45, η_p^2 = 0.66; Experiment 2 BF_{10} = 925.9, η_p^2 = 0.77), and this effect was stronger when four templates needed to be encoded (interaction effects Experiment 1 BF_{10} = 4.74 $\times 10^6$, η_p^2 = 0.85; Experiment 2 BF_{10} = 1.12 $\times 10^9$, η_p^2 = 0.82).

There was no main effect of template availability on total *search* duration in either experiment (BF_{10} = 0.26, η_p^2 = 0.001; see Supplementary Materials Figure 1), meaning that increased template sampling duration was the main cause of the increased trial completion times when templates could not be resampled.

Furthermore, participants fixated longer on individual templates when those templates could not be resampled, regardless of the number of templates that needed to be memorized (main effects of template availability Experiment 1 BF_{10} = 26.0, η_p^2 = 0.54; Experiment 2 BF_{10} = 2419.8, η_p^2 = 0.74; Figure 2.3b,d), which suggests that participants attempted to encode templates more deeply when they knew that they could not resample later.

Together, these findings show that participants spent more time encoding templates when they could not resample them later, which was linked to longer completion times. When templates could be resampled, it therefore seems that encoding fewer templates, and encoding them less deeply, was a relatively efficient strategy which compensated for the added time cost of making multiple gaze crossings.

2.4.2 Templates were encoded just-in-time or refreshed

Analysis

Data of Experiments 1 and 2 were combined (N = 32), using all trials from the conditions in which participants could resample (1-Unlimited and 4-Unlimited, including target-absent and incorrect trials). Outcomes were conceptually similar when including only target-present and correct trials.

We explored with two outcome variables whether resampling was used to just-in-time encode subset of templates, or whether it was used to refresh existing representations in VWM: **(1)** Onset of each gaze crossing to the template area, expressed as a percentage of trial duration. Onsets were defined as the onset of saccades which left the search area and landed in the template area. **(2)** The number of Unique Templates Fixated after each crossing. By definition, this value was always 1 in the 1-Unlimited condition, since only one template was present. In the 4-Unlimited condition, this value could range from 1 to 4.

We next calculated these outcome variables based on whether they described the 1^{st}, 2^{nd}, 3^{rd} or 4^{th} crossing within a trial. Too few 3^{rd} and 5^{th} crossings were made in the 1-Unlimited and 4-Unlimited conditions respectively, so those crossings and subsequent crossings are not reported. Crossings in which no templates were fixated were excluded (4.3%). 4 out of 32 participants did not make any 2^{nd} crossings in the 1-Unlimited conditions and were therefore excluded, leaving 28 remaining partic-

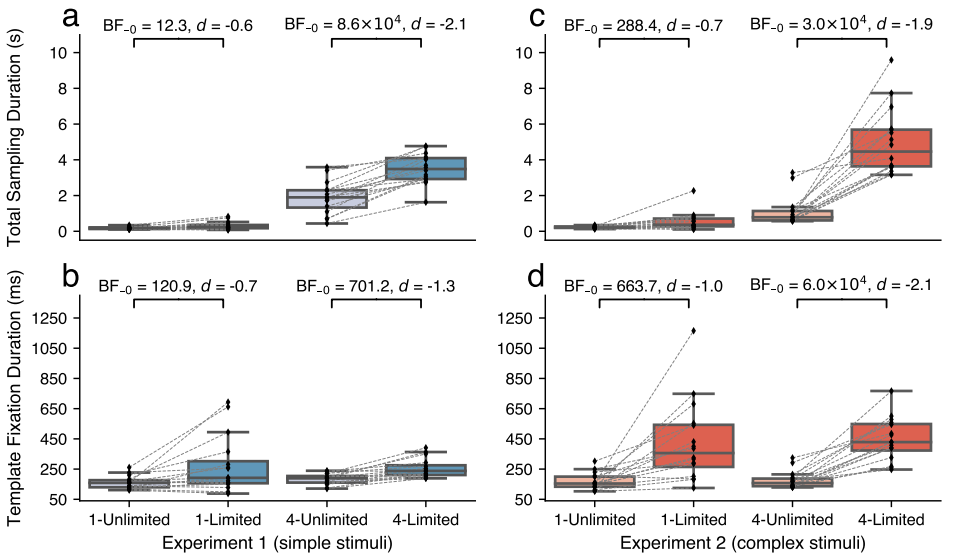


Figure 2.3: Increased dwell times and fixation durations when participants could not resample. **(a, c)** Time spent sampling the template area (in seconds; sum of fixation durations). **(b, d)** Median duration of fixations in the template area (in milliseconds) as a measure of the attempted depth of encoding of individual templates. Note: All panels visualize data of correctly answered and matching trials only. Diamond markers denote individual participants.

ipants for analysis of the 1-Unlimited condition. Additionally, one participant did not make any 3^{rd} or 4^{th} crossings in the 4-Unlimited conditions, and was therefore excluded, leaving 31 remaining participants for analysis of the 4-Unlimited condition. Per outcome variable, values beyond the overall 99^{th} percentile were excluded as outliers. The outcome variables were then aggregated by the mean per participant, per condition.

Results

In both conditions, participants made their first crossing almost immediately after trial onset (1-Unlimited *Mdn* = 8.4%, *MAD* = 2.64%; 4-Unlimited *Mdn* = 4.2%, *MAD* = 2.08%), and thus did not elaborately inspect the search array before crossing towards the template area (Figure 2.4a). When participants needed to memorize one template, secondary gaze crossings were made relatively late in the trial (*Mdn* = 70.0%, *MAD* = 10.65%), which suggests that this crossing often served to "double-check" whether the target was indeed found (or verifying that it was absent), by refreshing the template representation in VWM.

When four templates needed to be memorized, secondary crossings were made relatively earlier in the trial (*Mdn* = 39.9%, *MAD* = 6.63%) than in the one-template condition, BF_{10} = 4437.2, d = 1.6. Third crossings were made just past halfway through the trial (*Mdn* = 57.7%, *MAD* = 6.85%), and fourth crossings (*Mdn* = 69.4%, *MAD* = 5.90%) were made around the same time as secondary crossings in the one-template condition, BF_{10} = 0.2, d = -0.1.

The number of unique templates fixated in the 4-Unlimited condition (Figure 2.4b) suggests two principal strategies: Some participants fixated (i.e., attempted to encode) approximately one template per crossing, in all crossings, thus loading VWM minimally. Other participants rather fixated multiple templates in their first crossing, and fewer templates in subsequent crossings. Most of the latter group of participants (who averaged three or more fixated templates in their initial crossing) still fixated approximately two unique templates in their secondary crossing – which suggests that these participants tried to rely more on memory, but were not always successful in that attempt.

In sum, these findings suggest that resampling was used in two primary ways; either to double-check whether the target indeed matched the template, or as a means to only partially encode (a subset of) the templates in the initial crossing. Subsequent crossings could then be used to just-in-time encode remaining templates, or to strengthen existing VWM representations, if necessary.

2.4.3 Usefulness of template resampling

Analysis

Given that participants could use external sampling both to just-in-time encode subsets and to refresh existing representations in VWM, we investigated more specifically how these strategies were applied. **(1)** The Number of Gaze Crossings to the template area provides a measure of whether resampling was applied differently in target-absent versus target-present trials. In target-present trials, search could regularly terminate before all templates were encoded. Conversely, participants needed to

compare all templates against the search array when there was no target. A greater number of crossings in target-absent than target-present trials would therefore be expected in exhaustive search. This variable was aggregated by the mean number of crossings per participant, per condition. **(2)** Resampling could also serve to refresh existing template representations in VWM. We computed Gaze Ended on Templates; the percentage of trials in which the last fixation of the trial occurred in the template area – meaning that a response was given while (or directly after) fixating a template. Although not comprehensive, this outcome variable represents the majority of instances in which participants double-checked template representations in VWM.

The percentage of trials in which the gaze ended on the templates in the 1-Unlimited condition was analysed by performing a one-sample t-test against 0, because there were no such occurrences of double-checking in the incorrect trials.

Results

Participants made consistently more gaze crossings to the template area in target-absent trials than in target-present trials (main effect BF_{10} = 29317.57, η_p^2 = 0.77), in both the 1-Unlimited condition (BF_{+0} = 5.0, d = 0.4) and in the 4-Unlimited condition (BF_{+0} = 8.57 $\times 10^7$, d = 1.6; Figure 2.5a). In the 4-Unlimited condition, participants crossed nearly four times per trial (*Mdn* = 3.77, *MAD* = 1.29) in target-absent trials, which suggests that they inspected the templates more exhaustively in those trials, and applied resampling dynamically in order to verify that there was indeed no target.

Furthermore, participants' gaze ended in the template area (indicative of double-checking behaviour) more frequently in correctly-answered trials than in incorrect trials in both conditions (main effect of correctness BF_{10} = 429.84, η_p^2 = 0.49; Figure 2.5b), which suggests that double-checking was a useful strategy for achieving higher

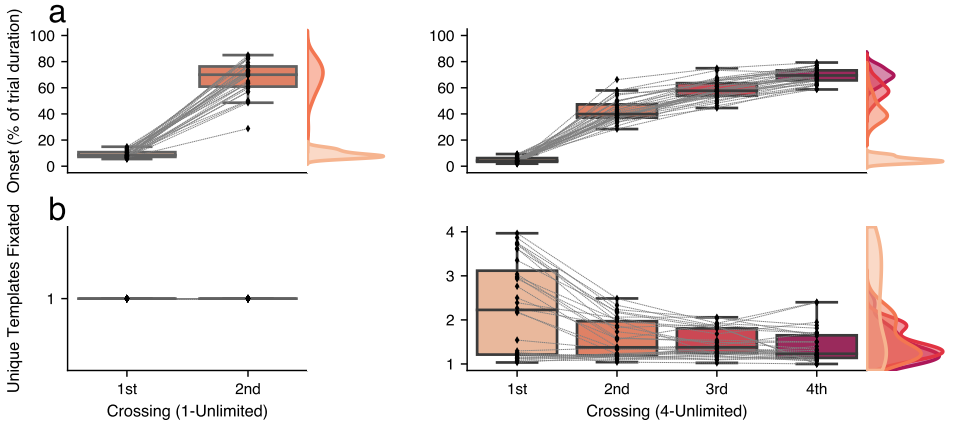


Figure 2.4: Resampling could be used to refresh the template representation in VWM or to encode templates just-in-time. **(a)** The onset of crossings towards the template area, expressed as a percentage of trial duration. **(b)** The number of unique templates that were fixated per crossing. In the 1-Unlimited condition, only one unique template could be fixated. Note: Outcome measures of Experiments 1 and 2 combined. Diamond markers denote individual participants. Not all participants made multiple crossings in all trials.

accuracy. There was no main effect of the number of templates (BF_{10} = 0.53, η_p^2 = 0.07), which indicates that participants used double-checking equally often in both one- and four-template conditions. Although the absolute percentages of correctly-answered trials in which this behaviour occurred were relatively low (1-Unlimited *Mdn* = 5.6%, *MAD* = 8.24%; 4-Unlimited *Mdn* = 9.3%, *MAD* = 8.93%), there was a clear link between double-checking at the end of trials and increased accuracy on the task.

In sum, participants dynamically used externally available templates to their advantage across conditions. For instance, they resampled more often in order to verify target absence, and used double-checking at the end of the trial to achieve higher accuracy.

2.5 General discussion

The role of VWM in visual search has been studied almost exclusively with templates which can only be memorized *before* starting search (e.g., Bahle et al., 2018; Olivers & Eimer, 2011; Van Moorselaar et al., 2014). The external world, however, frequently provides possibilities to offload memory to the environment or to refresh template representations in memory *during* search. Across two experiments, we investigated whether participants delayed the encoding of templates when external templates remained available, whether they refreshed existing template representations, and how this ultimately affected task performance. Results showed that participants used external templates in all conditions that allowed it - in particular by delaying encoding (in line with predictions from VWM research; Ballard et al., 1995; Draschkow

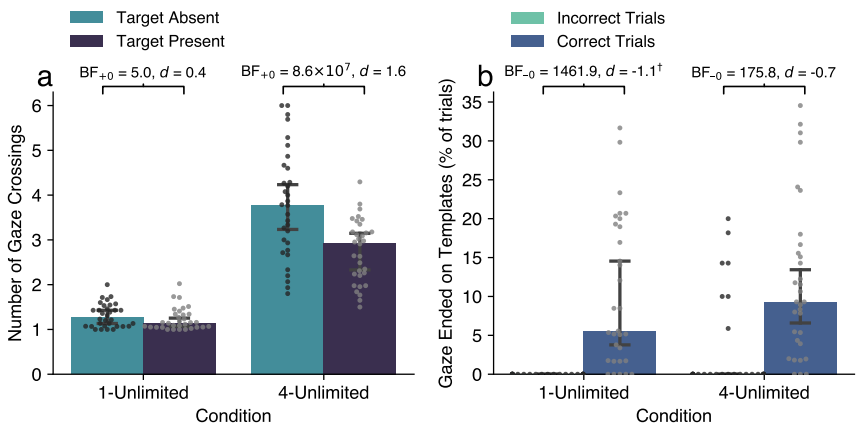


Figure 2.5: The ability to resample was used advantageously in both one- and four-template conditions, and in target-absent and target-present trials. **(a)** The number of gaze crossings per trial as an indicator of the degree of resampling. Split by condition and by target presence. **(b)** The percentage of trials in which the gaze ended in the template area, as an indicator of 'double-checking' behaviour. Split by condition and by incorrectly/correctly-answered trials. Note: Markers denote individual participants, aggregated over all trials. Bars display medians over participants with 95% CIs around the mean. † Results from a one-sample t-test against 0.

et al., 2021; Droll et al., 2005; O'Regan, 1992; Risko & Dunn, 2015; Risko & Gilbert, 2016; Somai et al., 2020; Van der Stigchel, 2020), or by refreshing their existing internal template representations (conform e.g., Alfandari et al., 2019).

Does resampling aid visual search - and if so, in what way? A benefit of the possibility to resample (i.e., on accuracy and completion time; Hulleman & Olivers, 2017; J. Palmer et al., 2000; Wolfe, 2021) was present in all but the easiest condition (with one relatively simple template), and this benefit scaled as search became more difficult (complex stimuli and more templates; Bethell-Fox & Shepard, 1988; Drew & Wolfe, 2014; Eng et al., 2005; Ort & Olivers, 2020; Van Moorselaar et al., 2014). Firstly, participants spent less time dwelling on the templates in their initial inspection when they could resample, compared to when they could not. Secondly, participants made shorter fixations on individual templates when they could resample, which suggests that they attempted to encode the fixated templates less deeply and thereby relied less on internal storage in VWM. We argue that participants spent less time encoding templates when they could resample, because potentially insufficient representations could simply be refreshed later in the trial. This reliance on external templates, relative to fully loading VWM, was therefore temporally efficient in such a way that it offset the cost of making additional saccades between the search- and template areas. As such, being able to resample templates allows for decreased VWM usage in terms of the number of encoded templates *and* depth of encoding, which in turn provides a clear time-benefit to search.

Resampling also provided participants with ways to boost confidence during search, thereby increasing accuracy on the task. Specifically, participants resampled the template area more often in target-absent trials than in target-present trials as a way of verifying that indeed no target was present. Furthermore, participants occasionally refixated the template area directly before giving a response, thereby refreshing existing template representations in VWM - which was linked to higher accuracy. Together, these findings highlight that resampling can benefit accuracy in multiple ways.

Interestingly, the fact that participants could fixate templates for verification at the end of trials must mean that not only template representations were encoded in VWM, but that some elements of the search array were also in memory – not only as elements which help guide search (as in Wolfe, 2021), but also as target templates. In instances of search where external templates remain available, templates and targets can therefore serve interchangeable roles throughout search and within VWM (reminiscent of hybrid search; Drew et al., 2017; Li et al., 2023).

Irrespective of strategy, almost all participants could perform the task at above-chance level (even at the highest difficulty), which suggests that resampling was generally not strictly necessary. However, there were individual differences regarding the number of templates that were fixated in the initial crossing; some participants encoded one template at a time, while other participants attempted to encode multiple templates in each crossing. These individual differences in strategies could in turn relate to individual differences in, for example, VWM capacity or executive functioning.

Furthermore, it is likely that not only the number of templates and stimulus set

influence the degree of external (re)sampling. Varying other aspects of the task, such as the number of distractors, stimulus size, and crowding, may further modulate how frequently templates are resampled. Because distractors could here occur multiple times within the search array (thereby decreasing search difficulty; J. Duncan & Humphreys, 1989), the degree of resampling may have been higher if distractors could not occur redundantly.

In the framework of Guided Search (and alternative models), the final step consists of comparing an attended item in the search array to the template in memory (J. Palmer et al., 2000; Wolfe, 2021). Extending this framework, we suggest that not only the search array can be refixated, but that template representations in VWM may also be resampled before a decision is made. In many instances of search, the external world can therefore not only provide us with the challenge (find a target), but can also ease the challenge (by allowing us to refresh the template or to delay encoding).

2.6 Conclusion

While visual search is commonly studied with to-be-memorized - and subsequently unavailable - search templates, many instances of search are clearly different. For instance, we might be desperate while trying to find that missing screw when assembling a new cupboard, but fortunately we can refresh the template representation by looking back at the manual. Participants frequently revisited templates during search when they were given the chance - more so when search was difficult. How participants used external sampling hereby differed; in some instances participants encoded only subsets of templates, in other instances participants double-checked - both of which benefited search performance. Given that we can resample templates in many instances of visual search, which is often beneficial to task performance, we strongly advise not to hide your Swedish furniture assembly instruction manual. These findings bear implications for influential models of visual search, which should consider the option that not only the search array, but also external templates, can be resampled.

Supplementary Materials and Data Availability
All data, together with analysis scripts and supplementary materials, may be retrieved via the Open Science Framework https://osf.io/ec7b6/.

Acknowledgements
This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 863732). The authors thank Laura van Zantwijk for assistance with data collection; Tessie Hamers, Eline van den Hondel and Mischa van Ouwerkerk for assistance with the pilot study.

Author contributions
AJH: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing. **CS**: Conceptualization, Writing – Original Draft, Writing – Review & Editing, Supervision. **TCWN**: Conceptualization, Writing – Review & Editing, Supervision. **SVdS**: Conceptualization, Writing – Review & Editing, Supervision, Funding Acquisition.

Chapter 3

Persistent resampling of external information despite twenty-five repetitions of the same visual search templates

Alex J. Hoogerbrugge
Christoph Strauch
Tanja C. W. Nijboer
Stefan Van der Stigchel

Published as:
Hoogerbrugge, A. J., Strauch, C., Nijboer, T. C. W., & Van der Stigchel, S. (2024). Persistent resampling of external information despite 25 repetitions of the same visual search templates. *Attention, Perception, & Psychophysics, 86*(1).
doi.org/10.3758/s13414-024-02953-z

Abstract

We commonly load visual working memory minimally when to-be-remembered information remains available in the external world. In visual search, this is characterised by participants frequently resampling previously encoded templates, which helps minimize cognitive effort and improves task performance. If all search templates have been rehearsed many times, they should become strongly represented in memory, possibly eliminating the benefit of reinspections. To test whether repetition indeed leads to less resampling, participants searched for sets of 1, 2, and 4 continuously available search templates. Critically, each unique set of templates was repeated 25 trials consecutively. Although the number of inspections and inspection durations initially decreased strongly when a template set was repeated, behaviour largely stabilised between the tenth and last repetition: Participants kept resampling templates frequently. In Experiment 2, participants performed the same task, but templates became unavailable after 15 repetitions. Strikingly, accuracy remained high even when templates could not be inspected, suggesting that resampling was not strictly necessary in later repetitions. We further show that seemingly 'excessive' resampling behaviour had no direct within-trial benefit to speed nor accuracy, and did not improve performance on long-term memory tests. Rather, we argue that resampling was partially used to boost metacognitive confidence regarding memory representations. As such, eliminating the benefit of minimizing working memory load does not eliminate the persistence with which we sample information from the external world – although the underlying reason for resampling behaviour may be different.

3.1 Introduction

Visual search is one of the most common tasks that we perform throughout the day (Wolfe, 2010). Frequently, we search for our friend in a crowd, for a symbol on our phone's keyboard, or for a screw while assembling furniture. Although these tasks may seem trivial at first, one visual search task requires the completion of several subtasks. For example, we must first know what we are searching for; we won't be able to find the screw that we need if we don't know what it looks like. To this end, we encode the search target (e.g., memorize a picture of the screw from an instruction manual), and maintain a template of its appearance in visual working memory (VWM). Alternatively, if we have done this task before, we can remobilize existing long-term memory (LTM) representations of our target (instead of visually sampling from the instruction manual) and hold it in VWM. This VWM template then helps us guide search towards possibly relevant locations, and allows us to decide at each of those locations whether what we see matches our internally represented template, or is something else (Olivers & Eimer, 2011; J. Palmer et al., 2000; Wolfe, 2021).

It is evident that VWM is an essential component within the guided visual search process, although *when* and *how much* VWM is loaded depends on the task constraints, and the limitations (or facilitation) provided by the environment. Specifically, eye movements are relatively effortless (Koevoet, Strauch, Naber, & Van der Stigchel, 2023; Theeuwes, 2012; Theeuwes et al., 1998), which is why participants generally prefer to make just-in-time eye movements towards relevant external information rather than to load up and maintain 'effortful' VWM representations (Kahneman, 1973; O'Regan, 1992; Van der Stigchel, 2020). This preference has been robustly shown across various tasks, with participants tending to minimize the number of items that they encode into memory (Böing et al., 2023; Draschkow et al., 2021; Droll et al., 2005; Hayhoe et al., 2003; Hoogerbrugge et al., 2023; Koevoet, Naber, et al., 2023; Melnik et al., 2018; Risko & Dunn, 2015; Risko & Gilbert, 2016; Sahakian et al., 2023; Somai et al., 2020; Triesch et al., 2003). Furthermore, participants often reinspect previously encoded items, suggesting that they also encode less elaborately, and prefer to use external information to refresh internal representations instead (Ballard et al., 1995; Hoogerbrugge et al., 2023; Koevoet, Naber, et al., 2023; Sahakian et al., 2023).

This minimization of VWM load also occurs in visual search tasks when templates can be resampled throughout. In those cases, participants mainly encode and search for one template at a time before encoding the next template (Hoogerbrugge et al., 2023; Li et al., 2023). Participants hereby frequently resample external information which they encoded earlier – not only when searching for four complex templates, but also when searching for just one simple template (Alfandari et al., 2019; Hoogerbrugge et al., 2023). This behaviour is beneficial to completion time, accuracy, and effort; participants can spend less time and fewer VWM resources to encode templates, and instead encode or refresh VWM contents only when needed. These findings highlight the relative benefit of dynamically minimizing VWM load with the help of the external world when we can, even on simple search tasks.

The aforementioned studies have commonly considered saccades to be relatively effortless compared to VWM maintenance, but they primarily investigated the short-term dynamics of external sampling versus internal maintenance, i.e., on a trial-

by-trial basis. We here instead investigated these dynamics on a more long-term scale, by examining resampling behaviour over the course of many trials. Say you are assembling a bookcase, and you need the same type of screw twenty-five times during the building process. In that case, it may become more time-efficient and less effortful to elaborately encode the visual search template into (long-term) memory, relative to repeatedly refreshing internal representations by making many saccades towards your instruction manual. For example, as we repeatedly search for the same target, we tend to build up an increasingly elaborate internal representation of that search target (likely in interplay with LTM; Carlisle et al., 2011; Ebbinghaus, 1885; Hout & Goldinger, 2010; Pashler et al., 2007; Woodman et al., 2001, 2007). Moreover, visual search becomes relatively easy and efficient, even for multiple items, when those items are stored in LTM (e.g., Drew & Wolfe, 2014; Drew et al., 2017; Wolfe, 2012; Woodman et al., 2001, although guided search is characterised by different limitations than hybrid search). In other words, on a longer-term scale, resampling of search templates could eventually become redundant – in which case resampling behaviour should eventually cease.

We here investigated whether the preference for (re)sampling external information is persistent, and whether the balance between storing in memory versus sampling externally is different on a long-term scale than on a short-term scale. In two experiments, participants searched for templates which remained available for inspection throughout trials. Critically, each unique template set was repeated twenty-five times consecutively. This should make it more (effort-)efficient to elaborately encode items (either into VWM or LTM), and thus to decrease external sampling behaviour. However, given the persistence of within-trial sampling behaviour (as observed in e.g., Hoogerbrugge et al., 2023), it is uncertain whether participants would opt for a longer-term optimum which may cost more effort in the short-term.

3.2 Experiment 1

3.2.1 Methods

All data together with analysis scripts and Supplementary Materials may be retrieved via the Open Science Framework https://osf.io/nr5qe/. Example videos of trials can be viewed at https://osf.io/hy9dm/. This study was not preregistered.

Participants and procedure

Fifteen participants (13 female, M_{age} = 22.5) performed the experiment. Sample size was based on previous studies using similar paradigms (e.g., Alfandari et al., 2019; Hoogerbrugge et al., 2023)

Prior to the experiment, participants read the information letter, signed an informed consent form, and indicated their age and gender. Participants received €7 per hour or course credits, with Experiment 1 taking approximately 90 minutes. The study was approved by the Faculty Ethics Review Board of Utrecht University (protocol number 21-0297).

Apparatus

Monocular gaze location was recorded with an EyeLink 1000+, at 1kHz. Stimuli were presented on a 27" 2560×1440 LCD monitor at 100Hz. Participants were seated and stabilized with a chin- and forehead rest at 67.5 cm from the monitor. The experiment was implemented with PyGaze (Dalmaijer et al., 2014).

All gaze metrics are reported in degrees of visual angle (°). Before the start of the experiment, and between each block, the eye tracker was calibrated and validated with a 9-dot grid, allowing a mean error of $0.5°$ and a maximum per-dot error of $1.0°$. The quality of calibration was automatically evaluated throughout the experiment while each pre-trial fixation cross was presented. If the gaze prediction error exceeded $1.5°$ for more than two consecutive trials, the eye tracker was re-calibrated.

Fixations were detected using I2MC in Python (Hessels et al., 2017). All fixation candidates shorter than 60 ms were removed, and fixation candidates which were separated by less than $1°$ distance were merged. This approach has been shown to remove variation in gaze event detection between eye trackers and fixation detection algorithms (Hooge et al., 2022).

Stimuli

Stimuli were a subset of complex shapes (introduced by Arnoult, 1956), which are commonly used in VWM research (e.g., Hoogerbrugge et al., 2023; Sahakian et al., 2023; Somai et al., 2020). The stimuli could be shown in four configurations ($90°$ rotations) and in 8 colors, equally spaced along a perceptually uniform color map (HSLuv). One of the original 30 stimuli was removed due to its high rotational symmetry, resulting in 928 unique stimuli. Stimuli were circa $1.5°$ in size.

Task and design

Participants performed a visual search task, in which the screen was divided into two sections by a vertical line; a smaller template area (left) and a larger search area (right; Figure 3.1). The template area occupied the leftmost quarter ($12.7°$) of the screen and contained either 1, 2, or 4 templates. Templates would only be shown when gaze was detected in the template area, such that participants could not peripherally attend templates and search items simultaneously. The search area occupied the rightmost three quarters ($38.1°$) of the screen and contained either one target and 16 distractors in target-present trials, or no target and 17 distractors in target-absent trials. A stimulus was considered a target only if it exactly matched one of the templates (shape, colour and rotation). 75% of trials were target-present trials.

Each trial would only start if a fixation was detected at a central fixation cross. Participants indicated for each trial whether one of the stimuli in the search area matched a template or not by pressing the 'z'-key or '/'-key, respectively. There was no time limit. Participants were instructed to be as fast and accurate as possible, and received feedback after their response ('Correct' or 'Incorrect' in blue or red text, respectively).

Conditions with 1, 2, and 4 templates were blocked and block order was counter-balanced according to a Latin square. Within each condition, participants searched for 6 unique template sets (thus resulting in 18 unique template sets in the whole

experiment), each of which was repeated 25 times consecutively. Across those "repetitions", the template set remained the same, but distractors and their locations were randomly drawn without replacement. A stimulus that had previously been used as a template could not be used as a template nor as a distractor. Participants were informed about the repetitions before the experiment, and were instructed on-screen whenever a new template set was introduced. The experiment was preceded by four repetitions of two template sets as practice trials.

Long-term memory test

After the main body of the experiment was finished, participants were given a five- to ten-minute break. Their recognition of the template sets they had encountered during the experiment was then probed. 18 Template sets from the experiment were presented in random order, interleaved with 6 foils (75%/25%). Due to an error in the code, the true template sets could be probed multiple times – repeated occurrences of a template set were discarded from analysis (this mistake was fixed after Experiment 1). None of the actual templates could occur within the foil sets. Participants indicated whether they recognized the template sets (*yes*/*no*). Participants were informed before the start of the experiment that there would be a long-term memory test.

Analysis

We report four key outcome variables: (a) Gaze Crossings to Templates: the number of times that participants moved their gaze from the search area to the template area as a measure of sampling behaviour; (b) Dwell Time on Templates: the sum of all fixation durations in the template area per repetition as an indicator of how much time participants spent encoding templates (in milliseconds); (c) Response Time: the response time for each repetition, measured from trial onset until keypress (in seconds); (d) Balanced Accuracy, which takes into account the unequal balance

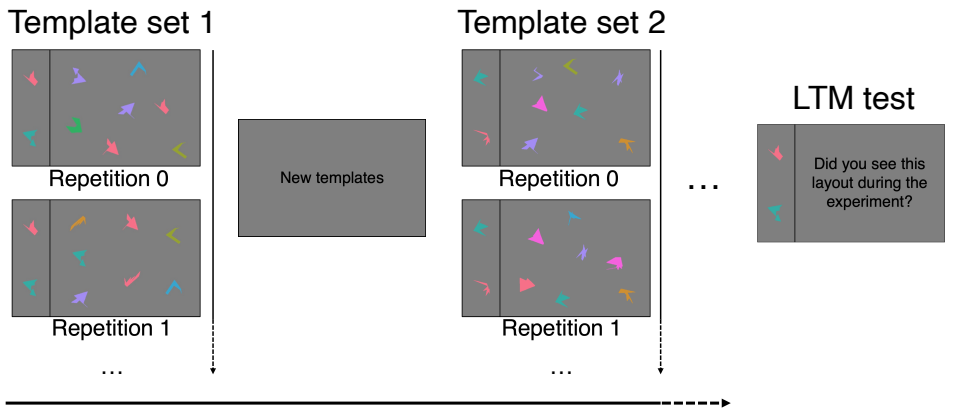


Figure 3.1: Experimental design. Participants searched whether one of the templates on the left-hand side of the vertical bar (12.7°) was present in the search array to the right-hand side of the vertical bar (38.1°). Participants repeated this task twenty-five times in a row, and search trials had no time limit. Across those twenty-five repetitions, the template set remained the same, and the search array was changed. After twenty-five repetitions, participants were shown a screen with the text "New templates". Five to ten minutes after the experiment, long-term memory of template sets was probed. Stimuli are not to scale.

of target-present and target-absent trials (calculated as the mean of sensitivity and specificity; 1.0 denotes perfect accuracy, 0.5 denotes chance-level accuracy; Brodersen et al., 2010).

For Gaze Crossings to Templates, we computed the mean over all six template sets per participant, per repetition. For Dwell Time on Templates and Response Time, the median was computed instead. Balanced Accuracy was also computed over all six template sets, per participant, per repetition.

For statistical analyses, we split each outcome measure into 5 equally sized bins (repetitions 0-4, 5-9, etc.; non-binned figures are reported in the Supplementary Figures 1 & 3). All trials were used for analyses, including target-absent trials and incorrectly-responded trials (unless stated otherwise). Analyses using only correctly-answered target-present trials provided conceptually similar results.

We computed repeated-measures ANOVAs (5 bins x 3 set sizes), and report main effects of bin and template set size, as well as interaction effects between bin and set size. If the assumption of sphericity was violated for an outcome variable, we report corrected ANOVAs (Greenhouse-Geisser if $\varepsilon < .75$, otherwise Huynh-Feldt; following Abdi, 2010).

3.2.2 Results

When a new template set was introduced, participants initially inspected the template area $M = 1.24$ ($SD = 0.36$), $M = 1.89$ ($SD = 0.49$), and $M = 3.21$ ($SD = 1.01$) times for 1, 2, and 4 templates, respectively. The number of inspections generally decreased as template sets were repeated (F (2.2, 30.4) = 76.07, $p < .001$, $\eta_p^2 = .85$; Figure 3.2A), and the degree of this decrease differed between set sizes: the number of gaze crossings decreased faster relative to the initial repetition for one template than for two templates, and faster for two templates than for four templates (interaction effect F (2.6, 36.2) = 4.21, $p = .015$, $\eta_p^2 = .23$; Supplementary Figure 2A). Notably, participants on average still inspected the template area almost twice in the very last repetition of each template set when searching for four templates ($M = 1.79$, $SD = 1.30$). They did so in half of the final repetitions of each template set ($M = 0.51$, $SD = 0.69$) when searching for two templates – highlighting the persistence of resampling behaviour. Only when searching for one template did participants almost stop resampling in the last repetition of each template set ($M = 0.09$, $SD = 0.14$).

Similarly, the amount of time spent inspecting templates per repetition strongly dropped when a template set was first repeated, and then slowly decreased over the course of repetitions. Overall, participants dwelled longer when more templates were presented (F (1.2, 17.4) = 36.99, $p < .001$, $\eta_p^2 = .73$), and dwelled shorter over the course of repetitions (F (1.6, 22.2) = 33.37, $p < .001$, $\eta_p^2 = .70$; Figure 3.2B). Additionally, the degree of decrease in dwell time was different per set size: dwell times decreased faster when searching for one and two templates than when searching for four templates (interaction effect F (2.5, 35.1) = 8.66, $p < .001$, $\eta_p^2 = .38$; Supplementary Figure 2B).

Participants needed longer to complete the search task when searching for greater set sizes, F (2, 28) = 67.0, $p < .001$, $\eta_p^2 = .83$ (Figure 3.2C). Furthermore, response times decreased as template sets were repeated (F (2.1, 29.8) = 26.37, $p < .001$, $\eta_p^2 =$

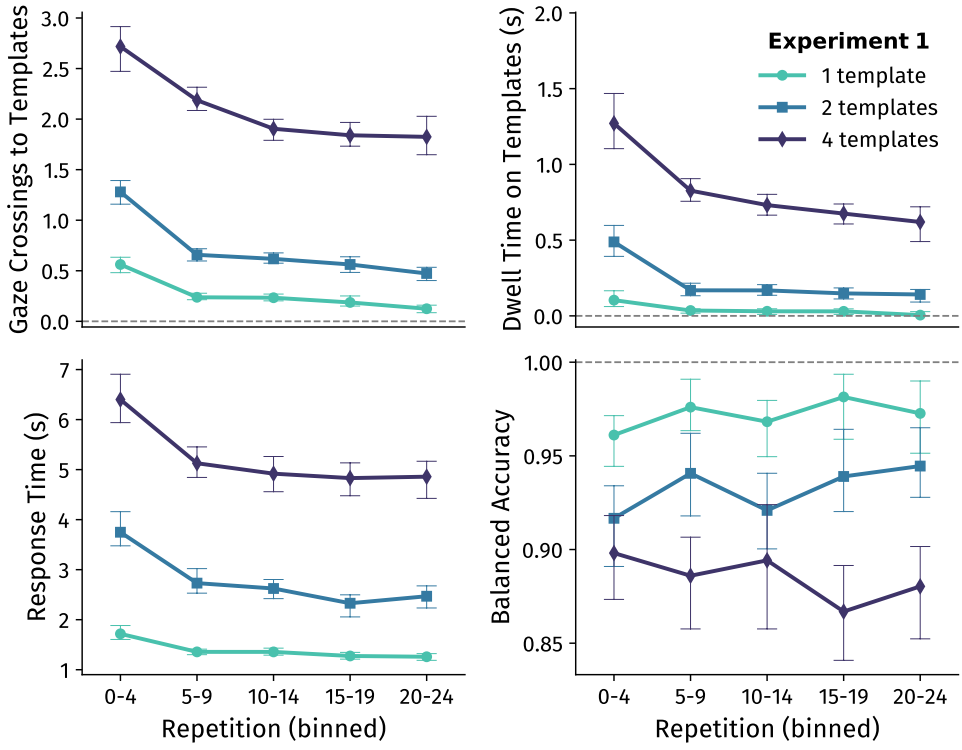


Figure 3.2: Experiment 1 outcome measures. Data was aggregated over all six template sets per participant, split per template set size and binned in sets of five repetitions. The subfigures show across-participant (N=15) averages, ± 95% within-participant confidence intervals (Morey, 2008).

.65), and this decrease differed between set sizes: response times decreased faster when searching for one or two templates than when searching for four templates (interaction effect F (4.6, 64.4) = 3.62, p < .001, η_p^2 = .21; Supplementary Figure 2C).

Participants achieved higher accuracy when searching for fewer templates F (1.3, 18.2) = 14.07, p < .001, η_p^2 = .50 (Figure 3.2D), but did not get better or worse as template sets were repeated (p = .28), nor was there an interaction between number of templates and number of repetitions (p = .32).

3.2.3 Interim discussion

Our findings show that resampling behaviour is quite persistent when visual search templates remain available. Even after twenty-five repetitions of the same templates, participants still inspected those templates almost twice per trial when searching for four templates, and once every two trials when searching for two templates. Only when searching for one single template did participants almost stop resampling. Over the course of those twenty-five repetitions, participants became quicker at completing the task, but accuracy remained stable over time.

These findings raise the question whether participants resampled templates because it helped them maintain high accuracy (i.e., template availability was *necessary* for accurate maintenance of template representations in memory). To test whether resampling was strictly necessary even after many repetitions, we ran a follow-up experiment and tested whether participants could maintain high accuracy if templates could not be inspected anymore. In Experiment 2, we allowed participants to build up memory representations by keeping templates available for inspection in the first fifteen repetitions of a template set, but removed the templates for the last ten repetitions.

3.3 Experiment 2

3.3.1 Methods

Experiment 2 followed the same design, procedure, and analysis as Experiment 1, except for the following: Fourteen participants performed the experiment (12 female, M_{age} = 21.6), one of which was an additional participant to replace a corrupted dataset. Participants searched for 2 or 4 templates. In each condition, 8 unique template sets were presented (resulting in 16 unique template sets across the experiment), each repeated 25 times. Importantly, templates could not be inspected in the last 10 repetitions of each template set. The template area was colored a darker shade of gray in those repetitions. Participants were made aware of this before the experiment, but were not provided with the actual repetition numbers. In the long-term memory test, all 16 template sets were probed in random order, interleaved with 16 foils.

The number of gaze crossings and dwell times were analyzed without taking into account the last 10 repetitions, given that templates could not be inspected in those repetitions.

3.3.2 Results

Similar to Experiment 1, when a new template set was introduced, participants initially inspected the template area *M* = 2.13 (*SD* = 0.81) and *M* = 4.38 (*SD* = 2.31) times for 2 and 4 templates, respectively. The number of inspections again decreased as template sets were repeated (F (1.2, 15.9) = 74.87, p < .001, η_p^2 = .85; Figure 3.3A), and the degree of this decrease differed between set sizes: the number of inspections decreased faster relative to the initial repetition when searching for two templates than when searching for four templates (interaction effect F (2, 26) = 14.53, p < .001, η_p^2 = .53; Supplementary Figure 4A). Comparing between Experiments 1 and 2 (only the repetitions in which templates were available for inspection), there was an interaction effect between experiment and the number of repetitions (F (1.3, 36.0) = 8.85, p = .003, η_p^2 = .25); participants decreased the number of inspections more quickly in Experiment 2 (Figure 3.4A). Notably, however, the total number of gaze crossings made in the first fifteen repetitions was similar between both experiments. When searching for two templates, participants made *M* = 12.78 (*SD* = 7.66) crossings in Experiment 1, and *M* = 10.92 (*SD* = 7.34) crossings in Experiment 2 (t (27) = 0.67, p = .51, d = .25). When searching for four templates, participants made *M* = 34.04 (*SD* = 15.61) crossings in Experiment 1, and *M* = 28.98 (*SD* = 12.13) crossings in Experiment 2 (t (27) = 0.97, p =

.34, d = .36). As such, participants had overall seen templates an equal amount of times in both experiments – but approached the first fifteen repetitions differently in Experiment 2 than in Experiment 1.

This finding is further reflected by dwell time in the first fifteen repetitions; participants initially dwelled longer than in Experiment 1 but decreased their dwell times more quickly (interaction effect F (1.2, 32.1) = 6.76, p = .011, η_p^2 = .20; Figure 3.4A). However, overall, participants spent an equal amount of total time inspecting templates in the first fifteen repetitions of both experiments. When searching for two templates, participants had dwelled M = 4.86 (SD = 3.11) seconds in Experiment 1, and M = 5.61 (SD = 3.47) seconds in Experiment 2 (t (27) = -0.61, p = .55, d = .-23). When searching for four templates, participants had dwelled M = 34.04 (SD = 15.61) seconds in Experiment 1, and M = 28.98 (SD = 12.13) seconds in Experiment 2 (t (27) = -0.81, p = .43, d = -.30). Within Experiment 2, the amount of time spent inspecting templates per repetition again initially dropped when a template set was repeated, and then slowly decreased further over the course of repetitions, F (1.2, 15.2) = 52.65, p < .001, η_p^2 = .80 (Figure 3.3B). Overall, there was a set size effect on dwell time (F (1, 13) = 25.52, p < .001, η_p^2 = .66), and the degree of decrease in dwell time across repetitions was different per set size: dwell times decreased faster when searching for two templates than when searching for four templates (interaction effect F (1.2, 16.1) = 14.24, p < .001, η_p^2 = .52; Supplementary Figure 4B) .

Participants again needed more time to complete the search task when searching for greater set sizes, F (1, 13) = 57.39, p < .001, η_p^2 = .82 (Figure 3.3C). Furthermore, response times decreased as template sets were repeated (F (2.0, 26.2) = 24.14, p < .001, η_p^2 = .65), and this decrease differed between set sizes: response times decreased faster when searching for two templates than when searching for four templates (interaction effect F (2.3, 29.8) = 5.0, p = .011, η_p^2 = .28; Supplementary Figure 4C). Importantly, participants did not become slower or faster after template removal: There was no significant difference in response time between the last five repetitions before template removal and the repetitions thereafter, nor the final ten repetitions without templates. Comparing between the two experiments (analyzing all repetitions, but only the two- and four-template conditions), there was no main effect of experiment on response time (F (1, 27) = 0.10, p = 0.749), nor were there interaction effects with template set size (p = .378) or the number of repetitions (p = .147; Figure 3.4A).

Participants again achieved higher accuracy when searching for fewer templates (F (1, 13) = 8.0, p = 0.014, η_p^2 = .38; Figure 3.3D). Generally, participants did not get better or worse over time (p = .083). Strikingly, accuracy was not significantly different between the last five repetitions before template removal and the five repetitions after removal (e.g., 4 templates; t (13) = 1.70, p = .11, d = .45). To further test for the absence of a drop in accuracy after template removal, we used Bayesian ANOVAs to test for the absence of effects (BF_{01}). Comparing bins 2, 3, 4 and 5 (i.e., repetitions 5-24), we found evidence for the absence of a main effect of bin on accuracy; BF_{01} = 2.61. Furthermore, there was moderate evidence for the absence of interaction effect between the number of templates and bin on accuracy; BF_{01} = 5.99. Post-hoc t-tests further indicated weak evidence for the absence of effects between bins 3-4 (BF_{01} = 1.87) and bins 3-5 (BF_{01} = 2.30). Besides the former results, we tested the 4-template condition separately,

since that is where the possible drop in accuracy seems most pronounced. Testing bins 2-5, again there was evidence for the absence of an effect of bin on accuracy; BF_{01} = 2.51. Although these tests do not provide conclusive evidence for the absence of effects, they were consistent regardless of how the effects were tested. These findings suggest that template representations in memory were strong enough to maintain high accuracy. Furthermore, comparing between the two experiments (all repetitions; 2 and 4 templates), there was no significant effect of experiment version on accuracy (p = .993), nor were there interaction effects with template set size (p = .962) or the number of repetitions (p = .473; Figure 3.4A).

3.3.3 Interim discussion

Experiment 2 conceptually replicated the findings of Experiment 1. Although participants seemed to use a different external sampling strategy in Experiment 2 than in Experiment 1 (they sampled more early on, and decreased this more quickly), they had actually visited the template area equally frequently and for the same duration at the end of the first fifteen repetitions as in Experiment 1. Together, our findings highlight that participants did not strictly need to keep resampling templates in order

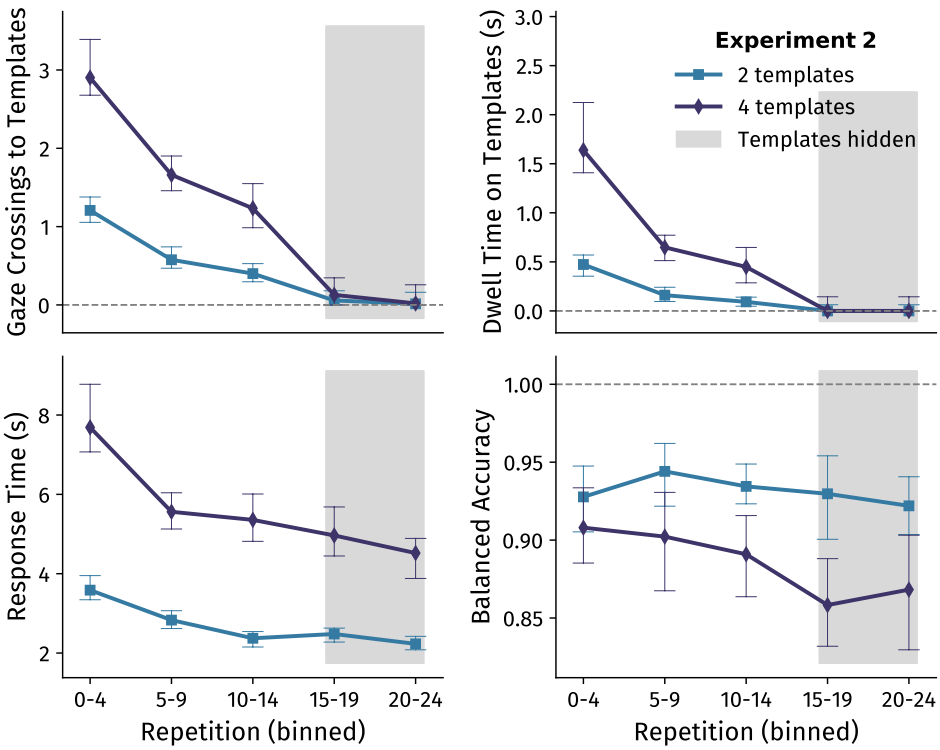


Figure 3.3: Experiment 2 outcome measures. Data was aggregated over all eight template sets per participant, split per template set size and binned in sets of five repetitions. The subfigures show across-participant (N=14) averages, ± 95% within-participant confidence intervals (Morey, 2008).

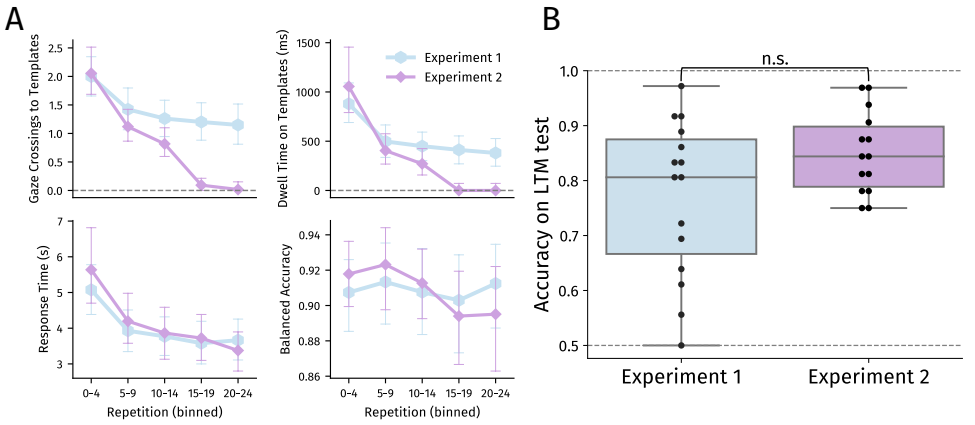


Figure 3.4: (A) Outcome measures as reported in Figure 3.2 and Figure 3.3, collapsed across set sizes 2 and 4, split per experiment. Repetitions were binned into groups of five to provide a clearer overview. Markers denote across-participant averages (N=15 and N=14, respectively), ± 95% within-participant confidence intervals (Morey, 2008). (B) Accuracy (calculated as balanced accuracy) on long-term memory tests. Accuracy of 0.5 denotes chance-level performance, 1.0 denotes perfect accuracy.

to boost speed or accuracy, supporting our notion that participants are persistent in their resampling behaviour, even if it does not aid standard performance metrics.

To understand why resampling was so persistent, we first examined within-trial outcomes to further elucidate whether there was indeed no short-term gain to resampling behaviour. Thereafter, we explored whether participants optimized for longer-term gain, in which case performance on long-term memory tests should improve as a result of more frequent and longer inspections during the task. Finally, we investigated whether participants resampled in order to boost metacognitive confidence rather than improve speed or accuracy.

3.4 The purpose of search template inspections

3.4.1 Short-term efficiency of template inspections

Considering a possible differentiation between short- and long-term optimization, we investigated whether there was an immediate within-trial benefit to making multiple template inspections (as observed in Hoogerbrugge et al., 2023). To analyze this, we combined data from Experiments 1 and 2 and excluded all repetitions from Experiment 2 in which templates could not be inspected. Furthermore, we excluded the first five repetitions of all template sets, given that we wanted to investigate the usefulness of resampling when templates have already been seen several times. Finally, we limited the data to trials with at most four gaze crossings in order to ensure sufficient data for analysis.

Linear Mixed-Effect (LME) models showed that there was a significantly positive effect of the number of gaze crossings to templates on response time, meaning that participants were generally more than a second slower at completing the task for

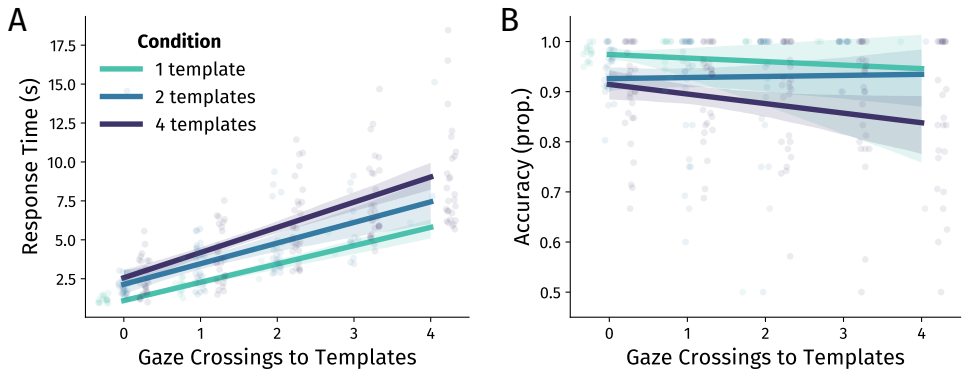


Figure 3.5: The relation between the number of gaze crossings to templates and standard performance metrics: (A) response time, (B) accuracy. Scatterplots show averages split by participant, template set size, and number of gaze crossings. Solid lines denote fitted linear regressions, with shaded areas around the lines denoting the 95% Confidence Interval. Analyses were performed using Linear Mixed Effect models, and as such the regression lines primarily serve illustrative purposes.

each additional crossing they made (β = 1.29, 95% CI [1.01, 1.59], p < .001; Figure 3.5A; see Supplementary Materials section 2 for a description of LMEs). Furthermore, making additional gaze crossings to the templates affected accuracy neither negatively nor positively (LME β = -0.01, 95% CI [-0.04, 0.02], p = .432; Figure 3.5B).

In sum, making additional template inspections was not immediately beneficial for speed. In the case of accuracy, we do not know whether a trial would have been responded to correctly if those additional crossings were not made. Therefore, additional crossings ($\geq$ 0) may have served to maintain a high level of accuracy instead of increasing it. Given that no direct within-trial benefit of resampling could be observed, we hereafter investigated possible long-term benefits.

3.4.2 Long-term efficiency of template inspections

After both experiments, all participants except one could recognize templates above chance level (M_{exp1} = 0.77, SD = 0.14; M_{exp2} = 0.85, SD = 0.08; Figure 3.4B). Thus, participants could draw at least a portion of search templates from LTM during the task – although it remains unclear whether they actually did so, or rather kept templates active in working memory.

Although average LTM accuracy was descriptively higher after Experiment 2, the difference between experiments was not significant (t (27) = -1.87, p = .073, d = -0.69). Therefore, having been able to inspect templates more often (Experiment 1) did not link to better or worse LTM representations than in Experiment 2, in which participants had to search without resampling for the last ten repetitions.

In order to investigate whether resampling templates was globally rather than locally optimized behaviour, we tested whether the amount and duration of inspections during the search task predicted later LTM test accuracy. We limited the data to trials with at most four gaze crossings in order to ensure sufficient data for analysis, and then compared the number of inspections made to LTM test accuracy, per template

set. Participants scored better on the LTM test for larger set sizes (LME β = 0.60, 95% CI [0.13, 1.08], p = .013). A greater amount of inspections of a template set during the experiment was associated with *worse rather than better* recognition of that template set afterwards (LME β = -0.63, 95% CI [-1.17, -0.09], p = .022). To illustrate: Template sets which were on average inspected once or less per trial were recognized with an accuracy of M = 0.76 (SD = 0.15). Inspecting template sets once up to three times was linked to similar accuracy, but with greater variance; M = 0.75 (SD = 0.26) for one-to-two crossings and M = 0.73 (SD = 0.39) for two-to-three crossings, respectively. Template sets which were on average inspected three to four times per trial were recognized worst, M = 0.57 (SD = 0.44). There was no effect of dwell time on templates.

The finding that making more inspections was linked to worse LTM recognition, and that dwell time had no effect on it, suggests that the quality of LTM representations was not a result of less elaborate encoding during the search task. Rather, it suggests that participants at least partially inspected template sets more often when they recognized that their LTM representations of those templates were worse. However, given that those additional inspections did not actually improve LTM performance, it is possible that participants may not have had the confidence to act on their template representations in memory. Therefore, these inspections may have served to boost metacognitive confidence rather than to boost the actual memory representations (in line with Desender et al., 2018; Sahakian et al., 2023).

3.4.3 Template inspections for confidence boosts?

If template re-inspections are used to boost metacognitive confidence, this should be reflected in an increased number of inspections in target-absent trials (similar to e.g., more fixations and longer response times in target-absent search; Gilchrist & Harvey, 2000; Wolfe et al., 2010). Specifically, when there is no target in the search array, one may doubt whether there was indeed no target or whether one has overlooked it. In that case, it reinspecting the templates may be a means of boosting confidence that the target was not overlooked. To this end, we combined data from both experiments (excluding the last 10 repetitions from Experiment 2 in which templates were not available), split per template set size and whether trials were target-present or -absent. Participants indeed made more gaze crossings to templates in target-absent trials overall (F (1, 14) = 45.32, p < .001, η_p^2 = .76; Figure 3.6A), and did this primarily when searching for two templates (t (28) = 6.75, p < .001, Cohen's d = 1.25) and when searching for four templates (t (28) = 9.56, p < .001, d = 1.77).

We additionally investigated behaviour which could not be caused by the experimental manipulation of target presence. Rather, we tested whether there was an increased number of inspections after giving an incorrect response, which is arguably a more natural cause of uncertainty regarding memory representations. Indeed, participants made more gaze crossings to templates after making a mistake in the previous trial, F (1, 14) = 28.27, p < .001, η_p^2 = .67 (Figure 3.6B). Specifically, participants inspected templates more frequently after a mistake when searching for one template (t (14) = 3.75, p = .002, d = .97) and when searching for four templates (t (28) = 4.54, p < .001, d = .84). Although an incorrect response could be caused by poor memory representations, we have already shown that additional inspections did not strongly

aid accuracy. It is therefore likely that a large portion of inspections after errors were used to boost confidence rather than to boost memory representations.

Finally, fixation duration on search targets may reflect the internal decision-making process regarding whether a target or a distractor is fixated (e.g., Becker, 2011; Hooge & Erkelens, 1996; Wolfe, 2021; Wolfe et al., 2022) - and as such, shorter fixations on targets may indicate a higher degree of confidence. We analysed target fixation durations as a function of the number of crossings in each trial. To ensure sufficient data, we included only trials with at most 1, 2, or 4 crossings for each template set size respectively. Participants indeed fixated targets 9 ms less for each additional crossing they made towards the templates (LME β = -9.10, 95% CI [-12.0, -6.21], p < .001; Figure 3.6C), further supporting the notion that decision time is decreased due to higher metacognitive confidence. It should be noted, however, that participants also fixated distractors longer with each additional crossing, although by less than 2 ms (LME β = 1.96, 95% CI [0.67, 3.24], p = .005).

3.5 General Discussion

When to-be-remembered information remains available for inspection throughout a trial, we often prefer to offload working memory in favour of sampling external information only when needed – largely as a means of limiting cognitive effort (Böing et al., 2023; Draschkow et al., 2021; Droll et al., 2005; Hayhoe et al., 2003; Hoogerbrugge et al., 2023; Koevoet, Naber, et al., 2023; Melnik et al., 2018; Risko & Dunn, 2015; Risko & Gilbert, 2016; Sahakian et al., 2023; Somai et al., 2020; Triesch et al., 2003). Besides limiting effort, being able to resample external information in visual search is also considered beneficial for task speed and accuracy, at least when every trial requires us to search for new templates (e.g., Hoogerbrugge et al., 2023; Li et al., 2023).

However, when we repeatedly search for the same templates, it may instead be less cognitively demanding to build strong memory representations of those templates early on, and decrease sampling behaviour as the search task repeats. In two experiments, we investigated whether resampling of external information was still the preferred strategy when participants searched for the exact same templates in twenty-five consecutive trials. In both experiments, when searching for a new, unfamiliar template set, participants made multiple gaze crossings towards the template area per trial – in line with aforementioned findings on short-term optimization. As template sets were repeated, participants sampled external information less, but did not actually stop sampling as long as templates remained available for inspection. In Experiment 2, we removed access to templates after fifteen repetitions, and showed that participants did not strictly need to resample templates in later repetitions in terms of task speed and accuracy. Accuracy remained high when templates were made unavailable, hence our statement that resampling behaviour is *persistent*.

In the short term (i.e., within trials), more (re)sampling was associated with longer trial completion times and no benefit to accuracy. This finding deviates from previous studies (e.g., of visual search; Hoogerbrugge et al., 2023; Li et al., 2023), which introduced new templates on every trial. We therefore investigated whether resampling in our repeated search task followed a different trade-off (storing in memory

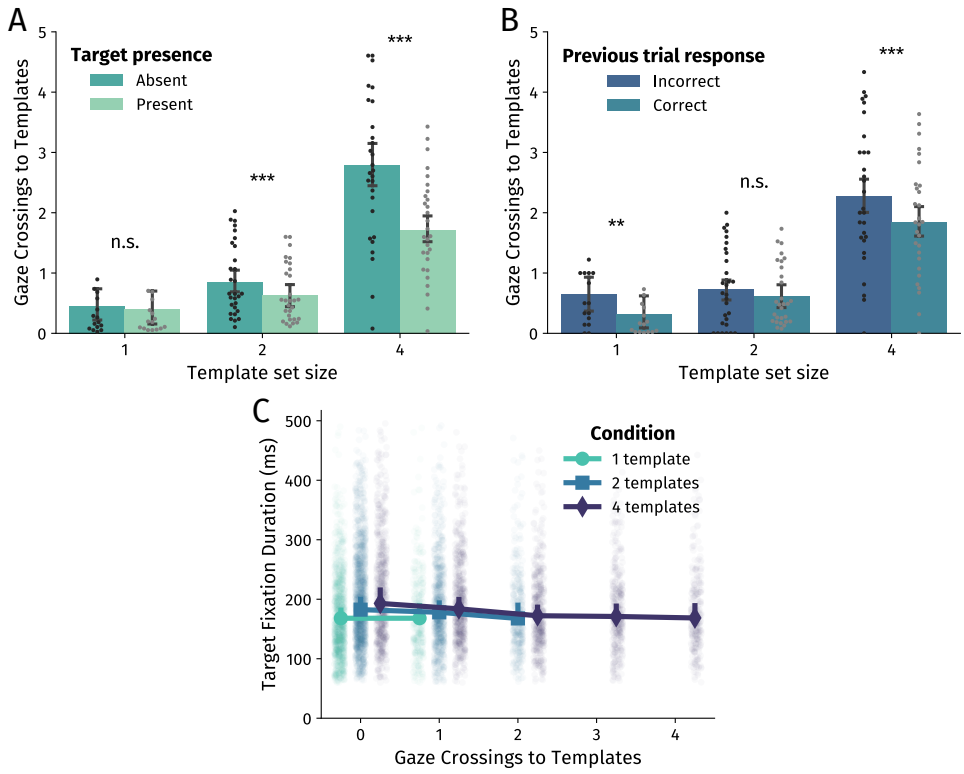


Figure 3.6: (A) Number of gaze crossings to templates, split per set size and whether trials contained a target or no target. (B) Number of gaze crossings to templates, split per set size and whether the previous trial was correct or incorrect. Bars in A and B denote across-participant averages (N=15 for set size 1, N=29 for set sizes 2 and 4), ± 95% within-participant confidence intervals (Morey, 2008). (C) Fixation duration on target, split per condition and the number of gaze crossings to templates. In order to ensure sufficient data for analysis, the number of crossings was cut-off at 1, 2, and 4 crossings for set sizes 1, 2, and 4, respectively. Scatterpoints show non-aggregated data (i.e., all trials without grouping). Larger markers denote medians over these trials, ± 95% within-participant confidence intervals. Note: paired samples t-tests *** $p < .001$; ** $p < .01$; n.s. $p > .05$.

versus sampling externally) than in those studies, and served long-term rather than short-term gain. Five to ten minutes after the main body of the experiment was completed, all but one participant could still recognise template sets that they had encountered at above chance level, meaning that most templates were fairly strongly represented in LTM. It should be noted that participants could have responded that they recognised a template set if they remembered only a subset of a probed set, or the inverse if they recognised none of the foils – this may therefore provide an inflated estimate of how well all templates were represented in LTM. Furthermore, merely being frequently exposed to a target can already strengthen its representation in long-term memory (Carlisle et al., 2011; Ebbinghaus, 1885; Greene & Soto, 2012; Hout & Goldinger, 2010; Pashler et al., 2007; Woodman et al., 2001, 2007), so our LTM test does not elucidate whether templates were already stored in LTM after the first few repetitions (and then consolidated through repeatedly activating those representations in memory), or if resampling in later repetitions still specifically

helped to strengthen template representations. Somewhat surprisingly, we found that more resampling during the search task was actually linked to worse recognition on the LTM test, and that sampling duration did not affect this outcome. This suggests that participants may have inspected templates more frequently when their memory representations of those templates was worse, although it did not seem to help on the later test.

Evidently, the trade-off between storing internally and sampling externally is different when we know that we will have to execute the same task multiple times rather than just once. Of course, some degree of (re)sampling was necessary to (1) initially encode templates and (2) improve the memory quality of those templates or counteract memory decay, but we observed a remarkable amount of seemingly 'excessive' sampling behaviour, given that it often did not contribute to short-term or long-term task performance. What could then explain this behaviour? Our search task contained relatively complex stimuli (polygons which could occur in multiple rotations and colours) which may have been easy to confuse with similar items during search – although the detrimental effect of stimulus complexity on one's ability to memorize and use them does diminish as a result of repeated exposure (Bethell-Fox & Shepard, 1988; Eng et al., 2005). Related, memory contents may be erroneous or may degrade over time (Baddeley & Hitch, 1974; Gold et al., 2005; Hardt et al., 2013; Van der Stigchel, 2020), perhaps even more so for complex stimuli. Besides limiting cognitive effort by offloading memory, resampling valid and stable external information may therefore be preferable over completely relying on 'fallible' memory. In other words, participants may not have had the confidence to act on their template representations in memory, or increased their threshold for which level of confidence they were willing to act on. This idea is in line with Sahakian et al. (2023), who reported that participants sometimes resample external information even when there is still sufficient information in working memory, and that this depends on the ease with which external information can be accessed. Similarly, Desender et al. (2018) showed that participants regularly chose to resample external information due to low subjective confidence, even if objective task accuracy was equal. As such, we explored whether resampling behaviour was partially used to boost metacognitive confidence rather than improving the quality of memory content. Participants inspected templates more often in target-absent trials and after they had given an incorrect response, which indicates that resampling behaviour is influenced by uncertainty stemming from both current and previous trials (Gilchrist & Harvey, 2000; Wolfe et al., 2010). They also made shorter fixations on targets when they had inspected templates more often in that trial, which may indicate that the decision process was faster due to a higher degree of confidence regarding the quality of memory content (Becker, 2011; Hooge & Erkelens, 1996; Wolfe, 2021) .

Perhaps, participants spent overall more effort during the task – encoding items more actively or executing the task more attentively – when they knew that templates would become unavailable, although the effect of increased effort on improvement of memory quality is debatable (Braver et al., 2007; Koevoet, Naber, et al., 2023; Master et al., 2023; Tyler et al., 1979; Zacks et al., 1983). Alternatively, participants resampled so often because they prioritized accuracy over speed. However, due to the nature of the task, if participants truly emphasized accuracy, one would expect almost no mistakes

(because they could resample ad infinitum). Accuracy in both experiments was high, but not perfect – suggesting that participants did give speeded responses. Future research may attempt to elucidate how the persistence of resampling behaviour is affected by speed-accuracy-effort trade-offs.

Making saccades towards templates may also be habitual behaviour. The template area became a darker shade of grey when templates could not be inspected in Experiment 2 – yet, participants sometimes made saccades towards the template area in the first repetition even when they were clearly unavailable (Figure 3.3A), which could point to habitual or reflexive saccadic behaviour. Making a saccade is relatively effortless, and reflexive eye movements have been frequently observed in previous studies (e.g., oculomotor capture; Theeuwes, 2012; Theeuwes et al., 1998), in which involuntary saccades were often made towards novel or salient (task-irrelevant) objects, and gaze could stay at those items for up to 150 ms. In the current study, templates only appeared on screen when gaze was detected in the appropriate location, which eliminates abrupt-onset saliency as factor. Furthermore, individual fixations on templates were generally longer than 150 ms, suggesting that participants mostly made gaze crossings to the template area to actively inspect items rather than out of habit or reflex.

Participants were not told specifically after which repetition templates would become unavailable, so they would be less inclined to postpone elaborate encoding of templates until the last repetition before removal. Participants did seem to change their behaviour somewhat over the course of the experiment, as they learned to estimate when templates would be removed. In later template sets, participants sampled slightly less often and for shorter amounts of time, but the behavioural pattern remained similar: Participants frequently sampled templates when possible. Furthermore, overall response times and accuracy did not change significantly over the course of the experiment. In Supplementary Materials section 3, we report on these possible learning effects in more detail. Moreover, the time-points at which participants inspected templates did not change over the course of trials; see Supplementary Materials section 4.

Participants may have persistently resampled external information for various underlying reasons – i.e., due to individual differences in their working memory capacity, their willingness to load memory, willingness to utilize memory content, and in baseline metacognitive confidence. Together, our findings suggest that at least three factors play a role in the persistent resampling of external information. Besides the unavoidable initial encoding of search templates, and a regular revisit to rehearse and enhance the quality of memory representations, we showed that the boosting of metacognitive confidence is at least one of the additional reasons for this behaviour. But this is not a comprehensive account and other factors may be at play. Moreover, we can only infer that these reasons are at play across the task but cannot estimate this for individual participants or trials.

We have here provided a novel paradigm which may provide further insights into visual search at the intersection of guided search and hybrid search paradigms. Generally, investigations into guided search employ either singletons or novel items on each trial (Liesefeld et al., 2024; Wolfe, 2021), whereas hybrid search investigates

the interplay between searching through an external array and searching through (activated long-term) memory for many items (Drew & Wolfe, 2014; Drew et al., 2017; Wolfe, 2012). Our paradigm allows to investigate whether participants transition from working-memory-guided search towards a more hybrid search approach as search templates are repeated and consolidated in long-term memory. For example, after external templates were hidden in Experiment 2, participants may have suddenly switched from maintaining template representations in VWM towards retrieving those templates from LTM. In that case, the transient (but non-significant) drop in accuracy at that point potentially reflects a switch cost associated with activating representations in, or the retrieval of information from, LTM (Mayr & Kliegl, 2000; Rogers & Monsell, 1995). Moreover, response times remained stable during this period, meaning that *if* a transition between VWM and LTM indeed occurred, one is not necessarily faster than the other.

Together, our findings illustrate the persistence with which external information is resampled, even after twenty-five consecutive searches for the same templates. We here showed that, when eliminating the need to offload working memory, participants still resample external information, but partially to boost metacognitive confidence rather than enhancing the quality of memory representations. As such, we argue that the commonly reported trade-off between storing in memory versus just-in-time sampling externally (which is considered to be an optimization of the expenditure of cognitive effort associated with working memory maintenance) should not only be investigated in a short-term time frame, but should also take into account longer-term optimizations.

Supplementary Materials and Data Availability

All data and Supplementary Materials may be retrieved via the Open Science Framework https://osf.io/nr5qe/. Example videos of trials can be viewed at https://osf.io/hy9dm/.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 863732). The authors thank Noa Hoevers for her assistance with data collection and Joost de Jong for suggesting one of the additional analyses.

Author contributions

AJH: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing. **CS**: Conceptualization, Writing – Original Draft, Writing – Review & Editing, Supervision. **TCWN**: Conceptualization, Writing – Review & Editing, Supervision. **SVdS**: Conceptualization, Writing – Review & Editing, Supervision, Funding Acquisition.

Chapter 4

Multi-target visual search flexibly switches between concurrent and sequential search modes

Alex J. Hoogerbrugge
Christoph Strauch
Noa Hoevers
Christian N. L. Olivers
Tanja C. W. Nijboer
Stefan Van der Stigchel

Under review as:
Hoogerbrugge A. J., Strauch, C., Hoevers, N., Olivers, C. N. L., Nijboer, T. C. W., & Van der Stigchel, S. Multi-target visual search flexibly switches between concurrent and sequential search modes.

Abstract

Investigations into people's ability to use multiple working memory representations to concurrently search for targets have led to mixed findings. Although the discourse has predominantly centered around capacity limits in multi-target search, we here propose that people can switch between sequential and concurrent search. In Experiment 1, manual responses and oculomotor behaviour revealed that participants could search sequentially, and concurrently for at least two targets, when instructed. In Experiments 2a and 2b, participants were free to choose how they searched. Trial-level modelling showed that participants primarily used sequential and concurrent search as specific modes, and flexibly adjusted between either mode dependent on template set size, template availability, stimulus properties, and individual preference. Our findings stress the dynamic and adaptive nature of visual search. Moreover, understanding that different search modes can be flexibly picked as 'tools from the toolbox' may reconcile inconsistencies in prior findings.

4.1 Introduction

When assembling a piece of furniture, you will often need several different screws to finish one set of instructions. In order to achieve this, you may look at the assembly manual, memorise one screw, and search for it among the pile of hardware before moving on to the next screw.

When humans perform such a search task, attention first needs to be directed at the image of the to-be-found item (an *external* target template; Hoogerbrugge et al., 2023), which must then be encoded into visual working memory (VWM) as an *internal* target template (Awh et al., 2006; Fougnie, 2008). The item must subsequently be actively maintained in VWM, such that its representation can be used to guide attention towards locations in the search array which are most likely to contain a matching target (Desimone & Duncan, 1995; Gunseli et al., 2014; Wolfe, 2021). Once a target candidate is selected and attention is shifted towards that candidate, the attended item must be compared to the internal template in VWM, and a target/no-target decision is made (Hout & Goldinger, 2015; Moore & Wolfe, 2001; Ort & Olivers, 2020; J. Palmer et al., 2000).

Alternatively, you may solve the furniture assembly task by attempting to memorise multiple screws and then search until you have found all of them before moving onto the next set of instructions. There are conflicting accounts regarding how attention is deployed during *multi*-target search, and whether multiple distinct templates in working memory can be used to guide attention concurrently. By some accounts, humans are able to hold multiple representations simultaneously activated in VWM, which can then be used to guide search (Beck & Hollingworth, 2017; Beck et al., 2012; Godwin et al., 2015; Grubert et al., 2024; R. S. Williams et al., 2023). Other studies have opposed this account of simultaneous control, arguing that, while multiple representations can be stored in VWM, only one can be used to guide attention in search at any moment in time (Houtkamp & Roelfsema, 2009; Ort et al., 2017, 2019; Van Moorselaar et al., 2014). As such, it is yet unclear whether or not fully concurrent search is possible.

Even if humans can search for multiple targets concurrently, they may not always choose to actually do so. It has been established that humans often prefer to rely on VWM as little as possible, and the degree of this reliance depends on various factors, such as the ease with which external information can be accessed (Ballard et al., 1995; Böing et al., 2023; Draschkow et al., 2021; Hoogerbrugge, Strauch, Böing, et al., 2024; Qing et al., 2024; Somai et al., 2020), the complexity of stimuli (Hoogerbrugge et al., 2023), metacognitive factors (Hoogerbrugge, Strauch, Nijboer, & Van der Stigchel, 2024; Sahakian et al., 2023, 2024), and individual differences in preferred working memory load (Hoogerbrugge et al., 2023; Meyerhoff et al., 2021). One may therefore ask how often humans would opt to search concurrently when also given the option to do it sequentially.

We posit that, even if humans *can* use multiple VWM representations to guide attention, they may not consistently utilize this capability, dependent on both task-related and individual factors – which could partially explain discrepant findings in the literature (cf. Frătescu et al., 2019). In the aforementioned furniture example, this could

either lead one to memorise a single screw at a time and search fully sequentially, or to memorise all items in a single inspection and then search concurrently, or even to memorise all items but still search for them sequentially (i.e., keep other items in accessory states; Olivers et al., 2011). In the present study, we aimed specifically at uncovering whether participants *can* and *do* use multiple VWM representations to guide attention across a search array when given the choice. Furthermore, machine-learning methods allowed us trial-level insights into whether participants searched sequentially or concurrently as specific search modes, or whether they employ a mixture of both modes on a trial level. Lastly, we asked to which degree the choice for specific search modes depends on the individual and on task specifics; namely VWM load, stimulus properties, and template availability.

4.2 Experiment 1: Instructed concurrent and sequential search

Experiment 1 tested whether participants can search sequentially or concurrently when explicitly instructed to do so. This also provided a reference profile for subsequent experiments in which participants were free to choose how to search.

4.2.1 Methods

Participants and procedure

Sixteen participants (15 female, M_{age} = 22.2; SD_{age} = 3.6) with normal vision and no colour-blindness were included in Experiment 1. Three of those 16 participants were replacements for participants who afterward indicated that they either misunderstood the provided task instructions or did not follow them in at least one of the conditions. The excluded datasets are available on OSF. Sample size was based on previous studies using similar paradigms (e.g., Hoogerbrugge, Strauch, Nijboer, & Van der Stigchel, 2024; Hoogerbrugge et al., 2023).

Prior to the experiment, participants read the information letter, signed an informed consent form, and indicated their age and gender. Participants received €7 per hour or course credits, with Experiment 1 taking 60-90 minutes. The study was approved by the Faculty Ethics Review Board of Utrecht University (protocol number 21-0297).

Apparatus

Monocular gaze location was recorded with an EyeLink 1000+, at 1 kHz. Stimuli were presented on a 27" 2560×1440 LCD monitor at 100 Hz. Participants were seated and stabilized with a chin- and forehead rest at 67.5 cm from the monitor. The experiment was implemented with PyGaze (Dalmaijer et al., 2014).

All gaze metrics are reported in degrees of visual angle (°). Before the start of the experiment, and between each block, the eye tracker was calibrated and validated with a 9-dot grid, allowing a mean error of 0.5° and a maximum per-dot error of 1.0°. Fixations were detected using I2MC in Python (Hessels et al., 2017). All fixation candidates shorter than 60 ms were removed, and fixation candidates which were separated by less than 1° distance were merged (cf. Hooge et al., 2022).

Stimuli

Stimuli were L, T and † shapes of circa 1.27° × 1.27° in size. The stimuli could be shown in any of four 90 degree rotations and in any of 7 canonical colours (blue, orange, green, red, purple, pink, yellow), thus providing a set of 84 unique stimuli in total. All stimuli were equally tall and wide, and contained an equal amount of coloured pixels. Stimuli in the search array were separated from each other by circa 2.54°. As such, it was difficult to peripherally distinguish shapes – making stimulus colour the primary feature to guide attention across the search array.

Task and design

Participants performed a visual search task, in which the screen was divided into two sections by a vertical line; a smaller template area and a larger search area (Figure 4.1). The template area occupied the leftmost third (16.8°) of the screen and contained either 2 or 4 equally spaced templates. Templates would only be shown when gaze was detected in the template area, such that participants could not covertly attend templates while their gaze was in the search area. The search area occupied the rightmost two-thirds (32.9°) of the screen and contained a grid of 42 equally-spaced stimuli.

Templates always differed from each other in colour, but their shapes and rotations were identical within trials. As such, in each trial participants effectively only had to memorise one stimulus in 2 or 4 colours. The search area contained 0, 1, or 2 target stimuli in conditions with two templates, or 0-4 targets in conditions with four templates; the other stimuli were distractors. Stimuli were only considered targets if their shape, rotation, and colour matched one of the templates. The same target never occurred multiple times in a trial layout. Depending on the template set size, the search area contained two or four *relevant* colours (matching the template colours) and one additional *irrelevant* colour (Figure 4.1 *inset box*). There were approximately an equal amount of stimuli for each colour. Distractor stimuli could occur multiple times in the search array.

Participants were instructed to click on all targets that they could find, and a circle appeared around each clicked location. When participants were satisfied that they had found all targets, they pressed the spacebar to end the trial.

Importantly, participants were instructed to search in a specific manner in each condition. In *Sequential* conditions, participants were instructed to start by encoding and searching for only the topmost template, while ignoring all other templates. Subsequently, they were told to encode and search the next template from the top while ignoring all others – and so on. As such, participants should have had only one target template in memory at any given time. In *Concurrent* conditions, participants were instructed to encode all templates before searching. Once they believed that they had memorised all templates, they were instructed to search for all targets at the same time. To encourage participants to memorise all items in the Concurrent conditions, templates would not reappear after search onset (when gaze first crossed from the template area to the search area).

Conditions consisted of 40 trials, which were blocked, and block order was counter-balanced according to a Latin square. Each trial started after fixation at a central

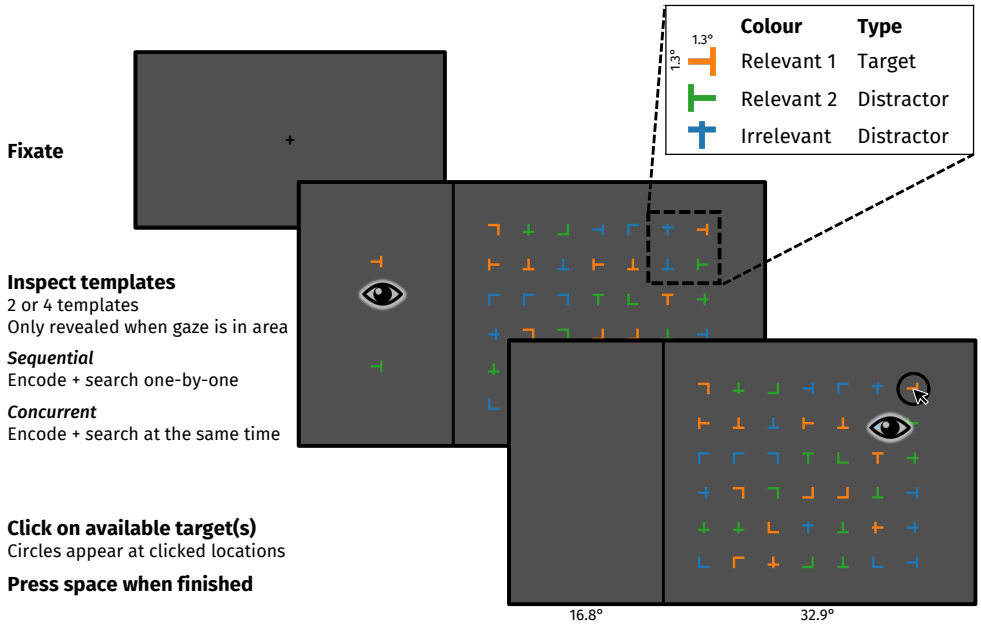


Figure 4.1: Experimental design. Participants searched whether one or more of the templates to the left-hand side of the vertical line were present in the search array. They clicked on each target (shape, rotation, and colour match) that they could find, and pressed the spacebar when they were satisfied that they had found all present targets. There were 2 or 4 templates, and likewise there could be 0-2 or 0-4 targets, respectively. Stimuli in the search area could be of a relevant colour (matching one of the template colours) or an irrelevant colour. In Experiment 1, participants were instructed to either search sequentially or concurrently. In Experiments 2a and 2b, participant were free to choose how to search. In Experiment 1 and 2a, all templates were of the same shape and rotation within trials (as shown in the figure). In Experiment 2b, templates were all of the same shape but could be of different rotations.

cross. The experiment was preceded by four practice trials.

4.2.2 Analysis

Target detection order

As a metric of sequential versus concurrent search, we report the order in which targets were clicked. For each trial, templates were assigned a *Relevant colour index*, from the top-most to bottom-most template. In the case of Figure 4.1, Relevant colour index 1 would therefore be orange, and Relevant colour index 2 would be green. This ordering of relevant colours matched the instructions given to participants.

All target clicks were assigned a proportion based on the total number of clicks in their respective trials, (i.e., 1st click out of 2, 2nd out of 3, etc.), then scaled between 0 and 1 per trial. As such, the first click received value 0 and the last click received value 1. Trials with fewer than two targets and fewer than two clicks were discarded because no slopes could be calculated.

We used linear regression models (Lm2 in Pymer4 0.8.1 Jolly, 2018) to estimate per-participant slopes for target click order, and report the average estimated slope (β),

±95% Confidence Intervals and *p*-values. In the 2-Sequential and 4-Sequential conditions, the expected slopes of a perfect observer would be 1.0 and 0.33, respectively. In the Concurrent conditions, the expected slopes would be 0.

Attentional guidance

Besides target click patterns, we report how frequently colours were fixated over the course of trials. If search occurred fully sequentially, all stimuli of one template colour should be fixated exhaustively (or until a target is found) before stimuli of the next colour are fixated, and so on. Contrary, in fully concurrent search, all template colours should be fixated approximately equally often across a whole trial. In all conditions, if search could be guided effectively based on colour, the irrelevant colour should be fixated at below-chance level.

We filtered all gaze data to retain only fixations within the search array, and each fixation was linked to the colour of the stimulus that was nearest to fixation. Subsequently, each fixation in the search array was assigned to its *relevant* colour index as described in the previous section. The residual colour in the search array (e.g., blue in Figure 4.1) was labelled as *irrelevant*. Because trials contained varying amounts of fixations and had varying response times, we report the normalised time course of a trial, computed with the proportion of fixations made in that trial (e.g., the 5^{th} fixation in a trial with 10 fixations occurs at proportion .5; the last fixation occurs at proportion 1.0). This measure was then split into 20 equally-sized bins. Within each of those bins, we computed the percentage of fixations made on each colour in the search array. Because participants gazed at a central fixation cross before the onset of each trial, the very first detected fixation was usually a residual from that central fixation – and therefore on a random colour.

To evaluate whether relevant colours were searched sequentially or concurrently, we tested whether each relevant colour was fixated more than the other relevant colours, using paired-samples t-tests at each binned time point. To evaluate whether attentional guidance could be used to suppress irrelevant colours, we tested whether the irrelevant colour was fixated below chance level using one-sample t-tests. All statistics were computed using SciPy 1.12 (Virtanen et al., 2020), and corrected for false discovery rate within colours at α = .05 using MNE 1.7.1 (Gramfort et al., 2013).

4.2.3 Results

The number of template inspections, response times and accuracy are reported in Supplementary Materials section 1.

Target detection order shows sequential and concurrent search under instruction

Participants could search sequentially when instructed. This is indicated by the order of targets clicked approximating the expected slopes of 1.0 and 0.33; (Figure 4.2A). When searching for two targets, Relevant colour 1 was clicked before Relevant colour 2, as reflected by a significant regression slope, β = .98, 95% CI [.935, 1.012], p < .001 (98.96% of trials). When searching for four targets, targets were clicked in the order in which they should be searched, β = .32, 95% CI [.294, .347], p < .001.

Participants could also search more concurrently when instructed. However, for the

two-target condition this was not perfect, as the first relevant colour was sometimes clicked earlier than the second colour (59.1% versus 40.9%) β = .17, 95% CI [.024, .326], p = .026), although this slope was closer to 0 (the expected slope for concurrent search) than to 1.0 (expected for sequential search). In four-target search, there was no consistent order in which targets were clicked, providing evidence for concurrent search, β = .03, 95% CI [-.008, .061], p = .128.

Participants could apply both search modes when instructed to do so; slopes were significantly different between sequential and concurrent conditions (2 templates t (15) = 11.12, p < .001; 4 templates t (15) = 13.46, p < .001).

Although target clicks are useful behavioural markers of sequential versus concurrent search, they capture only a few discrete points in time. Moreover, the reported target clicks do not inform us about guidance and may, for example, reflect a strategy in which participants non-selectively scanned the search array. As an extension of target clicks, we next analysed gaze behaviour as a more detailed marker of how participants searched.

Sequential and concurrent search are distinguishable from gaze behaviour

Fixation patterns showed that participants were able to search sequentially for two targets as well as for four targets, with colour as the main guiding feature. Participants encoded and searched each relevant colour sequentially, while successfully ignoring the other colours for which they were not searching at that time (as evidenced by consecutive peaks in fixation frequency; Figure 4.2B). Moreover, and importantly, participants were able to ignore the irrelevant colour throughout the whole trial: all but the first time bin were significantly below chance in both two- and four-target search. Together, these fixation patterns argue that observers were only looking for the target colours and not just any colour, and that attentional guidance was very effective when searching for only one target at a time.

When instructed to search concurrently, gaze behaviour was noticeably different than in sequential search. When searching for two targets, participants fixated both relevant colours equally often throughout trials – at no point was either relevant colour fixated significantly more than the other (Figure 4.2B). Moreover, the irrelevant colour was effectively ignored in all but the first time bin. Given that the first relevant target colour was sometimes clicked before the second relevant colour, we speculate that this colour may have been more strongly represented in VWM, causing the target to be found first on average, without affecting attentional guidance (see General Discussion).

Concurrent four-target search showed similar, but less pronounced, patterns. Relevant colours received an equal amount of fixations over the course of trials, except for in two time bins. Crucially, participants were less able to confine search to the relevant colours and included the irrelevant colour in their search more often than in concurrent search for two targets. The irrelevant colour was fixated significantly below chance in only 11 out of 20 time bins (compared to 19 out of 20 in 2-Concurrent), marking a decrease in the effectiveness of top-down attentional guidance.

In sum, we identified markers (click order and fixation patterns) of both sequential and concurrent search, and show that both modes of search are indeed possible

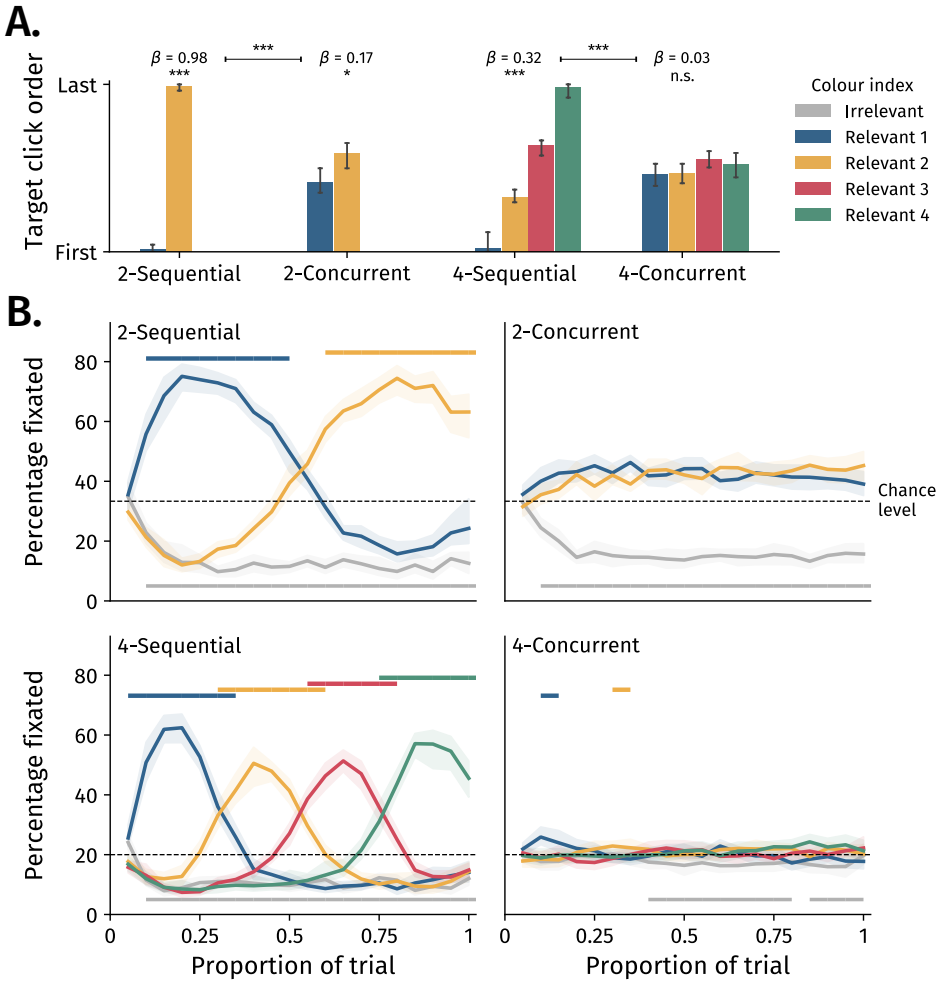


Figure 4.2: Outcomes of Experiment 1. **A.** Target click order. Bars denote across-participant averages, error bars denote ± 95% within-participant confidence intervals (Morey, 2008). β indicates average regression slope, asterisks indicate whether slopes were significantly greater than 0 and whether slopes differed significantly between conditions. *** $p < .001$, * $p < .05$, *n.s.* not significant. **B.** Fixation patterns. The percentages at each time point sum to 100. Given random eye movements, each colour's percentage fixated would be around 33% and 20% for 2- and 4-template conditions, respectively. Lines denote across-participant averages, shaded areas denote ± 95% within-participant confidence intervals. Dashes at the top of figures indicate a significantly greater percentage of fixations on that colour than on the other relevant colours; dashes at the bottom of figures indicate that the irrelevant colour was fixated significantly below chance (corrected $p < .05$).

when instructed, although top-down guidance was weaker when searching for four targets. Perhaps participants used a search mode which was not selective for colour in the latter condition; we speculate about this in the General Discussion.

4.3 Experiments 2a & 2b: Free-choice search

We investigated in Experiments 2a and 2b which mode of search participants actually opted to use when given free choice on how to search. Given consistent findings that humans tend to minimize simultaneous VWM load, one may expect that participants generally prefer sequential search (thereby keeping VWM load low), if those target templates could be reinspected during trials (in line with e.g., Hoogerbrugge, Strauch, Nijboer, & Van der Stigchel, 2024; Hoogerbrugge et al., 2023; Qing et al., 2024). Furthermore, what if participants were forced to memorise all target templates? When all templates are encoded in VWM, participants may opt to search for those items concurrently, or choose to sequentially prioritize representations in VWM while leaving other representations in accessory states (Lewis-Peacock et al., 2012; Olivers et al., 2011).

In Experiment 2, participants were able to reinspect templates in half of conditions (*Unlimited*), but restricted in the other half of conditions (*View-Once*). These two conditions allowed us to study whether participants searched differently when templates remained available in comparison to when VWM must be fully loaded.

4.3.1 Experiment 2a: Low search difficulty

Sixteen participants (12 female, M_{age} = 22.1; SD_{age} = 1.6) with normal vision and no colour-blindness performed Experiment 2a, which took approximately 90 minutes to complete.

Experiment 2a was identical to Experiment 1, except for the following: Participants were not instructed on which search strategy to use, but could either reinspect templates as often as they wished throughout the trial (*Unlimited* conditions), or only gaze in the template area once per trial (*View-Once* conditions). In the latter, they could inspect templates as long as they wanted, but the templates would not reappear after search onset. Blocked conditions consisted of 50 trials.

4.3.2 Experiment 2b: Increased search difficulty

Sixteen participants (12 female, M_{age} = 22.1; SD_{age} = 1.7) with normal vision and no colour-blindness performed Experiment 2b, of which 12 had also participated in Experiment 2a. Experiment 2b took approximately 100 minutes to complete.

Experiment 2b was identical to Experiment 2a, expect for one change: Templates could be shown in varying 90-degree rotations. Again, stimuli in the search array were only considered targets if their shape, rotation, and colour matched one of the templates. As such, participants had to remember more features than in Experiment 2a, making the task more difficult. Rotations of templates were randomly applied, such that not necessarily all templates were rotated differently.

4.3.3 Results

Target detection order reveals a mix of strategies

Target detection patterns in Experiment 2a provided evidence for a mix of sequential and concurrent search (Figure 4.3A). Participants searched at least somewhat sequentially both when templates could be reinspected (Unlimited), as well as when all templates had to be memorised before search onset (View-Once), but these patterns were less pronounced than the sequential patterns in Experiment 1. Notably, participants frequently searched in a sequential manner even when all templates were in memory, suggesting sequential prioritization in memory.

When templates could be reinspected, Relevant colour 1 was often clicked before Relevant colour 2, as reflected by a significant regression slope, β = .360, 95% CI [.140, .580], p = .003. Participants searched somewhat sequentially even when templates could not be reinspected; the first relevant colour was detected before the second colour, β = .331, 95% CI [.187, .475], p < .001. When searching for four items, targets were also clicked in a structured order – and this was the case for both Unlimited trials (β = .076, 95% CI [.026, .126], p = .005) and View-Once trials (β = .075, 95% CI [.025, .126], p = .006). Participants did not detect targets in a significantly different manner between Unlimited and View-Once conditions in Experiment 2a, as evidenced by paired samples t-tests (2 templates t (15) = .33, p = .74; 4 templates t (15) = .03, p = .98).

Neither the ability to reinspect, nor increased stimulus complexity, strongly affected in which order participants clicked on targets – only when searching for four complex targets did participants detect targets relatively more sequentially. In Experiment 2b (increased template complexity), regression slopes were descriptively larger than in Experiment 2a (Figure 4.3C). However, one-tailed paired samples t-tests between the 12 participants who performed both experiments showed that only the slopes in the 4-Unlimited condition were significantly larger in Experiment 2b than in Experiment 2a, t (11) = -3.59, p = .002, d = -1.04 (all other p > .05; the same results hold when performing independent samples t-tests on all participants of both experiments). Moreover, slopes in Experiment 2b were similar between 2-template conditions (t (15) = 1.07, p = .30), but significantly greater in the 4-Unlimited condition than in the 4-View-Once condition (t (15) = 1.13, p = .007).

Eye movements reveal consistent attentional guidance

Fixation patterns also provided evidence for a mix of both sequential and concurrent search when participants were not instructed whether to search sequentially or concurrently. When templates could be reinspected, gaze patterns were not as pronounced as in fully sequential search in Experiment 1, suggesting some degree of concurrent search. Vice versa, in conditions in which all templates were encoded before search onset, participants still exhibited some degree of sequential search.

In Experiment 2a, when searching for two targets, sequential patterns emerged, while the irrelevant colour was effectively ignored (Figure 4.3B). When searching for four targets, moderate peaks could be discerned in sequential order in both the 4-Unlimited and 4-View-Once conditions, but statistically the signal-to-noise ratio was very limited. The irrelevant colour was effectively ignored in all but the 4-Unlimited condition, indicating that participants were predominantly able to guide

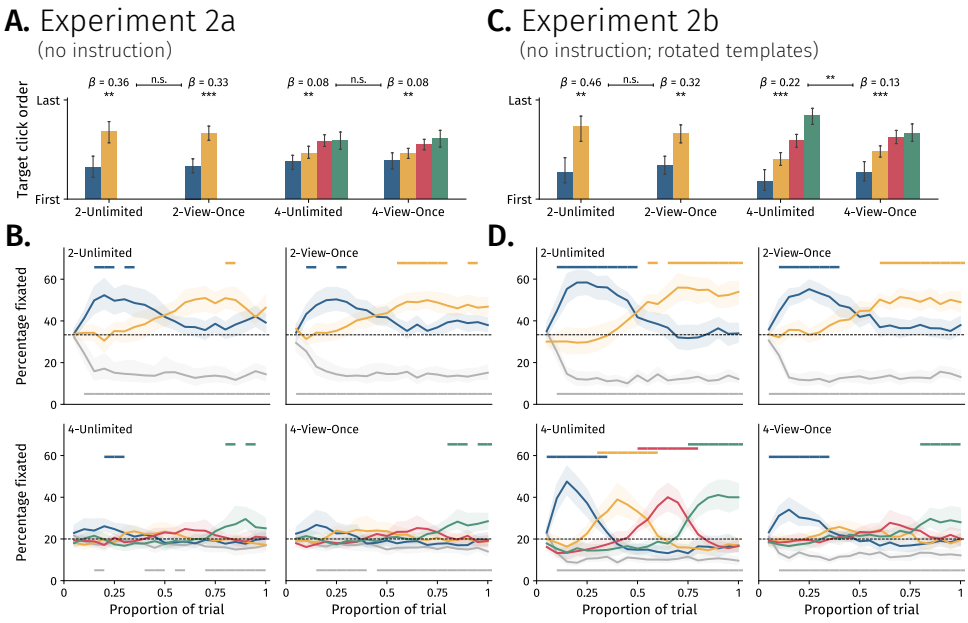


Figure 4.3: Outcomes of Experiments 2a and 2b. **A. & C.** Target click order. Bars denote across-participant averages, error bars denote ± 95% within-participant confidence intervals (Morey, 2008). *** $p < .001$, ** $p < .01$, *n.s.* not significant. **B. & D.** Fixation patterns. The percentages at each time point sum to 100. Given random eye movements, each colour's percentage fixated would be around 33% and 20% for 2- and 4-template conditions, respectively. Lines denote across-participant averages, shaded areas denote ± 95% within-participant confidence intervals. Dashes at the top of figures indicate a significantly greater percentage of fixations on that colour than on the other relevant colours; dashes at the bottom of figures indicate that the irrelevant colour was fixated significantly below chance (corrected $p < .05$).

top-down attention even when all items were in VWM. These findings suggest that participants searched with some degree of sequential prioritization, but often used another search mode. It is unclear why suppression of the irrelevant colour was worse in the 4-Unlimited condition than in other conditions, especially because it was not fully absent; participants still ignored the irrelevant colour in some stages of trials. We speculate on this in the General Discussion.

In Experiment 2b, stronger sequential patterns emerged in all conditions relative to Experiment 2a, further evidenced by more statistically significant differences (Figure 4.3D). Results from eye movements thus resemble those from the target click order; participants searched more sequentially when templates were complex, and there were limited differences in how participants searched between Unlimited and View-Once conditions. Only in the 4-template conditions in Experiment 2b was there a marked difference between Unlimited and View-Once conditions, where participants searched more sequentially when templates could be reinspected.

4.4 What drives concurrent versus sequential search?

Search behaviour in free-choice search showed a mix of search modes, with evidence for a balance between sequential and concurrent search. However, interpretation of our findings may be complicated by averaging artefacts; if half of trials were fully sequential and the other half were either fully concurrent or non-selective, target detection order would show moderate slopes and guidance patterns would show at least some bumps, obstructing clear interpretation of how participants searched. As such, we could not clearly state whether participants switched between strictly sequential and concurrent search between trials, or whether they used a mixed or non-selective search mode within trials. Moreover, even though participants were not instructed to do so, they often encoded and searched templates in our specified order of relevance (top to bottom). Nonetheless, there could be trials in which participants searched sequentially but did not follow this order. Those trials would then reduce the averaged slopes and gaze patterns, giving the impression that search was concurrent. As such, an analysis method was required which was order-agnostic and could indicate for individual trials, rather than on a group level, whether search was sequential or concurrent.

We here introduce a novel analysis method which can dissociate on a trial-by-trial basis whether search was sequential or concurrent. This method allowed us to investigate in more detail how often either mode of search was used within participants and conditions, as well as across participants and experimental manipulations.

We trained Random Forest classifiers on gaze data from Experiment 1, in which participants were instructed on which search strategy to use (*Step 1* in Figure 4.4). These instructions (sequential/concurrent) served as reference labels for our classifier. Importantly, the colour index labels were shuffled on each iteration and for each trial independently – thereby ensuring that the classifier learned to detect patterns instead of specific orders of colour labels. Models validated within Experiment 1 were highly accurate and robust; they classified trials at M_{AUC} = .855 (*SD* = .025) and M_{AUC} = .895 (*SD* = .024) on the 2- and 4-template conditions, respectively (*Step 2* in Figure 4.4 shows the obtained ROC-curves). Having validated that sequential and concurrent search could be clearly dissociated, we trained a new Random Forest model on all data from Experiment 1, and used it to predict search strategy in the data of Experiments 2a and 2b (*Step 3* in Figure 4.4). This process was bootstrapped 1000 times. We report how strongly the behaviour in each trial fits with sequential versus concurrent behaviour, expressed as the percentage of times that the trial was classified as sequential across all bootstrap iterations. Trials in which behaviour was more difficult to classify would thus receive a score close to 50%. We used mixed ANOVAs to discern which factors in our task influenced choice of strategy. For a detailed description of model implementation and analysis, see Supplementary Materials section 2.

4.4.1 Sequential and concurrent are distinct and dissociable 'modes' of search

Predictions from order-agnostic models showed that sequential and concurrent search are clearly dissociable strategies based on gaze alone.

Within conditions, we found that participants used three overarching search 'strategies' (Figure 4.4 shows aggregates across participants; Figure 4.5A shows an example participant; all individual participants and conditions are reported in Supplementary Materials section 3). Most commonly, participants were consistent in opting for one mode of search during the whole condition, evidenced by skewed classification distributions towards either consistently sequential or concurrent behaviour. Alternatively, participants sometimes switched between trials – but within conditions – from sequential to concurrent search modes or vice versa, evidenced by bimodal distributions with most trials classified confidently as either sequential or concurrent behaviour. Least commonly, participants sometimes appeared to employ a hybrid mode of search within trials, neither fully resembling sequential nor concurrent search behaviour. These modes of search were typified by Gaussian distributions of model predictions. Some participants even switched between these three strategy types across the experiment (Figure 4.5A).

For each participant, per condition, we tested whether model classifications across trials deviated from normal distributions using Shapiro-Wilk tests. In all conditions except one, the assumption of normality was violated for 75% to 100% of participants. Only in the 2-Unlimited condition in Experiment 2b, model predictions followed a normal distribution for 50% of participants, suggesting that some participants may have used a hybrid of sequential and concurrent search. In sum, search behaviour was strongly skewed towards either clearly sequential or clearly concurrent search as specific modes of search, although on some trials participants may have employed a mix of the two modes (e.g., concurrent search for both templates, but with one template more strongly represented than the other).

4.4.2 Search modes are used dynamically

Outcomes from our models revealed that participants applied sequential and concurrent search modes dynamically, not only within conditions but also between conditions and across experiments. Across all conditions in both experiments, 51.1% (SD = 33.5) of trials were classified as sequential search, and 48.9% as concurrent search.

Being able to reinspect templates was linked to more sequential search ($F(1, 30) = 13.70$, $p < .001$, $\eta_p^2 = .31$), and this effect was stronger in the more difficult Experiment 2b than in the easier Experiment 2a ($F(1, 30) = 5.30$, $p = .028$, $\eta_p^2 = .15$). Being able to reinspect templates also interacted with the number of templates in its effect on search mode; the higher prevalence of sequential search in Unlimited conditions compared to View-Once conditions was greater when searching for four templates than for two templates ($F(1, 30) = .73$, $p = .004$, $\eta_p^2 = .25$).

In Experiment 2a (Figure 4.4 *bottom left*), when participants searched for two templates, trials were equally likely to be classified as sequential and concurrent; 48.6% ($SD = 30.5$) of trials were classified as sequential when templates remained available, and 48.2% ($SD = 30.6$) when templates could only be viewed once. When participants searched for four templates, 42.7% ($SD = 35.3$, $p < .001$) of trials were classified as sequential when templates remained available, and 38.8% ($SD = 35.4$, $p < .001$) when templates could only be viewed once.

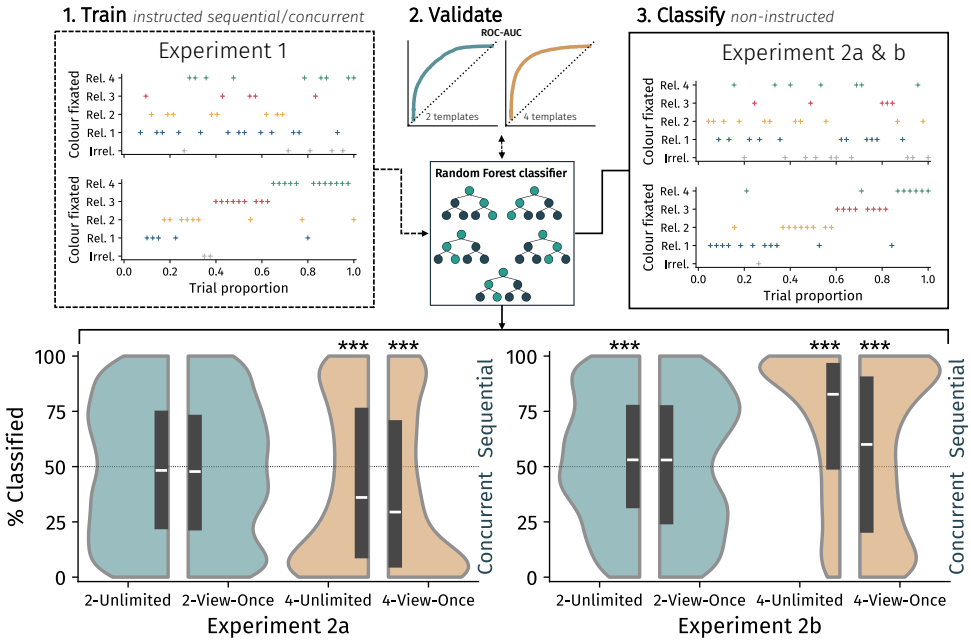


Figure 4.4: Methods and results for dissociating search strategy based on gaze. **1.** Random Forest classifiers were trained on 85% of instructed sequential/concurrent trials from Experiment 1, separately for 2- and 4-template conditions. This process was bootstrapped for 100 iterations. The two subfigures show fixation sequences in example trials from 4-Concurrent and 4-Sequential, respectively. **2.** Trained classifiers were validated using the remaining 15% of trials in each iteration. The two subfigures show averaged ROC-AUC curves from the 100 bootstrap iterations ± 95% range. **3.** New Random Forest classifiers were trained on all data from Experiment 1, and then used to classify data from Experiments 2a and 2b (the two subfigures show example trials). This process was bootstrapped for 1001 iterations, after which classifications were averaged for each trial. The bottom violin plots show the distribution of data (kernel density estimation), as well as the median and interquartile range over all trials. *** $p < .001$ (one-sample t-test against 50% chance level).

In Experiment 2b (Figure 4.4 *bottom right*), participants searched more often in a sequential manner than in Experiment 2a ($F(1, 30) = 10.18$, $p = .003$, $\eta_p^2 = .25$). Moreover, participants adjusted their behaviour relatively more when they were able to reinspect templates in Experiment 2b than in Experiment 2a ($F(1, 30) = 5.30$, $p = .028$, $\eta_p^2 = .15$). When participants searched for two templates, there was a higher prevalence of sequential search when templates could be reinspected (53.4%, $SD = 28.9$, $p < .001$). When templates could only be viewed once, around half of trials were classified as sequential, meaning that both search modes were used approximately equally often (51.0%, $SD = 30.6$). When participants searched for four templates, trials were significantly more often classified as sequential (4-Unlimited: 70.2%, $SD = 31.1$, $p < .001$; 4-View-Once: 55.4%, $SD = 35.7$, $p = .28$).

In sum, participants used both sequential and concurrent search modes dependent on template availability, the number of templates and template complexity.

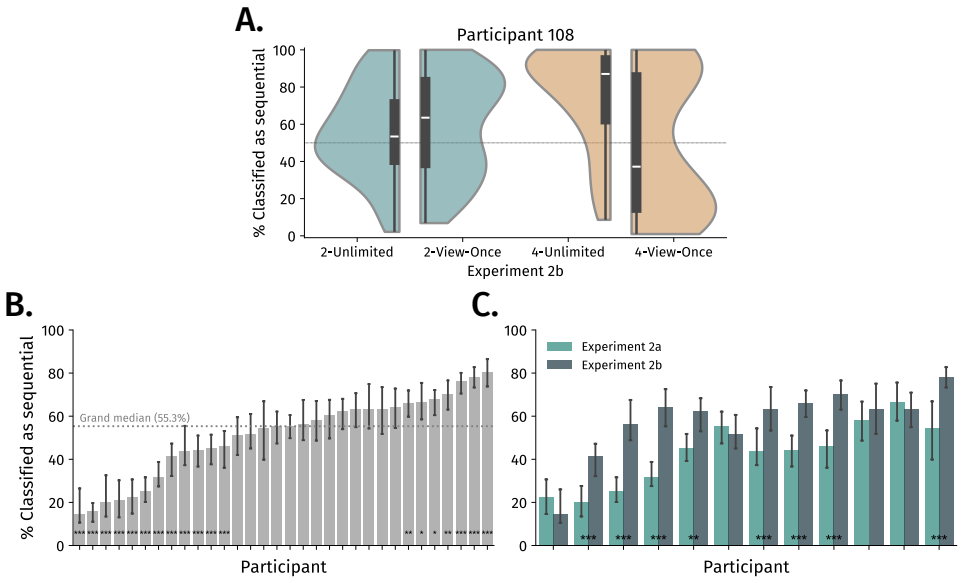


Figure 4.5: Individual differences in search behaviour. **A.** Violin plots of an example participant who showed a less definable mode of search (2-Unlimited), clearly defined sequential search (4-Unlimited), and bimodal distributions of either search mode (2-View-Once, 4-View-Once). Distributions show kernel density estimation, inner boxes show the median and interquartile range. **B.** Model predictions per participant, compared to the grand median of all predictions. Bar heights denote within-participant medians ± 95% confidence intervals. Asterisks denote significance of one-sample t-tests against the grand median. **C.** Model predictions for the 12 participants who participated in both Experiments 2a and 2b, split between experiments. Bar heights denote medians, errorbars denote ± 95% confidence intervals. Asterisks denote significance of t-tests. *** $p < .001$, ** $p < .01$, * $p < .05$.

4.4.3 Individuals use sequential versus concurrent search idiosyncratically

Using predictions from the models, we were further able to distinguish that participants employed sequential versus concurrent search to differing degrees. Overall, 12 out of 32 participants were significantly less often classified as sequential compared to the grand median over all predictions (as indicated by one-sample t-tests, corrected for false discovery rate). Their averages ranged from 14.5% to 46.2%. Conversely, 7 participants were significantly *more* often classified as sequential compared to the average (their averages ranged from 66.2% to 80.4%.; Figure 4.5B).

Not only was search behaviour a result of general individual preference, but individuals differed in the degree to which they changed search modes as a result of stimulus complexity. Eight out of 12 participants who participated in both Experiments 2a and 2b searched significantly more sequentially in Experiment 2b than in Experiment 2a (independent-samples t-tests), whereas 4 participants did not significantly change how they searched (Figure 4.5C).

4.5 General Discussion

Investigations into people's ability to use multiple working memory representations to concurrently search for targets have led to mixed findings. The results presented here provide evidence that people can search concurrently for at least two targets when instructed to do so, and that people use a mix of sequential and concurrent search when given free choice. We further revealed that sequential and concurrent can be considered specific and dissociable modes of search. Finally, we showed that the choice of search mode is flexibly adjusted as a result of task specifics and individual differences.

In Experiment 1, manual responses and oculomotor behaviour showed that participants were able to search concurrently for two targets when instructed to do so, but that attentional guidance suffered when searching concurrently for four items. In Experiments 2a and 2b, participants freely chose how they searched. Here, they exhibited a mix of sequential and concurrent search. A parsimonious model, using only fixation locations over time, was able to predict with high accuracy and robustness whether individual trials in Experiment 1 contained sequential or concurrent search. This model was then applied to make predictions of search strategies in Experiments 2a and 2b. Participants most often used sequential and concurrent search as specific search modes, although some trials were less dissociable, arguing for a hybrid mode of search within a limited number of trials. Our model further revealed that participants were flexible in which search mode they used – dependent on VWM load, stimulus properties, and template availability.

These findings highlight that whether we *can* do something does not mean that we *will* do it – and conversely, whether we don't do something does not mean that we can't. It has been established that humans are conservative in their willingness to expend more cognitive effort than is minimally necessary, as evidenced by e.g., the tendency to avoid simultaneous VWM load when possible (Ballard et al., 1995; O'Regan, 1992; Van der Stigchel, 2020; Wilson, 2002). However, this willingness can be modulated by task demands, and differs on an individual level (Draschkow et al., 2021; Hoogerbrugge et al., 2023; Meyerhoff et al., 2021; Qing et al., 2024; Sahakian et al., 2023; Somai et al., 2020). It follows that, if maintaining multiple attentional templates is cognitively effortful, humans might avoid doing so if the task does not explicitly require them to. In line with this idea, even when participants were forced to memorise all templates, they exhibited some degree of sequential search. On those trials, participants must therefore have sequentially prioritized templates in VWM, possibly leaving the others in accessory states (Olivers et al., 2011). The popular explanation for discrepant findings in the multi-target search literature is that concurrent search is capacity-limited (e.g., Houtkamp & Roelfsema, 2009; Ort & Olivers, 2020; Ort et al., 2017, 2019; Van Moorselaar et al., 2014), and we likewise found that concurrent search for four targets was limited. However, we suggest that the theory should be extended beyond capacity limitations to include the role of willingness to search concurrently, which in turn is related to environmental demands and individual factors.

When participants were instructed to search concurrently for four items, they exhibited behaviour in which they selectively fixated items of the relevant colours less distinctly

and ignored the irrelevant colour less strongly. In that condition, participants may have sometimes used other ways of searching which circumvented our instructions to search concurrently. In some cases, participants may have switched to a shape-search mode, of which only one needed to be remembered. Dropping the colour dimension from memory would still allow participants to find all targets while lowering memory load, but this would also increase the chance of false alarms if correct shapes – but incorrect colours – were clicked. Indeed, false alarm rates were highest in the 4-Concurrent condition in Experiment 1 and the 4-View-Once condition in Experiment 2a compared to other conditions (see Supplementary Materials Figure 1). In the 4-View-Once condition in Experiment 2b, participants could still drop the colour dimension, but even then they had to memorise four shapes. Hit rates in this condition were also markedly lower, which suggests that participants approached a general VWM capacity limit, not only a concurrent search limit. Moreover, participants may have used a memory-search mode, in which they non-selectively shifted attention across the search array (rather than selecting where to fixate next based on guiding templates in memory) and searched through all memory items at each fixation. Shape-search and memory-search may have been used individually, but could also be combined (drop the colour dimension *and* perform memory search). Gaze behaviour can provide support for both non-selective search and shape search (which would arguably decrease the size of the functional viewing field and thereby decrease saccade amplitudes and selectivity; Hulleman & Olivers, 2017; Wolfe, 2021; Wu & Wolfe, 2022). Namely, supplementary analyses showed that saccade amplitudes were indeed smallest and gaze behaviour was most systematic (i.e., scanning the array as if reading a book instead of selecting the best next option) in the 4-Concurrent condition compared to other conditions (although systematicity was not significantly different from the other conditions; see Supplementary Materials Figure 4). However, whether participants used shape-search or non-selective memory-search or both – and if so, to which degree – is difficult to exactly estimate from the current data. Interestingly, saccade amplitudes and systematicity showed much less pronounced effects when participants had free choice on how to search (Experiment 2) than in instructed search (Experiment 1), which may be another explanation for the observed mix of sequential and concurrent search modes; participants chose whichever mode allowed them to retain as much top-down guidance as possible. To avoid the possibility of participants using search modes that are non-selective for the 'intended' guidance dimension (colour in our case), it may be desirable to design multi-target search tasks in such a way that participants cannot reduce their template representation to a single dimension, and to keep track of how they move their gaze across the search array.

Interestingly, because targets occurred at most once per trial, template representations of found targets may be dropped from memory (Lewis-Peacock et al., 2018; Oberauer, 2001). This could in turn leave more cognitive overhead to actively guide search for the remaining items (Olivers et al., 2011). Participants were better able to ignore irrelevant colours in the second half of trials, providing initial evidence for this idea (Figure 4.2; 4-Concurrent). Thus, there may be environmental circumstances in which people use non-selective search, or simplify their search in such a way that only a single VWM representation is required, if possible. However, they may then also be able switch back to concurrently guided search as targets are found and

cognitive load decreases. As such, it would be interesting to study multi-target search with paradigms in which each target can occur multiple times (e.g., foraging tasks, cf. Á. Kristjánsson et al. (2014) and Wolfe (2013); although these may be difficult to reconcile with the aforementioned dimensionality issue).

We additionally suggest that concurrent search does not need to be equally balanced across templates. Our results from instructed search likely showed one such instance; in sequential search for two items, guidance seemed perfectly concurrent, although one item was often found before the other. Relatedly, participants were generally able to ignore the irrelevant colour in free-choice search, even with four template representations in memory. We therefore speculate that templates were not always equally strongly represented in VWM. Possibly, participants were able to have a stronger 'blue' than 'yellow' template (causing more likely detection of the blue target when attending it; Bays & Husain, 2008), while still allowing a similar amount of guidance and suppression from each of those templates (Yu et al., 2023; but see J. R. Williams et al., 2022). These findings highlight the added value of using multiple modalities (i.e., manual responses and gaze behaviour) in order to dissociate distinctive stages of search (Ort & Olivers, 2020) when investigating multi-target search.

In sum, we here report that sequential and concurrent multi-target visual search are both possible (although concurrent search has capacity limits), and that they can be considered two specific modes of search which are applied flexibly. Each of these modes of search may be considered as 'tools in the toolbox' of search strategies, which can be used depending on task demands. In the analogy of furniture assembly; in some cases one requires a steel hammer, and in other cases one requires a wooden mallet – both can do similar jobs but both are best suited for slightly different tasks. We argue that incorporating knowledge about this dynamic application of search modes may contribute towards a better understanding of multi-target search and be the key to reconcile inconsistencies in prior findings.

Supplementary Materials and Data Availability

All data and Supplementary Materials may be retrieved via the Open Science Framework https://osf.io/bqcm4/. Example videos of trials can be viewed at https://osf.io/f5t3e/.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 863732).

Author contributions

AJH: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing. **CS**: Conceptualization, Methodology, Writing – Original Draft, Writing – Review & Editing, Supervision. **NH**: Investigation, Writing – Review & Editing. **CNLO**: Methodology, Writing – Review & Editing. **TCWN**: Conceptualization, Writing – Review & Editing, Supervision. **SVdS**: Conceptualization, Methodology, Writing – Review & Editing, Supervision, Funding Acquisition.

Part II

Individual- and state-dependent influences on eye movements

Chapter 5

Saliency models perform best for women's and young adults' fixations

*Alex J. Hoogerbrugge**
*Christoph Strauch**
Gregor Baer
Ignace T. C. Hooge
Tanja C .W. Nijboer
Sjoerd M. Stuit
Stefan Van der Stigchel

Published as:
Strauch, C.*, Hoogerbrugge, A. J.*, Baer, G., Hooge, I. T. C., Nijboer, T. C. W., Stuit, S. M., & Van der Stigchel, S. (2023). Saliency models perform best for women's and young adults' fixations. *Communications Psychology, 1*(1). doi.org/10.1038/s44271-023-00035-8
* Authors contributed equally and share first-authorship

Abstract

Saliency models seek to predict fixation locations in (human) gaze behaviour. These are typically created to generalize across a wide range of visual scenes but validated using only a few participants. Generalizations across individuals are generally implied. We tested this implied generalization across people, not images, with gaze data of 1,600 participants. Using a single, feature-rich image, we found shortcomings in the prediction of fixations across this diverse sample. Models performed optimally for women and participants aged 18-29. Furthermore, model predictions differed in performance from earlier to later fixations. Our findings show that gaze behavior towards low-level visual input varies across participants and reflects dynamic underlying processes. We conclude that modeling and understanding gaze behavior will require an approach which incorporates differences in gaze behavior across participants and fixations; validates generalizability; and has a critical eye to potential biases in training- and testing data.

5.1 Introduction

The world provides us with rich potential visual input. The immense amount of available information, combined with the unevenly distributed photoreceptor cells of the retina and the limited processing capacity of the visual system, necessitates several steps of prioritization. The first, and perhaps most important, of these steps determines how gaze, and with it visual attention, is allocated across a given scene. The way the eyes are rotated affects which visual information falls on the eyes' highly or lowly resolving parts, and herein lies the foundation of how individuals see and perceive the world. Which aspects of a scene are most salient - and thus determine where observers will likely fixate - is therefore of crucial interest (Itti et al., 1998).

In order to computationally model visual saliency, topographic maps of image features such as local orientation, contrasts, spatial frequencies, or colors are integrated - in other words features that are of importance in early visual areas within the visual cortex. Initial applications of these models were seen in computer vision, namely the prioritization of highly informative locations to deal with limited processing capacity of computers (Itti et al., 1998). It was not long, however, until saliency maps were translated back to vision: Which locations of a scene, based on image features, will most likely attract covert and overt shifts of attention and thus be fixated to optimally use the brain's processing power (Itti & Koch, 2000). These models can be understood as spatial distribution maps that highlight salient over non-salient areas. In turn, such maps can be compared to actual gaze data from benchmarking data sets (Bylinskii et al., 2015; Kümmerer, Bylinskii, et al., 2022) and improvements can be made to iteratively adapt models to become ever closer to gaze data, thereby also improving understanding of how low-level visual input might drive this behavior.

While the initial and seminal models were constructed by considering visual features that are well known to be represented in the early visual areas of the brain (Itti & Koch, 2000), dozens if not hundreds of models have since been proposed, some following similar approaches, some semantically enhanced (e.g., Einhäuser et al., 2008), and some based on deep learning approaches instead (e.g., Kümmerer, Bethge, & Wallis, 2022). Researchers in turn benchmark their models on empirical data which usually contain fixations of a limited number of adult participants (e.g., Coutrot & Guyader, 2014; Judd et al., 2009, 2012) who view a very wide range of static images without instruction (see Kümmerer, Bylinskii, et al., 2022, for an overview). For instance, the influential and vast CAT2000 dataset (Borji & Itti, 2015) contains free viewing data of 120 participants (80 women, 40 men) aged 18-27 years with each of 4,000 images being viewed for 5 s by 24 participants. This approach guarantees generalization across (static) stimuli. Recent calls for more diverse samples in (psychological) research (Cheon et al., 2020; Jones, 2010; Rad et al., 2018) - beyond the overrepresented participant pools of most research universities (Cheon et al., 2020) - raise the question how well such findings may generalize across individuals rather than images. Just as much as demographic biases in training data might bias models, it may be asked whether saliency drives eye movements uniformly over time; that is, whether models are as predictive for the first as for intermediate or later fixations. Both questions can only be studied with massive samples, as fixation maps of single fixations, such as all second fixations, are otherwise too scarce to allow for inferences. With the

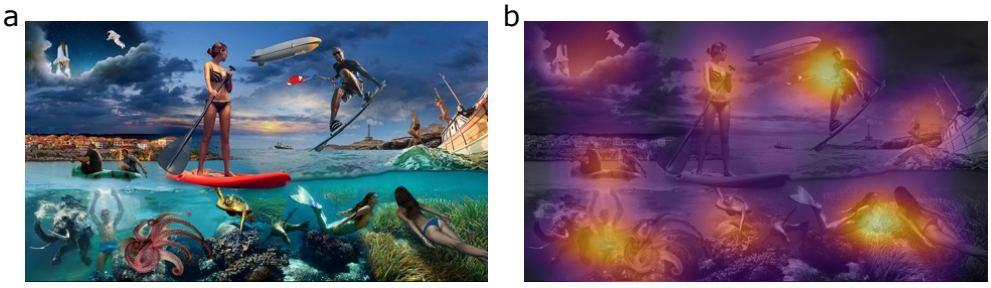


Figure 5.1: Stimulus and overlaid fixation map. Stimulus presented to assess free viewing data (**a**) and spatial distribution map of fixation locations overlaid (**b**) with brighter colors indicating more fixations. Usable gaze data were obtained from n = 2,607 participants using an eye tracker and monitor which were part of an installation at a science museum (see Supplementary Figure 1 for pictures of the setup). Collage made from licensed stock images from Shutterstock

emergence of massive online studies, such generalizability issues are increasingly addressed for other questions in psychology. Large samples, however, remain scarce in eye tracking research. Whether the field of saliency modeling is also affected by said sampling biases as one of the key challenges of present-day psychology (Rad et al., 2018) is therefore still unclear.

Here, we tested the generalizability of model predictions across people and across fixations, not stimuli, and set out to uncover possible biases in model predictions with regard to both gender, age, and across fixations. To this end, we evaluated performance of 21 saliency models, selected upon availability and their huge influence on the field to infer conclusions on saliency mapping as a general discipline. For each saliency model, performance was assessed relative to a spatial distribution map of fixation locations of n = 2,607 participants, including children, on a single image (see Figure 5.1). As further baselines, we employed (1) a central bias, based on the assumption that participants fixate the center more than the periphery (Tatler, 2007); (2) a meaning map (Henderson & Hayes, 2017), constructed of relative meaningfulness ratings for small patches of the stimulus; and 3) a single observer baseline, the averaged predictivity of each participant's fixation locations for fixations of all other participants. Furthermore, we evaluated model performance across fixations - in other words how well early, intermediate, and later viewing behavior could be modeled. For analyses that relate model performance to demographic details of participants, gaze behavior of n = 1,600 participants from 6 to 59 years of age and highest credibility of logged demographics were used.

5.2 Methods

The study was approved by the Ethics Review Board of the Faculty of Social Sciences at Utrecht University.

5.2.1 Participants and data exclusion

Overall, n = 2,607 valid free viewing gaze data sets were obtained, using an installation at the NEMO Science Museum, Amsterdam, which featured a metal box with a screen

Table 5.1: Demographics per age bin

		Age	Gender	
Bin	Sample size	Mean	Men (%)	Women (%)
6-11	58	9.97	48.3	51.7
12-17	149	14.28	64.4	35.6
18-23	249	20.44	55.4	44.6
24-29	431	26.10	54.5	45.5
30-35	242	32.26	50	50
36-41	164	38.60	51.2	48.8
42-47	175	44.36	60.6	39.4
48-53	97	50.08	55.7	44.3
54-59	34	55.94	82.4	17.6

and an eye tracker inside which participants looked into. All analyses not related to demographics were performed on all 2,607 data sets (M_{Age} = 28.79, men = 50.13%, women = 42.9%; non-binary = 6.97%). For analyses relating to demographics, data sets were only considered if no periods of more than 5 s of lost gaze position were recorded over the duration of the whole procedure (including entering demographics). For data sets adhering to this requirement, it is highly unlikely that the participants left the recording between free viewing and entering their demographics. Data sets were further excluded from any demographic-linked analyses if the default options (non-binary gender, year 2000 as year of birth) were not changed by the participant, resulting in n = 1,600 participants with demographics of high credibility (M_{Age} = 29.82, men = 55.6%, women = 44.4%; see Table 5.1 for detailed demographic information). N = 91 participants indicated non-binary gender across the 6-59 age range, but given that there was no option for 'prefer not to say', these data are only given in the Supplement and have to be interpreted with caution.

5.2.2 Statistics and Reproducibility

All statistical tests reported were two-sided. Tests are detailed alongside results. Bayesian tests use default JASP priors. The study was not preregistered.

5.2.3 Apparatus, Stimuli, and Procedure

Gaze was logged (asynchronously) at 60 Hz using a Tobii Eye Tracker 4C. This eye tracker is suited for this research question and setup. In general however, as the Tobii 4C filters the data for its intended use case (gaze interaction), Tobii advises against using it for research. A 27", 1920×1080 px monitor with a maximum luminance of 300 cd/m² was used for stimulus presentation, located at 80 cm distance from the eyes to the screen (50×24 degrees visual angle). A metal box around screen and tracker shielded the field of view from other visual stimulation. Participants could either stand, sit on a chair, or stand on a chair to be able to see the monitor and participate. Other than that, the setup was not height adjustable. Auditory information was given exclusively after free viewing via two loudspeakers positioned close to the participants' ears. See Supplementary Figure 1 for pictures of the setup.

Participants were required to look at a central circle that gradually filled to start the experiment and perform a five-point calibration of the eye tracker. Participants were presented with a full-screen collage image (Figure 5.1) for 10 s of free viewing without instruction. The image was constructed so that it would include a wide variety of objects, both inanimate and animate, both facing or not facing the beholder, as well as free spaces with low information (i.e., empty sea or sky). Participants could decide on whether or not to donate their data by fixating a laterally positioned 'yes' or 'no' button respectively.

Upon giving consent, participants were prompted to indicate their gender by gazing at a central (non-binary), left (man), or right (woman) circle. Subsequently, year of birth could be entered, with 2000 as default option. This year could be iteratively decreased or increased by gazing at a circle on the left or on the right, respectively.

5.2.4 Data quality and preprocessing

Eye tracking data quality can be assessed by precision, accuracy, and data loss (Dunn et al., 2024). While accuracy cannot be assessed with the current setup, precision, calculated as in Hooge et al. (2018), was *Mdn* = 0.68° (*SD* = 0.28°), loss was *M* = 0.8% (*SD* = 2.4%). These are reasonable values given the special nature of the setup (see SI Data quality for more information and Supplementary Figure 5). Neither precision nor loss have visibly driven results across demographic groups. Fixation candidates were detected from raw gaze data with an algorithm specifically built for noisy data (Hessels et al., 2020). Fixation candidates were discarded if shorter than 60 ms, or merged if intermittent saccade candidates were smaller than 1 degree of visual angle in amplitude. This procedure has been demonstrated to prevent event-detection related biases (Hooge et al., 2022). Given that participants needed to look at the center of the screen to start free-viewing, all fixations with onsets before the start of free-viewing were removed from the dataset. The following eighteen fixations were considered for analyses to account for differential fixation counts of participants (this equated to *M* = 6.906 s of free viewing). Participants with fewer than eighteen fixations (n = 117), thus deviating more than 1.5 median absolute deviation from the median, were excluded resulting in the total of 2,607 participants (Number of fixations per participant: *Mdn* = 25.0 *MAD* = 4.45). Fixations that were located outside of the bounds of the screen of the experimental setup were excluded.

5.2.5 Baseline spatial distribution maps

To evaluate the performance of the predictions obtained from 21 saliency models tested here, four different baselines were constructed. 1) A map of all actual fixation locations served as the upper bound for the performance of any model (comparison between the binary array of fixation locations and its smoothed counterpart). This **fixation map** was constructed from fixation locations by applying a Gaussian filter (*SD* = 1 degree of visual angle) to the fixation map, effectively making it continuous (Bylinskii et al., 2018; Le Meur & Baccino, 2013) - in other words, discrete fixation locations were blurred over with this kernel. This approach allows to construct regions rather than pixels for fixation determination and acts as regularisation for potential small scale measurement error (Bylinskii et al., 2018). 2) A **meaning-map**,

a model created from successive ratings of small patches of the image by n = 59 participants served as gold standard model (Henderson & Hayes, 2017), possibly best incorporating information about objects and semantics, as previously proposed for computational models (Einhäuser et al., 2008). To create the meaning map, the image was split into overlapping patches with diameters of 1.5, 3 and 7 degrees of visual angle. These patches were then rated for meaningfulness by n = 59 participants (Mdn_{Age} = 25 years, *SD* = 7.399 years; men: 39, women: 19, non-binary: 2), recruited via Prolific without restriction regarding demographics, using Gorilla in an online experiment. This experiment took about 15 minutes to complete during which participants had to rate 200 patches each. Participants were each rewarded with 9 euros. 3) A **Gaussian central bias** (Tatler, 2007), skewed to the aspect ratio of the screen (*SD* = screen half dimensions), served as the baseline performance that should be achieved by any model. Effectively, the central bias lets saliency be maximal at the center with declining saliency towards the edges of the screen. Central biases can outperform saliency maps (Tatler, 2007) and are therefore incorporated in many of the more recent, here evaluated, models (e.g., Kümmerer, Wallis, & Bethge, 2017; Linardos et al., 2021). 4) Lastly, a **single observer model** expressed how well one participant's gaze behavior matched gaze behavior of all other participants. This procedure was repeated over all participants (similar to leave-one-out cross-validation), and scores from all iterations were then averaged.

5.2.6 Evaluation metric

A multitude of evaluation metrics for saliency maps have been put forward (see Bylinskii et al., 2018; Le Meur & Baccino, 2013, for reviews) of which Normalized Scanpath Saliency (NSS) was used here. NSS correlates highly to other metrics and has generally favorable properties as it requires minimal prior assumptions (Bylinskii et al., 2018; Riche, Duvinage, et al., 2013). NSS was extracted per model by first z-standardizing the respective saliency map, and overlaying it with the binary map of discrete fixation locations. For each fixated pixel, the z-score of the corresponding pixel was taken from the saliency map and a grand mean was calculated over those values. All maps (baselines or model predictions) were evaluated against the discrete fixation locations. For the single observer model, discrete fixation locations of one participant were evaluated against the blurred fixation map of all other participants. As such, NSS accounts for the relative saliency of regions as predicted by a given saliency map, not absolute saliencies that differ between models (Bylinskii et al., 2016, 2018). False positives and false negatives are equally weighed, and (nonbound) positive NSS scores indicate above chance-level performance, whereas negative NSS scores indicate worse than chance performance. As NSS is reduced to one score, it does not indicate which regions drive better or worse than chance performance. For this reason, graphical representations of the delta between predicted models and the spatial distribution map of fixation locations are given in Supplementary Figure 6 and Supplementary Figure 7.

Table 5.2: Performance of visual saliency models and baselines. Model performance (NSS, Normalized Scanpath Saliency score) for baselines and models are given as rows. Negative numbers denote worse than chance performance. The *Improvement* column denotes relative performance between central bias (0%) and fixation map (100%) in % NSS.

Model	NSS	Authors	Improvement
Baselines			
Fixation map	0.709		100.00%
Central bias	0.014	Tatler (2007)	0.00%
Single observer	0.176		23.32%
Meaning map	0.382	Henderson and Hayes (2017)	52.97%
Models			
SALICON	0.462	Jiang et al. (2015)	64.45%
SalGAN	0.434	Pan et al. (2017)	60.42%
DeepGazeIIE	0.411	Linardos et al. (2021)	57.15%
DeepGazeII	0.408	Kümmerer, Wallis, and Bethge (2017)	56.78%
QSS	0.350	Schauerte and Stiefelhagen (2012)	48.37%
IMSIG	0.342	Hou et al. (2011)	47.20%
DeepGazeI	0.339	Kümmerer et al. (2014)	46.83%
DVA	0.307	Hou and Zhang (2008)	42.13%
SSR	0.280	Seo and Milanfar (2009)	38.35%
SAM	0.279	Cornia et al. (2018)	38.10%
ICF	0.269	Kümmerer, Wallis, et al. (2017)	36.70%
AIM	0.255	Bruce and Tsotsos (2005)	34.75%
IKN	0.206	Itti et al. (1998)	27.66%
RARE2012	0.200	Riche, Mancas, et al. (2013)	26.79%
BMS	0.194	J. Zhang and Sclaroff (2013)	25.88%
CAS	0.172	Goferman et al. (2011)	22.71%
GBVS	0.171	Harel et al. (2006)	22.65%
SUN	0.166	L. Zhang et al. (2008)	21.85%
FES	0.060	Rezazadegan Tavakoli et al. (2011)	6.72%
LDS	0.043	Fang et al. (2016)	4.29%
CVS	-0.076	Erdem and Erdem (2013)	-12.86%

5.3 Results

5.3.1 Descriptive model performance

Normalized scanpath saliency (NSS) scores were used to evaluate model performance, a measure that has generally favorable properties and correlates highly to other indicators of model performance (Bylinskii et al., 2018; Riche, Duvinage, et al., 2013). The continuous spatial distribution map of fixations of all participants yielded a NSS of 0.709 relative to the discrete fixation locations of all participants, effectively establishing the upper bound for any model's performance. Surprisingly, model performance was higher for many models (e.g., SALICON, SalGAN DeepGazeII) than for the meaning map (NSS = 0.382), which has been described to outperform regular saliency models (Henderson & Hayes, 2017). The central bias had a prediction very close to

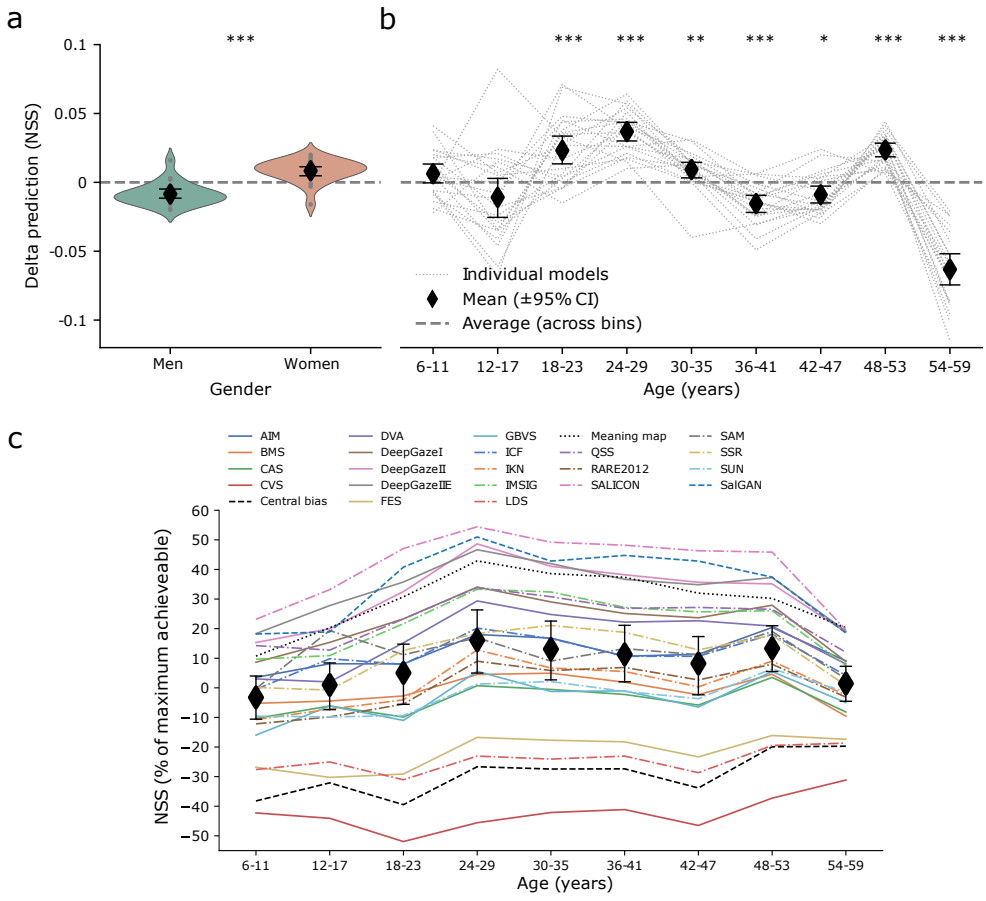


Figure 5.2: Model performance across demographic bins. Positive and negative values indicate better- and worse than average performance, respectively. Individual data points represent NSS deviations from average model performance across age bins. **a**: Deviations for men and women. **b**: Deviations across age bins. Gray dots and lines depict NSS deviations for individual models. ***: statistically significant at $p < 0.001$, ** at $p < 0.01$, * at $p < 0.05$. **c**: NSS scaled between single observers and fixation maps' NSS per age bin. Black diamonds represent average deviations alongside 95% confidence intervals. n = 1,600.

chance level (NSS = 0.014) and was outperformed by all but one of the evaluated saliency models; the single observer was outperformed by most models. Performance of individual models is given in Table 5.2 with absolute NSS and percentage deviations in model performance scaled between central bias and the overall fixation map.

5.3.2 Model performance across individuals

Models performed significantly better at predicting fixation locations of women (n = 710) than of men (n = 890; $BF_{10} = 244.904$, $t(20) = 4.779$, $p < 0.001$, *Cohen's d* = 1.043; see Figure 5.2, Supplementary Figure 2 for spatial distribution maps of fixation locations for men and women, and Supplementary Table 3; for results on participants who reported other gender, see Supplementary Table 4 and Supplementary Figure 4). The average difference in predictions across gender for all models (NSS = 0.017) closely

Table 5.3: Differences in model performance split by age groups. Deviations (NSS) per age bin from the average across age bins. Negative numbers indicate worse model performance compared to other age bins, whereas positive numbers indicate better performance compared to other age bins. Summary statistics (excluding baselines) are given in the bottom two rows (bold: $p < 0.05$; inferential t-tests in Supplementary Table 2, tests for contrasts between age bins in Table 5.4).

	Deviation in model performance relative to average model performance across age bins								
Model	6-11	12-17	18-23	24-29	30-35	36-41	42-47	48-53	54-59
Baselines									
Fixation map	0.186	0.028	0.003	-0.103	-0.105	-0.105	-0.065	-0.018	0.179
Central bias	-0.031	0.007	-0.02	0.033	0.014	-0.006	-0.029	0.039	-0.008
Single observer	0.071	0.026	0.034	-0.009	-0.021	-0.037	-0.016	-0.021	-0.029
Meaning map	-0.029	-0.024	0.042	0.047	0.015	-0.002	-0.006	-0.006	-0.037
Models									
RARE2012	-0.019	-0.035	0.003	0.044	0.015	0.006	0.003	0.034	-0.053
SalGAN	-0.007	-0.068	0.069	0.057	0.004	0.006	0.024	0.006	-0.087
DeepGazeIIE	0.004	0	0.048	0.043	0.009	-0.031	-0.014	0.015	-0.072
SALICON	-0.006	-0.012	0.071	0.039	0.003	-0.011	0.008	0.024	-0.115
DVA	-0.006	-0.061	0.028	0.054	0.019	-0.008	0.02	0.015	-0.06
FES	0.008	-0.029	-0.004	0.04	0.021	-0.001	-0.014	0.016	-0.036
QSS	0.037	-0.034	0.035	0.036	0.01	-0.024	0.004	0.009	-0.069
SSR	0.005	-0.046	0.045	0.026	0.031	0.005	-0.006	0.028	-0.089
CVS	0.02	0.011	-0.015	0.01	0.013	-0.004	-0.023	0.008	-0.024
IMSIG	0.013	-0.035	0.036	0.044	0.03	-0.011	0.006	0.017	-0.096
LDS	0.018	0.021	0	0.021	0.001	-0.013	-0.03	0.01	-0.031
ICF	0.001	0.024	0.017	0.037	0.007	-0.041	-0.018	0.031	-0.059
GBVS	-0.022	0.015	-0.005	0.051	0.001	-0.014	-0.026	0.044	-0.043
CAS	0.023	0.017	0.004	0.026	0.007	-0.019	-0.02	0.033	-0.068
SUN	0.018	-0.018	-0.004	0.018	0.011	-0.024	-0.018	0.041	-0.021
DeepGazeI	0.003	-0.008	0.044	0.046	0.009	-0.025	-0.008	0.027	-0.087
AIM	0.024	0.005	0.01	0.017	0	-0.049	-0.022	0.038	-0.025
SAM	-0.001	0.082	0.033	0.015	-0.04	-0.03	-0.019	0.033	-0.073
DeepGazeII	-0.009	-0.041	0.038	0.064	0.014	-0.012	0.001	0.011	-0.065
IKN	-0.014	-0.024	0.006	0.059	0.014	-0.008	-0.017	0.035	-0.053
BMS	0.041	0.007	0.027	0.026	0.017	-0.016	-0.02	0.019	-0.101
Summary statistics (models only)									
Mean deviation	-0.006	-0.011	**0.023**	**0.037**	**0.009**	**-0.015**	**-0.009**	**0.024**	**-0.063**
Bayes Factor$_{10}$	0.739	0.567	**92.727**	**>5,000**	**6.982**	**258.279**	**4.050**	**>5,000**	**>5,000**

corresponded to the difference for the central bias (NSS = 0.018, Supplementary Table 3).

Further differences in model performance were found across age groups. Data were binned in nine groups, each spanning 6 years of age from 6-59, which ensured at least 34 participants per bin (see Table 5.1 for demographics per bin). Model performance was then averaged across age bin averages to account for class imbalances in the grand average, and this was used as baseline for across-age comparisons. Therefore this average, unlike the values reported in Table 5.2, is not biased towards the groups with most participants. Age biases across models were fairly consistent (Figure 5.2; Table 5.3, see Table 5.4 for *t*-tests and respective effect sizes between bins). One age group revealed the clearest positive deviation from average model performance: Those of the arguably most oversampled population, corresponding to most college students and young academics, respectively (18-29). This was accompanied by generally less consistent or worse predictions for other age bins. Note that, due to smaller group sizes, the age bins 6-11 and 54-59 warrant more caution before interpretation

Table 5.4: Contrasts for percentage explained NSS of the explainable NSS between all age bins. Contrasts are for data as given in Figure 5.2c and Table 5.3 across age bins. Data distribution was assumed to be normal but this was not formally tested for these contrasts.

Comparison		Mean Difference	*t*	*p* (Bonferroni)	95% CI (Lower)	95% CI (Upper)
6-11	12-17	-3.889	-2.548	0.424	-8.854	1.077
	18-23	-8.191	-5.367	< .001	-13.157	-3.225
	24-29	-19.075	-12.499	< .001	-24.041	-14.110
	30-35	-16.039	-10.509	< .001	-21.004	-11.073
	36-41	-14.151	-9.272	< .001	-19.117	-9.185
	42-47	-11.345	-7.434	< .001	-16.311	-6.379
	48-53	-16.375	-10.730	< .001	-21.341	-11.409
	54-59	-3.811	-2.497	0.487	-8.777	1.154
12-17	18-23	-4.302	-2.819	0.195	-9.268	0.663
	24-29	-15.187	-9.951	< .001	-20.153	-10.221
	30-35	-12.150	-7.961	< .001	-17.116	-7.184
	36-41	-10.262	-6.724	< .001	-15.228	-5.296
	42-47	-7.457	-4.886	< .001	-12.423	-2.491
	48-53	-12.487	-8.182	< .001	-17.452	-7.521
	54-59	0.077	0.051	1	-4.889	5.043
18-23	24-29	-10.885	-7.132	< .001	-15.850	-5.919
	30-35	-7.848	-5.142	< .001	-12.813	-2.882
	36-41	-5.960	-3.905	0.005	-10.926	-0.994
	42-47	-3.154	-2.067	1	-8.120	1.811
	48-53	-8.184	-5.363	< .001	-13.150	-3.218
	54-59	4.379	2.870	0.168	-0.586	9.345
24-29	30-35	3.037	1.990	1	-1.929	8.003
	36-41	4.925	3.227	0.055	-0.041	9.891
	42-47	7.730	5.065	< .001	2.764	12.696
	48-53	2.700	1.769	1	-2.265	7.666
	54-59	15.264	10.002	< .001	10.298	20.230
30-35	36-41	1.888	1.237	1	-3.078	6.854
	42-47	4.693	3.075	0.089	-0.273	9.659
	48-53	-0.337	-0.221	1	-5.302	4.629
	54-59	12.227	8.012	< .001	7.261	17.193
36-41	42-47	2.805	1.838	1	-2.160	7.771
	48-53	-2.224	-1.458	1	-7.190	2.741
	54-59	10.339	6.775	< .001	5.373	15.305
42-47	48-53	-5.030	-3.296	0.044	-9.996	-0.064
	54-59	7.534	4.937	< .001	2.568	12.500
48-53	54-59	12.564	8.233	< .001	7.598	17.529

than the other age bins. Notably, variation in model performance was relatively large for under-aged (< 18 year-old) participants compared to other participants. Again, the central bias showed a largely similar tendency to the overall model biases, in line with findings reported earlier on a smaller selection of models and age groups (Açık et al., 2010; Krishna & Aizawa, 2017). Spatial distribution maps of fixation locations are given per age bin in Supplementary Figure 2; absolute NSS per model and baselines are given in Supplementary Table 1. Individual predicted maps per model (Supplementary Figure 6), and deviations between models' predictions and actual fixation locations (Supplementary Figure 7) are given in the supplementary information. The degree of biases across age became even more apparent with NSS scaled between single

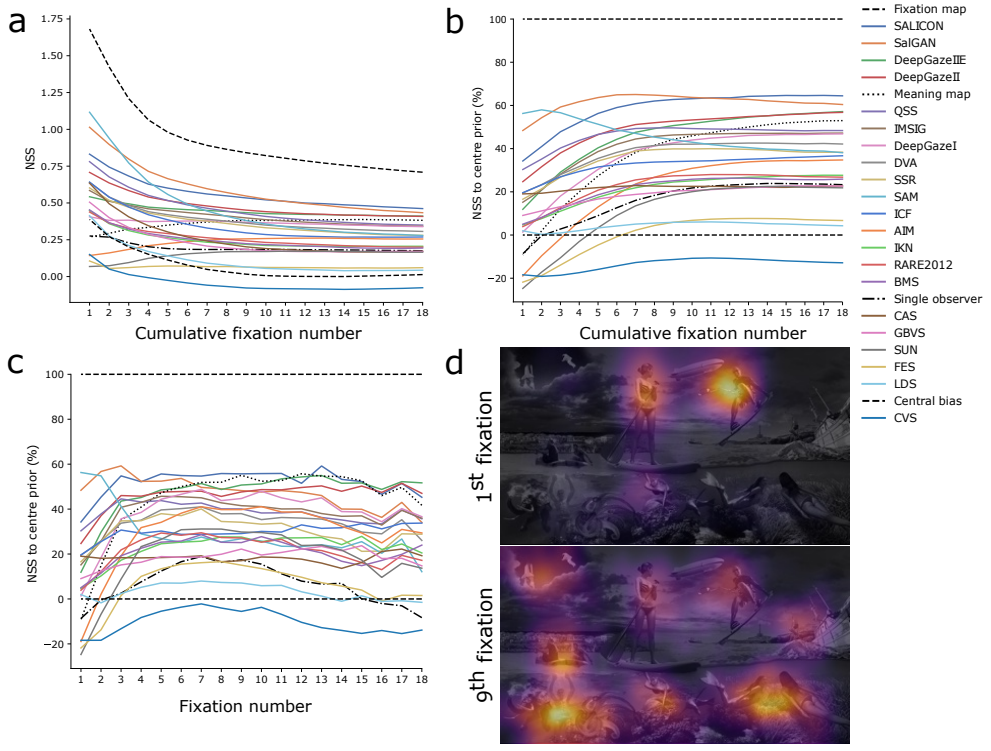


Figure 5.3: Model performances across fixations. **a**: Model (colors) and baseline (dashed/hatched/dotted) benchmarking across fixations with unscaled NSS. **b**: Model benchmarking across fixations with NSS scaled to maximally achievable NSS per fixation. Fixation maps are cumulative, i.e., righter points on the x-axis indicate model performance including all previous fixation data. The rightmost data corresponds to the benchmarking reported in Table 5.2. **c**: Same visualization as in b, but per fixation instead of cumulative fixations. **d**: Fixation map for only the 1st fixation shows a much more focal distribution of fixations than for the 9th fixation, which much more closely resembles the map after all 18 fixations (Figure 5.1 right). n = 2,607.

observers and fixation maps' NSS per age bin (Figure 5.2 c). Differences were striking here, again with best performances for participants in young adulthood (Figure 5.2). Differences as a function of age were substantial ($F(1.699,33.971) = 38.269$, $p < 0.001$, $\eta^2 = 0.657$) and were highly significant between most age bins. For instance, performance for children aged 6-11 differed from performance of all adults (all $p < 0.001$) except those in the oldest bin (see Table 5.4 for full post-hoc t-tests).

5.3.3 Model performance across fixations

Model performance was characterized by substantial variation in NSS, but revealed that fixations are predicted differently well over time. Here, the fixation map, reflecting an upper bound, showed much higher NSS scores early than later on, likely because fixations were much more focal for early viewing (see Figure 5.3 d). Models predicted the subset of early fixations generally better than when consecutive fixations were added, but only in absolute NSS scores (Figure 5.3 a; e.g., first vs. ninth fixation $t(20) = 5.881$, $p < 0.001$, 95% CI = [0.692, 1.857]). When scaled to the maximum achievable

NSS, many models improved relative to this maximum as consecutive fixations were added, given that the NSS scores of the fixation map and central bias also decreased (Figure 5.3 b, first vs. ninth fixation $t(20) = -6.634$, $p < 0.001$, 95% CI = [-2.057, -0.821]). Relative performances across models differed as a function of the cumulative fixations made - e.g., SAM (Cornia et al., 2018) performed very well on the first two fixations in comparison to the overall benchmark, which was best predicted by current, leading deep learning models such as SALICON, SalGAN, and DeepGazeIIE (Jiang et al., 2015; Linardos et al., 2021; Pan et al., 2017). This variability showed in different model rank orders between the first few fixations, relatively later fixations, or all 18 fixations, respectively (corresponding to the leftmost and the rightmost rank order in Figure 5.3 b). Of course, a map of cumulative fixations contains more information whereas a map for only the first fixation is arguably more sparse - despite already 2,607 fixations contributing to it. Are differences in performance across fixations therefore driven by data sparsity? Fixation maps for only the first and the ninth fixation, respectively, (Figure 5.3 d) compared with the overall map of all fixations (Figure 5.1), showed markedly different patterns as function of early- versus late viewing behavior: Fixation maps were highly focal for the first and much more spread out for later fixations, data sparsity cannot have driven these effects. The benchmark of models per fixation (Figure 5.3 c) further showed relatively worse performance for early fixations (first vs. ninth fixation $t(20) = -4.679$, $p < 0.001$, 95% CI = [-1.544, -0.482]). Around six to ten fixations marked the point of best performance, which roughly matches the overall number of fixation clusters observed. NSS scaled between central bias and fixation map then dropped again across models. This challenges the current approach of optimizing saliency models on just one fixation map per image - in fact, what attracts gaze early on might be substantially different from what attracts gaze after a few fixations or extended viewing.

5.4 Discussion

Here we set out to investigate how well saliency maps, models that have been proposed to predict the deployment of visual attention, and by extension fixation locations, generalize across individuals and the number of fixations. Using a sample of 2,607 participants and 21 highly influential saliency models, gender and age biases in model performance were found for the subset of 1,600 participants with credible demographics. Specifically, predictions were better for women and adults aged 18 to 29. These demographics (women, 18-29), perhaps incidentally, best represent those of the majority of participants in psychological research in general (predominantly younger adult women), as well as in the vast majority of benchmarking data (Borji & Itti, 2015; Kümmerer, Bylinskii, et al., 2022). Overall, a large portion of the biases in predictions across demographic groups followed relative differences in prediction of the central bias baseline, in line with previous work with smaller samples and more distinct age groups (Açık et al., 2010; Krishna & Aizawa, 2017; Krishna et al., 2018; Rider et al., 2018).

Besides demographic-based differences, model performance differed as a function of which fixation was to be predicted. Here, models only performed well for early fixations in absolute NSS, but actually worse when scaled with the maximum achievable

NSS. Our sample allowed to study how model performance evolves as a function of the number of fixations, as maps are not sparse; even if only the first or ninth fixation is used as a basis, it contains information from 2,607 fixations. Worse predictions for early and later fixations compared to intermediate (i.e., sixth to tenth) fixations highlight the importance of more closely defining whether earlier or later viewing behaviour is modeled. One possibility to account for the variability in prediction quality across fixations lies in adjusting how saliency models apply thresholds: For first fixations, only the few most focal locations could be emphasized by greedier thresholds to represent the focal distribution of fixation locations observed early on. For subsequent fixations, a more liberal approach might be employed, allowing more spread-out predictions - as observed for intermediate or later fixations. The here reported data further suggest that different models might capture different visuo-attentional processes - with some models being more predictive of early fixations (e.g., Cornia et al., 2018) and others being better at predicting later fixations (e.g., Jiang et al., 2015; Linardos et al., 2021). If the goal is to define which part of a scene is fixated first, particular models may perform best and thus be the method of choice. If the goal is to predict which parts of a scene will be fixated eventually, however, different models that weigh objects and semantic information more strongly are to be recommended. This view could resolve a number of outstanding debates on whether low-level features or semantics drive gaze behavior most strongly (Borji et al., 2013; Cerf et al., 2009; Einhäuser et al., 2008; Henderson & Hayes, 2017; Henderson et al., 2021; Pedziwiatr et al., 2021). We speculate that *both* accounts have their merit: The answer could depend on the viewing duration and the number of objects in a scene. Remarkably, fixation maps were very focal for the first few fixations. As soon as the number of fixations approached the number of bigger objects in our scene, model performance did not notably change further when using a cumulative fixation map, indicating that participants eventually fixated most objects, but in differing sequences. Very late fixations, in turn, could disperse even further and be captured worse as a consequence. Saliency models are optimized using cumulative fixation maps obtained from several seconds of free viewing. This practice might have introduced bias, as later fixations are disproportionately weighed in these maps relative to the initial two or three fixations. The findings and account put forward here could (partially) explain differences across benchmarking datasets and results (Kümmerer, Bylinskii, et al., 2022). A straightforward prediction would be that benchmarks which use gaze data on images with a short viewing duration favor models that prominently weigh relevant low-level features (possibly such that are common in faces; Cerf et al., 2009; Einhäuser et al., 2008), whilst benchmarks with gaze data obtained from viewing behavior over a longer time favor models that weigh semantic content more strongly.

Taken together, the here reported findings put the current approach of evaluating and improving models into question, which is predominantly to design and benchmark models around fixation maps constructed from several seconds of free viewing by college student participants. More generally, the present findings reveal that models of psychological processes - even as fundamental as low-level visual behavior - can be affected by systematic and substantial biases introduced via training and benchmarking datasets. Proper modeling and understanding of human spatial gaze behavior will require an approach that incorporates diverse samples, if the aim of said models is to predict behavior of more than just college students. Furthermore, saliency models

could improve further by addressing effects of early versus later fixations, or this issue should be explicitly addressed in limitations of models. Different models might be used to account for these differences or models could incorporate adjustable options: For whom and for when should fixations be predicted? Knowledge about developmental differences in sampling behavior and saliency computations (Açık et al., 2010; Franchak et al., 2016; Gottlob & Madden, 1999; Mitchell & Neville, 2004; Rider et al., 2018) could be incorporated here, as well as findings into differences in fixations across viewing duration (e.g., Ossandón et al., 2014; Pannasch et al., 2008).

Most models outperformed the central bias, in line with several benchmarking results (Kümmerer, Bylinskii, et al., 2022). However, many models already incorporate a central bias and the image at hand features many spread-out objects. Generally, more recent models performed best at predicting the overall fixation map (i.e., including data of all participants and fixations), some of which are currently also leading in benchmarks across stimuli (e.g., Linardos et al., 2021). Indeed, deep learning-based models (Jiang et al., 2015; Kümmerer, Wallis, & Bethge, 2017; Linardos et al., 2021; Pan et al., 2017) generally outperformed more traditional and interpretable models which are foremost centered around low-level image feature computations. However, these deep-learning models still suffered from qualitatively similar limitations regarding generalizability across individuals and fixations. Furthermore, the meaning map was outperformed by several models, in contrast to initial findings (Henderson & Hayes, 2017), reminiscent of recent criticism (Pedziwiatr et al., 2021), but see Henderson et al. (2021). Meaning maps could possibly be advanced further by ensuring high correspondence between the demographics of raters and participants whose fixations are to be predicted.

5.4.1 Limitations

Naturally, the sample put forward here, while vast, is in turn affected by sampling biases. For example, children need to be willing to wait and focus, older adults need to have sufficient vision to participate. Which further differences between people, beyond gender and age, are relevant when it comes to saliency maps remains to be determined. With this first step, we hope to stimulate research into this, including into non-western populations in the field of saliency mapping, as suggested in many areas of psychology (Henrich et al., 2010; Jones, 2010; Rad et al., 2018). As a way forward, to overcome the biases uncovered here, we provide the present database to help improve across-participant and across-fixation generalizations and will supply it with additional data as long as the exhibition remains active. Replications of presently reported findings across multiple images would be desirable, for instance using other large scale data collections such as described here. Other limitations to be kept in mind lie in the less controlled setup than used in most common benchmarking data sets, or limited possible inferences on gaze behavior of participants with non-binary gender. The practical implications here are simple: no gender should be made the default option and 'prefer not to say' options should be given to not obscure findings. Whilst we did not find qualitative differences in findings for two different smoothing kernel sizes to create the fixation maps, more extreme choices for kernels or flexible central biases might affect the here reported bounds and thus findings.

5.4.2 Conclusions

Past findings (e.g., on the central bias outperforming many models; Tatler, 2007) could have been taken to cast doubt on the general usefulness of saliency modeling. Such criticisms, however, have led to even increased efforts to develop more powerful models, e.g., by incorporating the central bias. The here identified systematic challenges to saliency modeling - generalizability across individuals and which fixation is to be predicted - might require new approaches in turn, for instance by incorporating information besides low-level image features (see also De Haas et al., 2019, for related calls) in order to make the next leap forward happen. If models can only generalize to overall fixation maps across images (Kümmerer, Bylinskii, et al., 2022), but fail at generalizing across earliest fixations or work best only for certain demographic samples, different models might be needed to predict different (groups of) individuals and visuo-attentional processes. Or, even better, models might incorporate which fixations are to be predicted - distinctions for the demographics and the number of the fixation could hereby set the path for much more powerful models. We argue that visual saliency should be considered a dynamic, interactive, and integrative result of low-level image features (e.g., Itti et al., 1998), as well as semantic information (e.g., De Haas et al., 2019; Einhäuser et al., 2008) or meaning maps (Henderson & Hayes, 2017) under the consideration of the individual differences that have been associated with gaze behavior (e.g., Franchak et al., 2016; Mitchell & Neville, 2004).

Supplementary Materials and Data Availability
All code and data are available via the Open Science Framework https://osf.io/sk4fr/.

Acknowledgements
This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 863732). This research was funded by a VICI Grant 211-011 from the Netherlands Organization for Scientific Research to Stefan Van der Stigchel. We thank the NEMO Museum Amsterdam for their help with data collection and all participants who donated their data. We thank Tobii for consulting over data quality.

Author contributions
CS: Conceptualization, Methodology, Formal Analysis, Data Curation, Funding Acquisition, Writing – Original Draft, Writing – Review & Editing. **AJH**: Methodology, Formal Analysis, Data Curation, Software, Writing – Original Draft, Writing – Review & Editing. **GB**: Methodology, Formal Analysis, Data Curation, Software, Writing – Review & Editing. **ITCH**: Methodology, Writing – Review & Editing. **TCWN**: Writing – Review & Editing. **SMS**: Methodology, Writing – Review & Editing. **SVdS**: Conceptualization, Methodology, Funding Acquisition, Writing – Review & Editing.

Chapter 6

Seeing the Forrest through the trees: Oculomotor metrics are linked to heart rate

Alex J. Hoogerbrugge
Christoph Strauch
Zoril A. Oláh
Edwin S. Dalmaijer
Tanja C. W. Nijboer
Stefan Van der Stigchel

Published as:
Hoogerbrugge, A. J., Strauch, C., Oláh, Z. A., Dalmaijer, E. S., Nijboer, T. C. W., & Van der Stigchel, S. (2022). Seeing the Forrest through the trees: Oculomotor metrics are linked to heart rate. *PLOS ONE*, *17*(8). doi.org/10.1371/journal.pone.0272349

Abstract

Fluctuations in a person's arousal accompany mental states such as drowsiness, mental effort, or motivation, and have a profound effect on task performance. Here, we investigated the link between two central instances affected by arousal levels, heart rate and eye movements. In contrast to heart rate, eye movements can be inferred remotely and unobtrusively, and there is evidence that oculomotor metrics (i.e., fixations and saccades) are indicators for aspects of arousal going hand in hand with changes in mental effort, motivation, or task type. Gaze data and heart rate of 14 participants during film viewing were used in Random Forest models, the results of which show that blink rate and duration, and the movement aspect of oculomotor metrics (i.e., velocities and amplitudes) link to heart rate–more so than the amount or duration of fixations and saccades. We discuss that eye movements are not only linked to heart rate, but they may both be similarly influenced by the common underlying arousal system. These findings provide new pathways for the remote measurement of arousal, and its link to psychophysiological features.

6.1 Introduction

Remotely and unobtrusively detecting fluctuations in arousal is of wide interest to researchers in fields such as human-computer interaction, psychology, and ergonomics. This interest is due to the fact that changes in arousal are not only related to physical exertion, but also to psychological concepts for which arousal is often assessed as an objective approximation, such as the degree of excitedness, drowsiness, or mental effort during a given task. Given that arousal levels are related to task performance following an inverted U-shape function (Teigen, 1994; Yerkes & Dodson, 1908), they have a profound effect on task performance – for instance on various critical tasks, arousal may affect the safety of operators and other people who rely on those operators (Williamson et al., 2011). Although fluctuations in arousal can be detected from various objective sources such as electroencephalography (EEG), functional Magnetic Resonance Imaging (fMRI), heart rate, or skin conductance (Hart & Staveland, 1988), these methods require direct physical interaction with measurement devices or can be quite obtrusive. Only few parameters can be assessed remotely, such as oculomotor metrics obtained via video-based eye-tracking.

In the current study we investigate how well heart rate – one of the best investigated central indicators of arousal – can be predicted from remotely accessible oculomotor metrics as alternative peripheral indicators of arousal. A link between these indicators is plausible given the extensive support for correlations between oculomotor metrics and various psychological concepts, such as mental effort. For instance, it has been shown that the degree of pupil dilation can provide an accurate indication of participants' mental effort in both controlled and naturalistic viewing tasks (Beatty, 1982; Palinko et al., 2010). Furthermore, it has been shown that the peak velocity of saccades decreases as mental effort increases (Di Stasi et al., 2010, 2013) and increases as motivation increases (Muhammed et al., 2020). Similarly, mental effort has been shown to covary with heart rate and with several derivatives of heart rate measures (Charles & Nixon, 2019). Additionally, it has been shown that changes in arousal are paired with an altered rate of eyeblinks (Maffei & Angrilli, 2019; Wood & Hassett, 1983), and that spontaneous eyeblinks occur in tandem with an increase in heart rate variability (Nakano & Kuriyama, 2017).

While oculomotor measures are fairly robust, they can be influenced by the environmental circumstances under which they were obtained. For instance, pupil dilation is impacted by the luminance of the scene that is being watched, and microsaccades and peak velocities of saccades can only be reliably measured by expensive high-speed, low-noise trackers. Additionally, among eye tracking scientists there is no unified concept of how fixations and saccades should be defined – and thus the application of differing fixation- and saccade detection techniques may result in differing outcomes, even if they are applied to the same dataset (Hessels et al., 2018). As such, incorporating several metrics which can be independently extracted (e.g., pupil size, oculomotor movement, blinks) would improve robustness of the model, as it reduces dependence on one single extraction technique. This also applies to cases in which pupil dilation measurements are unreliable or missing, or the eye tracker's sampling rate is too low to extract peak saccade velocities.

The benefits of relating oculomotor metrics to heart rate are two-fold. Firstly, scruti-

nizing the links between oculomotor metrics and heart rate can foster our theoretical understanding of common underlying mechanisms and thereby our definition of arousal. Secondly, ever since the seminal works of Buswell (1935) and Yarbus (1967), we have been aware that eye movements are foremost driven by task type (top-down) and visual saliency (bottom-up). Later on, it has been shown more reliably that task type – such as search or free-viewing – influences gaze behaviour (Henderson & Hollingworth, 1998; Le-Hoa Võ & Wolfe, 2015; Mills et al., 2011), and that oculomotor metrics besides pupil dilation (Strauch et al., 2021) and peak saccade velocity (e.g., saccade amplitude, fixation duration) can provide sufficient information for machine learning algorithms to predict task type at above chance level (Hanke et al., 2016; Kootstra et al., 2020). In this manuscript, we describe that, besides top-down and bottom-up mechanisms, arousal – estimated by the link to heart rate – also contributes to eye movements.

To this end, we use data from the studyforrest dataset (Hanke et al., 2016). This dataset contains eye tracking and pulse oximetry measurements from participants while they watched the 1994 motion picture Forrest Gump. We investigate whether oculomotor metrics can provide sufficient information for regression models and machine learning models to accurately predict high or low heart rates of participants in this naturalistic viewing task. Furthermore, we investigate how strongly each oculomotor feature contributes to the correct prediction of heart rate, thereby providing insight into how specific aspects of oculomotor movement is driven by a common measurement of arousal, such as heart rate.

6.2 Methods

All analyses were performed with Python 3.8.10, using SciPy version 1.6.2 (Virtanen et al., 2020) and scikit-learn version 0.24.2 (Pedregosa et al., 2011). All code and outcomes can be retrieved from https://osf.io/skcd8/.

6.2.1 Raw data

Eye tracking data and pulse oximetry data were obtained from the studyforrest dataset, which contains data of fourteen participants that were measured while being presented with the 2-hour film Forrest Gump (Hanke et al., 2014, 2016). The raw eye tracking data was measured with an Eyelink 1000 at a frequency of 1 kHz and pulse oximetry measurement was applied to record heart rate data at an effective frequency of 100 Hz. A full description of the recordings and anomalies can be found in (Hanke et al., 2016) and at https://studyforrest.org.

6.2.2 Oculomotor feature detection

Fixations and saccades were extracted based on the algorithm proposed in Hessels et al. (2020), which operationalizes fixations and saccades as phases of slow and fast eye movements, respectively. Firstly, the raw 1 kHz x and y gaze signals were smoothed by applying a Savitzky-Golay filter. We then applied an adaptive velocity threshold algorithm to this smoothed signal, thereby obtaining candidate fixation phases, with everything in between being candidate saccade phases. Thereafter, we applied

two basic merging criteria. Firstly, saccade candidates with amplitude < 1.0° were removed, thereby merging neighbouring fixation candidates. Subsequently, all fixation candidates with duration < 60ms were removed. This procedure successfully removes large differences in oculomotor event classification between different algorithms (Hooge et al., 2022). Gaze amplitudes and velocities in pixels were converted to degrees of visual angle by multiplying their values by 0.0186 (Hanke et al., 2016). Lastly, blinks were detected by finding periods in which no pupil data was measured. All events which lasted less than 30ms, or more than 3 seconds, were removed. These thresholds were set so that neither brief nor longer periods of data loss would be incorrectly detected as blinks.

6.2.3 Data pre-processing

After extraction of oculomotor metrics, data of each participant was split into 240 chunks of 30 seconds each. However, the last chunk was often shorter than 30 seconds, and some chunks had too much data loss. As such, these chunks were discarded, resulting in 3327 data points. Then, heart rate detection was performed over the raw pulse oximetry signal within these chunks, using HeartPy (Van Gent et al., 2018). Thirty seconds were selected as chunk size because it provides a balance between sufficient data per chunk (> 20 fixations on average, and sufficient time for accurate heart rate detection), and a sufficient number of chunks for machine learning purposes.

For each chunk, twelve features were extracted: (1, 2, 3) the duration of each fixation, saccade, and blink event; (4, 5) the amplitude of each fixation and saccade event; (6, 7) the peak velocity of each fixation and saccade event; (8, 9) the mean velocity of each fixation and saccade event; and (10, 11, 12) the count of fixation, saccade, and blink events in that chunk.

We took a two-fold approach to testing whether oculomotor metrics can be sufficient predictors of heart rate. Firstly, we posed that the prediction of heart rate could be considered a regression problem, in which we aimed to predict heart rate on a continuous scale. Secondly, we posed that the prediction of heart rate could also be considered a binary classification problem (above some threshold or below some threshold). This approach can be useful when the aim is to only predict whether someone is either excessively or insufficiently aroused.

To prepare our dependent variable for binary classification, the heart rate of each chunk was expressed as a z-score; the number of standard deviations from the median heart rate of that respective participant over the full film. Each z-score was then converted to a binary variable – namely low if $z < -.5$, and high if $z > .5$. All other chunks were considered neutral and discarded. Since the distributions of heart rate were often skewed, and due to slightly differing amounts of data loss, our binarization did not result in equally large samples of high and low labels. As a result, 513 chunks were below the threshold, and 607 chunks were above the threshold. A total of 1120 data points remained after binarizing the heart rate data. Distributions of each feature, split per label, are reported in Figure 6.1.

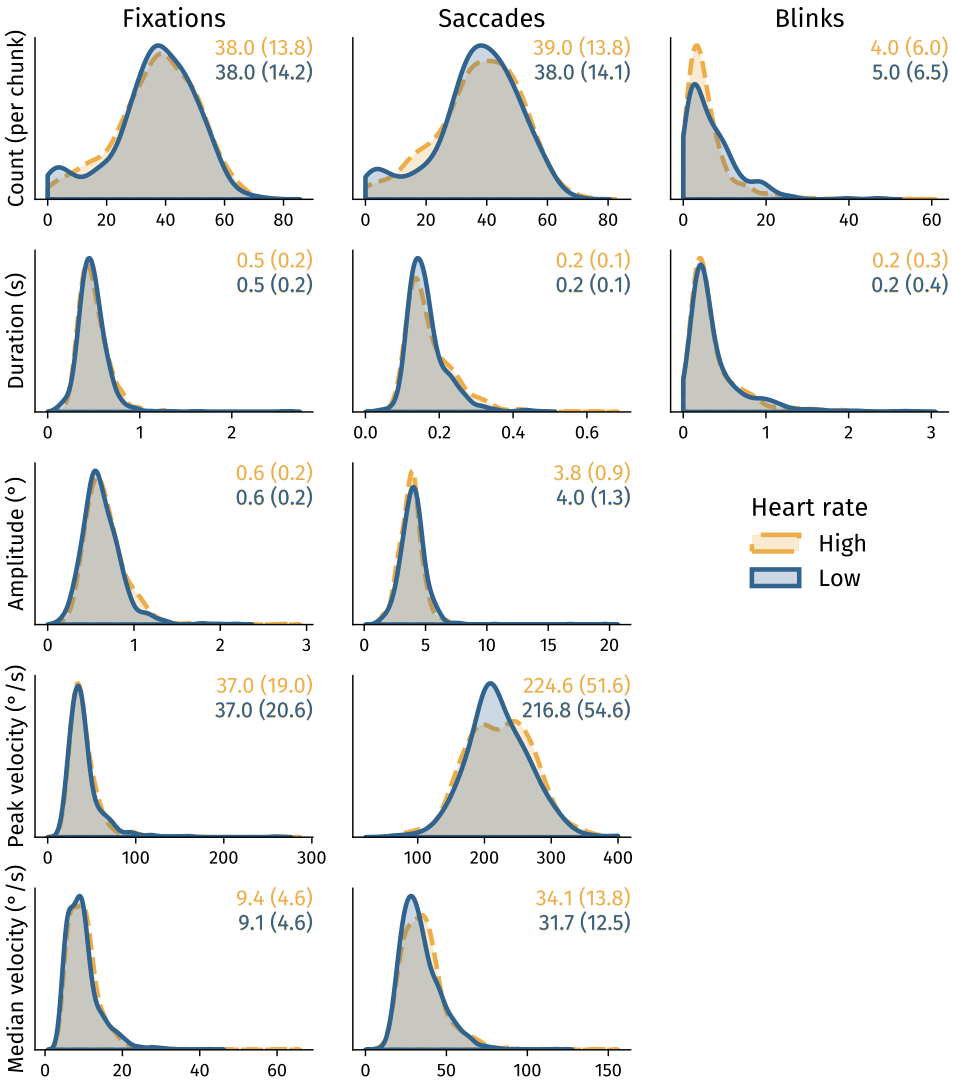


Figure 6.1: Distributions (kernel density estimation) for each of the twelve features, per label (high or low heart rate). The distributions are computed over all chunks for all participants, thus 1120 data points per feature (513 low, 607 high). Orange and blue values indicate median and standard deviation of each of the high and low heart rate distributions, respectively.

6.2.4 Feature pre-processing

As is common in machine learning pipelines, our classifier required an equally long set of features per chunk of data, and the described feature set did not comply with this requirement. For example, if 30 saccades were made within one chunk, and 40 saccades were made in another chunk, the peak saccade velocity variable would contain 30 and 40 values for each of those chunks, respectively. Therefore, our data needed to be aggregated. Three methods were explored, as outlined in the next

subsections.

Averaging

Within each chunk, the average of each of the twelve features was computed, providing one value per feature for each chunk. This approach provides the most intuitive insight into the amount of information contained within each feature, which in turn contributes towards correct classification.

Feature explosion

It could be argued that simply calculating the mean value over features would discard relevant information, since, for instance, the mean saccade velocity across chunks may be equal, but the variance across chunks could be different. Similar to the approach of Kootstra et al. (2020), a set of 13 statistical descriptors (e.g., mean, variance, uniformity) was employed to describe the distribution of each of the features 1-9 within each chunk (see S1 Table for a full list of the statistical descriptors). Through this method, the dataset was thus 'exploded' and contained 3 count features + (9 features $\times$ 13 descriptors) = 119 features.

Feature explosion and dimensionality reduction

To aid interpretation of these 119 features, each of the oculomotor metrics was to be described in at most two variables. To this end, each of the nine exploded features was reduced from a description of dimensionality 13 to a description of dimensionality 2 by taking the two components with the highest explained variance from Principal Component Analysis (PCA). This resulted in a set of 3 count features + (9 features $\times$ 2 descriptors) = 21 features. On average, the first two components taken from PCA provided an explained variance of 98.98% for the nine features.

6.2.5 Regression pipeline

We fitted a multiple linear regression with heart rate per chunk as the dependent variable, and either of the features obtained by the methods outlined above as independent variables. In addition, a similar but polynomial regression was fitted, to identify possible non-linear links. All regression models were fit to the train set and R^2 was evaluated on the test set.

6.2.6 Machine learning pipeline

Logistic Regression, K-Nearest Neighbours and Random Forest Classifier were used to predict high versus low heart rate from oculomotor metrics. Each type of model was run independently 50 times, with a new 80/20% stratified train/test split for each run, and with the default set of parameters as provided by scikit-learn. On average over those 50 runs, and across the three different pre-processing approaches, the Random Forest classifier performed best of the three models, and thus this model was selected for further optimization (see Table 6.2).

Subsequently, hyperparameter optimization of the Random Forest classifier was implemented over the number of trees (range 10-200; step size 1) and the maximum depth per tree (range 1-30 + unlimited depth). All other hyperparameters were kept

Table 6.1: Outcomes (R^2) of the linear regression models, per pre-processing approach. Features were either derived directly from the pre-processing approach, or with added second-degree polynomials for each feature. Models were fit to the 80% train set and evaluated on the 20% test set.

	R^2	R^2 (with 2nd degree polynomials)
Averaging	.18	.30
Explosion	.21	<. 01
Explosion + reduction	.17	.12

Table 6.2: AUCs of the model pre-selection process (averaged over 50 independent model runs). [a]The average (SD) outcome on the test set over 50 runs of the optimized model is reported.

	Log. Reg.	K-NN	Rand. Forest	Rand. Forest + optim.[a]
Averaging	.622	.617	**.696**	**.698** *(.04)*
Explosion	.590	.588	.660	.664 *(.05)*
Explosion + reduction	.614	.585	.666	.678 *(.04)*

as default. We then constructed 500 candidate combinations of hyperparameters by randomly sampling from their specified distributions. Each candidate combination was assigned the same 80% training set and was evaluated on that set using 5-fold stratified cross-validation and Area Under the Curve (AUC) as performance metric. An AUC of 0.50 constitutes classification at chance level and 1.0 constitutes complete accuracy. The model and parameter combination that led to the best cross-validation result was then tested on the 20% holdout set. To compensate for randomness effects in the sampling of the training- and test sets, and in the sampling of hyperparameters, this search process was repeated 50 times and means and standard deviations are reported.

Finally, the contributions of all features towards correct classification were extracted from the best-performing model using permutation importance (Altmann et al., 2010). For each feature, a one-sample t-test was performed to test whether that feature's importance differed significantly from the overall mean (higher importance is better; t-test α = .05).

6.3 Results

6.3.1 Regression

R^2 for regression models ranged between < .01 and .30 (see Table 6.1 for full results), indicating that oculomotor metrics provide limited information towards prediction of heart rate as a continuous variable.

6.3.2 Classification

Overall, the averaging pre-processing approach provided the best performance at classifying whether a participant had a high- versus low heart rate within a chunk (AUC = .696). The model pre-selection results and the results of optimization are reported in Table 6.2.

The best-performing model performed consistently above chance and achieved an

Table 6.3: AUCs and parameters of the best-performing models and runner-up models resulting from hyperparameter search (on the averaging pre-processing approach). Model ranks were defined based on cross-validated classification performance. All values are averages over 50 runs. In each run, only the best model was tested against the test set. [a]Includes at least one model where the maximum depth was unlimited

	Model rank 1	Model rank 2	Model rank 3
Cross-validation performance (AUC)	.703	.701	.700
20% holdout set performance (AUC)	.698	-	-
Average number of trees	126.5	135.0	127.2
Average maximum depth per tree	20.2[a]	19.3[a]	19.5[a]

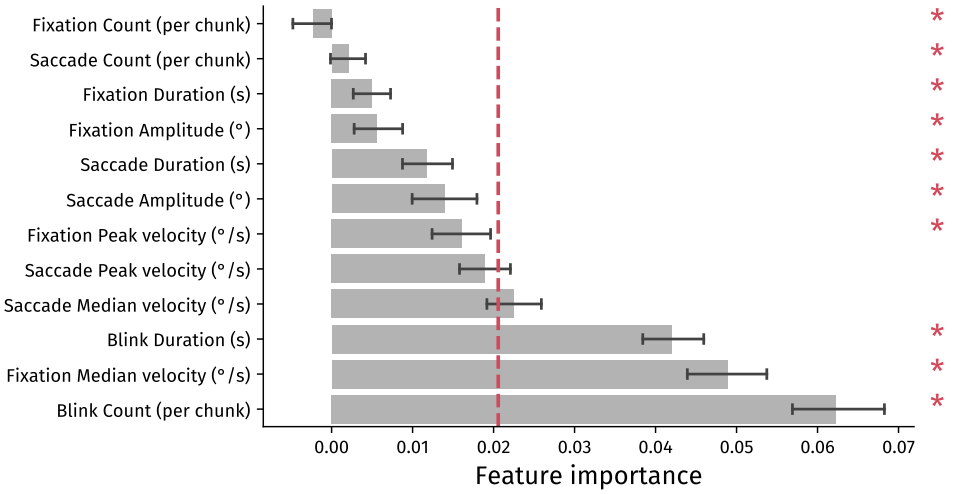


Figure 6.2: Mean (± 95% CI) feature importances as extracted from the best-performing model of each of the 50 runs. Higher values imply a higher degree of information within the variable. The vertical dashed line represents the overall mean of all importance values. The asterisks represent where feature importances differed significantly from the overall mean.

average AUC of .703 (SD = .02) on the cross-validation sets, and an average AUC of .698 (SD = .04) on the test sets over 50 independent runs. An overview of the best models and the runner-up models is reported in Table 6.3.

The extraction of feature importances revealed blink rate, duration, and features associated with oculomotor movement to be most predictive of heart rate ([fixation and saccadic] median velocity, saccadic peak velocity; Figure 6.2). All other features were found to contribute worse-than-average towards classification.

6.4 Discussion

In the current study, we investigated how well oculomotor metrics may predict heart rate and which of these features drive this prediction predominantly. To this end, we used a public dataset of participants whose physiological data were obtained while watching the 1994 Forrest Gump motion picture. Although oculomotor metrics

provided limited predictive value for linear and polynomial regressions (up to R^2 of .30), a Random Forest model could predict high- versus low heart rate consistently at above-chance level. In this model, the features which contributed most strongly towards correct classification were blink rate, blink duration, and the median velocity within fixations and saccades, and the saccadic peak velocity.

Interestingly, each of the features that contributed most strongly pertains to either information regarding blinks, or regarding oculomotor movement (velocities and amplitudes), and not so much to durations or counts of fixations and saccades. The importance of blink rate and blink duration provides support for the suggested link between an altered rate of eyeblinks and changes in arousal (Maffei & Angrilli, 2019; Wood & Hassett, 1983) and changes in heart rate metrics (Nakano & Kuriyama, 2017). At first sight, the relative importance of fixation velocity might be surprising, since fixations are spatially stable. However, differences in fixation velocities may be the result of physiological drift or microsaccades, sometimes referred to as fixational drift or fixational eye movements (Rolfs, 2009). The occurrence of microsaccades has been found to be positively coupled to heartbeat, and may thus explain the amount of information captured in the fixation velocity variable (Ohl et al., 2016). The peak and median velocity of saccades are fourth and fifth in the list of informative features, which aligns with earlier literature which suggested that saccadic peak velocity indicates mental effort (Di Stasi et al., 2010, 2013) and motivation (Muhammed et al., 2020) – two cognitive processes closely linked to modulations in arousal.

Feature importance, however, does not indicate specifically which aspect of a distribution provides the most information towards correct classification. This makes it difficult to speculate about the direction of the effect of the included features, further complicated by inconsistencies in the literature. For instance, microsaccades occur more frequently with high mental effort in some tasks, but not in others (Pastukhov & Braun, 2010; Siegenthaler et al., 2014), suggesting that the modulation of eye movement and heart rate by the arousal system is highly task-dependent. This is further evidenced by the fact that we find increased saccadic- and fixational velocities in high heart rate periods, whereas it is usually found that saccadic and fixational velocity are negatively correlated with arousal (Di Stasi et al., 2013; Siegenthaler et al., 2014). While, except within velocity, no consistently different medians within features were found between low- and high heart rate periods, it is remarkable that standard deviations were consistently equal or higher when heart rate was low, as compared to when it was high (with the exception of median saccade velocity). High arousal levels could be associated with a reduction in variability in oculomotor behaviour, as is the case with heart rate (Kazmi et al., 2016).

Based on these findings, we speculate that heart rate is not only linked to fixational eye movements (Ohl et al., 2016), but to oculomotor movements in general. This link might come into place due to changes in the common underlying arousal system, or merely as an effect of changes in blood pressure during the heartbeat cycle. Our findings therefore suggest that a substantial portion of oculomotor behaviour is linked to heart rate, and not only by top-down goals of the beholder (Kootstra et al., 2020), or bottom-up visual features of the scene (Itti et al., 1998), as is commonly assumed. To this end, other physiological indicators could be compared to oculomotor metrics in their ability to predict heart rate. Because there is no unified definition of

arousal, investigating the links between the aforementioned indicators would allow to isolate more specific subcomponents of arousal, and improve our definition of the term.

Speculating about neural underpinnings for a link between the oculomotor features described here and heart rate, we see a potential role for the locus coeruleus, a sympathetic center in the brain that acts antagonistically to parasympathetic activation associated with heart rate variability (Mather et al., 2017). The noradrenergic locus coeruleus affects oculomotor behavior mainly via its inputs to the superior colliculus that is crucial in bringing about several oculomotor behaviours (Strauch et al., 2022). Note that locus coeruleus-centered and superior colliculus-centered circuits have been associated with differential attentional functions at the level of the brain stem, including alerting and orienting (Strauch et al., 2022).

Another putative candidate might be the hypothalamus (though bidirectionally linked to the locus coeruleus; Nakano & Kuriyama, 2017; Strauch et al., 2022) which modulates activity in the autonomous nervous system. Its link to the basal ganglia (and changes in the dopamine system) might explain the relation between blinks and heart rate, as changes in dopamine levels in the basal ganglia are monitored with changes in blink rate (Nakano & Kuriyama, 2017). Although a relation between heart rate and oculomotor features and these two brain regions seems plausible, it is important to stress that this currently mere speculation and should be the subject of future research.

The current study is limited in its comprehensiveness of oculomotor features. For instance, pupil dilation has been shown to encode aspects of arousal (Beatty, 1982; Palinko et al., 2010). However, these measurements are distorted when gaze position changes (Hayes & Petrov, 2016) and could therefore not be reliably measured. Another step could be to link eye movements to on-screen movements in order to obtain smooth pursuits. This might be meaningful, as deviations in smooth pursuit trajectories have been found to be indicative of mental effort – and thus by proxy arousal (Kosch et al., 2018). Currently, smooth pursuits are likely to be captured within fixations and saccades at the high- and low ends of their respective velocity distributions. Lastly, as a next step, microsaccades could be investigated in detail (Duchowski et al., 2020; Engbert & Kliegl, 2003). This would require robust detection algorithms that work without static scenes and with monocular data.

Different parameter sets and pre-processing approaches all lead to similar model performances, as shown in Table 1, 2, and 3. For example, the pre-processing approaches of averaging and feature explosion without reduction led to very similar outcomes in classification accuracy. However, the averaging approach required less processing time and can be interpreted more intuitively. Furthermore, classification accuracy could be improved by using more complex models, but again at the cost of interpretability.

The current study does not directly provide a method for the real-time prediction of heart rate from oculomotor metrics, since the proposed Random Forest classification pipeline requires that a baseline heart rate is established from which to derive low or high heart rate labels as the dependent variable. However, future research may attempt to establish a baseline heart rate measurement before the start of a

given task and subsequently investigate whether the prediction of heart rate can be conducted in real-time.

6.5 Conclusion

In conclusion, oculomotor metrics obtained during a naturalistic viewing task contain sufficient information to predict high versus low heart rates above chance during that same task. These findings not only establish oculomotor metrics as unobtrusively measurable predictors of heart rate, but open new pathways for investigation of the link between oculomotor metrics and various indicators of arousal.

Supplementary Materials and Data Availability
All code and outcomes can be retrieved from the Open Science Framework https://osf.io/skcd8/.

Acknowledgements
This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 863732). The authors would like to thank Roy Hessels for his input regarding fixation detection.

Author contributions
AJH: Conceptualization, Methodology, Formal Analysis, Data Curation, Software, Writing – Original Draft, Writing – Review & Editing. **CS**: Methodology, Writing – Original Draft, Writing – Review & Editing. **ZAO**: Conceptualization, Methodology, Writing – Review & Editing. **ESD**: Conceptualization, Methodology, Writing – Review & Editing. **TCWN**: Methodology, Writing – Review & Editing. **SVdS**: Conceptualization, Methodology, Funding Acquisition, Writing – Review & Editing.

Epilogue

General Discussion

Summary of findings

Chapter 1 – Being able to inspect external information just-in-time during a copying task is an important factor for people's choice to offload visual working memory; any disruption to the continuous availability of external information is the predominant driver of increased visual working memory usage, whereas the predictability of delay and the delay itself are less relevant (Hoogerbrugge, Strauch, Böing, et al., 2024).

Chapter 2 – Being able to rely on external target templates in a visual search task is beneficial to task speed, accuracy, and effort (Hoogerbrugge et al., 2023).

Chapter 3 – Reliance on the external world is persistent; people frequently reinspect external search templates even after many repetitions of the same templates. Persistent reinspection can help people boost metacognitive confidence regarding representations in memory, on top of boosting the objective quality of those representations (Hoogerbrugge, Strauch, Nijboer, & Van der Stigchel, 2024).

Chapter 4 – In multi-target visual search, people can flexibly opt to use either sequential (one-by-one) or concurrent (simultaneous) search as specific modes, dependent on template availability, task difficulty, and individual preference (Hoogerbrugge et al., in preparation).

Chapter 5 – Where people look during free-viewing differs between age groups; e.g., 18-29 year-old adults look at an image differently than younger children aged 6-17 (Strauch et al., 2023).

Chapter 6 – How (fast) we move our eyes during free-viewing is linked to heart rate, via underlying physiological arousal mechanisms (Hoogerbrugge et al., 2022).

Visual working memory in context

Visual working memory (VWM) is a capacity-limited system for the short-term maintenance, manipulation, and comparison of visual information (Baddeley & Herring, 1983; Baddeley & Hitch, 1974). Much of the discourse around VWM has centred around its capacity limits. For example, some researchers have argued that the average capacity is around four discrete items (Luck & Vogel, 2013), whereas others have argued that VWM resources can be (re)allocated as needed, allowing people to remember few items with high fidelity or more items with lower fidelity (Ma et al., 2014, in this dissertation I am somewhat agnostic about this debate). Regardless, VWM capacity

is strongly limited. Although investigating VWM in the context of its capacity has been incredibly useful to the field, it has been argued (and I reiterate in Part I) that VWM should also be studied in a more naturalistic context (Van der Stigchel, 2020). Namely, in many daily situations, external information remains available, such as when you have an instruction manual to hand. In those cases, humans tend to use VWM minimally, but rather use the world as an *external storage* (O'Regan, 1992; Van der Stigchel, 2020; Wilson, 2002). Instead of memorising and maintaining multiple item representations (which is relatively costly; Beatty, 1982; Kahneman, 1973), people prefer to make eye movements (which are relatively cheap; Koevoet, Strauch, Naber, & Van der Stigchel, 2023; Theeuwes, 2012) and encode external information only if and when needed for the task at hand. This baseline behaviour has been shown consistently in various paradigms, and can be manipulated by altering the cost of memorising versus the cost of inspecting information (see Qing et al., 2024 for a meta-analysis of copying tasks, and see e.g., Gajewski and Henderson, 2005; Inamdar and Pomplun, 2003; Risko and Gilbert, 2016 for other paradigms).

Which factors constitute 'cost' have mainly been limited to two to avenues of investigation. First, increasing the distance between two pieces of information increases the cost of eye- or head-movements, and can cause people to rely more on memory (Ballard et al., 1995; Draschkow et al., 2021; Inamdar & Pomplun, 2003). Second, the introduction of a brief delay before external information can be viewed also causes people to rely more on memory (Böing et al., 2023; Gray et al., 2006; Melnik et al., 2018; Sahakian et al., 2023, 2024; Somai et al., 2020). In Chapter 1, I discuss the latter and highlight that increased reliance on VWM with increased time-costs should at least partially be seen as a result of people's inability to inspect external information just-in-time. Participants performed a copying task, in which the example layout was intermittently occluded (see Chapter 1 Figure 1.1). In Experiment 1, the layout was always available, or appeared and disappeared for several seconds within a six-second loop. In Experiment 2, the layout either immediately appeared when participants looked at it, or after a constant or variable delay. Across both experiments, participants relied more on memory whenever constant availability of the example layout was disrupted, but there were limited additional effects of how long or how variable delays were on how often participants inspected the example. That is to say: Increased VWM usage does not just seem to be the result of having to wait, but also a result of disruptions to our ability of accessing information *as and when* we need it. Constant availability of information allows us to e.g., prepare for upcoming episodes of encoding (Koevoet, Naber, et al., 2023), take as much time as we need to encode (Bays et al., 2011; de Jong et al., 2023), and avoid task-switching costs (Nieuwenhuis & Monsell, 2002; Rogers & Monsell, 1995). The fact that VWM is used in a just-in-time manner had been shown by Mary Hayhoe and her colleagues (Droll and Hayhoe, 2007; Hayhoe et al., 2003; Triesch et al., 2003; but see e.g., Xu et al., 2025), but I here showed that disruptions to this ability also affect the trade-off between storing and sampling. This may underlie why it can be frustrating when you build furniture together with someone else whilst they keep taking away your instruction manual, or when web pages take longer than expected to load.

The trade-off between storing internally versus sampling externally is thus a matter of cost (i.e., effort and time) and continuous availability. In Chapter 2, participants

Suggestion box 1: What happens to no-longer-relevant VWM contents?
Representations in VWM can become irrelevant as a (sub)task is completed. Many such instances have occurred throughout this dissertation; every time that an encoded template was not found, it became irrelevant, and could be safely discarded from memory. Or could it? Exactly how no-longer-relevant representations are discarded from VWM, and when this occurs, is not yet entirely clear – but some mechanism must exist, otherwise VWM would fill up and no new information could ever be stored (Oberauer, 2018). There is evidence to suggest that people can use discard mechanisms to actively remove information from VWM (DeRosa et al., 2024; Ecker et al., 2014; Lewis-Peacock et al., 2018), but also that irrelevant VWM representations can linger in some cases (Bae & Luck, 2019a; Oberauer, 2018). It is quite possible that whether information is removed from VWM depends on the task at hand, and in the paradigms presented in Part I of this dissertation we are presented with a unique situation: Participants could – within trials – decide for themselves whether VWM contents were no longer relevant. When a target-absent decision is made for a specific template, the trial does not necessarily end if not all templates have been searched yet. Therefore, in the continuation of the trial, we may still encounter that target which we thought was absent. Furthermore, we are aware that our memory is fallible – and thus each target-absent decision in our visual search paradigm was likely associated with a certain decision confidence (Chun & Wolfe, 1996). This raises interesting questions: Are previously-encoded templates removed whenever a target-absent decision is made for that template, but the trial does not end? And are we more likely to remove VWM contents after making a high-confidence decision? Paradigms in which participants can choose to discard VWM contents (and even resample those discarded templates if necessary) provide a particularly promising setting to further study how VWM load is kept low in naturalistic settings.

performed a search task in which one or four external search templates could be inspected throughout the entire trial or only before search onset (see Chapter 2 Figure 2.1). In Experiment 1, I used relatively simple stimuli (Landolt C's with a gap in eight possible locations), in Experiment 2 I increased stimulus complexity by using complex polygons (I also used these stimuli in Chapters 1 & 3 and in the General Introduction Figure 4). Here, I showed that findings from copying task paradigms also replicate in a visual search task: Participants relied on the external world whenever it remained available. I further showed that availability of external information is also highly practical; people are generally faster and more accurate at the task when they are able to reinspect external templates and therefore have to rely less on VWM. This is both a result of being able to delay the encoding of templates and the ability to reinspect templates later on in trials. The latter I called 'refreshing' or 'double checking' memory contents. Moreover, this beneficial effect is exaggerated when the cost of memorising is higher (i.e., when stimuli are more complex). Participants additionally approached the task in two primary ways; some participants fixated all or most of the templates in their fist inspection of the template area and decreased this over the course of subsequent inspections, whereas other participants fixated one unique template and did so consistently across multiple inspections (see Chapter 2 Figure 1.4). Böing et al. (in prep.) called these 'low loaders' and 'high loaders', in the sense that some people tend to rely more on VWM than others do (see also Lin & Leber, 2024). In sum, being able to (re)sample external information is used frequently and is beneficial to task performance and effort – hence I titled the chapter "Don't hide the instruction manual" and argued that influential models of visual search (e.g., Guided Search; Wolfe, 2021) should be extended to include the ability to reinspect external search templates.

People often (re)sample external information when this information remains available, because it costs relatively little effort and has practical benefits. In Chapter 3, I

showed that resampling behaviour is incredibly persistent, even after eliminating its benefits to effort and performance. Here, participants did a similar task as in Chapter 2, but I repeated each template set twenty-five trials consecutively (see Chapter 3 Figure 3.1). As we repeatedly encode the same stimuli (or even if we are just exposed to them or *think* of them), we tend to build up increasingly stronger memory representations of those stimuli (Carlisle et al., 2011; Ebbinghaus, 1885; Hout & Goldinger, 2010; Pashler et al., 2007; Souza et al., 2015; Woodman et al., 2001, 2007; Xu et al., 2025). Resampling external information should therefore eventually become redundant. However, even after twenty-five repetitions, participants reinspected templates frequently (circa twice per trial when searching for four templates). To test whether this seemingly excessive resampling was necessary for performance, in Experiment 2 I removed the ability to reinspect templates after fifteen repetitions. It turns out that participants could still perform the task with high accuracy in the latter ten repetitions in which templates were unavailable. Further analyses showed that, after the first few repetitions, there was no within-trial (short-term) benefit of reinspecting, nor was there a benefit to longer-term memory performance. Rather, I showed that participants used template reinspections to boost their confidence regarding existing memory representations besides boosting the actual representations themselves. These findings in Chapter 3 are in line with e.g., Sahakian et al. (2023, 2024) who posited that VWM is not always 'depleted' before people reinspect external information, and that introducing a penalty for making mistakes made participants even more reluctant to use information in VWM. I discuss that this intuitively makes sense: Our memory contents are susceptible to errors and decay, and we are aware of this. If reinspecting the image of a screw only requires an eye movement toward the instruction manual and a little bit of time, it is relatively cheap to double check and to boost our confidence, compared to making a mistake and having to *dis*assemble and *re*assemble the furniture again. I therefore note that, when templates are repeated many times (but perhaps even at the single-trial level), people inspect those templates for three primary reasons: (1) Templates need to be initially encoded in the first few repetitions; (2) Templates may need to be refreshed in order to counteract decay and interference (Camos et al., 2018; Souza et al., 2018); (3) Templates are reinspected in order to boost one's metacognitive confidence regarding existing representations in VWM. Although other studies have found that VWM representations from previous trials are sometimes carried over (Bae & Luck, 2019a; Shan & Postle, 2022), the data here suggest that, in Chapters 1, 2 and 4, stimuli could be flexibly removed from VWM after trials (Ecker et al., 2014; Lewis-Peacock et al., 2018; Oberauer, 2001), whereas they could be maintained in Chapter 3. Alternatively, there was sufficient interference – for example from distractors or new templates – to ameliorate any between-trial effects in those other chapters (Clapp et al., 2010; Oberauer & Lin, 2017). Future research may point out whether, and to which degree, intertrial effects and interference from distractors play a role in these paradigms. In sum, when external information remains available, people initially offload memory by encoding information just-in-time – which helps to limit effort expenditure and boost performance. When the same external information is used more frequently and offloading memory becomes redundant, the performance (accuracy and speed) benefits of resampling information become less pronounced, but rather shift toward more metacognitive benefits.

As such, whether people *can* do something does not mean that they *will* do so.

Suggestion box 2: From working memory to long-term memory
When repeatedly searching for the same stimuli, we build up increasingly stronger memory representations of those stimuli, and these representations are likely to eventually enter long-term memory (LTM; Carlisle et al., 2011; Hout & Goldinger, 2010; Pashler et al., 2007; Souza et al., 2015; Woodman et al., 2001, 2007). In Chapter 3, I let participants search for the same templates twenty-five times, and indeed showed that these templates had (at least partially) entered LTM after the search experiment had ended. Notably, Wolfe (2012) took a different approach, instead ensuring that participants had learned up to 100 unique items before the search experiment started. Wolfe showed that searching for many targets through LTM is relatively faster than searching for targets through the external world. Given that LTM search is so efficient, it is likely that participants would eventually transition from VWM- to LTM-based search, but I could not elucidate specifically if and when this indeed occurred. Why would we want to figure this out? As an example, some police departments employ super-recognisers (SRs) whose job it is to scan hours of video footage in search of specific faces, such as those of suspected criminals (Bate et al., 2019; Robertson et al., 2016). During the first few minutes of searching for novel faces, it is to be expected that the SRs predominantly search from VWM, but after several hours this may transition to LTM-based search. How efficiently can we expect SRs to search for novel faces in comparison to well-established faces, and for how many faces can they search simultaneously in an efficient manner? Having a better understanding regarding the transition from VWM- to LTM-based search – and its contributing factors – is therefore of particular relevance to applied settings.

Moreover, the degree to which people rely on the external world is highly dynamic – dependent on both environmental demands and affordances, internal trade-offs, and individual preferences. In Chapter 4, I investigated whether these mechanisms may explain mixed findings regarding whether people are able to search for multiple targets concurrently. I adjusted the visual search paradigm from Chapters 2 and 3, and made template colour the primary feature for attentional guidance, although targets were defined by both colour and shape. In Experiment 1, participants were instructed to either search sequentially (encode one template, search for it, repeat) or concurrently (encode all templates and then search for them at the same time). I found that participants could indeed search both sequentially and concurrently when instructed, as evidenced by fixations on currently-relevant colours and suppression of currently-irrelevant colours, although attentional guidance was impeded when searching concurrently for four items. In Experiment 2a and 2b, I did not instruct participants whether to search sequentially or concurrently, which resulted in participants using a mix of both types between trials. Using Random Forest classifiers, I was able to determine from gaze behaviour, at the single-trial level, that participants used sequential and concurrent search as specific and dissociable search modes – only a small portion of trials were more difficult to classify. Interestingly, participants flexibly adjusted which of the two search modes they used between trials and between conditions, based on the number of templates, their complexity (between Experiments 2a and 2b), and whether templates could be reinspected or not. Moreover, some participants were more inclined to search sequentially, whereas others searched more concurrently overall – and the degree to which participants changed their behaviour between Experiments 2a and 2b differed idiosyncratically. I further note that participants may have made the choice to use either search mode while factoring in whether they could maintain effective attentional guidance or not. This may also have implications for e.g., the copying task paradigm, in which participants first have to search for each item that they wish to place (Draschkow et al., 2021; Hoogerbrugge, Strauch, Böing, et al., 2024; Kumle et al., 2024). If attentional

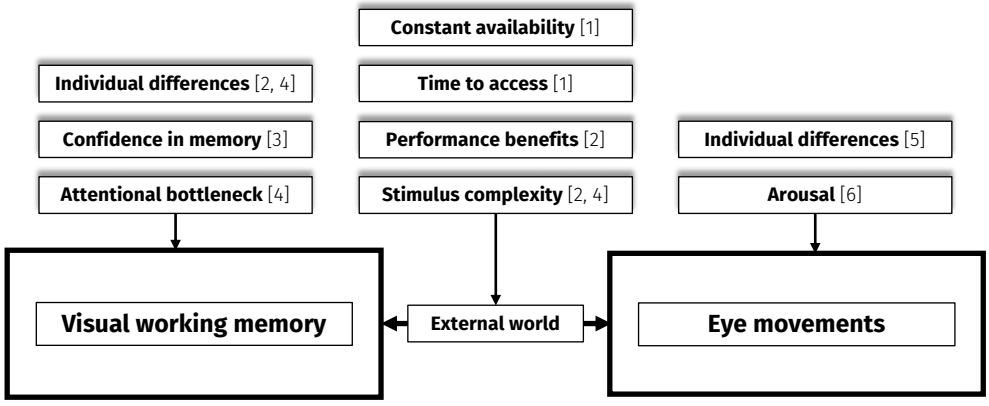


Figure 6: How we make use of the (visual) external world, limited to factors which have been shown in this dissertation. Numbers in square brackets indicate chapters. See General Discussion Supplementary Figure 1 for a more complete overview including factors which were discussed but not directly shown in this dissertation.

guidance is impeded while searching concurrently for multiple items, participants have an additional reason in the copying task to keep VWM load low in order to facilitate search for these items. From Chapter 4, it thus becomes clear that people can use both sequential and concurrent search modes, and that they flexibly adjust which of these they use dependent on task demands and individual preferences. It is therefore quite possible that some studies found a lack of concurrent search due to these flexible strategy choices. Given these findings I highlight that it is critical to investigate visual search behaviour at the trial level, and reiterate that VWM is used dynamically in interaction with eye movements. When assembling furniture, whether you memorize and search for multiple screws sequentially or concurrently may thus depend on the availability of the manual, the visual complexity of each screw, as well as your baseline preference.

In sum, beyond the previously studied time- and distance costs, based on the first four chapters I argue that researchers who study the interaction between VWM and eye movements should also take into account that this interaction is dependent on (1) being able to sample information just-in-time; (2) stimulus complexity; (3) objective benefits, i.e., speed and accuracy; (4) subjective benefits, i.e., metacognitive confidence; (5) individual differences; (6) whether and how guidance is affected (see Figure 6).

Individual- and state-dependent influences on eye movements

In Part II of this dissertation, I investigated where and how we move our eyes when there is no specific task for participants. In these instances, generally called free-viewing, eye movements are thought to be more strongly driven by bottom-up visual input than by top-down goals (Henderson & Hollingworth, 1998; Mills et al., 2011). For example, a bright red object on an otherwise dark background will generally capture

Suggestion box 3: Looked-but-failed-to-see errors
The paradigms presented in Part I of this dissertation provide an interesting opportunity, in that participants can choose how elaborately they encode each template representation in VWM. Although looked-but-failed-to-see (LBFTS) errors can occur even when we have perfect information about a search target (Wolfe et al., 2022), it is quite likely that less-detailed representations will cause us to miss targets more often (J. R. Williams et al., 2022). Therefore, especially in free-choice paradigms, participants must constantly make a decision: How elaborately do I want to encode each template, and how much does it matter if I miss a target? When assembling furniture, missing a search target is frustrating, but missing a target can have especially costly consequences for e.g., radiologists and their patients. I have shown in this dissertation that participants often 'double check' template representations when possible, and that this is beneficial to both accuracy and confidence in memory. Can searchers be induced to encode template representations more elaborately? Can LBFTS errors be reduced by giving professional searchers an instruction-manual-like cheat sheet with images of the to-be-found targets? Finding efficient ways to reduce LBFTS errors is highly valuable to furniture-builders and radiologists alike.

our attention and draw eye movements more strongly than the background (Itti et al., 1998). That is not to say that these eye movements are purely reflexive; attentional capture by saliency can be attenuated by previous experience (i.e., statistical learning) and in some cases by top-down goals (D. H. Duncan & Theeuwes, 2024; Theeuwes, 2024; Theeuwes et al., 1998). Although the aforementioned effect of statistical learning is usually studied in the relatively short term, it may extend to long-term effects as well (e.g., de Lange et al., 2018; Gayet & Peelen, 2022). Given the modulatory role that previous experience has on attentional capture by saliency, it stands to reason that what constitutes saliency, or to which degree people are affected by saliency, differs per person. Interestingly, in one of the earliest published studies on free-viewing, Buswell (1935) already noted that differences in fixation durations were more strongly linked to individual differences than to differences between pictures. Given these possible idiosyncrasies in attentional bias and gaze behaviour, it is worth mentioning how saliency models are developed. Predictive saliency models are generally inspired by neuroscientific principles and/or iteratively improved by training on behavioural data. After the design phase, the models' efficacy is usually validated by comparing them to actual gaze behaviour from a participant sample (Kümmerer et al., 2018). However, participant samples are often relatively small, and often consist of university students between the ages of 18 and 35. These are logical and practical limitations, but as a result saliency predictions may not be equally accurate at predicting where people look when tested outside of this usual age range.

In Chapter 5, I tested 21 saliency models on their ability to predict where observers of different ages and gender would look at an image. To this end, we collected gaze behaviour and demographic data from a uniquely large sample of 1,600 visitors (6-59 years old) to the NEMO Science Museum Amsterdam, where they free-viewed a single collage image for ten seconds. Saliency models were indeed worse at predicting where e.g., children would look as compared to young adults[1]. Notably, Buswell mentioned that there were no consistent differences in gaze patterns between children and adults, although he analysed most of his data based on visual inspection (Buswell,

[1]I have included the original publication in this dissertation, which states that there was a gender bias. After publication, we received an updated sample from the NEMO Museum Amsterdam with 4,173 valid participants. In that sample, the gender bias was no longer present, and I therefore do not describe this in the General Introduction and General Discussion sections. All other reported findings were replicated, see https://osf.io/sk4fr.

1935; Wade, 2020). I report fixation heatmaps for different age groups in Chapter 5 Supplementary Figure 2, which do indeed appear to be visually comparable, but were actually significantly different between children and adults as evidenced by the main outcomes of this chapter. Taking Buswell's approach, qualitative visual inspection suggests to me that children looked more at the animals in our image, whereas adults looked more at faces and bodies. Furthermore, it seems to me that children's gaze patterns were more entropic than those of adults, which were more focal – perhaps children explore more with their gaze (this was unlikely to be caused by underlying eye tracking differences; see Chapter 5 Supplementary Figure 5). Although these differences between children and adults are speculative, more detailed analyses may reflect how people's personal history and biases can attenuate oculomotor capture – not just by saliency, but also by semantics (some saliency models do incorporate semantics but are relatively sparse regarding individual factors, or have small sample sizes; e.g., Açık et al., 2010; Henderson and Hayes, 2017; Krishna and Aizawa, 2017; Krishna et al., 2018; Kümmerer, Wallis, and Bethge, 2017; Rider et al., 2018). Additionally, the subtleness of these across-age differences in where we look at images are reflected in small but significantly changing vertical and horizontal biases (pseudoneglect) across ages (I wrote about this with my colleagues in Strauch et al., 2024), strengthening the notion that *where* we look does differ idiosyncratically between people. There is no demographic information of our sample outside of age and gender, but it seems likely that other factors such as education level and cultural background would further influence how people look at images (as is the case for e.g., working memory and attention; Cowan, 2014; Gómez-Pérez & Ostrosky-Solís, 2006). Overall, Chapter 5 highlights how subtle, but important, individual differences can be when trying to predict where people will look at an image.

Besides *where* we move our eyes, *how* we move our eyes is affected by various mechanisms. In fact, the oculomotor system is tightly coupled to cognitive and physiological factors, as it is part of the central nervous system (Hoar, 1982). Pupil size, saccade velocities and microsaccade rates are affected by cognitive load (Beatty, 1982; Di Stasi et al., 2010; Koevoet, Strauch, Van der Stigchel, et al., 2023; Pastukhov & Braun, 2010; Siegenthaler et al., 2014; Strauch et al., 2022); and blinks and microsaccades have been linked to variations in heart rate (Maffei & Angrilli, 2019; Nakano & Kuriyama, 2017; Ohl et al., 2016). Moreover, it is likely that this coupling is not just a result of arousal but is functional; (micro)saccade rates and velocities may allow us to explore the environment when arousal is high or vice versa, blinks may occur during breakpoints of attentional processing, increased pupil size may improve visual processing, et cetera (Di Stasi et al., 2013; Eckstein et al., 2017; Nakano & Kuriyama, 2017; Vilotijević & Mathôt, 2024). It is thus clear that the oculomotor system is tightly coupled to arousal, but there was limited integrative evidence of how all of these oculomotor metrics are linked to – and whether they can be used to remotely measure – arousal. In Chapter 6, I used twelve oculomotor features from the *studyforrest* dataset (Hanke et al., 2014) in order to classify whether participants had low- or high heart rate (as a readout of arousal; Azarbarzin et al., 2014; Grassi et al., 1998; Mather et al., 2017) during movie free-viewing. Oculomotor metrics (primarily blink rates and eye movement velocities) provided sufficient information to consistently predict heart rate at above-chance level, in line with aforementioned studies. I did not investigate pupil size in this study, because the reliable and 'pure' measurement of pupil size as a readout of arousal

is inconvenienced by e.g., stimulus brightness and pupil foreshortening as a result of gaze location (Hayes & Petrov, 2016; Strauch et al., 2022). Given the established informativeness of pupil size changes for many cognitive and physiological processes, it is likely an important oculomotor feature to include for the integrative prediction of arousal – and recent advancements may allow future investigations to do so (Cai et al., 2024).

In sum, in Chapters 5 and 6 I showed that *where* and *how* we move our eyes is dependent on individual (demographic) differences and arousal states (Figure 6).

Visual working memory and eye movements in context

In the previous sections I first discussed how visual working memory and eye movements are linked in search- and copying tasks, and then how eye movements during free-viewing are affected by internal factors. These are obviously two very different task types, and it is known that task instructions influence where and how we move our eyes (e.g., Le-Hoa Võ & Wolfe, 2015; Mills et al., 2011; Yarbus, 1967). Can these two aspects of the dissertation be integrated? In part, yes. Regardless of task, they are two sides of the same coin, providing a more complete picture of how we interact with the external world around us.

First, a plethora of studies have shown that salient objects can capture attention during search tasks, and that the degree to which this occurs is modulated by previous experience (see Theeuwes, 2024 for a review). Although 'previous experience' is generally defined as short-term (implicit) statistical learning, there is evidence that the resulting attentional effects transfer between tasks (Van Moorselaar & Theeuwes, 2024). This is reminiscent of how contextual expectations learned over a lifetime affect visual search and perception (de Lange et al., 2018; Gayet & Peelen, 2022; Gayet et al., 2024); a toothbrush is more often found near a sink than on a dining table, and airport baggage screeners are more likely to detect a weapon than I am. Relatedly, saliency and the memorability of objects and scenes are tightly linked (Bainbridge, 2019; Constant & Liesefeld, 2021; Isola et al., 2011). Extending this reasoning, I would here speculate that idiosyncratic development as a result of individual experiences will affect what constitutes saliency, and thereby how attention is deployed and which objects or scenes are more memorable. I once received an email from a student who attended one of my guest lectures and who was an avid knitter. In that email, she told me that, when she was still a novice, finding each mistake in her knitting took quite long; she had to search through her entire work and compare each individual stitch to the 'external' pattern template. Over time, she felt like this became substantially easier and quicker; in her experience, she could spot mistakes almost immediately and she could search for multiple kinds of mistakes (twisted- and dropped stitches) concurrently. In other words, her mistakes became more salient, and pattern templates became easier to accurately remember, as she became more skilled. Beyond that anecdote, it is quite likely that differential deployment of attention as a result of idiosyncratic development exerts influence on how people perform VWM-based tasks and thus interact with the world around them.

Besides the aforementioned idiosyncrasies, the modulatory role of age (at the group

level) in how attention is captured may translate to VWM-based tasks. When items are held in VWM, they are susceptible to interference from novel incoming sensory information (such as distractors in a search task; Bae & Luck, 2019b; Clapp et al., 2010; Gresch et al., 2021; Hakim et al., 2020) as well as from long-term memory representations (such as from previously-task-relevant information; Jonides & Nee, 2006; Oberauer & Lin, 2017). Notably, younger children and older adults appear to be more susceptible to visual distraction than young adults (Hommel et al., 2004; McGinnis, 2012; Wetzel & Schröger, 2007), and this increased distractibility may extend to increased interference in VWM (Emery et al., 2008; Lustig and Jantz, 2015; but see Maniglia and Souza, 2020). Especially in Chapters 1-4, there were a few possible sources of interference, such as distractors in the search tasks and repeated stimulus occurrence (both as relevant- and irrelevant information). Age differences in susceptibility to distraction and interference may therefore lead those younger and older groups to be less reliant on VWM when external information can be reinspected. Moreover, neurotypical individuals generally exhibit small gaze biases (visuospatial pseudoneglect) towards the left- and upper parts of the visual field in many tasks. However, my colleagues and I showed that these biases shift over the course of the lifespan, trending more towards the right and further upwards (Strauch et al., 2024). Although the range of these biases is limited to a few degrees of visual angle, they may lead younger people to approach each visual search scene slightly differently than older individuals. I consider the possibility that these attentional differences between age groups translate to *how* VWM is applied in daily life a promising avenue for future investigation.

Furthermore, the effects of saliency and arousal on how we move our eyes in non-VWM tasks likely transfer to VWM tasks and affect the storage-sampling trade-off. For example, oblique saccades require more effort than those along the cardinal axes (Koevoet, Strauch, Naber, & Van der Stigchel, 2023; Koevoet et al., 2024) – and these differences make external sampling from some locations inherently more costly than from others, even if they are equidistant from the current fixation. Koevoet and colleagues reported that oblique saccades are therefore made less often, both during a task in which participants simply made saccades to one of two dots, as well as during guided visual search tasks with low- and high cognitive demand. Especially in their high-demand task, participants made even fewer costly saccades, again showing that arousal levels influence where and how we move our eyes – also in guided search tasks. Moreover, saccade costs still affect where we move our eyes when trial displays contain strong saliency content (Koevoet et al., under review). As such, if highly salient (and likely relevant) to-be-remembered information in a VWM task requires a very costly saccade, we may be inclined to first fixate and memorise a less-salient (and likely less-relevant) object that is in a more affordable location (reminiscent of central biases in free-viewing; Tatler, 2007; Tatler et al., 2006). Conversely, representations in VWM partially *determine* what is salient; and VWM contents are therefore likely to influence attentional capture even in the absence of particular saliency content (Gayet et al., 2013; Olivers et al., 2006). Thus, saliency and arousal likely influence how we interact with the external world during VWM tasks. Vice versa, VWM contents can modulate what constitutes saliency and VWM load inherently affects our arousal, thereby influencing where and how we move our eyes to sample from the external world.

In sum, I posit that there is sufficient cause to expect that findings from goal-directed and VWM-related tasks (Part I) will largely translate to free-viewing (Part II), and vice versa. However, I suggest that there is also merit in taking the two different aspects discussed in this dissertation at face value. A better understanding of how VWM and eye movements are used in interaction with the external world is, besides theoretically informative, practically useful for e.g., clinical applications (Böing et al., 2023, 2025), education (Baddeley & Hitch, 1974), visualisation considerations (Bylinskii et al., 2017), and user experience (UX) design (such as instruction manuals; Sauter et al., 2020; Sullivan et al., 2021).

Conclusion

While walking through the forest, while at a sports match, or while assembling Swedish furniture, we constantly visually sample our surroundings. Luckily, sampling relevant visual information from the external world is not just a challenge that we need to solve – the external world itself can also ease that challenge, because much of the visual information remains available and allows us to sample from it in a just-in-time manner. When and how we make use of visual working memory and eye movements to sample from the external world is therefore incredibly dynamic and contextual, effectuated by a trade-off which weighs almost every aspect of the challenges and affordances provided by the external world. In this dissertation, I have therefore argued that it is of great importance to investigate how we make use of the external world by studying both visual working memory and eye movements in their respective and combined contexts.

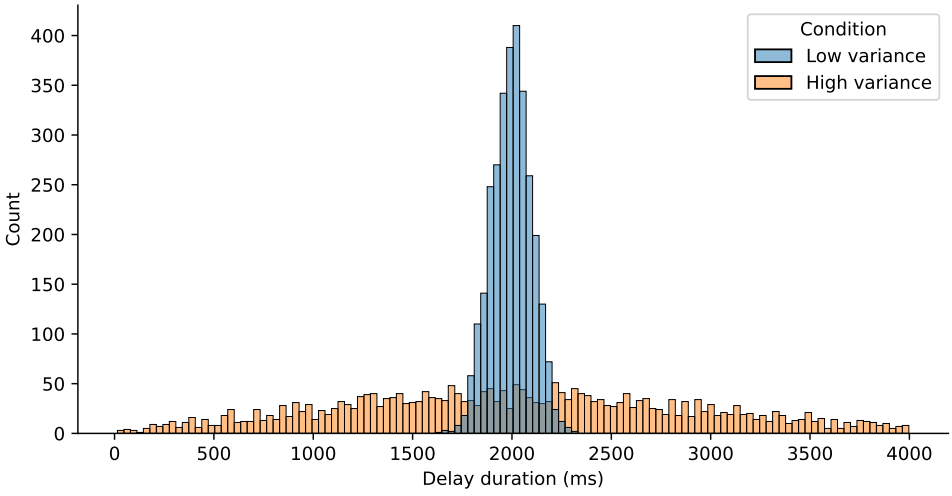
Supplementary materials

Chapter 1: Just-in-time encoding into visual working memory is contingent upon constant availability of external information

Supplementary Table 1: Outcome measures used for analysis of both experiments.

	Outcome variable	Description
A	Example grid inspections	Calculated by counting how many times within a trial the participant made a saccade across the centre of the screen from the right side to the left side. In effect, this variable represents how often participants sampled externally by looking toward the example grid after focusing on the working- and resource area. We did not count crossings in which only the hourglass was fixated, and assumed that short fixations would be unlikely to allow for meaningful encoding (e.g., Bays et al., 2011). Therefore an inspection would only be counted if the example grid was viewed for at least 120ms before the participant crossed back towards the working- and resource area.
B	Fixations per inspection	Computed by dividing the number of fixations within the boundaries of the example grid by the number of useful inspections. This variable approximates how much information participants attempted to take in each time they placed their overt attention on the example grid.
C	Items placed per inspection	Computed by dividing the number of correctly placed items per trial by the number of useful inspections made in that trial. It is an estimate of how many items participants (accurately) encoded during each inspection.
D	Completion time (s)	Calculated from the start of the trial until all items were placed correctly, or until the 42-second timer was reached. Because the periods during which the example grid was occluded were not useless to participants (i.e., they could still place items during that time), only the time spent gazing at the hourglass in the location of the occluded example grid was subtracted from the completion time.
E	Errors per trial	An error constituted the attempted placement of any item in an incorrect slot in the working grid. A greater number of errors may reflect that items were encoded less accurately (Koevoet, Naber, et al., 2023; van den Berg et al., 2012) or that participants had more liberal thresholds for the quality of memory representations that they were willing to act on (Sahakian et al., 2023).
$\mathbf{F}_{Exp1}$	Proportion spent waiting	Expressed as the duration that participants spent gazing at the hourglass, divided by the actual duration with which the example grid was occluded during that trial. This measure effectively reflects the proportion of a trial that participants spent unproductively waiting. For example: In a trial in the Low condition, if the example grid was occluded for 12 seconds in total and a participant spent 600 ms gazing at the hourglass, the proportion spent waiting is 0.05. In the High condition, if the grid was occluded for 6 seconds in total and a participant spent 300 ms gazing at the hourglass, the proportion spent waiting is also 0.05. As such, the proportion that participants spent waiting was standardized between 0 and 1 and could be compared between conditions.

	Outcome variable	Description
$\mathbf{F}_{Exp2}$	Time spent waiting (s)	Represents how long participants gazed at the hourglass while the example grid was occluded, and provides an indication whether overall delay durations were similar between conditions in which a delay was present.



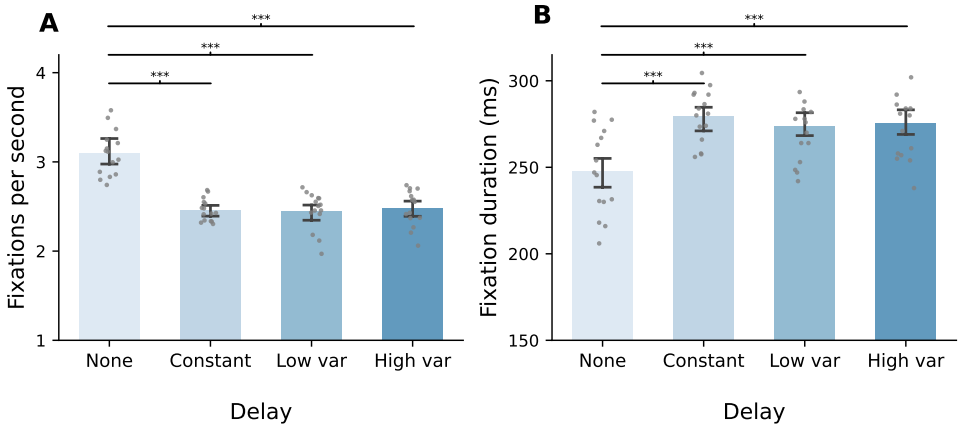
Supplementary Figure 1: Distribution of generated delay durations in Experiment 2.

Supplementary Table 2: Outcomes of Bayesian Repeated-Measures ANOVAs, between the three delay conditions in Experiment 2 (constant, low variance, high variance).

	Outcome variable	BF_{10}
A	Example grid inspections	1.581
B	Fixations per inspection	0.278
C	Items placed per inspection	0.516
D	Completion time (s)	0.227
E	Errors per trial	0.163
F	Time spent waiting (s)	1.288

Supplementary Table 3: Statistical outcomes for Bayesian paired samples t-tests between the three delay conditions in Experiment 2, corrected for three comparisons.

	Outcome variable	Condition	Comparison	BF_{10}
A	Example grid inspections	Constant	Low variance	1.673
		Constant	High variance	0.380
		Low variance	High variance	0.274
B	Fixations per inspection	Constant	Low variance	0.211
		Constant	High variance	0.172
		Low variance	High variance	0.233
C	Items placed per inspection	Constant	Low variance	0.465
		Constant	High variance	0.170
		Low variance	High variance	0.293
D	Completion time (s)	Constant	Low variance	0.174
		Constant	High variance	0.243
		Low variance	High variance	0.168
E	Errors per trial	Constant	Low variance	0.154
		Constant	High variance	0.154
		Low variance	High variance	0.154
F	Time spent waiting (s)	Constant	Low variance	1.022
		Constant	High variance	0.414
		Low variance	High variance	0.251

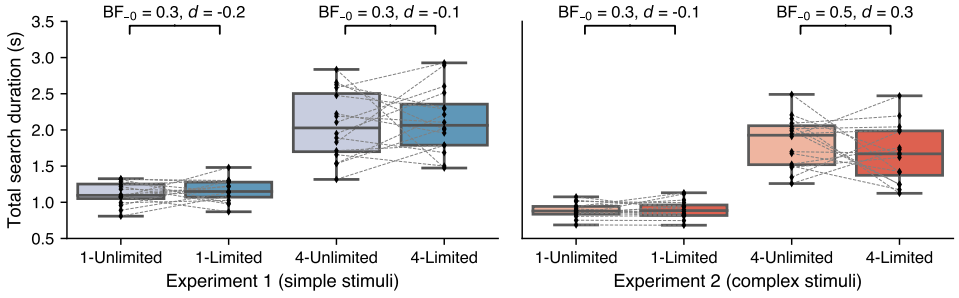


Supplementary Figure 2: Barplots (mean ± 95% within-subjects CI) for each variable, per condition. Individual points represent within-participant aggregates. **A.** The number of fixations per second per trial. **B.** The median fixation duration in milliseconds per trial. **Note.** Post-hoc paired samples t-tests (Bonferroni corrected); *** $p < .001$, ** $p < .01$, * $p < .05$.

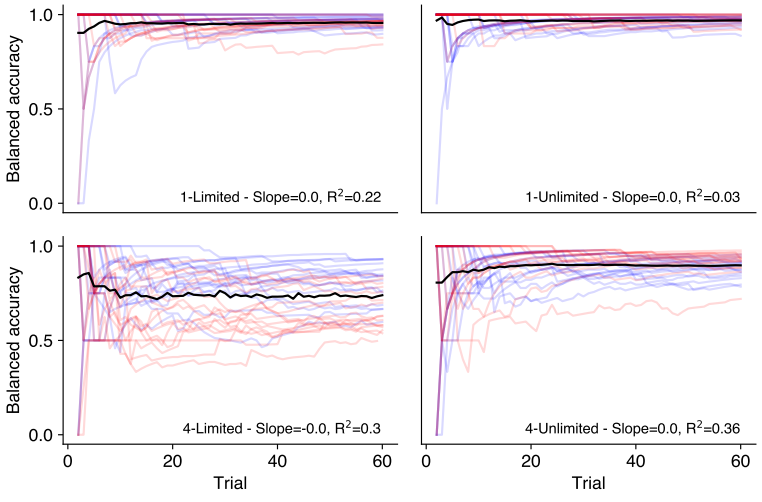
Supplementary Table 4: Outcomes of Linear Mixed Effect model (LME) with completion time as dependent variable. Data from Experiment 2. Conditions were grouped into *no delay* (baseline condition) and *delay* (constant, low variance, high variance). LME models were run with the Lmer function from lme4 (Bates et al., 2015, version 1.1-35.1) using pymer4 (Jolly, 2018, version 0.8.1). Formula used: `Completion time ~ Inspections + Fixations + Errors + Condition + (Inspections + Fixations + Errors + Condition | ID)`.

	β	2.5% CI	97.5% CI	SE	df	t	p	Sig.
(Intercept)	9.449	7.057	11.840	1.220	13.517	7.744	<.001	***
Example grid inspections	1.840	1.239	2.440	0.306	11.627	6.003	<.001	***
Fix. per insp.	0.492	0.326	0.658	0.085	12.539	5.803	<.001	***
Errors per trial	2.190	1.879	2.502	0.159	12.499	13.777	<.001	***
Condition (no delay - delay)	-4.151	-5.750	-2.552	0.816	13.785	-5.089	<.001	***

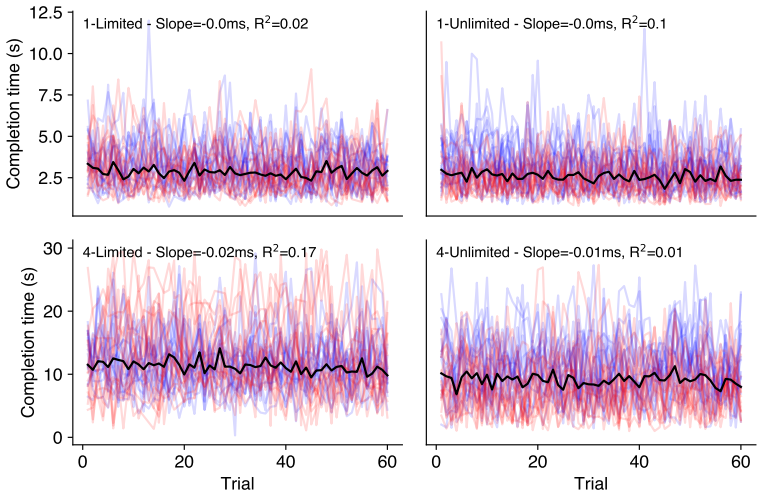
Chapter 2: Don't hide the instruction manual: A dynamic trade-off between using internal and external templates during visual search



Supplementary Figure 1: No differences in total search duration between Unlimited- and Limited conditions. Search duration was computed as the sum duration of all fixations in the search array. Note: Both panels visualize data of correctly answered and matching trials only. Diamond markers denote individual participants.



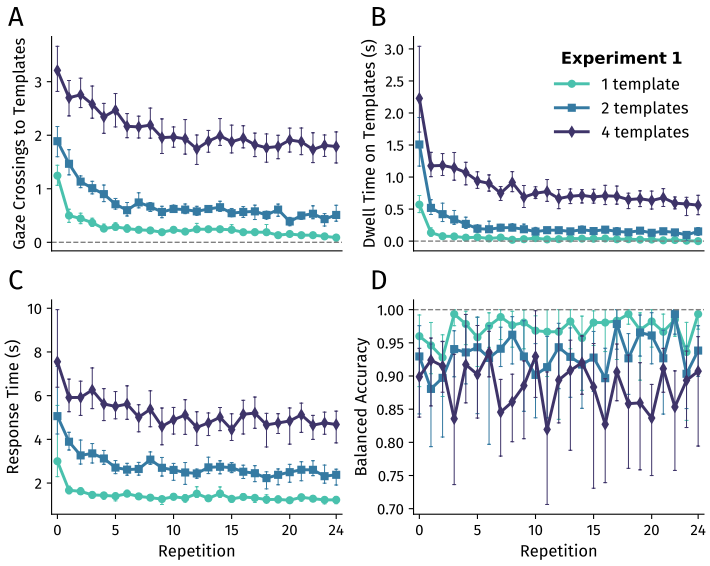
Supplementary Figure 2: Balanced accuracy over trials. The accuracy of the first trial was always 0 or 1. Balanced accuracy converges and fluctuations become smaller as additional trials are taken into account. There was no clear learning effect (increase in accuracy) within blocks. Note: Blue lines visualize participants in Experiment 1, red lines visualize participants in Experiment 2. The thick black line shows the median over all participants.



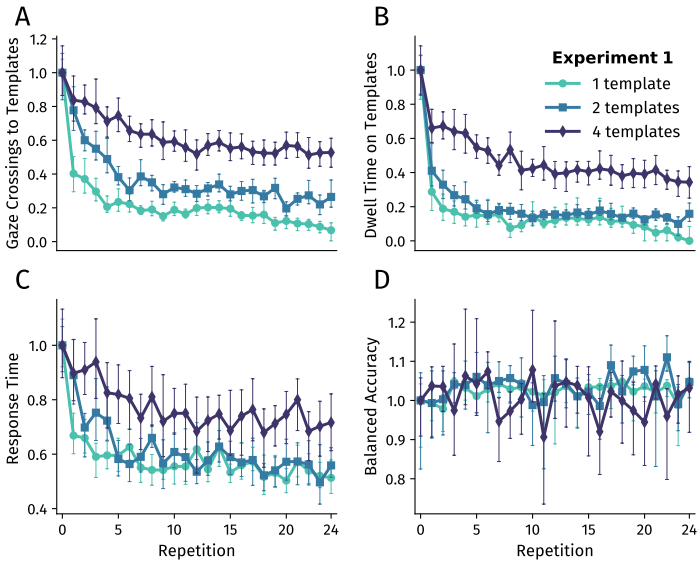
Supplementary Figure 3: Completion time over trials. The accuracy of the first trial was always 0 or 1. Balanced accuracy converges and fluctuations become smaller as additional trials are taken into account. There was no clear learning effect (decrease in completion time) within blocks. Note: Blue lines visualize participants in Experiment 1, red lines visualize participants in Experiment 2. The thick black line shows the median over all participants.

Chapter 3: Persistent resampling of external information despite twenty-five repetitions of the same visual search templates

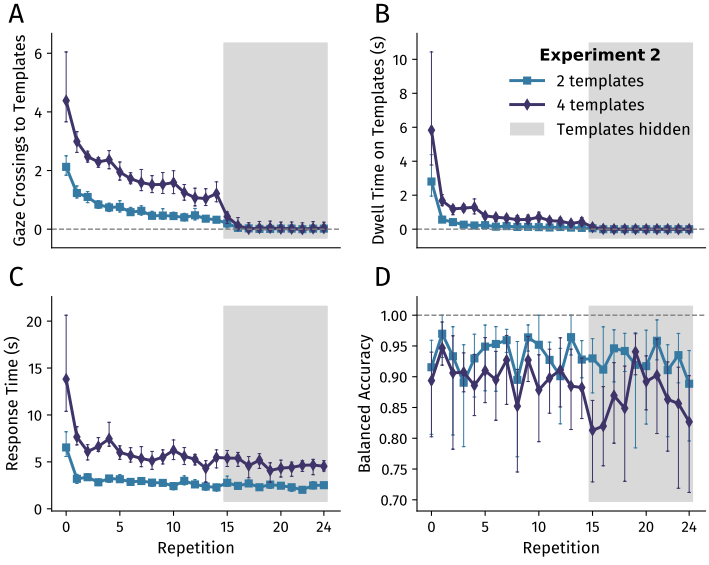
Supplementary Figures of Main Outcomes



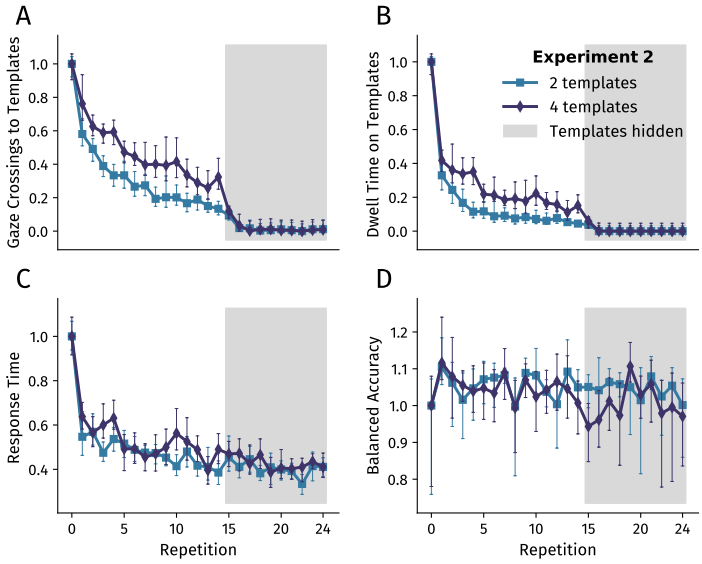
Supplementary Figure 1: Experiment 1 outcome measures. Data was aggregated over all six template sets per participant, split per template set size. The subfigures show across-participant (N=15) averages, ± 95% within-participant confidence intervals (Morey, 2008).



Supplementary Figure 2: Experiment 1 outcome measures, relative to the initial (0^{th}) repetition. Data was aggregated over all six template sets per participant, split per template set size and repetition. The subfigures show across-participant (N=15) averages, ± 95% within-participant confidence intervals.



Supplementary Figure 3: Experiment 2 outcome measures. Data was aggregated over all eight template sets per participant, split per template set size. The subfigures show across-participant (N=14) averages, ± 95% within-participant confidence intervals.



Supplementary Figure 4: Experiment 2 outcome measures, relative to the initial (0^{th}) repetition. Data was aggregated over all eight template sets per participant, split per template set size and repetition. The subfigures show across-participant (N=15) averages, ± 95% within-participant confidence intervals.

The purpose of search template inspections: LMEs

Linear Mixed Effect (LME) models were run with the Lmer function from lme4 (Bates et al., 2015, version 1.1-35.1) using pymer4 (Jolly, 2018, version 0.8.1). We report the formulae and output. Long-term memory scores were binary (correct or incorrect), and were therefore tested with a binomial model family. For each model, we attempted to control for as many relevant variables as possible, provided that the model would still converge. All data together with analysis scripts and supplementary materials may be retrieved via the Open Science Framework https://osf.io/nr5qe/.

Short-term benefits (paragraph 4.1)

Response Time

Response Time (seconds) ~ Crossings * Bin + Condition + (Crossings * Bin + Condition | ppID)

	β	2.5 % CI	97.5 % CI	*SE*	df	t	*p*	sig.
(Intercept)	0.678	0.343	1.012	0.171	22.277	3.973	0.001	***
No. crossings	1.299	1.010	1.588	0.147	12.232	8.812	< .001	***
Bin	0.056	0.010	0.101	0.023	78.224	2.369	0.020	*
Condition	0.504	0.351	0.657	0.078	27.196	6.445	< .001	***
No. crossings * Bin	0.120	-0.004	0.243	0.063	7.547	1.890	0.098	

Accuracy

Correct ~ Crossings * Bin + Condition + Version + (Crossings * Bin + Condition + Version | ppID)

	β	2.5 % CI	97.5 % CI	*SE*	df	t	*p*	sig.
(Intercept)	0.996	0.929	1.064	0.034	25.212	29.030	< .001	***
No. crossings	-0.013	-0.044	0.019	0.016	39.008	-0.794	0.432	
Bin	-0.002	-0.019	0.016	0.009	33.685	-0.170	0.866	
Condition	-0.020	-0.033	-0.007	0.006	30.594	-3.094	0.004	**
Version	-0.019	-0.063	0.025	0.023	19.182	-0.843	0.410	
No. crossings * Bin	-0.001	-0.011	0.010	0.005	9.101	-0.140	0.891	

LTM performance (paragraph 4.2)

LTM correct ~ Crossings + Dwell Time + Condition + (Crossings + Dwell Time + Condition | ppID); family = binomial

	β	2.5 % CI	97.5 % CI	*SE*	OR	Prob.	z	*p*	sig.
(Intercept)	0.132	-0.885	1.150	0.519	1.142	0.533	0.255	0.799	
No. crossings	-0.627	-1.165	-0.089	0.274	0.534	0.348	-2.284	0.022	*
Dwell time	0.001	-.000	0.001	.000	1.001	0.500	1.169	0.243	
Condition	0.602	0.126	1.078	0.243	1.826	0.646	2.480	0.013	*

Fixation duration (paragraph 4.3)

Targets

Target fixation dur. (ms) ~ Crossings + Condition + (Crossings | ppID)

	β	2.5 % CI	97.5 % CI	*SE*	df	t	*p*	sig.
(Intercept)	180.16	169.84	190.48	5.27	37.36	34.22	< .001	***
No. crossings	-9.10	-11.98	-6.21	1.47	33.13	-6.18	< .001	***
Condition	7.73	5.95	9.51	0.91	6231.61	8.51	< .001	***

Distractors

Distractor fixation dur. (ms) ~ Crossings + Condition + (Crossings | ppID)

	β	2.5 % CI	97.5 % CI	*SE*	df	t	*p*	sig.
(Intercept)	156.48	149.20	163.77	3.72	32.23	42.10	< .001	***
No. crossings	1.96	0.67	3.24	0.66	36.63	2.98	0.005	**
Condition	5.47	4.59	6.36	0.45	5897.14	12.10	< .001	***

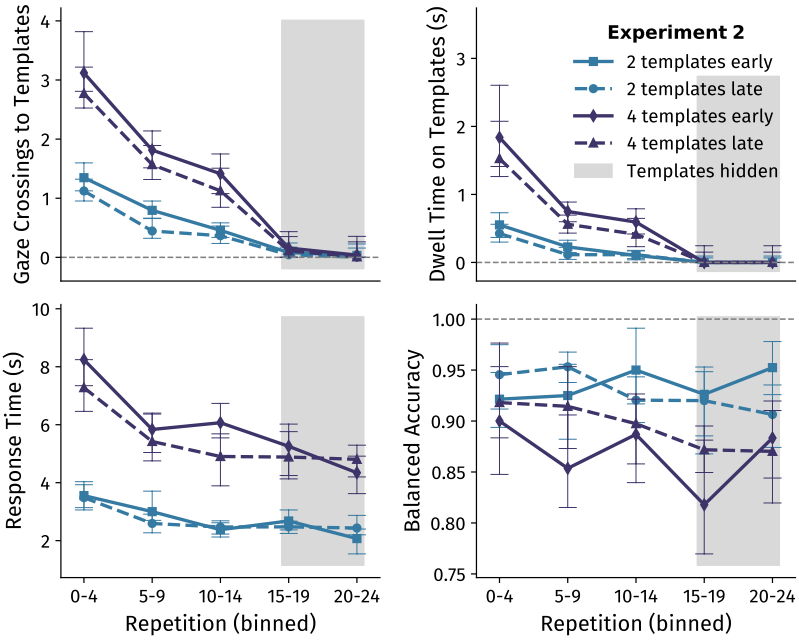
Possible learning effects

To test for possible learning effects, we split the data based on the first three and last three template sets of each block, respectively. We used repeated-measure ANOVAs to test for main effects of early-versus-late template sets, as well as the interaction with the number of templates and the repetition bin (Supplementary Table 1 and Supplementary Figure 5). Participants changed their strategy over the course of six template sets, as evidenced by two significant main effects; in later template sets they inspected templates less often and dwelled shorter. This did not affect our main conclusions, however: Although significant interaction effects between early/late template sets and repetition bin indicate that response times and accuracy were different over the course of the twenty-five repetitions, overall response times and accuracy remained unaffected by learning, as evidenced by non-significant main effects.

Supplementary Table 1: Outcomes of repeated-measure ANOVAs (*p*-value / η_p^2) for the first three template sets (Early) versus the last three template sets (Late) of each block in Experiment 2. Tests where *p*<.05 are highlighted in bold font.

RM ANOVA	Gaze crossings	Dwell time	Response Time	Accuracy
Main effect of Early/Late	**.002 / .530**	**.040 / .285**	.125 / .172	.261 / .096
Templates * EarlyLate	.614 / .020	.268 / .094	.124 / .172	.415 / .052
Repetition bin * EarlyLate	**.016 / .238**	**.042 / .170**	**.017 / .203**	**.049 / .164**
Templates * Rep. bin * EarlyLate	.422 / .068	.493 / .062	.164 / .116	.731 / .038

Furthermore, we split the data based on the counterbalanced block order; participants either performed the 2-template condition first and the 4-template condition last



Supplementary Figure 5: Experiment 2 outcome measures. Data was aggregated over all eight template sets per participant, split per template set size and binned in sets of five repetitions. Data was furthermore split between early (first 3) template sets and late (last 3) template sets per condition. The subfigures show across-participant (N=14) averages, ± 95% within-participant confidence intervals.

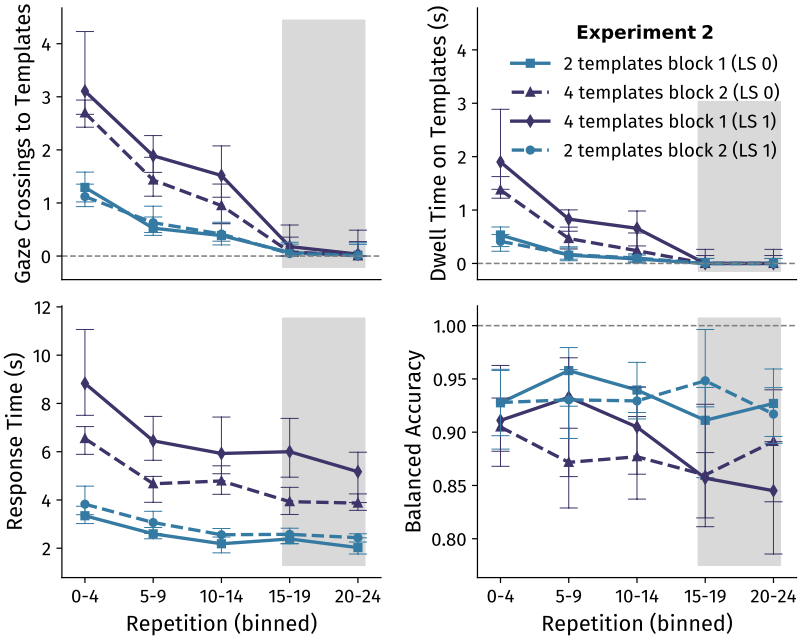
(Latin Square 0), or vice versa (Latin Square 1). We used mixed-design ANOVAs to test for main effects of block order, as well as the interaction with the number of templates and the repetition bin. In Supplementary Figure 6, it appears that block order descriptively affected our outcome measures. However, none of the statistical tests (reported in Supplementary Table 2) were significant.

Supplementary Table 2: Outcomes of mixed-design ANOVAs (p-value / η_p^2). The effect of block order (Latin Square) was a between-subjects factor.

ANOVA	Gaze crossings	Dwell time	Response Time	Accuracy
Main effect of Latin Square	.467 / .045	.306 / .087	.156 / .160	.894 / .002
Templates * LS	.083 / .230	.077 / .238	.083 / .230	.745 / .009
Repetition bin * LS	.636 / .033	.602 / .054	.636 / .037	.378 / .082
Templates * Rep. bin * LS	.343 / .086	.162 / .125	.555 / .050	.123 / .138

Timing of gaze crossings

In Hoogerbrugge et al. (2023) we analysed when crossings to templates occurred, and how often participants ended with their gaze on templates. We repeated those analyses here, though taking into account the time course of multiple repetitions. To investigate whether there was a pattern in the timing of gaze crossings toward



Supplementary Figure 6: Experiment 2 outcome measures. Data was aggregated over all eight template sets per participant, split per template set size and binned in sets of five repetitions. Data was furthermore split based on Latin Square block order. The subfigures show across-participant (N=14) averages, ± 95% within-participant confidence intervals.

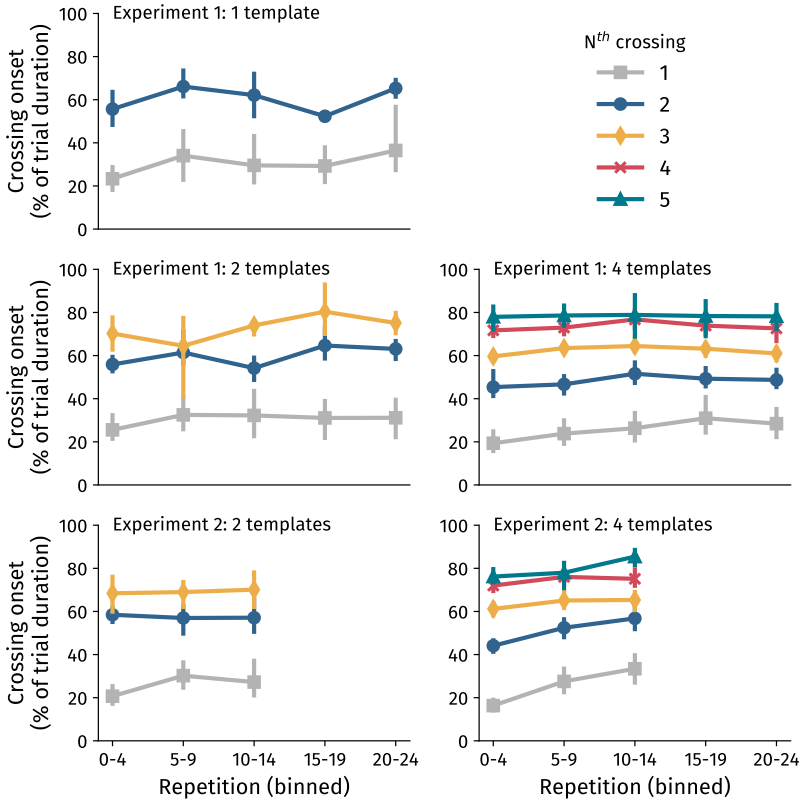
templates, we split our data based on the 'Nth' crossing within each trial (i.e., 1st, 2nd, …, 5th crossings). To retain sufficient data for analysis we only used data up to 2, 3, and 5 crossings for 1, 2, and 4 templates, respectively. However, note that the amount of data is still limited, given that many trials contained no crossings at all.

Generally, 1st crossings occurred fairly early within trials (Supplementary Figure 7) – but not immediately after trial onset. If crossings occurred immediately after trial onset, we would expect to see percentages below 10%, as in Hoogerbrugge et al. (2023). This suggests that participants first searched briefly before (re)inspecting templates. Note that this data does not include trials in which no gaze crossings were made at all. The onset of 'final' crossings was generally between 60% and 80%, which is in line with our previous work (ca. 70% for both 1- and 4-template conditions). We suspect that some of this is already double-checking or 'refreshing' behaviour, as described in Hoogerbrugge et al. (2023), but participants generally still took some time (the last 20-30% of trials) to search for the double-checked item(s). Upon visual inspection, 1st crossings seemed to occur earlier in the first five repetitions than in later repetitions. Using a Linear Mixed Effect model (Supplementary Table 3), we statistically tested whether crossing onsets differed between repetition bins, between Nth crossings, and between conditions. There was no main effect of the repetition bin on whether crossings occurred, nor were there significant interactions with the Nth crossing or number of templates. Thus, inspection timing was stable over the course of the 25 repetitions, which in turn suggests that participants' strategy regarding

when to resample templates remained stable over time.

Supplementary Table 3: LME outcomes for the onset of crossings (as a percentage of trial duration): Onset ~Bin * Nth crossing * No. templates + (Bin * Nth crossing * No. templates | ppID)

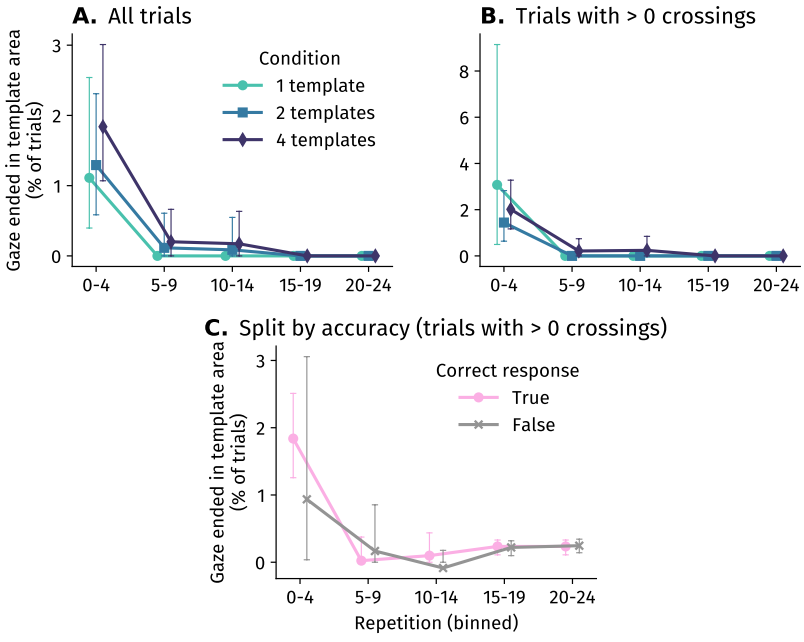
	β	2.5%CI	97.5%CI	*SE*	*df*	*T*	*p*
(Intercept)	-4.47	-21.61	12.68	8.75	55.41	-0.51	.612
Repetition bin	-1.26	-8.71	6.19	3.80	34.03	-0.33	.742
Nth crossing	**30.21**	**20.52**	**39.90**	**4.94**	**66.02**	**6.11**	**<.001**
No. templates	3.18	-1.58	7.94	2.43	48.75	1.31	.197
Bin * Nth crossing	2.56	-1.63	6.75	2.14	32.69	1.20	.239
Bin * No. templates	1.26	-0.72	3.24	1.01	29.82	1.25	.221
Nth crossing * No. templates	**-3.71**	**-6.19**	**-1.23**	**1.27**	**62.97**	**-2.93**	**.005**
Bin * Nth crossing * No. templates	-0.78	-1.84	0.28	0.54	32.15	-1.45	.157



Supplementary Figure 7: Onsets of crossings towards the template area as a percentage of trial duration, across repetition bins. Split by Nth crossing and by condition. Points denote averages, errorbars denote ± 95% within-participant confidence intervals.

Furthermore, in our previous work, we found that participants ended with their gaze on templates in 5-10% of trials on average (ranging between 0% and 35%;

Figure 5B in Hoogerbrugge et al. (2023)). Moreover, ending a trial while inspecting templates was linked to higher accuracy. In the current study, we found that trials in which gaze ended on the templates occurred considerably less often overall (Supplementary Figure 8A). In order to obtain a closer comparison to our previous work – in which each trial contained new templates and therefore at least one crossing was always necessary – we included only trials in which any crossings were made at all (Supplementary Figure 8B). The percentage of trials in which gaze ended on templates was still lower than in Hoogerbrugge et al. (2023). Finally, we split this data based on whether a correct or incorrect response was given after 'double-checking' (Supplementary Figure 8C). From this, there is no evidence that double-checking at the end of trials was linked to a benefit for accuracy – but the number of occurrences is small, thus we do not wish to over-interpret this outcome measure.



Supplementary Figure 8: **A.** Percentage of trials in which gaze ended in the template area, split per condition. Points denote across-participant averages ± 95% within-participant confidence intervals. **B.** Percentage of trials in which gaze ended in the template area, split per condition. Only for trials in which more than zero crossings were made. Points denote across-participant averages ± 95% within-participant confidence intervals. **C.** Percentage of trials in which gaze ended in the template area, aggregated over conditions, split by correct/incorrect response. Only for trials in which more than zero crossings were made. Points denote across-participant averages ± 95% within-participant confidence intervals.

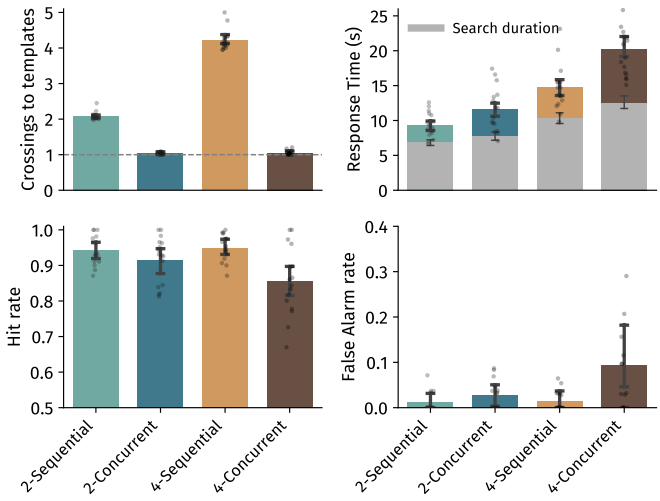
There is a possible explanation for the discrepancy with Hoogerbrugge et al. (2023). There, participants would encode templates, then search, then after searching decide whether a double check was necessary. In the current study however, participants could already decide soon after trial onset (or even in between trials) whether their memory representations were 'good enough' to start searching. Thus, double checks at the end of trials were probably less beneficial than in our previous study – again

emphasizing the need to study resampling behaviour in the context of longer-term optimizations.

Chapter 4: Multi-target visual search flexibly switches between concurrent and sequential search modes

Template inspections, response times and accuracy

We report four outcome variables (Figure Supplementary Figure 1): (1) Crossings to templates: the number of times that participants moved their gaze from the search area to the template area as a measure of external sampling behaviour. Participants had to cross to the template area for inspection at least once per trial. (2a) Response Time (seconds): the response time for each trial, measured from trial onset until the spacebar was pressed. (2b) Search duration (seconds): the sum of fixation durations within the search array. (3) Hit rate: the proportion of targets which was found. For example, if 3 out of 4 targets were clicked, the hit rate is 0.75. Clicks were counted as correct if they were within a 2° radius of the center of the target stimulus. Target-absent trials were excluded for the computation of this metric. (4) False Alarm (FA) rate: clicks on non-target stimuli as a proportion of the total number of clicks. For example, if one of three clicks was on a distractor, the FA rate is 0.33.



Supplementary Figure 1: (*Upper left*) Crossings to templates: the number of times that participants moved their gaze from the search area to the template area as a measure of external sampling behaviour. Participants had to cross to the template area at least once per trial. (*Upper right*) Response Time in seconds: the response time for each trial, measured from trial onset until the spacebar was pressed. (*Upper right; overlaid in gray*) Search duration in seconds: the sum of fixation durations within the search array. (*Lower left*) Hit rate: the proportion of targets which was found. Target-absent trials were excluded for the computation of this metric. (*Lower right*) False Alarm rate: clicks on non-target stimuli as a proportion of the total number of clicks. Bar heights denote across-participant averages, error bars denote ± 95% within-participant confidence intervals (Morey, 2008). For ease of comparison, y-axes for each outcome measure are identical between experiments.

Experiment 1

For Experiment 1, we computed repeated measures ANOVAs (in JASP 0.18.3; JASP Team, 2022), and report main effects of template set size (2/4 templates), search instruction (Sequential/Concurrent), and interaction effects for Reponse Time (Table Supplementary Table 1), Hit rate (Table Supplementary Table 2) and False Alarm rate (Table Supplementary Table 3).

Supplementary Table 1: Experiment 1: Response time

	Sum of Squares	df	Mean Square	F	p	η_p^2
Templates	788.361	1	788.361	196.900	< .001 ***	0.929
Instruction	238.631	1	238.631	60.794	< .001 ***	0.802
Templates * Instruction	44.921	1	44.921	13.754	0.002 **	0.478

Supplementary Table 2: Experiment 1: Hit rate

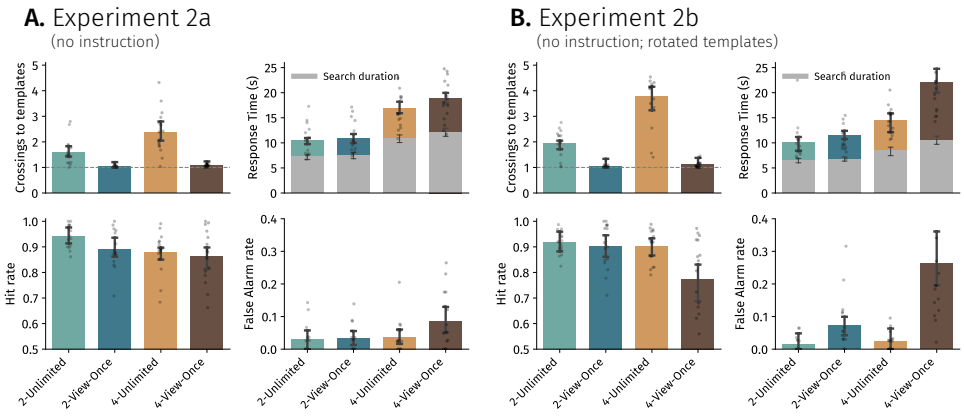
	Sum of Squares	df	Mean Square	F	p	η_p^2
Templates	0.011	1	0.011	3.268	0.091	0.179
Instruction	0.059	1	0.059	19.176	< .001 ***	0.561
Templates * Instruction	0.018	1	0.018	5.262	0.037 *	0.260

Supplementary Table 3: Experiment 1: False alarm rate

	Sum of Squares	df	Mean Square	F	p	η_p^2
Templates	0.021	1	0.021	4.323	0.055	0.224
Instruction	0.036	1	0.036	6.789	0.020 *	0.312
Templates * Instruction	0.016	1	0.016	3.890	0.067	0.206

Experiments 2a and 2b

For Experiments 2a and 2b (Figure Supplementary Figure 2), we computed mixed-measures ANOVAs, and report main effects of Experiment (2a/2b; between-subjects) template set size (2/4 templates), template availability (Unlimited/View-Once), and search difficulty (between-subjects), as well as interaction effects for Reponse Time (Table Supplementary Table 4), Hit rate (Table Supplementary Table 5) and False Alarm rate (Table Supplementary Table 6)



Supplementary Figure 2: (*Upper left*) Crossings to templates: the number of times that participants moved their gaze from the search area to the template area as a measure of external sampling behaviour. Participants had to cross to the template area at least once per trial. (*Upper right*) Response Time in seconds: the response time for each trial, measured from trial onset until the spacebar was pressed. (*Upper right; overlaid in gray*) Search duration in seconds: the sum of fixation durations within the search array. (*Lower left*) Hit rate: the proportion of targets which was found. Target-absent trials were excluded for the computation of this metric. (*Lower right*) False Alarm rate: clicks on non-target stimuli as a proportion of the total number of clicks. Bar heights denote across-participant averages, error bars denote ± 95% within-participant confidence intervals (Morey, 2008). For ease of comparison, y-axes for each outcome measure are identical between experiments.

Supplementary Table 4: Experiment 2: Response time

	Sum of Squares	df	Mean Square	F	p	η_p^2
Experiment (between-subjects)	1.367	1	1.367	0.027	0.871	$< .01$
Templates	1743.006	1	1743.006	253.734	$< .001$ ***	0.894
Templates * Experiment	0.650	1	0.650	0.095	0.761	0.003
Availability	275.397	1	275.397	36.522	$< .001$ ***	0.549
Availability * Experiment	85.526	1	85.526	11.342	0.002 **	0.274
Templates * Availability	124.920	1	124.920	28.524	$< .001$ ***	0.487
Templates * Availability * Experiment	46.140	1	46.140	10.536	0.003 **	0.260

Random Forest classifier

First, we removed trials with more than 100 fixations (1.3%). Each trial was then split into 100 equally-sized bins (similar to guidance analyses), and each bin which contained a fixation was filled with the corresponding colour index.

Supplementary Table 5: Experiment 2: Hit rate

	Sum of Squares	df	Mean Square	F	p	η_p^2
Experiment (between-subjects)	0.012	1	0.012	0.694	0.411	0.023
Templates	0.110	1	0.110	14.388	< .001 ***	0.324
Templates * Experiment	0.006	1	0.006	0.796	0.379	0.026
Availability	0.087	1	0.087	15.375	< .001 ***	0.339
Availability * Experiment	0.013	1	0.013	2.283	0.141	0.071
Templates * Availability	0.013	1	0.013	3.617	0.067	0.108
Templates * Availability * Experiment	0.047	1	0.047	13.085	0.001 **	0.304

Supplementary Table 6: Experiment 2: False Alarm rate

	Sum of Squares	df	Mean Square	F	p	η_p^2
Experiment (between-subjects)	0.072	1	0.072	5.995	0.020 *	0.167
Templates	0.134	1	0.134	30.180	< .001 ***	0.501
Templates * Experiment	0.039	1	0.039	8.691	0.006 **	0.225
Availability	0.244	1	0.244	28.398	< .001 ***	0.486
Availability * Experiment	0.121	1	0.121	14.069	< .001 ***	0.319
Templates * Availability	0.104	1	0.104	23.819	< .001 ***	0.443
Templates * Availability * Experiment	0.037	1	0.037	8.432	0.007 **	0.219

The dataset in Experiment 1 was split between 2- and 4-template conditions (1276 and 1258 trials, respectively). We trained a Random Forest classifier on 85% of trials to predict whether trials came from the Sequential or Concurrent condition (using the default parameters in scikit-learn 1.4; Pedregosa et al., 2011). The remaining 15% of trials were used to internally validate model accuracy, measured with Area-Under-the-Curve (AUC). To obtain a better estimate of classification accuracy and robustness, this process was iterated 100 times, each time resetting the model and sampling with replacement from the available trials. Colour indices of each trial were randomly permuted on each bootstrap iteration. As a result, models only learned the structure of fixation patterns and not a standardized order of fixations. We also tested Logistic Regression and Support Vector Machine classifiers, but these were less accurate.

After model validation, we trained new Random Forest classifiers on all data from Experiment 1. These models were then used to classify behaviour from Experiments 2a and 2b. The datasets in Experiments 2a and 2b contained 3168 and 3176 trials, respectively. Again, we took into account possible variance in data-subsets by iterating the above process 1000 times, each time resetting the model, sampling with replacement from the available trials, and permuting the colour values. We report how strongly the behaviour in each trial fits with sequential versus concurrent behaviour, expressed as the percentage of times that the trial was classified as sequential across all 1000 bootstrap iterations.

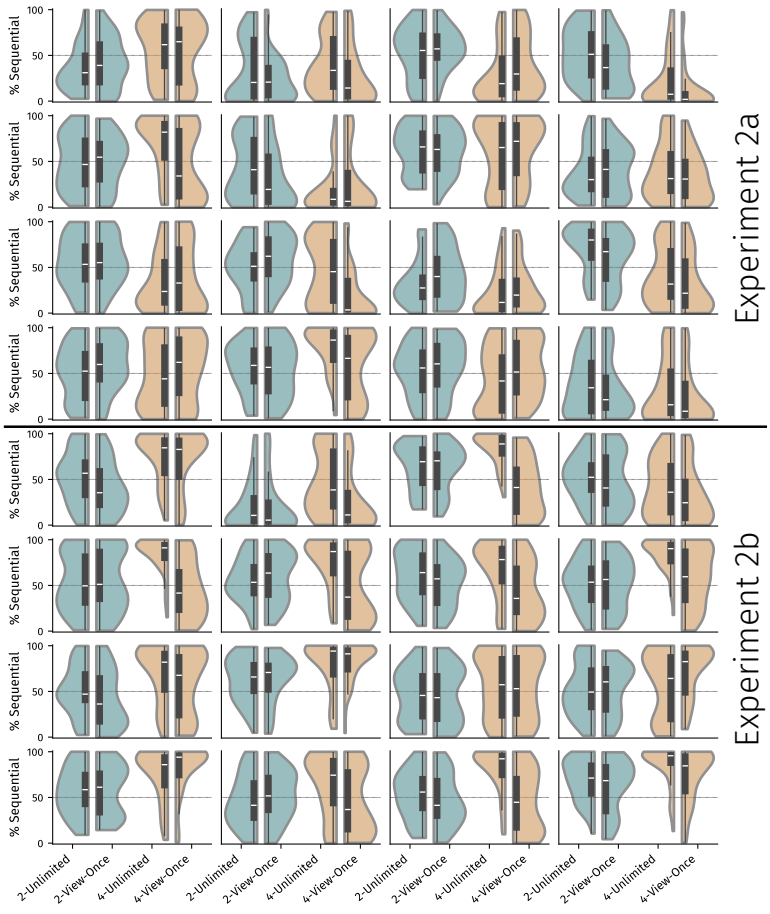
We ran mixed-measures ANOVAs to test whether model classifications were influenced by main effects of template set size (2/4 templates), template availability (Unlimited/View-Once), and template complexity (between-subjects), as well as interaction effects between those factors (Table Supplementary Table 7).

We report classification outcomes for all participants individually in Figure Supplementary Figure 3.

Supplementary Table 7: Classifier outcomes: Classification as sequential

	Sum of Squares	df	Mean Square	F	p	η_p^2
Experiment (between-subjects)	10271.988	1	10271.988	10.176	0.003	0.253
Templates	64.012	1	64.012	0.195	0.662	0.006
Templates * Experiment	5652.239	1	5652.239	17.225	< .001 ***	0.365
Availability	1742.613	1	1742.613	13.698	< .001 ***	0.313
Availability * Experiment	674.451	1	674.451	5.302	0.028 *	0.150
Templates * Availability	1312.550	1	1312.550	9.730	0.004 **	0.245
Templates * Availability * Experiment	424.400	1	424.400	3.146	0.086	0.095

Individual participants



Supplementary Figure 3: A full overview of all classification distributions from Experiment 2a (upper half) and 2b (lower half). Classifications are expressed as the percentage of times that a trial was classified as sequential across all bootstrap iterations that the trial occurred in.

Oculomotor metrics

We report three outcome variables (Figure Supplementary Figure 4): (1) Saccade amplitude in degrees of visual angle. Amplitudes were first averaged over trials, then over participants and conditions. (2) Systematicity (Kendall's Tau) for reading-style gaze movement (left-to-right, top-to-bottom). Values are adjusted against the τ obtained from shuffling all fixation locations within trials. 0 indicates low systematicity (equal to quasi-random patterns), 1 indicates very high systematicity. (3) The proportion of fixations made on irrelevant-coloured items, adjusted for chance level (33% and 20% for 2- and 4-template conditions, respectively). 0 indicates that no fixations were made on irrelevant colours, > 1 indicates that more fixations were made than would be expected at pure chance.

For Experiment 1, we computed repeated measures ANOVAs, and report main effects of template set size (2/4 templates), search instruction (Sequential/Concurrent), and interaction effects for Saccade amplitude (Table Supplementary Table 8), Systematicity (Table Supplementary Table 9) and Irrelveant colour fixation rate (Table Supplementary Table 10).

For Experiments 2a and 2b, we computed mixed-measures ANOVAs, and report main effects of Experiment (2a/2b; between-subjects) template set size (2/4 templates), template availability (Unlimited/View-Once), and search difficulty (between-subjects), as well as interaction effects for Saccade amplitude (Table Supplementary Table 11), Systematicity (Table Supplementary Table 12) and Irrelevant colour fixation rate (Table Supplementary Table 13).

Supplementary Table 8: Experiment 1: Saccade amplitude

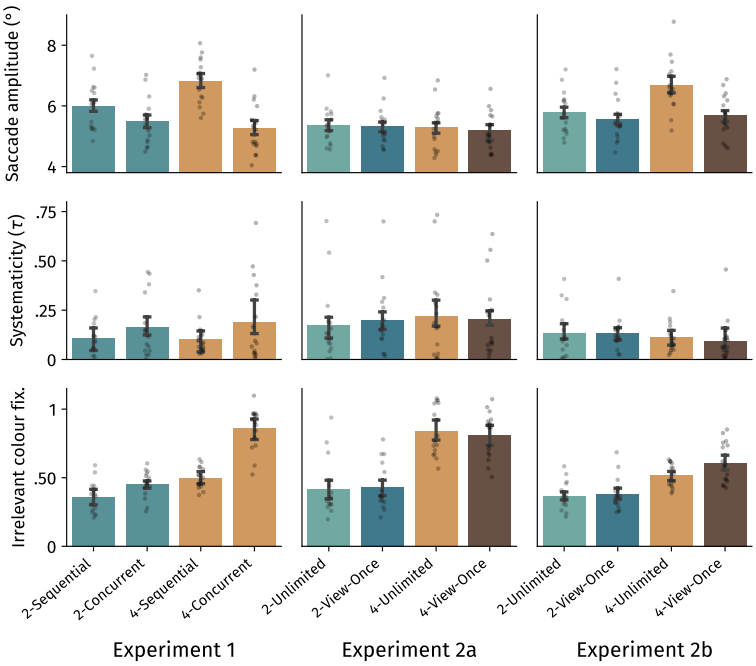
	Sum of Squares	df	Mean Square	F	p	η_p^2
Templates	1.270	1	1.270	6.961	0.019 *	0.317
Instruction	17.213	1	17.213	94.426	< .001 ***	0.863
Templates * Instruction	4.263	1	4.263	48.622	< .001 ***	0.764

Supplementary Table 9: Experiment 1 systematicity

	Sum of Squares	df	Mean Square	F	p	η_p^2
Templates	0.002	1	0.002	0.356	0.560	0.023
Instruction	0.084	1	0.084	3.480	0.082	0.188
Templates * Instruction	0.005	1	0.005	0.747	0.401	0.047

Supplementary Table 10: Experiment 1 irrelevant colour fixations

	Sum of Squares	df	Mean Square	F	p	η_p^2
Templates	1.219	1	1.219	155.739	< .001 ***	0.912
Instruction	0.836	1	0.836	54.471	< .001 ***	0.784
Templates * Instruction	0.289	1	0.289	57.345	< .001 ***	0.793



Supplementary Figure 4: Oculomotor metrics split per experiment (columns) and conditions. (*Top row*) Mean saccade amplitude in degrees of visual angle. (*Middle row*) Systematicity (Kendall's Tau) for reading-style gaze movement (left-to-right, top-to-bottom). Values are adjusted against the τ obtained from shuffling all fixation locations within trials. 0 indicates low systematicity (equal to quasi-random patterns), 1 indicates very high systematicity. (*Bottom row*) The proportion of fixations made on irrelevant-coloured items, adjusted for chance level (33% and 20% for 2- and 4-template conditions, respectively). 0 indicates that no fixations were made on irrelevant colours, > 1 indicates that more fixations were made than would be expected at pure chance. Bar heights denote across-participant averages, error bars denote ± 95% within-participant confidence intervals (Morey, 2008).

Supplementary Table 11: Experiment 2a and 2b saccade amplitude

	Sum of Squares	df	Mean Square	F	p	η_p^2
Experiment (between-subjects)	13.323	1	13.323	8.079	0.008 **	0.212
Templates	1.441	1	1.441	8.956	0.005 **	0.230
Templates * Experiment	2.966	1	2.966	18.433	< .001 ***	0.381
Availability	3.516	1	3.516	33.543	< .001 ***	0.528
Availability * Experiment	2.389	1	2.389	22.792	< .001 ***	0.432
Templates * Availability	1.417	1	1.417	12.802	0.001 **	0.299
Templates * Availability * Experiment	1.075	1	1.075	9.712	0.004 **	0.245

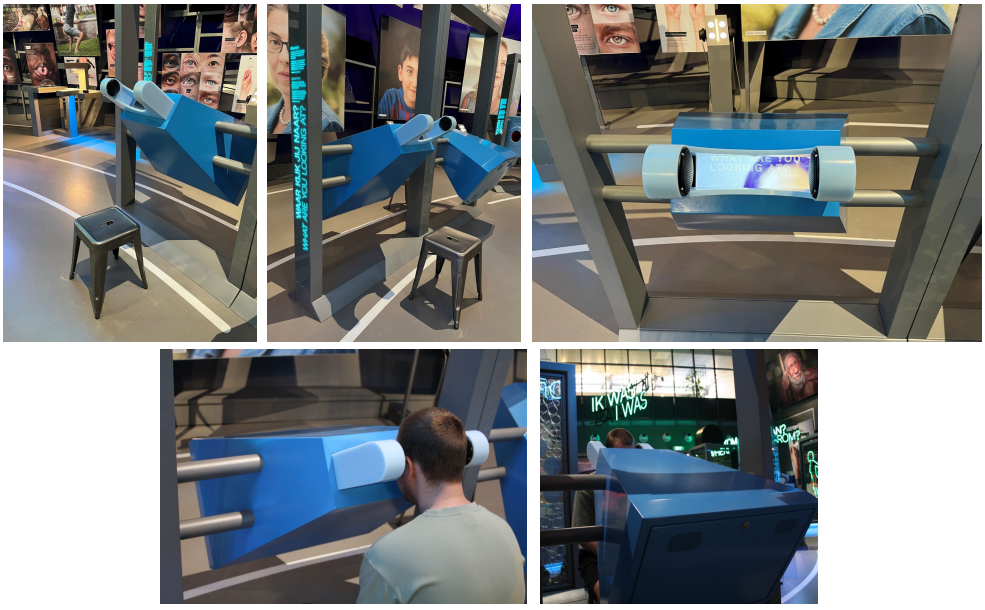
Supplementary Table 12: Experiment 2a and 2b systematicity

	Sum of Squares	df	Mean Square	F	p	η_p^2
Experiment (between-subjects)	0.213	1	0.213	2.658	0.113	0.081
Templates	$< .001$	1	1.274×10^{-4}	0.013	0.910	$< .01$
Templates * Experiment	0.023	1	0.023	2.334	0.137	0.072
Availability	$< .001$	1	9.558×10^{-5}	0.029	0.865	$< .01$
Availability * Experiment	0.002	1	0.002	0.573	0.455	0.019
Templates * Availability	0.007	1	0.007	1.181	0.286	0.038
Templates * Availability * Experiment	0.002	1	0.002	0.262	0.612	0.009

Supplementary Table 13: Experiment 2a and 2b irrelevant colour fixations

	Sum of Squares	df	Mean Square	F	p	η_p^2
Experiment (between-subjects)	0.784	1	0.784	12.820	0.001 **	0.299
Templates	2.785	1	2.785	133.859	$< .001$ ***	0.817
Templates * Experiment	0.350	1	0.350	16.818	$< .001$ ***	0.359
Availability	0.012	1	0.012	2.336	0.137	0.072
Availability * Experiment	0.030	1	0.030	5.731	0.023 *	0.160
Templates * Availability	0.003	1	0.003	0.602	0.444	0.020
Templates * Availability * Experiment	0.033	1	0.033	7.205	0.012 *	0.194

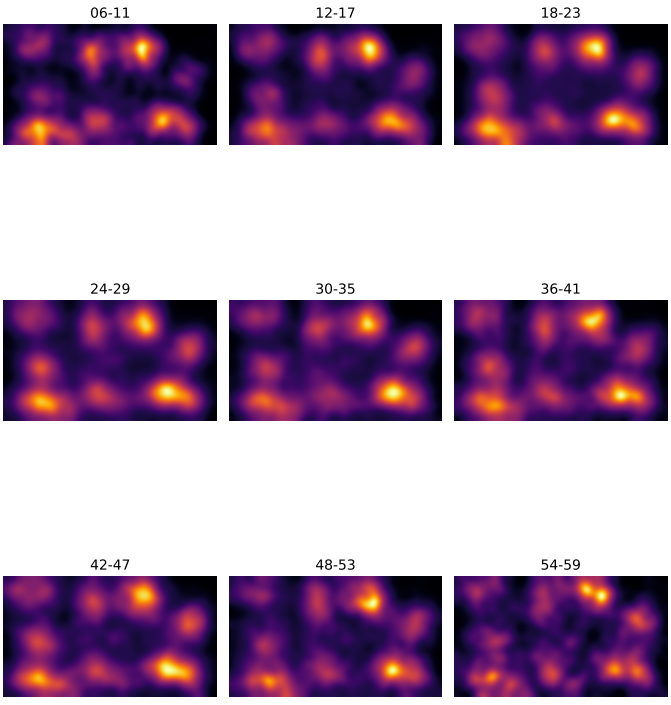
Chapter 5: Saliency models perform best for women's and young adults' fixations



Supplementary Figure 1: Setup at the science museum. Upper left and upper center: Metal box containing the eye tracker and monitor. Upper right: view inside the metal box, 'horns' are loudspeakers used after free viewing to ask for data donation. Bottom left: starting screen. Bottom center and right: participant taking part in the study. Participants sat on a small chair.

Further analyses across demographic groups

Supplementary Table 1 gives absolute NSS scores per model and baseline across age bins. Supplementary Figure 2 shows gaze distribution maps for each age bin.



Supplementary Figure 2: Spatial distribution maps of fixation locations, separated per age bin. n = 1,600.

Inferential statistics

Inferential statistics are given in Supplementary Table 2, including effect sizes and respective 95% confidence intervals for comparisons of NSS deviations across age bins from average across age bins (Figure 2, right). Assumed normality was not violated in any of the age bins' relative deviations to model performance, which is why one sample t-tests were conducted (Shapiro-Wilk tests, all $p > 0.073$) with the exception of the bin for 12-17 year olds, for which a Wilcoxon test was calculated ($V_{\text{Wilcoxon}} = 190.000$, $p = 0.007$, rank-biserial correlation = 0.810).

Detailed model performance across gender

Supplementary Table 3 gives model predictions in NSS. Supplementary Figure 3 depicts the spatial distribution maps of fixation locations. Note that the average difference between women and men (NSS = 0.0166) was highly similar to the difference of the central bias (NSS = 0.017). Supplementary Figure 3 depicts spatial distribution maps of fixation locations for women and men. Results for participants reporting

Supplementary Table 1: Absolute NSS values per baseline and model across age bins. M_{sample} is the average performance across all participants analyzed here, M_{bins} is the average NSS across age bin averages per baseline/model.

Model	M_{sample}	M_{bins}	6-11	12-17	18-23	24-29	30-35	36-41	42-47	48-53	54-59
Baselines											
Fixation map	0.71	0.834	1.02	0.862	0.837	0.731	0.729	0.729	0.769	0.816	1.013
Central bias	0.001	-0.005	-0.036	0.002	-0.025	0.028	0.009	-0.011	-0.034	0.034	-0.013
Single observer	0.174	0.185	0.256	0.211	0.219	0.176	0.164	0.148	0.169	0.164	0.156
Meaning map	0.382	0.367	0.338	0.343	0.409	0.414	0.382	0.365	0.361	0.361	0.33
Models											
RARE2012	0.194	0.182	0.163	0.147	0.185	0.226	0.197	0.188	0.185	0.216	0.129
SalGAN	0.42	0.402	0.395	0.334	0.471	0.459	0.406	0.408	0.426	0.408	0.315
DeepGazeIIE	0.405	0.392	0.396	0.392	0.44	0.435	0.401	0.361	0.378	0.407	0.32
SALICON	0.455	0.439	0.433	0.427	0.51	0.478	0.442	0.428	0.447	0.463	0.324
DVA	0.3	0.285	0.279	0.224	0.313	0.339	0.304	0.277	0.305	0.3	0.225
FES	0.052	0.043	0.051	0.014	0.039	0.083	0.064	0.042	0.029	0.059	0.007
QSS	0.338	0.328	0.365	0.294	0.363	0.364	0.338	0.304	0.332	0.337	0.259
SSR	0.265	0.252	0.257	0.206	0.297	0.278	0.283	0.257	0.246	0.28	0.163
CVS	-0.085	-0.087	-0.067	-0.076	-0.102	-0.077	-0.074	-0.091	-0.11	-0.079	-0.111
IMSIG	0.333	0.317	0.33	0.282	0.353	0.361	0.347	0.306	0.323	0.334	0.221
LDS	0.031	0.027	0.045	0.048	0.027	0.048	0.028	0.014	-0.003	0.037	-0.004
ICF	0.26	0.251	0.252	0.275	0.268	0.288	0.258	0.21	0.233	0.282	0.192
GBVS	0.166	0.156	0.134	0.171	0.151	0.207	0.157	0.142	0.13	0.2	0.113
CAS	0.161	0.154	0.177	0.171	0.158	0.18	0.161	0.135	0.134	0.187	0.086
SUN	0.167	0.165	0.183	0.147	0.161	0.183	0.176	0.141	0.147	0.206	0.144
DeepGazeI	0.334	0.319	0.322	0.311	0.363	0.365	0.328	0.294	0.311	0.346	0.232
AIM	0.26	0.259	0.283	0.264	0.269	0.276	0.259	0.21	0.237	0.297	0.234
SAM	0.258	0.255	0.254	0.337	0.288	0.27	0.215	0.225	0.236	0.288	0.182
DeepGazeII	0.4	0.382	0.373	0.341	0.42	0.446	0.396	0.37	0.383	0.393	0.317
IKN	0.202	0.188	0.174	0.164	0.194	0.247	0.202	0.18	0.171	0.223	0.135
BMS	0.186	0.175	0.216	0.182	0.202	0.201	0.192	0.159	0.155	0.194	0.074

Supplementary Table 2: Inferential statistics across age bins for one sample t-tests on deviation in NSS across models, separated per age bin. Confidence intervals for t-tests and Wilcoxon test (bin 30-35).

				95% CI	
Age bin (years)	$t(20)$	p	*Cohen's d*	Lower	Upper
6-11	1.662	0.112	0.363	-0.013	0.729
12-17	-1.453	0.162	-0.317	-0.681	0.055
18-23	4.308	$< .001$	0.94	0.497	1.364
24-29	10.756	$< .001$	2.347	1.626	3.035
30-35		0.007		0.555	0.925
36-41	-4.805	$< .001$	-1.048	-1.488	-0.589
42-47	-2.725	0.013	-0.595	-0.978	-0.198
48-53	9.099	$< .001$	1.986	1.345	2.597
54-65	-10.785	$< .001$	-2.354	-3.043	-1.631

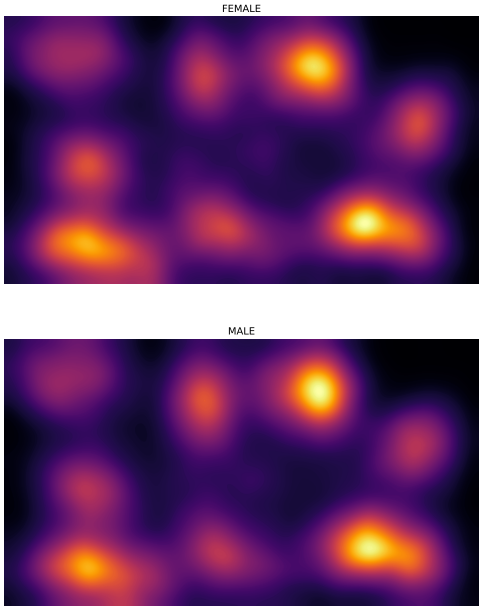
non-binary gender (n = 91) are given in Supplementary Table 4 and Supplementary Figure 4 - yet caution is necessary before interpreting these results as the non-binary option was the default setting and may therefore not have been selected intentionally but could have simply been left as-is by the participant. Furthermore, there was no 'prefer not to say' option, rendering interpretation difficult.

Supplementary Table 3: Model performance for men and women. NSS for the prediction of fixations for the full sample, men, and women. Relative deviations (NSS) between predictions for men and women are given in the fifth column. Positive numbers (blue) indicate better performance on women, negative numbers (red) indicate better performance on men.

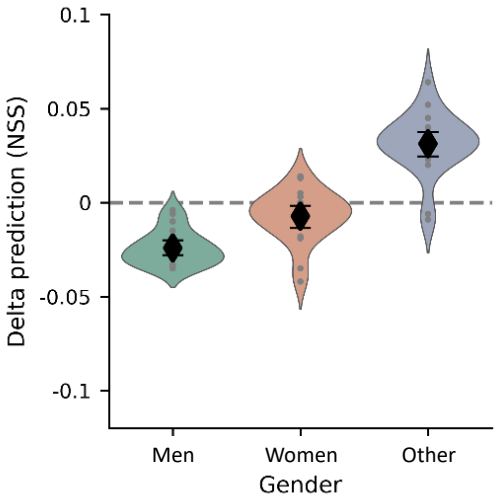
Model	Mean	Men	Women	Difference
Fixation map	0.723	0.71	0.736	0.026
Central bias	0.002	-0.007	0.011	0.018
Single observer	0.177	0.171	0.182	0.011
Meaning map	0.384	0.376	0.392	0.016
RARE2012	0.196	0.185	0.206	0.021
SalGAN	0.422	0.425	0.419	-0.006
DeepGazeIIE	0.409	0.39	0.427	0.037
SALICON	0.458	0.445	0.471	0.026
DVA	0.303	0.289	0.317	0.028
FES	0.054	0.042	0.066	0.024
QSS	0.34	0.335	0.344	0.009
SSR	0.267	0.253	0.282	0.029
CVS	-0.086	-0.084	-0.087	-0.003
IMSIG	0.335	0.33	0.34	0.01
LDS	0.031	0.027	0.035	0.008
ICF	0.262	0.253	0.27	0.017
GBVS	0.167	0.159	0.175	0.016
CAS	0.162	0.156	0.168	0.012
SUN	0.169	0.16	0.178	0.018
DeepGazeI	0.337	0.323	0.35	0.027
AIM	0.262	0.252	0.271	0.019
SAM	0.258	0.274	0.242	-0.032
DeepGazeII	0.402	0.391	0.414	0.023
IKN	0.204	0.188	0.22	0.032
BMS	0.188	0.168	0.208	0.04

Supplementary Table 4: Model performance for men, women, and participants of other gender. NSS for the prediction of fixations for the full sample, men, women, and participants with other gender. Note that results for participants with other gender have to be interpreted cautiously as this represented the default option and likely contains a substantial amount of data that is not from participants identifying as non-binary.

	Model	Mean	Men	Women	Other
0	Fixation map	0.766	0.710	0.736	0.852
1	Central bias	0.020	-0.007	0.011	0.057
2	Single observer	0.176	0.171	0.182	0.174
3	Meaning map	0.387	0.376	0.392	0.395
4	RARE2012	0.206	0.185	0.206	0.226
5	SalGAN	0.454	0.425	0.419	0.518
6	DeepGazeIIE	0.424	0.390	0.427	0.454
7	SALICON	0.472	0.445	0.471	0.501
8	DVA	0.323	0.289	0.317	0.363
9	FES	0.071	0.042	0.066	0.106
10	QSS	0.363	0.335	0.344	0.408
11	SSR	0.283	0.253	0.282	0.316
12	CVS	-0.069	-0.084	-0.087	-0.037
13	IMSIG	0.353	0.330	0.340	0.390
14	LDS	0.048	0.027	0.035	0.081
15	ICF	0.279	0.253	0.270	0.315
16	GBVS	0.179	0.159	0.175	0.202
17	CAS	0.176	0.156	0.168	0.202
18	SUN	0.164	0.160	0.178	0.155
19	DeepGazeI	0.351	0.323	0.350	0.381
20	AIM	0.258	0.252	0.271	0.252
21	SAM	0.284	0.274	0.242	0.336
22	DeepGazeII	0.425	0.391	0.414	0.470
23	IKN	0.221	0.188	0.220	0.254
24	BMS	0.203	0.168	0.208	0.232



Supplementary Figure 3: Spatial distribution maps of fixation locations, separated for women (upper) and men (lower). n = 1,600.

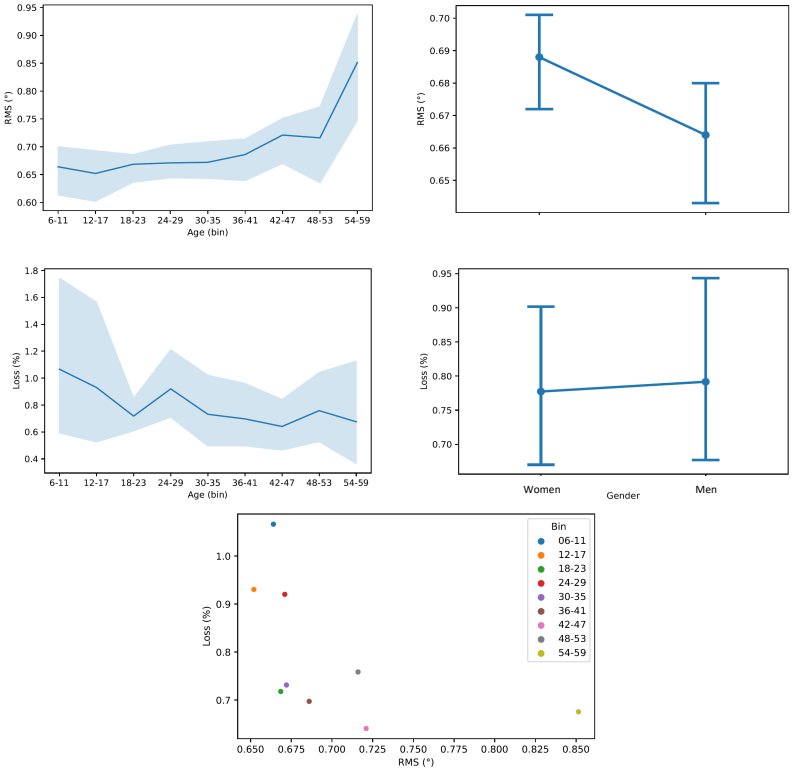


Supplementary Figure 4: Relative deviations in NSS across models for men, women, and other gender. n = 1,600 (men, women). n = 91 (other).

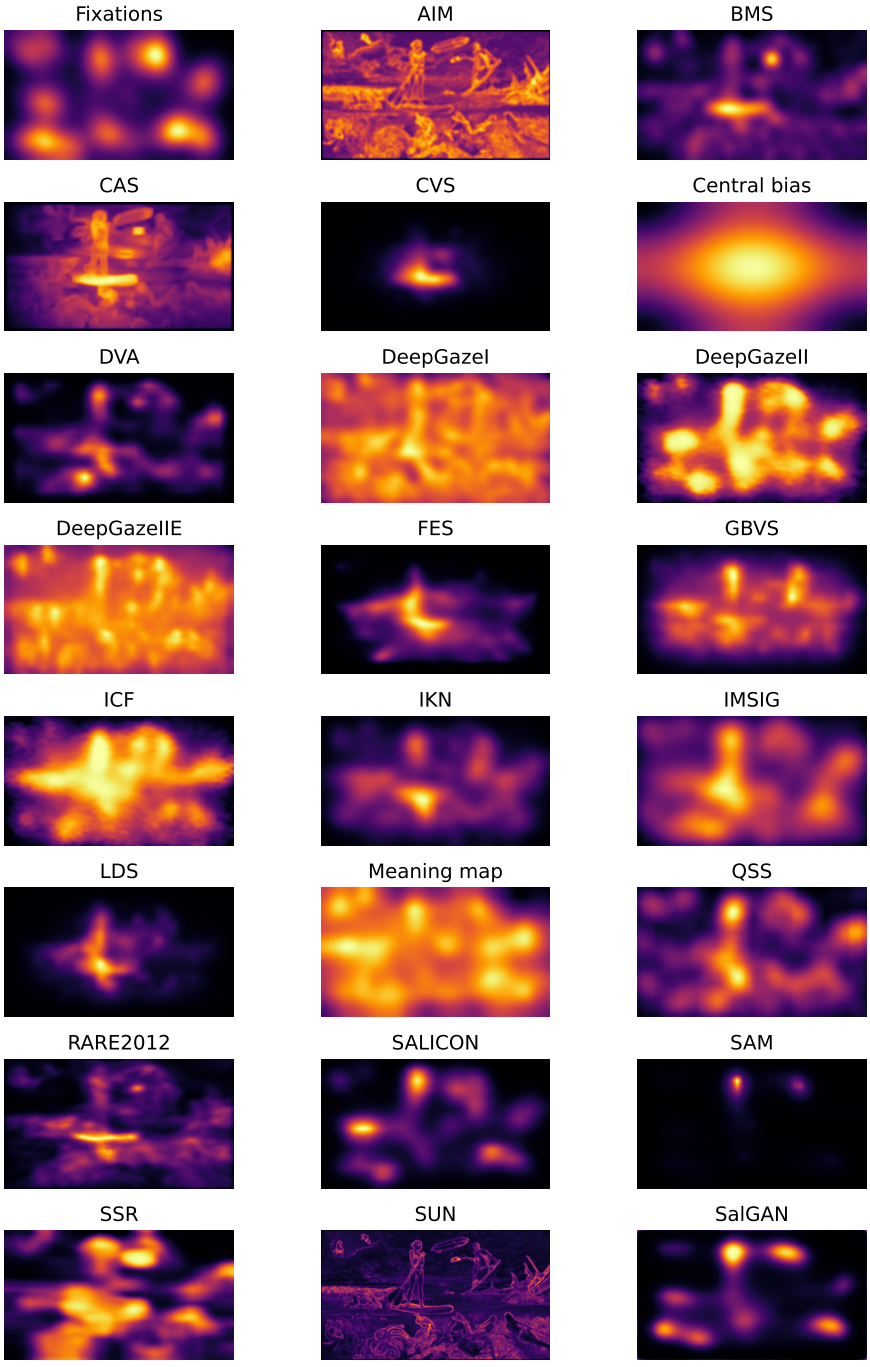
Data quality

Eye tracking data quality can be assessed by precision, accuracy, and data loss (Dunn et al., 2024). In the current study, eye tracking data quality was operationalized by precision and data loss only because the experimental protocol and set up did not allow to estimate accuracy. Precision was calculated by the sample-to-sample RMS deviation (s2s-RMSd), following Hooge et al. (2018). This method allows for estimating the precision without removing the saccades from the gaze signal. The s2s-RMSd was determined in a window of 200 ms that was slided through the gaze signal with steps of 32 milliseconds (two samples). For each participant, the median of the s2s-RMSd over all windows was calculated. The latter value was averaged over all participants.

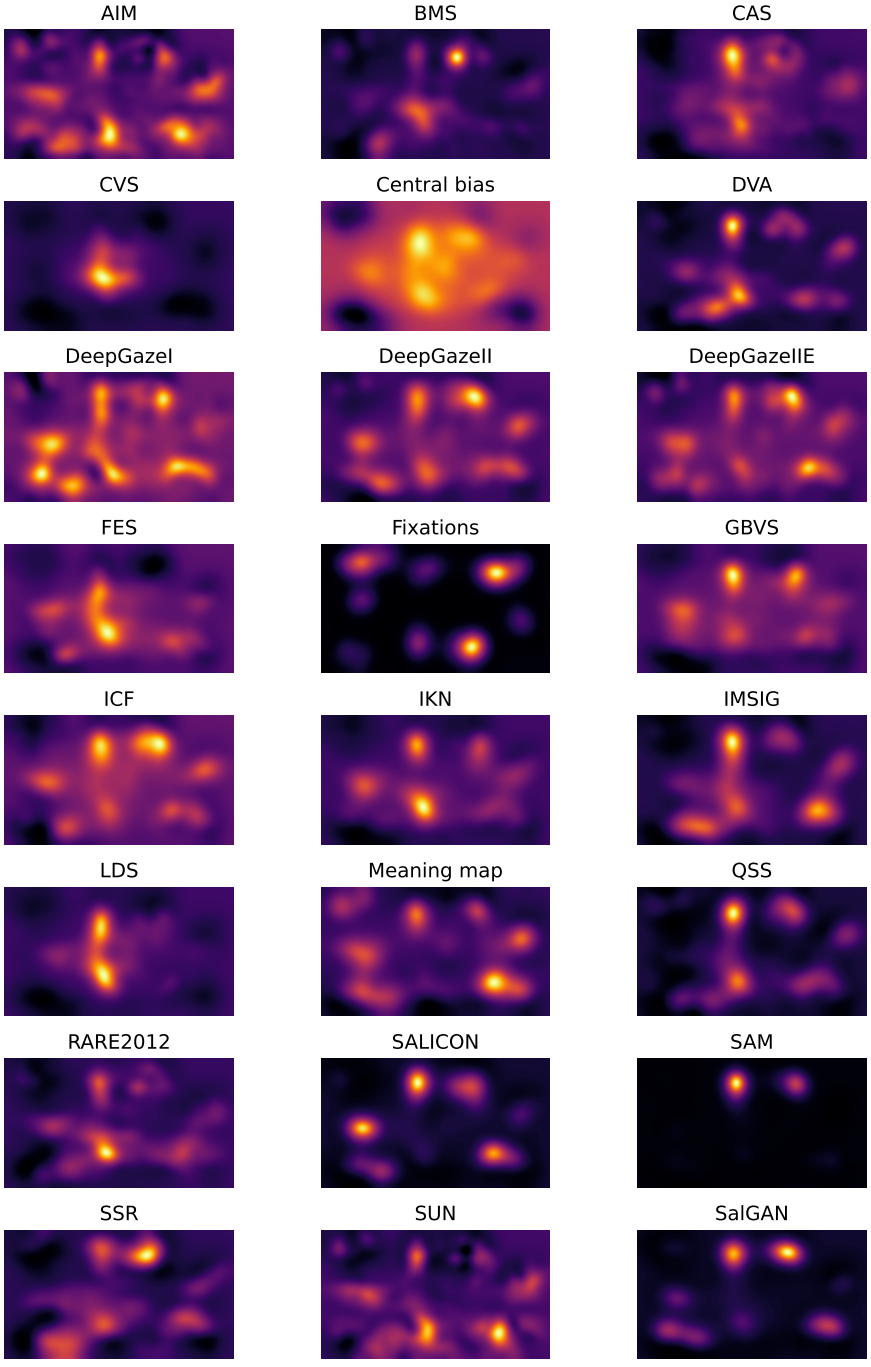
Median precision was 0.68° with a standard deviation of 0.28°. Data loss was *M* = 0.8%, *SD* = 2.4%. Women had an average data loss of 0.78% and men had an average data loss of 0.79%. Supplementary Figure 5 depicts data quality (precision in RMS and percentage data loss) across age bins and indicated gender. From the overall pattern of data quality indicators, data quality seems unlikely to have driven the main biases reported in the manuscript.



Supplementary Figure 5: Precision across age bins and gender in RMS (upper) and data loss across age bins and gender (middle row). Error bars indicate 95% confidence intervals. Scatter of precision against data loss across age bins (bottom). n = 1,600.



Supplementary Figure 6: All spatial distribution maps. Fixations, central bias, and meaning map are baselines. Other images depict predicted saliency maps per saliency model. Fixation map: n = 2,607.



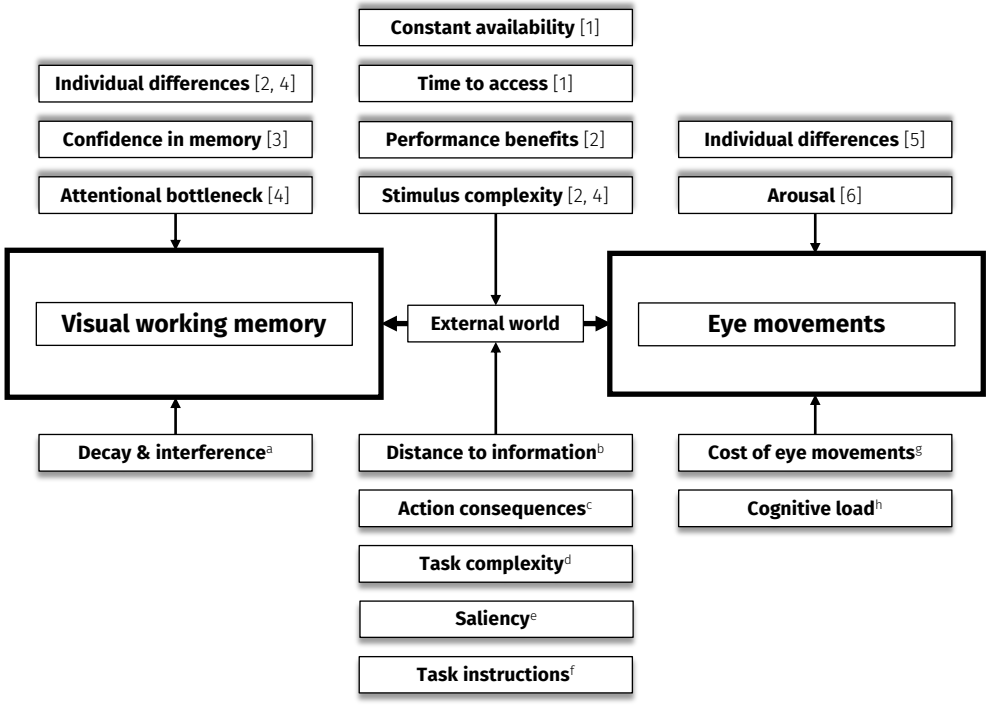
Supplementary Figure 7: Differences between spatial distribution maps of actual fixation locations and predicted saliency per model. 'Fixations' (third row, second column) represents discrete fixation locations against the smoothed fixation map, i.e., the upper bound. n = 2,607.

Chapter 6: Seeing the Forrest through the trees: Oculomotor metrics are linked to heart rate

List of statistical descriptors.

1. Mean
2. Variance
3. Skew
4. Kurtosis
5. Range
6. 10^{th} Percentile
7. 90^{th} Percentile
8. Interquartile Range
9. Absolute Mean Deviation
10. Energy
11. Root Mean Square
12. Entropy
13. Uniformity

General Discussion



Supplementary Figure 1: A more detailed (yet still incomplete) overview of how we make use of the (visual) external world. Superscripted letters link to non-exhaustive example references. [a]Camos et al., 2018; Souza et al., 2018. [b]Ballard et al., 1995; Draschkow et al., 2021; Inamdar and Pomplun, 2003. [c]Sahakian et al., 2024. [d]Kumle et al., 2024. [e]Itti et al., 1998. [f]Buswell, 1935; Mills et al., 2011; Yarbus, 1967. [g]Koevoet, Strauch, Naber, and Van der Stigchel, 2023; Koevoet et al., 2024. [h]Di Stasi et al., 2013; Koevoet et al., 2024; Pastukhov and Braun, 2010; Siegenthaler et al., 2014.

References

Abdi, H. (2010). The greenhouse-geisser correction. *Encyclopedia of research design*, *1*(1), 544–548.

Açık, A., Sarwary, A., Schultze-Kraft, R., Onat, S., & König, P. (2010). Developmental changes in natural viewing behavior: Bottom-up and top-down differences between children, young adults and older adults. *Frontiers in psychology*, *1*, 207. https://doi.org/10.3389/fpsyg.2010.00207

Adam, K. C., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology*, *97*, 79–97. https://doi.org/10.1016/j.cogpsych.2017.07.001

Alfandari, D., Belopolsky, A. V., & Olivers, C. N. L. (2019). Eye movements reveal learning and information-seeking in attentional template acquisition. *Visual Cognition*, *27*(5-8), 467–486. https://doi.org/10.1080/13506285.2019.1636918

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Anderson, J. R. (1996). A Simple Theory of Complex Cognition. *American Psychologist*, *51*(4), 355–365. https://doi.org/10.1037/0003-066X.51.4.355

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

Awh, E., Vogel, E. K., & Oh, S.-H. (2006). Interactions between attention and working memory. *Neuroscience*, *139*(1), 201–208. https://doi.org/10.1016/j.neuroscience.2005.08.023

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences*, *16*(8), 437–443. https://doi.org/10.1016/j.tics.2012.06.010

Azarbarzin, A., Ostrowski, M., Hanly, P., & Younes, M. (2014). Relationship between Arousal Intensity and Heart Rate Response to Arousal. *Sleep*, *37*(4), 645–653. https://doi.org/10.5665/sleep.3560

Baddeley, A. D., & Herring, S. R. (1983). Working memory. *Philosophical Transactions of the Royal Society of London. Biological Sciences*, *B 302*(1110), 311–324.

Baddeley, A. D., & Hitch, G. (1974). Working memory. The psychology of learning and motivation. *New York, NY: Academicp*.

Bae, G.-Y., & Luck, S. J. (2019a). Reactivation of Previous Experiences in a Working Memory Task. *Psychological Science*, *30*(4), 587–595. https://doi.org/10.1177/0956797619830398

Bae, G.-Y., & Luck, S. J. (2019b). What happens to an individual visual working memory representation when it is interrupted? *British Journal of Psychology*, *110*(2), 268–287. https://doi.org/10.1111/bjop.12339

Bahill, A. T., Clark, M. R., & Stark, L. (1975). The main sequence, a tool for studying human eye movements. *Mathematical Biosciences*, *24*(3), 191–204. https://doi.org/10.1016/0025-5564(75)90075-9

Bahle, B., Beck, V. M., & Hollingworth, A. (2018). The architecture of interaction between visual working memory and visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(7), 992–1011. https://doi.org/10.1037/xhp0000509

Bainbridge, W. A. (2019, January). Chapter One - Memorability: How what we see influences what we remember. In K. D. Federmeier & D. M. Beck (Eds.), *Psychology of Learning and Motivation* (pp. 1–27, Vol. 70). Academic Press. https://doi.org/10.1016/bs.plm.2019.02.001

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

Baloh, R. W., Sills, A. W., Kumley, W. E., & Honrubia, V. (1975). Quantitative measurement of saccade amplitude, duration, and velocity. *Neurology*, *25*(11), 1065–1065. https://doi.org/10.1212/WNL.25.11.1065

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Portch, E., Murray, E., & Dudfield, G. (2019). The consistency of superior face recognition skills in police officers. *Applied Cognitive Psychology*, *33*(5), 828–842. https://doi.org/10.1002/acp.3525

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bays, P. M. (2014). Noise in Neural Populations Accounts for Errors in Working Memory. *Journal of Neuroscience*, *34*(10), 3632–3645. https://doi.org/10.1523/JNEUROSCI.3204-13.2014

Bays, P. M., Gorgoraptis, N., Wee, N., Marshall, L., & Husain, M. (2011). Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision*, *11*(10), 6. https://doi.org/10.1167/11.10.6

Bays, P. M., & Husain, M. (2007). Spatial remapping of the visual world across saccades. *NeuroReport*, *18*(12), 1207. https://doi.org/10.1097/WNR.0b013e328244e6c3

Bays, P. M., & Husain, M. (2008). Dynamic Shifts of Limited Working Memory Resources in Human Vision. *Science*, *321*(5890), 851–854. https://doi.org/10.1126/science.1158023

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Beck, V. M., & Hollingworth, A. (2017). Competition in saccade target selection reveals attentional guidance by simultaneously active working memory representations. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(2), 225–230. https://doi.org/10.1037/xhp0000306
Beck, V. M., Hollingworth, A., & Luck, S. J. (2012). Simultaneous Control of Attention by Multiple Working Memory Representations. *Psychological Science*, *23*(8), 887–898. https://doi.org/10.1177/0956797612439068
Becker, S. I. (2011). Determinants of Dwell Time in Visual Search: Similarity or Perceptual Difficulty? *PLOS ONE*, *6*(3), e17740. https://doi.org/10.1371/journal.pone.0017740
Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 12–23. https://doi.org/10.1037/0096-1523.14.1.12
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), 3630. https://doi.org/10.3390/jcm12113630
Böing, S., Ten Brink, A. F., Ruis, C., Schielen, Z. A., Van den Berg, E., Biesbroek, J. M., Nijboer, T. C. W., & Van der Stigchel, S. (2025). Inspecting the external world: Memory capacity, but not memory self-efficacy, predicts offloading in working memory. *Journal of Clinical and Experimental Neuropsychology*, *0*(0), 1–23. https://doi.org/10.1080/13803395.2024.2447263
Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision*, *14*(3), 29–29. https://doi.org/10.1167/14.3.29
Borji, A., & Itti, L. (2015). Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*. https://doi.org/10.48550/arXiv.1505.03581
Borji, A., Sihite, D. N., & Itti, L. (2013). Objects do not predict fixations better than early saliency: A re-analysis of einhäuser et al.'s data. *Journal of vision*, *13*(10), 18–18. https://doi.org/10.1167/13.10.18
Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5), 4–4. https://doi.org/10.1167/11.5.4
Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*(1), 85–109. https://doi.org/10.1037/a0030779
Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. In *Variation in working memory* (pp. 76–106). Oxford University Press, USA.
Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The Balanced Accuracy and Its Posterior Distribution. *2010 20th International Conference on Pattern Recognition*, 3121–3124. https://doi.org/10.1109/ICPR.2010.764
Bruce, N., & Tsotsos, J. (2005). Saliency based on information maximization. *Advances in neural information processing systems*, *18*.
Buswell, G. T. (1935). How people look at pictures: A study of the psychology and perception in art.
Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (2015). Mit saliency benchmark.
Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2018). What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, *41*(3), 740–757. https://doi.org/10.1109/TPAMI.2018.2815601
Bylinskii, Z., Kim, N. W., O'Donovan, P., Alsheikh, S., Madan, S., Pfister, H., Durand, F., Russell, B., & Hertzmann, A. (2017). Learning Visual Importance for Graphic Designs and Data Visualizations. *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, 57–69. https://doi.org/10.1145/3126594.3126653
Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., & Durand, F. (2016). Where Should Saliency Models Look Next? In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 809–824). Springer International Publishing. https://doi.org/10.1007/978-3-319-46454-1_49
Cai, Y., Strauch, C., Van der Stigchel, S., & Naber, M. (2024). Open-DPSM: An open-source toolkit for modeling pupil size changes to dynamic visual inputs. *Behavior Research Methods*, *56*(6), 5605–5621. https://doi.org/10.3758/s13428-023-02292-1
Cain, M. S., Adamo, S. H., & Mitroff, S. R. (2013). A taxonomy of errors in multiple-target visual search. *Visual Cognition*, *21*(7), 899–921. https://doi.org/10.1080/13506285.2013.843627
Camos, V., Johnson, M., Loaiza, V., Portrat, S., Souza, A., & Vergauwe, E. (2018). What is attentional refreshing in working memory? *Annals of the New York Academy of Sciences*, *1424*(1), 19–32. https://doi.org/10.1111/nyas.13616
Carlisle, N. B., Arita, J. T., Pardo, D., & Woodman, G. F. (2011). Attentional templates in visual working memory. *Journal of Neuroscience*, *31*(25), 9315–9322. https://doi.org/10.1523/JNEUROSCI.1097-11.2011
Cerf, M., Harel, J., Huth, A., Einhäuser, W., & Koch, C. (2009). Decoding what people see from where they look: Predicting visual stimuli from scanpaths. *Attention in cognitive systems: 5th international workshop on attention in cognitive systems, WAPCV 2008 fira, santorini, greece, may 12, 2008 revised selected papers 5*, 15–26. https://doi.org/10.1007/978-3-642-00582-4_2
Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*, *74*, 221–232. https://doi.org/10.1016/j.apergo.2018.08.028
Cheon, B. K., Melani, I., & Hong, Y.-y. (2020). How USA-centric is psychology? An archival study of implicit assumptions of generalizability of findings to human nature based on origins of study samples. *Social Psychological and Personality Science*, *11*(7), 928–937. https://doi.org/10.1177/1948550620927269
Chun, M. M., & Wolfe, J. M. (1996). Just Say No: How Are Visual Searches Terminated When There Is No Target Present? *Cognitive Psychology*, *30*(1), 39–78. https://doi.org/10.1006/cogp.1996.0002
Clapp, W. C., Rubens, M. T., & Gazzaley, A. (2010). Mechanisms of Working Memory Disruption by External Interference. *Cerebral Cortex*, *20*(4), 859–872. https://doi.org/10.1093/cercor/bhp150
Constant, M., & Liesefeld, H. R. (2021). Massive Effects of Saliency on Information Processing in Visual Working Memory. *Psychological Science*, *32*(5), 682–691. https://doi.org/10.1177/0956797620975785

Corneil, B. D., Van Wanrooij, M., Munoz, D. P., & Van Opstal, A. J. (2002). Auditory-Visual Interactions Subserving Goal-Directed Saccades in a Complex Scene. *Journal of Neurophysiology*, *88*(1), 438–454. https://doi.org/10.1152/jn.2002.88.1.438
Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2018). Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, *27*(10), 5142–5154. https://doi.org/10.1109/TIP.2018.2851672
Coutrot, A., & Guyader, N. (2014). How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of vision*, *14*(8), 5–5. https://doi.org/10.1167/14.8.5
Cowan, N. (2014). Working Memory Underpins Cognitive Development, Learning, and Education. *Educational Psychology Review*, *26*(2), 197–223. https://doi.org/10.1007/s10648-013-9246-y
Cowan, N. (2016). *Working memory capacity: Classic edition*. Psychology press.
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
de Jong, J., van Rijn, H., & Akyürek, E. G. (2023). Adaptive Encoding Speed in Working Memory. *Psychological Science*, 09567976231173902. https://doi.org/10.1177/09567976231173902
de Lange, F. P., Heilbron, M., & Kok, P. (2018). How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002
De Haas, B., Iakovidis, A. L., Schwarzkopf, D. S., & Gegenfurtner, K. R. (2019). Individual differences in visual salience vary along semantic dimensions. *Proceedings of the National Academy of Sciences*, *116*(24), 11687–11692. https://doi.org/10.1073/pnas.1820553116
DeRosa, J., Kim, H., Lewis-Peacock, J., & Banich, M. T. (2024). Neural Systems Underlying the Implementation of Working Memory Removal Operations. *The Journal of Neuroscience*, *44*(2), e0283232023. https://doi.org/10.1523/JNEUROSCI.0283-23.2023
Desender, K., Boldt, A., & Yeung, N. (2018). Subjective Confidence Predicts Information Seeking in Decision Making. *Psychological Science*, *29*(5), 761–778. https://doi.org/10.1177/0956797617744771
Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*(Volume 18, 1995), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205
Di Stasi, L. L., Catena, A., Cañas, J. J., Macknik, S. L., & Martinez-Conde, S. (2013). Saccadic velocity as an arousal index in naturalistic tasks. *Neuroscience and Biobehavioral Reviews*, *37*(5), 968–975. https://doi.org/10.1016/j.neubiorev.2013.03.011
Di Stasi, L. L., Renner, R., Staehr, P., Helmert, J. R., Velichkovsky, B. M., Cañas, J. J., Catena, A., & Pannasch, S. (2010). Saccadic peak velocity sensitivity to variations in mental workload. *Aviation Space and Environmental Medicine*, *81*(4), 413–417. https://doi.org/10.3357/ASEM.2579.2010
Dodge, R. (1903). FIVE TYPES OF EYE MOVEMENT IN THE HORIZONTAL MERIDIAN PLANE OF THE FIELD OF REGARD. *American Journal of Physiology-Legacy Content*, *8*(4), 307–329. https://doi.org/10.1152/ajplegacy.1903.8.4.307
Dodge, R., & Cline, T. S. (1901). The angle velocity of eye movements. *Psychological Review*, *8*(2), 145–157. https://doi.org/10.1037/h0076100
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
Drew, T., Boettcher, S. E., & Wolfe, J. M. (2017). One visual search, many memory searches: An eye-tracking investigation of hybrid search. *Journal of Vision*, *17*(11), 1–10. https://doi.org/10.1167/17.11.5
Drew, T., & Wolfe, J. M. (2014). Hybrid search in the temporal domain: Evidence for rapid, serial logarithmic search through memory. *Attention, Perception, & Psychophysics*, *76*(2), 296–303. https://doi.org/10.3758/s13414-013-0606-y
Droll, J. A., & Hayhoe, M. M. (2007). 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
Droll, J. A., Hayhoe, M. M., Triesch, J., & Sullivan, B. T. (2005). Task Demands Control Acquisition and Storage of Visual Information. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 1416–1438. https://doi.org/10.1037/0096-1523.31.6.1416
Duchowski, A. T., Krejtz, K., Zurawska, J., & House, D. H. (2020). Using Microsaccades to Estimate Task Difficulty during Visual Search of Layered Surfaces. *IEEE Transactions on Visualization and Computer Graphics*, *26*(9), 2904–2918. https://doi.org/10.1109/TVCG.2019.2901881
Duhamel, J.-R., Colby, C. L., & Goldberg, M. E. (1992). The Updating of the Representation of Visual Space in Parietal Cortex by Intended Eye Movements. *Science*, *255*(5040), 90–92. https://doi.org/10.1126/science.1553535
Duncan, D. H., & Theeuwes, J. (2024). Secondary capture: Salience information persistently drives attentional selection. *Journal of Experimental Psychology: Human Perception and Performance*, *50*(9), 942–951. https://doi.org/10.1037/xhp0001223
Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*, 433–458. https://doi.org/10.1037/0033-295X.96.3.433
Dunn, M. J., Alexander, R. G., Amiebenomo, O. M., Arblaster, G., Atan, D., Erichsen, J. T., Ettinger, U., Giardini, M. E., Gilchrist, I. D., Hamilton, R., Hessels, R. S., Hodgins, S., Hooge, I. T. C., Jackson, B. S., Lee, H., Macknik, S. L., Martinez-Conde, S., Mcilreavy, L., Muratori, L. M., … Sprenger, A. (2024). Minimal reporting guideline for research involving eye tracking (2023 edition). *Behavior Research Methods*, *56*(5), 4351–4357. https://doi.org/10.3758/s13428-023-02187-1
Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot.
Ecker, U. K. H., Lewandowsky, S., & Oberauer, K. (2014). Removal of information from working memory: A specific updating process. *Journal of Memory and Language*, *74*, 77–90. https://doi.org/10.1016/j.jml.2013.09.003
Eckstein, M. K., Guerra-Carrillo, B., Miller Singley, A. T., & Bunge, S. A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, *25*, 69–91. https://doi.org/10.1016/j.dcn.2016.11.001

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of vision*, *8*(14), 18–18. https://doi.org/10.1167/8.14.18
Emery, L., Hale, S., & Myerson, J. (2008). Age differences in proactive interference, working memory, and abstract reasoning. *Psychology and Aging*, *23*(3), 634–645. https://doi.org/10.1037/a0012577
Eng, H. Y., Chen, D., & Jiang, Y. (2005). Visual working memory for simple and complex visual stimuli. *Psychonomic Bulletin & Review*, *12*(6), 1127–1133. https://doi.org/10.3758/BF03206454
Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*(9), 1035–1045. https://doi.org/10.1016/S0042-6989(03)00084-1
Erdem, E., & Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of vision*, *13*(4), 11–11. https://doi.org/10.1167/13.4.11
Fang, S., Li, J., Tian, Y., Huang, T., & Chen, X. (2016). Learning discriminative subspaces on random contrasts for image saliency analysis. *IEEE transactions on neural networks and learning systems*, *28*(5), 1095–1108. https://doi.org/10.1109/TNNLS.2016.2522440
Findlay, J. M. (1997). Saccade Target Selection During Visual Search. *Vision Research*, *37*(5), 617–631. https://doi.org/10.1016/S0042-6989(96)00218-0
Findlay, J. M., & Gilchrist, I. D. (2003, August). *Active Vision*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198524793.001.0001
Fougnie, D. (2008). The relationship between attention and working memory. *New research on short-term memory*, *1*.
Franchak, J. M., Heeger, D. J., Hasson, U., & Adolph, K. E. (2016). Free viewing gaze behavior in infants and adults. *Infancy : the official journal of the International Society on Infant Studies*, *21*(3), 262–287. https://doi.org/10.1111/infa.12119
Frătescu, M., Van Moorselaar, D., & Mathôt, S. (2019). Can you have multiple attentional templates? Large-scale replications of Van Moorselaar, Theeuwes, and Olivers (2014) and Hollingworth and Beck (2016). *Attention, Perception, & Psychophysics*, *81*(8), 2700–2709. https://doi.org/10.3758/s13414-019-01791-8
Gajewski, D., & Henderson, J. M. (2005). Minimal use of working memory in a scene comparison task. *Visual Cognition*, *12*(6), 979–1002. https://doi.org/10.1080/13506280444000616
Gayet, S., Battistoni, E., Thorat, S., & Peelen, M. V. (2024). Searching near and far: The attentional template incorporates viewing distance. *Journal of Experimental Psychology: Human Perception and Performance*, *50*(2), 216–231. https://doi.org/10.1037/xhp0001172
Gayet, S., Paffen, C. L. E., & Van der Stigchel, S. (2013). Information Matching the Content of Visual Working Memory Is Prioritized for Conscious Access. *Psychological Science*, *24*(12), 2472–2480. https://doi.org/10.1177/0956797613495882
Gayet, S., & Peelen, M. V. (2022). Preparatory attention incorporates contextual expectations. *Current Biology*, *32*(3), 687–692.e6. https://doi.org/10.1016/j.cub.2021.11.062
Gilchrist, I. D., & Harvey, M. (2000). Refixation frequency and memory mechanisms in visual search. *Current Biology*, *10*(19), 1209–1212. https://doi.org/10.1016/S0960-9822(00)00729-6
Godwin, H. J., Walenchok, S. C., Houpt, J. W., Hout, M. C., & Goldinger, S. D. (2015). Faster than the speed of rejection: Object identification processes during visual search for multiple targets. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(4), 1007–1020. https://doi.org/10.1037/xhp0000036
Goferman, S., Zelnik-Manor, L., & Tal, A. (2011). Context-aware saliency detection. *IEEE transactions on pattern analysis and machine intelligence*, *34*(10), 1915–1926. https://doi.org/10.1109/TPAMI.2011.272
Gold, J. M., Murray, R. F., Sekuler, A. B., Bennett, P. J., & Sekuler, R. (2005). Visual Memory Decay Is Deterministic. *Psychological Science*, *16*(10), 769–774. https://doi.org/10.1111/j.1467-9280.2005.01612.x
Gómez-Pérez, E., & Ostrosky-Solís, F. (2006). Attention and Memory Evaluation Across the Life Span: Heterogeneous Effects of Age and Education. *Journal of Clinical and Experimental Neuropsychology*, *28*(4), 477–494. https://doi.org/10.1080/13803390590949296
Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature*, *391*(6666), 481–484. https://doi.org/10.1038/35135
Gottlob, L. R., & Madden, D. J. (1999). Age differences in the strategic allocation of visual attention. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *54*(3), P165–P172. https://doi.org/10.1093/geronb/54B.3.P165
Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. S. (2013). MEG and EEG Data Analysis with MNE-Python. *Frontiers in Neuroscience*, *7*(267), 1–13. https://doi.org/10.3389/fnins.2013.00267
Grassi, G., Vailati, S., Bertinieri, G., Seravalle, G., Stella, M. L., Dell'Oro, R., & Mancia, G. (1998). Heart rate as marker of sympathetic activity. *Journal of Hypertension*, *16*(11), 1635–1639. https://doi.org/10.1097/00004872-199816110-00010
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. https://doi.org/10.1037/0033-295X.113.3.461
Greene, C. M., & Soto, D. (2012). Neural Repetition Effects in the Medial Temporal Lobe Complex are Modulated by Previous Encoding Experience. *PLOS ONE*, *7*(7), e40870. https://doi.org/10.1371/journal.pone.0040870
Gresch, D., Boettcher, S. E. P., van Ede, F., & Nobre, A. C. (2021). Shielding working-memory representations from temporally predictable external interference. *Cognition*, *217*, 104915. https://doi.org/10.1016/j.cognition.2021.104915
Grubert, A., Wang, Z., Williams, E., Jimenez, M., Remington, R., & Eimer, M. (2024, August). The capacity limitations of multiple-template visual search during task preparation and target selection. https://doi.org/10.22541/au.172449983.31891805/v1
Gunseli, E., Meeter, M., & Olivers, C. N. L. (2014). Is a search template an ordinary working memory? Comparing electrophysiological markers of working memory maintenance for visual search and recognition. *Neuropsychologia*, *60*, 29–38. https://doi.org/10.1016/j.neuropsychologia.2014.05.012

Hakim, N., Feldmann-Wüstefeld, T., Awh, E., & Vogel, E. K. (2020). Perturbing Neural Representations of Working Memory with Task-irrelevant Interruption. *Journal of Cognitive Neuroscience*, *32*(3), 558–569. https://doi.org/10.1162/jocn_a_01481
Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F. R., Nigbur, R., Waite, A. Q., Baumgartner, F., & Stadler, J. (2016). A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. *Scientific Data*, *3*(1), 1–15. https://doi.org/10.1038/sdata.2016.92
Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W., & Stadler, J. (2014). A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific Data*, *1*(1), 1–18. https://doi.org/10.1038/sdata.2014.3
Hansen, J. P., Mardanbegi, D., Biermann, F., & Bækgaard, P. (2018). A gaze interactive assembly instruction with pupillometric recording. *Behavior Research Methods*, *50*(4), 1723–1733. https://doi.org/10.3758/s13428-018-1074-z
Hardt, O., Nader, K., & Nadel, L. (2013). Decay happens: The role of active forgetting in memory. *Trends in Cognitive Sciences*, *17*(3), 111–120. https://doi.org/10.1016/j.tics.2013.01.001
Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in neural information processing systems*, *19*. https://doi.org/10.7551/mitpress/7503.003.0073
Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, *52*(100), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9
Hayes, T. R., & Petrov, A. A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods*, *48*(2), 510–527. https://doi.org/10.3758/s13428-015-0588-x
Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*(1), 6. https://doi.org/10.1167/3.1.6
Henderson, J. M., & Ferreira, F. (2013, May). *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. Psychology Press.
Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature human behaviour*, *1*(10), 743–747. https://doi.org/10.1038/s41562-017-0208-0
Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2021). Meaning maps capture the density of local semantic features in scenes: A reply to pedziwiatr, kümmerer, wallis, bethge & teufel (2021). *Cognition*, *214*, 104742. https://doi.org/10.1016/j.cognition.2021.104742
Henderson, J. M., & Hollingworth, A. (1998, January). Eye Movements During Scene Viewing. In *Eye Guidance in Reading and Scene Perception* (pp. 269–293). Elsevier. https://doi.org/10.1016/b978-008043361-5/50013-4
Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29. https://doi.org/10.1038/466029a
Hessels, R. S., Niehorster, D. C., Kemner, C., & Hooge, I. T. C. (2017). Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC). *Behavior Research Methods*, *49*(5), 1802–1823. https://doi.org/10.3758/s13428-016-0822-1
Hessels, R. S., Niehorster, D. C., Nyström, M., Andersson, R., & Hooge, I. T. (2018). Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. *Royal Society Open Science*, *5*(8). https://doi.org/10.1098/rsos.180502
Hessels, R. S., van Doorn, A. J., Benjamins, J. S., Holleman, G. A., & Hooge, I. T. (2020). Task-related gaze control in human crowd navigation. *Attention, Perception, and Psychophysics*, *82*(5), 2482–2501. https://doi.org/10.3758/s13414-019-01952-9
Hjortskov, N., Rissén, D., Blangsted, A. K., Fallentin, N., Lundberg, U., & Søgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology*, *92*(1), 84–89. https://doi.org/10.1007/s00421-004-1055-z
Hoar, R. M. (1982). Embryology of the eye. *Environmental Health Perspectives*. https://doi.org/10.1289/ehp.824431
Hommel, B., Li, K. Z. H., & Li, S.-C. (2004). Visual Search Across the Life Span. *Developmental Psychology*, *40*(4), 545–558. https://doi.org/10.1037/0012-1649.40.4.545
Hooge, I. T. C., & Erkelens, C. J. (1996). Control of fixation duration in a simple search task. *Perception & Psychophysics*, *58*(7), 969–976. https://doi.org/10.3758/BF03206825
Hooge, I. T. C., Niehorster, D. C., Nyström, M., Andersson, R., & Hessels, R. S. (2018). Is human classification by experienced untrained observers a gold standard in fixation detection? *Behavior Research Methods*, *50*(5), 1864–1881. https://doi.org/10.3758/s13428-017-0955-x
Hooge, I. T., Niehorster, D. C., Nyström, M., Andersson, R., & Hessels, R. S. (2022). Fixation classification: How to merge and select fixation candidates. *Behavior Research Methods*, (2001). https://doi.org/10.3758/s13428-021-01723-1
Hoogerbrugge, A. J., Sahakian, A., Brouwer, R., Klauss, G., Strauch, C., Nijboer, T., & Van der Stigchel, S. (2024, August). Three unpublished, publicly available visual search datasets with 151 participants. https://doi.org/10.31219/osf.io/kgpq4
Hoogerbrugge, A. J., Strauch, C., Böing, S., Nijboer, T. C. W., & Van der Stigchel, S. (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
Hoogerbrugge, A. J., Strauch, C., Nijboer, T. C. W., & Van der Stigchel, S. (2023). Don't hide the instruction manual: A dynamic trade-off between using internal and external templates during visual search. *Journal of Vision*, *23*(7), 14. https://doi.org/10.1167/jov.23.7.14
Hoogerbrugge, A. J., Strauch, C., Nijboer, T. C. W., & Van der Stigchel, S. (2024). Persistent resampling of external information despite 25 repetitions of the same visual search templates. *Attention, Perception, & Psychophysics*, *86*, 2301–2314. https://doi.org/10.3758/s13414-024-02953-z
Hoogerbrugge, A. J., Strauch, C., Oláh, Z. A., Dalmaijer, E. S., Nijboer, T. C. W., & Van der Stigchel, S. (2022). Seeing the forrest through the trees: Oculomotor metrics are linked to heart rate. *PLOS ONE*, *17*(8), e0272349. https://doi.org/10.1371/journal.pone.0272349

Hou, X., Harel, J., & Koch, C. (2011). Image signature: Highlighting sparse salient regions. *IEEE transactions on pattern analysis and machine intelligence*, *34*(1), 194–201. https://doi.org/10.1109/TPAMI.2011.146
Hou, X., & Zhang, L. (2008). Dynamic visual attention: Searching for coding length increments. *Advances in neural information processing systems*, *21*.
Hout, M. C., & Goldinger, S. D. (2010). Learning in repeated visual search. *Attention, Perception, & Psychophysics*, *72*(5), 1267–1282. https://doi.org/10.3758/APP.72.5.1267
Hout, M. C., & Goldinger, S. D. (2015). Target templates: The precision of mental representations affects attentional guidance and decision-making in visual search. *Attention, Perception, & Psychophysics*, *77*(1), 128–149. https://doi.org/10.3758/s13414-014-0764-6
Houtkamp, R., & Roelfsema, P. R. (2009). Matching of visual input to only one item at any one time. *Psychological Research PRPF*, *73*(3), 317–326. https://doi.org/10.1007/s00426-008-0157-3
Huang, L., & Awh, E. (2018). Chunking in working memory via content-free labels. *Scientific Reports*, *8*(1), 23. https://doi.org/10.1038/s41598-017-18157-5
Hulleman, J., & Olivers, C. N. L. (2017). The impending demise of the item in visual search. *Behavioral and Brain Sciences*, *40*, e132. https://doi.org/10.1017/S0140525X15002794
Inamdar, S., & Pomplun, M. (2003). Comparative search reveals the tradeoff between eye movements and working memory use in visual tasks. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *25*, 599–604.
Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? *CVPR 2011*, 145–152. https://doi.org/10.1109/CVPR.2011.5995721
Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489–1506. https://doi.org/10.1016/S0042-6989(99)00163-7
Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. https://doi.org/10.1109/34.730558
James, W. (1890). The principles of psychology volume II by william james (1890).
JASP Team. (2022). JASP (Version 0.16.3)[Computer software]. https://jasp-stats.org/
Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). Salicon: Saliency in context. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1072–1080. https://doi.org/10.1109/CVPR.2015.7298710
Jolly, E. (2018). Pymer4: Connecting R and Python for Linear Mixed Modeling. *Journal of Open Source Software*, *3*(31), 862. https://doi.org/10.21105/joss.00862
Jones, D. (2010). A WEIRD view of human nature skews psychologists' studies.
Jonides, J., & Nee, D. E. (2006). Brain mechanisms of proactive interference in working memory. *Neuroscience*, *139*(1), 181–193. https://doi.org/10.1016/j.neuroscience.2005.06.042
Jonides, J. (1983). Further toward a model of the Mind's eye's movement. *Bulletin of the Psychonomic Society*, *21*(4), 247–250. https://doi.org/10.3758/BF03334699
Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. *MIT technical report*. https://doi.org/10.5220/0005678701340142
Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. *2009 IEEE 12th international conference on computer vision*, 2106–2113. https://doi.org/10.1109/ICCV.2009.5459462
Kahneman, D. (1973). *Attention and effort*. Prentice-Hall, Inc.
Kazmi, S. Z. H., Zhang, H., Aziz, W., Monfredi, O., Abbas, S. A., Shah, S. A., Kazmi, S. S. H., & Butt, W. H. (2016). Inverse correlation between heart rate variability and heart rate demonstrated by linear and nonlinear analysis. *PLoS ONE*, *11*(6), e0157557. https://doi.org/10.1371/journal.pone.0157557
Kingsley, H. L. (1932). An experimental study of'search'. *The American Journal of Psychology*, *44*(2), 314–318. https://doi.org/10.2307/1414831
Koch, C., & Ullman, S. (1987). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. In *Matters of Intelligence* (pp. 115–141). Springer, Dordrecht. https://doi.org/10.1007/978-94-009-3833-5_5
Koevoet, D., Naber, M., Strauch, C., Somai, R. S., & Van der Stigchel, S. (2023). Differential aspects of attention predict the depth of visual working memory encoding: Evidence from pupillometry. *Journal of Vision*, *23*(6), 9. https://doi.org/10.1167/jov.23.6.9
Koevoet, D., Strauch, C., Naber, M., & Van der Stigchel, S. (2023). The Costs of Paying Overt and Covert Attention Assessed With Pupillometry. *Psychological Science*, 09567976231179378. https://doi.org/10.1177/09567976231179378
Koevoet, D., Strauch, C., Van der Stigchel, S., Mathôt, S., & Naber, M. (2023). Revealing visual working memory operations with pupillometry: Encoding, maintenance, and prioritization. *WIREs Cognitive Science*, *n/a*(n/a), e1668. https://doi.org/10.1002/wcs.1668
Koevoet, D., Zantwijk, L. V., Naber, M., Mathôt, S., Stigchel, S. V. d., & Strauch, C. (2024). Effort Drives Saccade Selection. *eLife*, *13*. https://doi.org/10.7554/eLife.97760.1
Kootstra, T., Teuwen, J., Goudsmit, J., Nijboer, T., Dodd, M., & Van der Stigchel, S. V. (2020). Machine learning-based classification of viewing behavior using a wide range of statistical oculomotor features. *Journal of Vision*, *20*(9), 1–15. https://doi.org/10.1167/jov.20.9.1
Kosch, T., Hassib, M., Woźniak, P. W., Buschek, D., & Alt, F. (2018). Your eyes tell: Leveraging smooth pursuit for assessing cognitive workload. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*. https://doi.org/10.1145/3173574.3174010
Krishna, O., & Aizawa, K. (2017). Age-adapted saliency model with depth bias. *Proceedings of the ACM symposium on applied perception*, 1–8. https://doi.org/10.1145/3119881.3119885
Krishna, O., Helo, A., Rämä, P., & Aizawa, K. (2018). Gaze distribution analysis and saliency prediction across age groups. *PloS one*, *13*(2), e0193149. https://doi.org/10.1371/journal.pone.0193149
Kristjánsson, Á., Jóhannesson, Ó. I., & Thornton, I. M. (2014). Common Attentional Constraints in Visual Foraging. *PLOS ONE*, *9*(6), e100752. https://doi.org/10.1371/journal.pone.0100752

Kristjánsson, T., Thornton, I. M., & Kristjánsson, Á. (2018). Time limits during visual foraging reveal flexible working memory templates. *Journal of Experimental Psychology: Human Perception and Performance*, *44*, 827–835. https://doi.org/10.1037/xhp0000517
Kumle, L., Võ, M. L.-H., Nobre, A. C., & Draschkow, D. (2024). Multifaceted consequences of visual distraction during natural behaviour. *Communications Psychology*, *2*(1), 1–13. https://doi.org/10.1038/s44271-024-00099-0
Kümmerer, M., Theis, L., & Bethge, M. (2014). Deep gaze I: Boosting saliency prediction with feature maps trained on ImageNet. *International conference on learning representations (ICLR 2015)*, 1–12. https://doi.org/10.48550/arXiv.1411.1045
Kümmerer, M., Bethge, M., & Wallis, T. S. A. (2022). DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, *22*(5), 7. https://doi.org/10.1167/jov.22.5.7
Kümmerer, M., Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (2022). MIT/tübingen saliency benchmark. https://saliency.tuebingen.ai/
Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2018). Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (pp. 798–814, Vol. 11220). Springer International Publishing. https://doi.org/10.1007/978-3-030-01270-0_47
Kümmerer, M., Wallis, T. S., Gatys, L. A., & Bethge, M. (2017). Understanding low-and high-level contributions to fixation prediction. *Proceedings of the IEEE international conference on computer vision*, 4789–4798. https://doi.org/10.1109/ICCV.2017.513
Kümmerer, M., Wallis, T., & Bethge, M. (2017). Deepgaze ii: Predicting fixations from deep features over time and tasks. *Journal of Vision*, *17*(10), 1147–1147. https://doi.org/10.1167/17.10.1147
Latour, P. L. (1962). Visual Threshold During Eye Movements. *Vision research*, *2*(3), 261–262. https://doi.org/10.1016/0042-6989(62)90031-7
Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior research methods*, *45*(1), 251–266. https://doi.org/10.3758/s13428-012-0226-9
Le-Hoa Võ, M., & Wolfe, J. M. (2015). The role of memory for visual search in scenes. *Annals of the New York Academy of Sciences*, *1339*(1), 72–81. https://doi.org/10.1111/nyas.12667
Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Neural Evidence for a Distinction between Short-term Memory and the Focus of Attention. *Journal of Cognitive Neuroscience*, *24*(1), 61–79. https://doi.org/10.1162/jocn_a_00140
Lewis-Peacock, J. A., Kessler, Y., & Oberauer, K. (2018). The removal of information from working memory. *Annals of the New York Academy of Sciences*, *1424*(1), 33–44. https://doi.org/10.1111/nyas.13714
Li, A., Chen, Z., Wolfe, J. M., & Olivers, C. N. L. (2023). How do people find pairs? *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0001390
Liesefeld, H. R., Lamy, D., Gaspelin, N., Geng, J. J., Kerzel, D., Schall, J. D., Allen, H. A., Anderson, B. A., Boettcher, S., Busch, N. A., Carlisle, N. B., Colonius, H., Draschkow, D., Egeth, H., Leber, A. B., Müller, H. J., Röer, J. P., Schubö, A., Slagter, H. A., … Wolfe, J. (2024). Terms of debate: Consensus definitions to guide the scientific discourse on visual distraction. *Attention, Perception, & Psychophysics*. https://doi.org/10.3758/s13414-023-02820-3
Lin, Y.-t., & Leber, A. B. (2024). Individual variation in encoding strategy optimization in visual working memory: Evidence from a change detection task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, No Pagination Specified–No Pagination Specified. https://doi.org/10.1037/xlm0001398
Linardos, A., Kümmerer, M., Press, O., & Bethge, M. (2021). DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. *Proceedings of the IEEE/CVF international conference on computer vision*, 12919–12928. https://doi.org/10.48550/arXiv.2105.12441
Liversedge, S., Gilchrist, I., & Everling, S. (2011, August). *The Oxford Handbook of Eye Movements*. OUP Oxford.
Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in cognitive sciences*, *17*(8), 391–400. https://doi.org/10.1016/j.tics.2013.06.006
Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience & Biobehavioral Reviews*, *62*, 100–108. https://doi.org/10.1016/j.neubiorev.2016.01.003
Lustig, C., & Jantz, T. (2015). Questions of age differences in interference control: When and how, not if? *Brain Research*, *1612*, 59–69. https://doi.org/10.1016/j.brainres.2014.10.024
Ma, W. J., Husain, M., Bays, P. M., & de Soissons, P. (2014). Changing concepts of working memory. *Nature neuroscience*, *17*(3), 347. https://doi.org/10.1016/S0031-9406(10)63634-6
Macaluso, E., Frith, C. D., & Driver, J. (2000). Modulation of Human Visual Cortex by Crossmodal Spatial Attention. *Science*, *289*(5482), 1206–1208. https://doi.org/10.1126/science.289.5482.1206
Maffei, A., & Angrilli, A. (2019). Spontaneous blink rate as an index of attention and emotion during film clips viewing. *Physiology and Behavior*, *204*, 256–263. https://doi.org/10.1016/j.physbeh.2019.02.037
Maniglia, M. R., & Souza, A. S. (2020). Age Differences in the Efficiency of Filtering and Ignoring Distraction in Visual Working Memory. *Brain Sciences*, *10*(8), 556. https://doi.org/10.3390/brainsci10080556
Master, S. L., Li, S., & Curtis, C. E. (2023, December). Trying harder: How cognitive effort sculpts neural representations during working memory. https://doi.org/10.1101/2023.12.07.570686
Mather, M., Joo Yoo, H., Clewett, D. V., Lee, T. H., Greening, S. G., Ponzio, A., Min, J., & Thayer, J. F. (2017). Higher locus coeruleus MRI contrast is associated with lower parasympathetic influence over heart rate variability. *NeuroImage*, *150*, 329–335. https://doi.org/10.1016/j.neuroimage.2017.02.025
Mayr, U., & Kliegl, R. (2000). Task-set switching and long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1124–1140. https://doi.org/10.1037/0278-7393.26.5.1124
McGinnis, D. (2012). Susceptibility to Distraction During Reading in Young, Young-Old, and Old-Old Adults. *Experimental Aging Research*, *38*(4), 370–393. https://doi.org/10.1080/0361073X.2012.699365
Meghanathan, R. N., Nikolaev, A. R., & van Leeuwen, C. (2019). Refixation patterns reveal memory-encoding strategies in free viewing. *Attention, Perception, & Psychophysics*, *81*(7), 2499–2516. https://doi.org/10.3758/s13414-019-01735-2
Meghanathan, R. N., van Leeuwen, C., & Nikolaev, A. R. (2015). Fixation duration surpasses pupil size as a measure of memory load in free viewing. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.01063

Melcher, D. (2007). Predictive remapping of visual features precedes saccadic eye movements. *Nature Neuroscience*, *10*(7), 903–907. https://doi.org/10.1038/nn1917
Melcher, D., & Colby, C. L. (2008). Trans-saccadic perception. *Trends in Cognitive Sciences*, *12*(12), 466–473. https://doi.org/10.1016/j.tics.2008.09.003
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
Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. *Neuron*, *100*(2), 463–475. https://doi.org/10.1016/j.neuron.2018.09.023
Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of vision*, *11*(8), 17. https://doi.org/10.1167/11.8.17
Mitchell, T. V., & Neville, H. J. (2004). Asynchronies in the development of electrophysiological responses to motion and color. *Journal of Cognitive Neuroscience*, *16*(8), 1363–1374. https://doi.org/10.1162/0898929042304750
Moore, C. M., & Wolfe, J. M. (2001). Getting beyond the serial/parallel debate in visual research: A hybrid approach. In *The limits of attention: Temporal constraints in human information processing* (pp. 178–198). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198505150.003.0009
Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*(2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061
Muhammed, K., Dalmaijer, E., Manohar, S., & Husain, M. (2020). Voluntary modulation of saccadic peak velocity associated with individual differences in motivation. *Cortex*, *122*, 198–212. https://doi.org/10.1016/j.cortex.2018.12.001
Nakano, T., & Kuriyama, C. (2017). Transient heart rate acceleration in association with spontaneous eyeblinks. *International Journal of Psychophysiology*, *121*, 56–62. https://doi.org/10.1016/j.ijpsycho.2017.09.003
Neisser, U., & Becklen, R. (1975). Selective looking: Attending to visually specified events. *Cognitive Psychology*, *7*(4), 480–494. https://doi.org/10.1016/0010-0285(75)90019-5
Ngiam, W. X. Q. (2023). Mapping visual working memory models to a theoretical framework. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-023-02356-5
Nieuwenhuis, S., & Monsell, S. (2002). Residual costs in task switching: Testing the failure-to-engage hypothesis. *Psychonomic Bulletin & Review*, *9*(1), 86–92. https://doi.org/10.3758/BF03196259
Oberauer, K. (2001). Removing irrelevant information from working memory: A cognitive aging study with the modified Sternberg task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(4), 948–957. https://doi.org/10.1037/0278-7393.27.4.948
Oberauer, K. (2018). Removal of irrelevant information from working memory: Sometimes fast, sometimes slow, and sometimes not at all. *Annals of the New York Academy of Sciences*, *1424*(1), 239–255. https://doi.org/10.1111/nyas.13603
Oberauer, K. (2019). Working Memory and Attention – A Conceptual Analysis and Review. *Journal of Cognition*, *2*(1). https://doi.org/10.5334/joc.58
Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958. https://doi.org/10.1037/bul0000153
Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21–59. https://doi.org/10.1037/rev0000044
Ohl, S., Wohltat, C., Kliegl, R., Pollatos, O., & Engbert, R. (2016). Microsaccades are coupled to heartbeat. *Journal of Neuroscience*, *36*(4), 1237–1241. https://doi.org/10.1523/JNEUROSCI.2211-15.2016
Olivers, C. N. L., Meijer, F., & Theeuwes, J. (2006). Feature-based memory-driven attentional capture: Visual working memory content affects visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1243–1265. https://doi.org/10.1037/0096-1523.32.5.1243
Olivers, C. N. L., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences*, *15*(7), 327–334. https://doi.org/10.1016/j.tics.2011.05.004
Olivers, C. N. L., & Roelfsema, P. R. (2020). Attention for action in visual working memory. *Cortex*, *131*, 179–194. https://doi.org/10.1016/j.cortex.2020.07.011
Olivers, C. N., & Eimer, M. (2011). On the difference between working memory and attentional set. *Neuropsychologia*, *49*(6), 1553–1558. https://doi.org/10.1016/j.neuropsychologia.2010.11.033
O'Regan, J. K. (1992). Solving the" real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *46*(3), 461. https://doi.org/10.1037/h0084327
Ort, E., Fahrenfort, J. J., & Olivers, C. N. L. (2017). Lack of Free Choice Reveals the Cost of Having to Search for More Than One Object. *Psychological Science*, *28*(8), 1137–1147. https://doi.org/10.1177/0956797617705667
Ort, E., Fahrenfort, J. J., Ten Cate, T., Eimer, M., & Olivers, C. N. L. (2019). Humans can efficiently look for but not select multiple visual objects. *eLife*, *8*. https://doi.org/10.7554/eLife.49130
Ort, E., & Olivers, C. N. L. (2020). The capacity of multiple-target search. *Visual Cognition*, *28*(5-8), 330–355. https://doi.org/10.1080/13506285.2020.1772430
Ossandón, J. P., Onat, S., & König, P. (2014). Spatial biases in viewing behavior. *Journal of Vision*, *14*(2), 20. https://doi.org/10.1167/14.2.20
Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. *Eye Tracking Research and Applications Symposium (ETRA)*, 141–144. https://doi.org/10.1145/1743666.1743701
Palmer, E. M., Van Wert, M. J., Horowitz, T. S., & Wolfe, J. M. (2019). Measuring the time course of selection during visual search. *Attention, Perception, & Psychophysics*, *81*(1), 47–60. https://doi.org/10.3758/s13414-018-1596-6

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research*, *40*(10), 1227–1268. https://doi.org/10.1016/S0042-6989(99)00244-8
Pan, J., Ferrer, C. C., McGuinness, K., O'Connor, N. E., Torres, J., Sayrol, E., & Giro-i-Nieto, X. (2017). Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*. https://doi.org/10.48550/arXiv.1701.01081
Pannasch, S., Helmert, J. R., Roth, K., Herbold, A.-K., & Walter, H. (2008). Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. *Journal of Eye Movement Research*, *2*(2). https://doi.org/10.16910/jemr.2.2.4
Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, *14*(2), 187–193. https://doi.org/10.3758/BF03194050
Pastukhov, A., & Braun, J. (2010). Rare but precious: Microsaccades are highly informative about attentional allocation. *Vision Research*, *50*(12), 1173–1184. https://doi.org/10.1016/j.visres.2010.04.007
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825–2830.
Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S., Bethge, M., & Teufel, C. (2021). There is no evidence that meaning maps capture semantic information relevant to gaze guidance: Reply to henderson, hayes, peacock, and rehrig (2021). *Cognition*, *214*, 104741. https://doi.org/10.1016/j.cognition.2021.104741
Posner, M. I. (1980). Orienting of Attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25. https://doi.org/10.1080/00335558008248231
Qing, T., Strauch, C., Van Maanen, L., & Van der Stigchel, S. (2024). Shifting reliance between the internal and external world: A meta-analysis on visual-working memory use. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-024-02623-z
Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115
Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To See or not to See: The Need for Attention to Perceive Changes in Scenes. *Psychological Science*, *8*(5), 368–373. https://doi.org/10.1111/j.1467-9280.1997.tb00427.x
Rezazadegan Tavakoli, H., Rahtu, E., & Heikkilä, J. (2011). Fast and efficient saliency detection using sparse sampling and kernel density estimation. *Scandinavian conference on image analysis*, 666–675. https://doi.org/10.1007/978-3-642-21227-7_62
Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition*, *76*(3), 269–295. https://doi.org/10.1016/S0010-0277(00)00084-6
Riche, N., Duvinage, M., Mancas, M., Gosselin, B., & Dutoit, T. (2013). Saliency and Human Fixations: State-of-the-Art and Study of Comparison Metrics, 1153–1160. Retrieved February 10, 2025, from https://openaccess.thecvf.com/content_iccv_2013/html/Riche_Saliency_and_Human_2013_ICCV_paper.html
Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013). Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, *28*(6), 642–658. https://doi.org/10.1016/j.image.2013.03.009
Rider, A. T., Coutrot, A., Pellicano, E., Dakin, S. C., & Mareschal, I. (2018). Semantic content outweighs low-level saliency in determining children's and adults' fixation of movies. *Journal of Experimental Child Psychology*, *166*, 293–309. https://doi.org/10.1016/j.jecp.2017.09.002
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
Rizzolatti, G., Riggio, L., Dascola, I., & Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, *25*(1, Part 1), 31–40. https://doi.org/10.1016/0028-3932(87)90041-8
Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face Recognition by Metropolitan Police Super-Recognisers. *PLOS ONE*, *11*(2), e0150036. https://doi.org/10.1371/journal.pone.0150036
Robinson, D. A. (1965). The mechanics of human smooth pursuit eye movement. *The Journal of Physiology*, *180*(3), 569–591. Retrieved November 22, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1357404/
Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*, 207–231. https://doi.org/10.1037/0096-3445.124.2.207
Rolfs, M. (2009). Microsaccades: Small steps on a long way. *Vision Research*, *49*(20), 2415–2441. https://doi.org/10.1016/j.visres.2009.08.010
Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 763–797. https://doi.org/10.1037/0096-1523.27.4.763
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
Sahakian, A., Gayet, S., Paffen, C. L. E., & Van der Stigchel, S. (2024). Action consequences guide the use of visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, No Pagination Specified–No Pagination Specified. https://doi.org/10.1037/xlm0001326
Salway, A. F. S., & Logie, R. H. (1995). Visuospatial working memory, movement control and executive demands. *British Journal of Psychology*, *86*(2), 253–269. https://doi.org/10.1111/j.2044-8295.1995.tb02560.x
Sauter, M., Stefani, M., & Mack, W. (2020). Towards Interactive Search: Investigating Visual Search in a Novel Real-World Paradigm. *Brain Sciences*, *10*(12), 927. https://doi.org/10.3390/brainsci10120927

Schauerte, B., & Stiefelhagen, R. (2012). Quaternion-based spectral saliency detection for eye fixation prediction. *European conference on computer vision*, 116–129. https://doi.org/10.1007/978-3-642-33709-3_9
Schiller, P. H. (1998). The Neural Control of Visually Guided Eye Movements. In *Cognitive Neuroscience of Attention*. Psychology Press.
Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*(11), 1156–1172. https://doi.org/10.1038/s41562-020-00938-0
Schütz, A. C., Braun, D. I., & Gegenfurtner, K. R. (2011). Eye movements and perception: A selective review. *Journal of Vision*, *11*(5), 9. https://doi.org/10.1167/11.5.9
Seo, H. J., & Milanfar, P. (2009). Nonparametric bottom-up saliency detection by self-resemblance. *2009 IEEE computer society conference on computer vision and pattern recognition workshops*, 45–52. https://doi.org/10.1109/CVPRW.2009.5204207
Shan, J., & Postle, B. R. (2022). The Influence of Active Removal from Working Memory on Serial Dependence. *5*(1), 31. https://doi.org/10.5334/joc.222
Siegenthaler, E., Costela, F. M., Mccamy, M. B., Di Stasi, L. L., Otero-Millan, J., Sonderegger, A., Groner, R., Macknik, S., & Martinez-Conde, S. (2014). Task difficulty in mental arithmetic affects microsaccadic rates and magnitudes. *European Journal of Neuroscience*, *39*(2), 287–294. https://doi.org/10.1111/ejn.12395
Simons, D. J. (1996). In Sight, Out of Mind: When Object Representations Fail. *Psychological Science*, *7*(5), 301–305. https://doi.org/10.1111/j.1467-9280.1996.tb00378.x
Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*(7), 261–267. https://doi.org/10.1016/S1364-6613(97)01080-2
Smith, N. D., Crabb, D. P., & Garway-Heath, D. F. (2011). An exploratory study of visual search performance in glaucoma: Study of visual search performance in glaucoma. *Ophthalmic and Physiological Optics*, *31*(3), 225–232. https://doi.org/10.1111/j.1475-1313.2011.00836.x
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. https://doi.org/10.1016/j.cortex.2018.12.017
Souza, A. S., Rerko, L., & Oberauer, K. (2015). Refreshing memory traces: Thinking of an item improves retrieval from visual working memory. *Annals of the New York Academy of Sciences*, *1339*(1), 20–31. https://doi.org/10.1111/nyas.12603
Souza, A. S., Vergauwe, E., & Oberauer, K. (2018). Where to attend next: Guiding refreshing of visual, spatial, and verbal representations in working memory. *Annals of the New York Academy of Sciences*, *1424*(1), 76–90. https://doi.org/10.1111/nyas.13621
Stokes, D., & Biggs, S. (2014). The Dominance of the Visual. In D. Stokes, M. Matthen, & S. Biggs (Eds.), *Perception and Its Modalities*. Oxford University Press. Retrieved December 27, 2024, from https://philarchive.org/rec/STOTDO-4
Strauch, C., Hirzle, T., Van der Stigchel, S., & Bulling, A. (2021). Decoding binary decisions under differential target probabilities from pupil dilation: A random forest approach. *Journal of Vision*, *21*(7), 1–13. https://doi.org/10.1167/jov.21.7.6
Strauch, C., Hoogerbrugge, A. J., Baer, G., Hooge, I. T. C., Nijboer, T. C. W., Stuit, S. M., & Van der Stigchel, S. (2023). Saliency models perform best for women's and young adults' fixations. *Communications Psychology*, *1*(1), 1–10. https://doi.org/10.1038/s44271-023-00035-8
Strauch, C., Hoogerbrugge, A. J., & Ten Brink, A. F. (2024). Gaze data of 4243 participants shows link between leftward and superior attention biases and age. *Experimental Brain Research*. https://doi.org/10.1007/s00221-024-06823-w
Strauch, C., Wang, C.-A., Einhäuser, W., Van der Stigchel, S., & Naber, M. (2022). Pupillometry as an integrated readout of distinct attentional networks. *Trends in Neurosciences*. https://doi.org/10.1016/j.tins.2022.05.003
Sullivan, B., Ludwig, C. J. H., Damen, D., Mayol-Cuevas, W., & Gilchrist, I. D. (2021). Look-ahead fixations during visuomotor behavior: Evidence from assembling a camping tent. *Journal of Vision*, *21*(3), 13. https://doi.org/10.1167/jov.21.3.13
Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of vision*, *7*(14), 4–4. https://doi.org/10.1167/7.14.4
Tatler, B. W. (2009). Current understanding of eye guidance. *Visual Cognition*, *17*(6-7), 777–789. https://doi.org/10.1080/13506280902869213
Tatler, B. W., Baddeley, R. J., & Vincent, B. T. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, *46*(12), 1857–1862. https://doi.org/10.1016/j.visres.2005.12.005
Teigen, K. H. (1994). Yerkes-Dodson: A Law for all Seasons. *Theory & Psychology*, *4*(4), 525–547. https://doi.org/10.1177/0959354394044004
Theeuwes, J. (2012). Automatic Control of Visual Selection. In M. D. Dodd & J. H. Flowers (Eds.), *The Influence of Attention, Learning, and Motivation on Visual Search* (pp. 23–62). Springer. https://doi.org/10.1007/978-1-4614-4794-8_3
Theeuwes, J. (2024). Attentional Capture and Control. https://doi.org/10.1146/annurev-psych-011624-025340
Theeuwes, J., Kramer, A. F., Hahn, S., & Irwin, D. E. (1998). Our Eyes do Not Always Go Where we Want Them to Go: Capture of the Eyes by New Objects. *Psychological Science*, *9*(5), 379–385. https://doi.org/10.1111/1467-9280.00071
Titchener, E. B. (1924). The Overlooking of Familiar Objects. *The American Journal of Psychology*, *35*(2), 304–305. https://doi.org/10.2307/1413844
Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. *Journal of Vision*, *3*(1), 9. https://doi.org/10.1167/3.1.9
Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(6), 607–617. https://doi.org/10.1037/0278-7393.5.6.607
van Lieshout, L. L., de Lange, F. P., & Cools, R. (2020). Why so curious? Quantifying mechanisms of information seeking. *Current Opinion in Behavioral Sciences*, *35*, 112–117. https://doi.org/10.1016/j.cobeha.2020.08.005
van Zoest, W., Van der Stigchel, S., & Donk, M. (2017). Conditional control in visual selection. *Attention, Perception, & Psychophysics*, *79*(6), 1555–1572. https://doi.org/10.3758/s13414-017-1352-3
van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785. https://doi.org/10.1073/pnas.1117465109

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
Van Gent, P., Farah, H., Van Nes, N., & Van Arem, B. (2018). Heart Rate Analysis for Human Factors: Development and Validation of an Open Source Toolkit for Noisy Naturalistic Heart Rate Data Reducing congestion at sags View project From Individual Automated Vehicles to Cooperative Traffic Management-Predicting the. *Proceedings of The 6th HUMMANIST Conference*, *13*(June), 13–14. http://resolver.tudelft.nl/uuid:5c638e14-d249-4116-aa05-2e566cf3df02
Van Moorselaar, D., & Theeuwes, J. (2024). Transfer of statistical learning between tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *50*(7), 740–751. https://doi.org/10.1037/xhp0001216
Van Moorselaar, D., Theeuwes, J., & Olivers, C. N. L. (2014). In competition for the attentional template: Can multiple items within visual working memory guide attention? *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1450–1464. https://doi.org/10.1037/a0036229
Vanyukov, P. M., Warren, T., Wheeler, M. E., & Reichle, E. D. (2012). The emergence of frequency effects in eye movements. *Cognition*, *123*(1), 185–189. https://doi.org/10.1016/j.cognition.2011.12.011
Vilotijević, A., & Mathôt, S. (2024). Functional benefits of cognitively driven pupil-size changes. *WIREs Cognitive Science*, *15*(3), e1672. https://doi.org/10.1002/wcs.1672
Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). {SciPy} 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2
Vogel, E. K., & Awh, E. (2008). How to Exploit Diversity for Scientific Gain: Using Individual Differences to Constrain Cognitive Theory. *Current Directions in Psychological Science*, *17*(2), 171–176. https://doi.org/10.1111/j.1467-8721.2008.00569.x
Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), 748–751. https://doi.org/10.1038/nature02447
Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(6), 1436–1451. https://doi.org/10.1037/0096-1523.32.6.1436
Wade, N. (2020). Looking at Buswell's pictures. *Journal of Eye Movement Research*, *13*. https://doi.org/10.16910/jemr.13.2.4
Wetzel, N., & Schröger, E. (2007). Cognitive control of involuntary attention and distraction in children and adolescents. *Brain Research*, *1155*, 134–146. https://doi.org/10.1016/j.brainres.2007.04.022
Williams, J. R., Brady, T. F., & Störmer, V. S. (2022). Guidance of attention by working memory is a matter of representational fidelity. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(3), 202–231. https://doi.org/10.1037/xhp0000985
Williams, R. S., Ferber, S., & Pratt, J. (2023). The specificity of feature-based attentional guidance is equivalent under single- and dual-target search. *Journal of Experimental Psychology: Human Perception and Performance*, *49*(11), 1430–1446. https://doi.org/10.1037/xhp0001157
Williamson, A., Lombardi, D. A., Folkard, S., Stutts, J., Courtney, T. K., & Connor, J. L. (2011). The link between fatigue and safety. *Accident Analysis and Prevention*, *43*(2), 498–515. https://doi.org/10.1016/j.aap.2009.11.011
Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, *9*(4), 625–636. https://doi.org/10.3758/BF03196322
Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238. https://doi.org/10.3758/BF03200774
Wolfe, J. M. (1998). What Can 1 Million Trials Tell Us About Visual Search? *Psychological Science*, *9*(1), 33–39. https://doi.org/10.1111/1467-9280.00006
Wolfe, J. M. (2010). Visual search. *Current Biology*, *20*(8). https://doi.org/10.1016/j.cub.2010.02.016
Wolfe, J. M. (2012). Saved by a Log: How Do Humans Perform Hybrid Visual and Memory Search? *Psychological Science*, *23*(7), 698–703. https://doi.org/10.1177/0956797612443968
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*(3), 10. https://doi.org/10.1167/13.3.10
Wolfe, J. M. (2021). Guided Search 6.0: An updated model of visual search. *Psychonomic Bulletin and Review*, *28*(4), 1060–1092. https://doi.org/10.3758/s13423-020-01859-9
Wolfe, J. M., Kosovicheva, A., & Wolfe, B. (2022). Normal blindness: When we Look But Fail To See. *Trends in Cognitive Sciences*, *26*(9), 809–819. https://doi.org/10.1016/j.tics.2022.06.006
Wolfe, J. M., Palmer, E. M., & Horowitz, T. S. (2010). Reaction time distributions constrain models of visual search. *Vision Research*, *50*(14), 1304–1311. https://doi.org/10.1016/j.visres.2009.11.002
Wood, J. G., & Hassett, J. (1983). Eyeblinking During Problem Solving: The Effect of Problem Difficulty and Internally vs Externally Directed Attention. *Psychophysiology*, *20*(1), 18–20. https://doi.org/10.1111/j.1469-8986.1983.tb00893.x
Woodman, G. F., Luck, S. J., & Schall, J. D. (2007). The Role of Working Memory Representations in the Control of Attention. *Cerebral Cortex*, *17*(suppl_1), i118–i124. https://doi.org/10.1093/cercor/bhm065
Woodman, G. F., Vogel, E. K., & Luck, S. J. (2001). Visual search remains efficient when visual working memory is full. *Psychological Science*, *12*(3), 219–224. https://doi.org/10.1111/1467-9280.00339
Wu, C.-C., & Wolfe, J. M. (2022). The Functional Visual Field(s) in simple visual search. *Vision Research*, *190*, 107965. https://doi.org/10.1016/j.visres.2021.107965
Xu, L., Sahakian, A., Gayet, S., Paffen, C. L., & Van der Stigchel, S. (2025). Latent memory traces for prospective items in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*. https://doi.org/10.1037/xhp0001257
Yarbus, A. L. (1967). Eye Movements During Perception of Complex Objects. In *Eye Movements and Vision* (pp. 171–211). Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-5379-7_8

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*(5), 459–482. https://doi.org/10.1002/cne.920180503
Yu, X., Zhou, Z., Becker, S. I., Boettcher, S. E. P., & Geng, J. J. (2023). Good-enough attentional guidance. *Trends in Cognitive Sciences*, *27*(4), 391–403. https://doi.org/10.1016/j.tics.2023.01.007
Zacks, R. T., Hasher, L., Sanft, H., & Rose, K. C. (1983). Encoding effort and recall: A cautionary note. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 747–756. https://doi.org/10.1037/0278-7393.9.4.747
Zelinsky, G. J., & Bisley, J. W. (2015). The what, where, and why of priority maps and their interactions with visual working memory. *Annals of the New York Academy of Sciences*, *1339*(1), 154–164. https://doi.org/10.1111/nyas.12606
Zelinsky, G. J., Loschky, L. C., & Dickinson, C. A. (2011). Do object refixations during scene viewing indicate rehearsal in visual working memory? *Memory & Cognition*, *39*(4), 600–613. https://doi.org/10.3758/s13421-010-0048-x
Zhang, J., & Sclaroff, S. (2013). Saliency detection: A boolean map approach. *Proceedings of the IEEE international conference on computer vision*, 153–160. https://doi.org/10.1109/ICCV.2013.26
Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A bayesian framework for saliency using natural statistics. *Journal of vision*, *8*(7), 32–32. https://doi.org/10.1167/8.7.32

Appendix

Nederlandse samenvatting

Deel I: Visueel werkgeheugen in context

Inleiding

Wanneer je meubilair van een bekende Zweedse winkel in elkaar schroeft heb je als het goed is de handleiding naast je liggen waarnaar je steeds terug kan kijken. Je kunt dan een keuze maken; óf je onthoudt nu alvast hoe alle schroeven eruit zien, óf je onthoudt er nu maar eentje en wacht voor de resterende schroeven af tot het moment dat je ze ook echt nodig hebt. Wanneer je dit soort visuele informatie tijdelijk in geheugen houdt om je taak te voltooien, maak je gebruik van visueel werkgeheugen. Het *ont*houden en *onder*houden van die representaties van schroeven in visueel werkgeheugen vereist aanhoudende neurale activiteit, wat het een relatief energie-vereisend proces maakt om dingen in werkgeheugen op te slaan. Het is daarbij belangrijk om te weten dat onze hersenen ontzettend energiezuinig zijn, en dat we daarom graag zo min mogelijk energie gebruiken voor zelfs zulke simpele cognitieve taken. Bovendien heeft visueel werkgeheugen maar een beperkte maximum capaciteit; dit wordt geschat op gemiddeld twee tot vier representaties. Die beperkte capaciteit en hoge kosten maken het niet verleidelijk om ons visueel werkgeheugen veel te belasten.

Gelukkig is de buitenwereld voor het grootste deel stabiel, en kunnen we er daarom van uit gaan dat de handleiding met plaatjes van schroeven niet zomaar verdwijnt. In plaats van informatie in ons (interne) visuele werkgeheugen op te slaan, kunnen we daarom de omgeving voor ons laten werken als *externe opslag*. We doen dit in interactie met oogbewegingen. Iedere keer dat we besluiten om relevante informatie niet per direct te onthouden maken we de beslissing om er later alsnog een keer naar te kijken – en er dus een oogbeweging naar te maken. Wat blijkt? Oogbewegingen vereisen over het algemeen relatief weinig energie vergeleken met dingen in werkgeheugen onderhouden, dus we kiezen er vaak voor om maar één ding tegelijk te onthouden.

In de eerste vier hoofdstukken van dit proefschrift onderzocht ik hoe de afweging gemaakt wordt tussen dingen onthouden in visueel werkgeheugen, versus het gebruikmaken van de omgeving als externe opslag in interactie met oogbewegingen.

Hoofdstuk 1

In Hoofdstuk 1 gaf ik mensen een soort puzzel, waarin ze een voorbeeld van zes vormpjes moesten onthouden en namaken op een andere plek op het scherm (dit

noemen we een kopiëertaak oftewel "copying task"; zie *General Introduction Figuur 3*). Eerder werk uit ons lab en andere labs heeft aangetoond dat mensen van nature bij voorkeur maar één of twee vormpjes per keer in werkgeheugen opslaan, die plaatsen, en dan weer naar het voorbeeld kijken. Mensen kijken in zo'n kopiëertaak dus liever meerdere keren heen en weer tussen het voorbeeld en de plek waar het voorbeeld nagemaakt moet worden. Echter, die eerdere onderzoeken hebben tevens aangetoond dat mensen dit doen omdat de hersenen een afweging maken tussen (1) de kosten geassocieerd met het opslaan in werkgeheugen, versus (2) de kosten van het vergaren van externe informatie. Deze afweging is van nature in balans, maar die balans valt te verstoren. Het is, zoals benoemd, bekend dat we werkgeheugen graag pas belasten wanneer het nodig is – maar het was nog niet bekend wat er gebeurt wanneer de voorspelbaarheid van toegang tot externe informatie verhinderd is. Stel je voor dat je in de auto rijdt en de zon schijnt onregelmatig in je ogen, of dat er iemand bij een concert steeds voor je langs beweegt en daarmee je zicht van het podium ontneemt. Vertrouwen we in zulke gevallen meer op geheugen omdat het onzeker is of we informatie kunnen vergaren precies wanneer we dat willen? Dat onderzocht ik in Hoofdstuk 1. In Experiment 1 van de kopiëertaak was het voorbeeld altijd zichtbaar, of verscheen het voorbeeld voor enkele seconden en verdween het daarna weer voor enkele seconden, onafhankelijk van wat de participant deed. In Experiment 2 verscheen het voorbeeld alleen wanneer de participant ernaar keek – en dit kon onmiddelijk zijn of na een (onvoorspelbare) vertraging. Ik vond dat de verstoring van constante beschikbaarheid van informatie de grootste aandrijver was om meer vormpjes per keer te onthouden. Verrassend genoeg maakten vertragingen of onvoorspelbaarheid hier verder weinig verschil in. In Hoofdstuk 1 liet ik dus zien dat de mate waarin we op werkgeheugen vertrouwen relatief zwaar leunt op de mogelijkheid om externe informatie te vergaren op het *exacte* moment dat we die informatie nodig hebben. Als je een nachtkastje in elkaar zet, samen met iemand die telkens op onhandige momenten je handleiding wegneemt en weer teruggeeft, vertrouw je dus waarschijnlijk meer op je werkgeheugen dan wanneer de handleiding continu beschikbaar is.

Hoofdstuk 2

In Hoofdstuk 2 liet ik ten eerste zien dat eerdere bevindingen uit verscheidene variaties van de kopiëertaak ook vertalen naar zoektaken. Hier liet ik participanten zoeken naar één of vier vormpjes. De vormpjes die gezocht moesten worden stonden aan de linkerkant van het scherm en moesten teruggezocht worden aan de rechterkant van het scherm (zie *General Introduction Figuur 4*). In een deel van de experimentele condities konden participanten zo vaak als ze wilden terugkijken naar de voorbeeld-vormpjes, terwijl ze in andere condities de vormpjes maar één keer konden bekijken voordat ze begonnen met zoeken. In Experiment 1 gebruikte ik simpele vormpjes (een **c** die in acht rotaties kon staan), in Experiment 2 gebruikte ik meer complexe vormpjes die moeilijker te onthouden waren. Ik vond dat de mogelijkheid om terug te kijken naar vormpjes bevorderlijk was voor de snelheid en accuratesse waarmee participanten de taak konden voltooien – en dat dit voordeel groter was wanneer participanten meer, en complexere, vormpjes zochten. Bovendien bood deze beschikbaarheid participanten de kans om (1) vormpjes één voor één te onthouden en (2) tijdens het zoeken af en toe terug kijken om hun werkgeheugen als het ware

op te frissen. Daarom noemde ik het hoofdstuk (vertaald) *Verstop de handleiding niet*, en ik beargumenteerde dat mijn bevindingen van Hoofdstuk 2 verwerkt dienen te worden in vooraanstaande modellen van zoekgedrag.

Hoofdstuk 3

In Hoofdstuk 3 vroeg ik me af hoe ver ik dit vertrouwen op de omgeving kon rekken. Stel je voor dat je voor de Zweedse meubelgigant werkt en op één dag vijfentwintig nachtkastjes in elkaar moet zetten. Het is bekend dat, hoe vaker we een object zien en onthouden, hoe sterker dat object gerepresenteerd wordt in werkgeheugen of zelfs in langetermijngeheugen. Als je vijfentwintig keer hetzelfde schroefje moet onthouden is het daarom te verwachten dat het op een gegeven moment niet meer nodig is om de handleiding te bekijken – je hebt het schroefje inmiddels immers zo vaak gezien dat je erover zou kunnen dromen. Ik testte in Hoofdstuk 3 of mensen in zo'n situatie ook daadwerkelijk stoppen met het bekijken van de handleiding. In Experiment 1 deed ik een vergelijkbare taak als in Hoofdstuk 2, maar liet ik participanten vijfentwintig keer achter elkaar naar exact dezelfde vormpjes zoeken. Opmerkelijk genoeg keken participanten zelfs aan het eind van die vijfentwintig herhalingen nog steeds regelmatig naar de voorbeeld-vormpjes! In Experiment 2 testte ik in hoeverre dit herhaaldelijk vergaren van externe informatie nou echt nodig was; ik verstopte de vormpjes na de eerste vijftien herhalingen, dus participanten moesten de laatste tien herhalingen puur op geheugen zien te doen. Ik vond dat participanten die laatste tien herhalingen goed konden voltooien zonder slechter te worden. Waarom keken participanten dan zo vaak terug? Ik liet met een aantal analyses zien dat participanten op een gegeven moment schakelden: in plaats van dat ze hun geheugen daadwerkelijk verfristen, gebruikten ze de voorbeeld-vormpjes vooral om het zelfvertrouwen in hun eigen werkgeheugen op te krikken. In Hoofdstuk 3 concludeerde ik daarom dat de mate waarin we in ons geheugen (ten opzichte van de omgeving) vertrouwen deels afhankelijk is van onze zelfverzekerdheid, en dat dit vooral op de langere termijn zichtbaar wordt.

Hoofdstuk 4

In Hoofdstuk 4 maakte ik gebruik van alle hiervoor benoemde bevindingen over de flexibiliteit van hoe we werkgeheugen gebruiken, in een poging om inconsistente bevindingen uit andere onderzoeken te verenigen. Er is namelijk discussie over de kwestie of het mogelijk is om naar meerdere objecten *tegelijk* te zoeken. Het is duidelijk dat we meerdere representaties (e.g., vormpjes) tegelijk in visueel werkgeheugen kunnen opslaan, maar het is nog onduidelijk of we meerdere van deze representaties tegelijk kunnen gebruiken om onze aandacht te sturen. In andere woorden, als je meerdere schroefjes uit de handleiding hebt onthouden, kun je daar dan tegelijk naar zoeken, of moet dat één voor één? Ik beredeneerde in Hoofdstuk 4 dat de vraag niet alleen moet zijn of mensen het *kunnen*, maar ook of ze het überhaupt *doen* als er niet expliciet naar gevraagd wordt. In Experiment 1 gaf ik participanten de instructie om twee of vier vormpjes sequentieel (één voor één) danwel tegelijk te zoeken. Ik vond dat participanten beide instructies konden volgen, en dat naar twee vormpjes tegelijk zoeken wel degelijk mogelijk is. Echter, wanneer participanten naar vier vormpjes tegelijk moesten zoeken konden ze hun aandacht minder goed

sturen, wat aanduidt dat ze tegen een limiet aan zaten. In Experiment 2a en 2b liet ik participanten zelf bepalen hoe ze zochten. Met behulp van een relatief simpele maar verfijnde analyse trof ik aan dat participanten bij vrije keuze zowel sequentieel als simultaan konden zoeken, en dat ze op een flexibele manier voor ieder van de twee kozen. Deze keuze was afhankelijk van verscheidene factoren, zoals of ze konden terugkijken naar de voorbeeld-vormpjes, het aantal vormpjes en de complexiteit van die vormpjes, maar ook op basis van individuele voorkeuren. Ik beargumenteerde daarom in Hoofdstuk 4 ten eerste dat deze flexibiliteit meegenomen moet worden in onderzoek naar hoe men zoektaken uitvoert. Ten tweede beargumenteer ik dat deze flexibiliteit in hoe mensen zoeken kan bijdragen aan het verenigen van inconsistente bevindingen in de literatuur. Het is daarbij tevens belangrijk dat participanten duidelijk geïnstrueerd worden over hoe ze moeten zoeken; in dit veld werd tot dusver vooral gevraagd of mensen tegelijk naar dingen *kunnen* zoeken, maar misschien waren eerdere bevindingen gemengd omdat mensen er soms voor kiezen om het niet te *doen*.

Deel II: Individuele en status-afhankelijke invloeden op oogbewegingen

Inleiding

Het is al minstens sinds de vroege twintigste eeuw wetenschappelijk bekend dat oogbewegingen ontzettend informatief zijn voor onderliggende cognitieve processen. Waar we naar kijken, en hoe (snel) we onze ogen bewegen is niet alleen afhankelijk van de informatie die beschikbaar is in de omgeving (zoals besproken in Deel I), maar ook van het soort taak dat we uitvoeren en hoe gestimuleerd we zijn – zowel door de taak als door spontane fluctuaties in activatie van het centrale zenuwstelsel. Daarbovenop komt het feit dat iedereen net op een andere manier oogbewegingen maakt en anders naar dingen kijkt. In de laatste twee hoofdstukken onderzocht ik *waarnaartoe* en *hoe* we oogbewegingen maken wanneer dit niet per se gepaard gaat met een werkgeheugentaak.

Hoofdstuk 5

In Hoofdstuk 5 testte ik in hoeverre verschillende mensen op dezelfde manier naar hetzelfde plaatje kijken. Er zijn heel veel populaire rekenmodellen waar je een plaatje in kan voeren, en die vervolgens proberen te voorspellen naar welke plekken in dat plaatje mensen zullen kijken. Dit soort modellen worden niet alleen gebruikt om wetenschappelijke kennis te vergaren, maar worden ook toegepast in commerciële settings – bijvoorbeeld om de effectiviteit van reclameborden te voorspellen voordat ze naar de drukkerij gaan. Ik testte een selectie van eenentwintig van dit soort modellen in hun vermogen om die voorspellingen te maken voor mensen van verschillende leeftijden. Mijn collega's en ik hebben een interactieve installatie staan op de bovenste verdieping van het NEMO Science Museum te Amsterdam. Bezoekers aan het museum krijgen in die installatie een collage-plaatje te zien zonder verdere instructies (zie *General Introduction Figuur 5*). Terwijl de bezoeker naar het plaatje kijkt houdt de installatie bij waar diegene op het plaatje kijkt, en na afloop wordt

om hun gender en leeftijd gevraagd. Dit leverde een uitzonderlijk grote dataset op van 1600 participanten die allemaal hetzelfde plaatje hebben bekeken – en waarvan we de demografische gegevens hebben! Vervolgens heb ik die dataset gebruikt om te laten zien dat bijna alle modellen goed waren in het voorspellen waar jonge volwassenen (18-29 jaar) zouden kijken, maar significant minder goed waren in het voorspellen van het kijkpatroon van bijvoorbeeld jonge kinderen (6-11 jaar). Vanwege de populariteit van dit soort modellen in de wetenschap en in de praktijk stel ik in Hoofdstuk 5 dat het belangrijk is om deze modellen voor *iedereen* te laten werken, niet alleen voor kleine groepen mensen.

Hoofdstuk 6

In Hoofdstuk 6 introduceerde ik een methode die op basis van oogbewegingsstatistieken kan voorspellen of mensen hoog- of laag-geactiveerd (*aroused*) zijn. Daarvoor maakte ik gebruik van een openbare dataset waarin participanten naar de film Forrest Gump (1994) keken terwijl hun oogbewegingen en hartslag werden bijgehouden. Gedurende de circa twee uur durende film zijn er genoeg spannende, rustige, grappige, en wellicht saaie momenten – en deze fases gaan gepaard met afwisselingen in hoe aroused we zijn. Die mate van arousal is feitelijk een manier om activatie van het centrale zenuwstelsel uit te drukken, en is af te lezen aan hartslag; hogere arousal gaat vaak gepaard met hogere hartslag en vice versa. In Hoofdstuk 6 liet ik zien dat een set van twaalf oogbewegings-karakteristieken (het aantal, de snelheid en de grootte van oogbewegingen, en het aantal keer dat participanten knipperden) genoeg informatie bevatten om consistent te kunnen voorspellen of participanten een hoge of lage hartslag hadden. Omdat hartslag en arousal zo nauw met elkaar verweven zijn, stelde ik in dit hoofdstuk dat *hoe* we onze ogen bewegen verandert als gevolg van fluctuaties in arousal, en bovendien dat oogbewegingen goede indirecte voorspellers zijn van activatie van het centrale zenuwstelsel.

Conclusie

Of we nou door het bos wandelen, naar een basketbalwedstrijd kijken, of Zweeds meubilair aan het bouwen zijn, we vergaren continu actief relevante visuele informatie uit de omgeving. Gelukkig is dit vergaren van relevante informatie niet alleen een uitdaging; de omgeving biedt ons ook handvatten omdat het merendeel van informatie gewoon beschikbaar blijft. Zodoende kunnen we in veel situaties informatie vergaren *wanneer* het nodig is zonder al te veel cognitieve energie te hoeven verbruiken. Wanneer en in welke mate we gebruik maken van werkgeheugen is daarom ontzettend flexibel on context-afhankelijk, bewerkstelligd door de hersenen die een afweging maken van alle voordelen en nadelen die werkgeheugen, oogbewegingen, en de omgeving te bieden hebben. In dit proefschrift heb ik zodoende beargumenteerd dat het ontzettend belangrijk is om, wanneer we onderzoek doen naar hoe we werkgeheugen gebruiken en oogbewegingen maken, rekening te houden met zowel de respectievelijke als gecombineerde context van de wereld waar we in leven.

Acknowledgments (Dankwoord)

Dit proefschrift zou er heel anders uit hebben gezien zonder de expliciete danwel impliciete bijdragen van heel veel bijzondere mensen.

Stefan, zonder jou was ik hier überhaupt niet geweest. Jij was het die aan mij vroeg of ik toevallig een PhD wilde doen in jouw lab. Ik wist direct dat ik dat wilde, maar ik moest nog lang twijfelen of het niet eens tijd werd om mijn horizon te verbreden buiten Utrecht. Na veel rondkijken was het toch duidelijk; Stefan is verreweg de persoon waar ik het liefst mijn PhD mee zou willen doen. En dat heeft uitstekend uitgepakt. Bedankt voor de kansen die je me hebt gegeven, voor je begrip in lastige (professionele alswel persoonlijke) situaties, voor je geduld, voor je advies, en voor de ontzettend productieve en gezellige sfeer die je hebt geïncubeerd in het lab. Je bent nog niet van me af!

Tanja, het begon allemaal toen jij een nieuwe onderzoeksassistent zocht voor de VR supermarkt – en vanaf daar is het balletje gaan rollen. Het PhD-traject is uiteindelijk wat anders gelopen dan verwacht; meer zoektaken, minder patiëntstudies. Toch hebben we er naar mijn mening een mooi proefschrift van gemaakt. Heel erg bedankt dat jij me net als Stefan een kans hebt gegeven en op weg hebt geholpen in de wetenschap.

Christoph, ik maakte er vaak grapjes over, maar ik ben heel blij dat ik jou erbij heb gevraagd als (destijds kersverse) supervisor. Het moest toch even gezegd worden. In het begin vroeg ik af en toe of jij m'n toetsenbord wilde overnemen omdat ik vastliep op een paragraaf, later moest ik af en toe m'n toetsenbord weer van je los worstelen. Ik ben gegroeid als aio, maar jij bent ook gegroeid als supervisor – mooi om samen deze ontwikkelingen door te hebben kunnen maken. Bovendien bedankt voor de jaarlijkse oliebol, race-fietsie-fietsen, en alle andere gezellige momenten. *Herr Strauch*, I must admit that the last years kinda slapped!

Andre (*An*/*Annepan*), **Damian** (*Deem*/*Daamiejan*), **Sanne** (*San*/*Sannie*), we hebben te veel dingen meegemaakt, te veel inside jokes, en te veel sterretjes om in één dankwoord te vatten. Dus ik houd het kort.

An, jij was niet uit het veld te slaan, behalve die ene keer dat het niet siggi was en je een existentiële crisis had. Wil je trouwens feedback op je plot?

> *"Hmm... wow... ik heb er nooit bij stilgestaan dat de radio 's nachts ook doorgaat."*

Deem, over all jouw gewoontes en memorabele uitspraken zou ik een los proefschrift kunnen schrijven. Moge je novi's voor altijd doodsig zijn en je pupil traces voor altijd op kromme palmbomen blijven lijken. Bovendien: paling paling paling paling paling!

> *"Ik ga nog ff snel een croissantje halen. Ik ga geen kroketten op een lege maag eten."*

San, jij zat altijd tegen mijn voeten aan te schoppen onder het bureau, keihard op je toetsenbord te rammen, of luidruchtig je bleekselderij te eten. Vervolgens keek ik je even aan langs onze schermen, moesten we allebei lachen, en was het weer goed.

> 17:02 hoe was het op knaoor
> 17:02 kantiir
> 17:02 knatoor
> 17:02 joe

An, Deem, San, Plankie, King Siggi, bedankt voor alles <3

Beleke (*B*/*Beeee*/*Beli*), echt hef dat ik je niet meer iedere dag om 3 uur ga zien voor snackietime, ik ga dat (en jou) heel erg missen. **Dominique** (*Doom*/*Domi*), bedankt voor de warme vriendschap die bijna onmiddellijk ontstond vanaf het moment dat we voor het eerst een online vrijmibo hadden – en bedankt voor je aanstekelijke schaterlach die over heel de gang te horen was. **Evi** (*F.L.Evi*/*Evilientje*), ik moet soms nog steeds lachen om alle onmogelijke zit(?)posities waarin ik jou heb zien werken, je nieuwsbrieven, je lunchkeuzes; maar ik ben ook super impressed om wat je allemaal hebt bereikt – you rock (and rap)! **Kabir** (*Mr. Cookieman*/*Snackspert*), one of the smartest and most helpful people around, there will be no more sprinting for the microwave – I snost I lost. **Laura** (*laulau*), ik ben zo blij dat ik zoveel tijd met je heb kunnen doorbrengen, en ik ben trots op hoe je je hebt ontwikkeld als persoon en onderzoeker – Ganz viel Liebe! **Noa** (ik durf je niet meer *Nootje* te noemen), het was echt slay om met jou tussen het kletsen door af en toe ook wat data te verzamelen – ik rate onze samenwerking 9 uit 11.

De drie Musketiers – **Marinke** (*Rinkiedinkie*), **Renee** (*Reneecitaatje*) en **Zoë** (*Z*/*Zo-EE*) – het was een titanenstrijd, geschikt voor de geschiedenisboeken, maar uiteindelijk hebben de Angels gezegevierd; de Gof is toch echt beter dan de Mams. Alle andere buren van H0.19 (**Chris K.**, **Famke**, **Isolde**, **Lara**, **Liselotte**), het was me een waar genoegen om elkaar elke dag opeens onverwacht door het raam aan te staren. Zoë, jou wil ik in het bijzonder bedanken voor je oneindige empathie (*"ja maar lieverd, luister nou, ..."*) – één, twee – je bijzondere manier van dingen opsommen, en – drie – je enthousiasme voor galgje (ik baal nog steeds van *yoga*).

Surya, **Teuni** (*Tjoenie*), jullie zijn mijn academische rolmodellen – wat was het fijn om met jullie in een "team" te zitten. Surya; the lord of the bootstrap, koning van de scherpe opmerkingen, en mister Caplab himself. Teuni; die razendsnel denkt en werkt, hilarisch is, en fantastische lasagne maakt. Bedankt dat jullie altijd klaar stonden om jullie encyclopedische kennis te delen en als mentors te fungeren. **Sam** (*Samuel*) the prankster, thanks for the game nights and the book recommendations;

no thanks for that time when you jumpscared me so much that I spilled coffee all over myself. **Sjoerd** (*Short Stud*), bedankt voor de vele (soms hilarische, soms bizarre, soms onverstaanbaar gemompelde, maar nooit oppervlakkige) gesprekken. Ik heb je nooit vergeven voor de surprise party die je voor me had georganiseerd. **Chris P.** (*Paffen*), bedankt voor je humor, het vrolijke gefluit in de gangen, en sorry voor het pesten.

To the "young" PhD candidates, et al. (**Daan** (*Daab*), **Evan**, **Jasmijn**, **Koert**, **Lasse**, **Marleen**, **Martin** (*Mr. TechSupport*), **Rick**, **Sterre**, **Tessie**), I wish you the best of luck with your promising careers, and I hope that I managed to provide you with at least some guidance during the time that we overlapped. Aan de "oude" aio's (**Martijn**, **Gijs**), op mijn beurt wil ik *jullie* dan weer bedanken voor jullie advies in mijn vroege jaren. **Dan**, **Liangyou**, **Luzi**, **Yuqing**, you are rockstars; I aspire to be as courageous as you are. **Chris J.**, **Krista**, bedankt voor jullie aanmoediging en begeleiding toen ik nog een jong bachelorstudentje was. **Janet**, **Jeanita** (*Big J*/*Zjaan*), **Eline**, **Eric**, **Karin** en **Nyazi**, bedankt voor jullie eeuwige optimisme en onuitputbare inzet om alles draaiende te houden. Datzelfde geldt voor het TechSupport team, in het bijzonder **Jim**, **Mark**, **Michael** en **Son**.

Thank you to all the *flexplek*-students who were often up for yapping; while I didn't want to do work, while I was waiting outside of Christoph's office, or when I simply just wanted to yap – a special shoutout in that regard to Bernd, Eva, Judith, Kelsi, Lotte and Marlies (and honorable member Kabir). To **all colleagues and students**; thank you for the chats in the hallway, cake on birthdays, joining the Friday drinks, coming to the NVP diversity event, and all other small acts of kindness that you have shown which have made my time at EP very enjoyable.

Jet, zo leuk om keer op keer met je op congressen te kunnen hangen, exchanges te doen en bijvoorbeeld een diversity event te organiseren. Wij zijn de toekomst zeiden we tegen elkaar; dus *op de toekomst!* **Ana**, **Veera**; **Cate**, **Chris**, **Docky**, **Jasper**, **Yayla**; it's been awesome getting to know you, and I hope we will keep running into each other – be it in Egmond aan Zee, St. Pete's, or elsewhere.

Lauriane, **Isabel**, **Veerle**, **Timo**, heel tof om mijn academische carrière al samenwerkend met jullie te zijn begonnen; ik herinner me vooral meetings in de vissenkom die niet altijd even efficiënt waren omdat we het eigenlijk té gezellig hadden.

Ray en **Nick** van de Gutenberg, het kostte me ongeveer twee jaar om te realiseren dat jullie ook benen hebben daar achter de bar – bedankt voor het af en toe kletsen (over o.a. koffie en muziek), de hilarische interacties (zoals een latte-art middelvinger en een koffiebeker met harige ballen erop), en natuurlijk voor de lekkere koffie (*Gof*).

Edwin Dalmaijer, **Chris Olivers**, **Ignace**, **Roy**, **Teuni**, bedankt voor de productieve samenwerkingen, jullie aanmoedigingen, en jullie woorden van wijsheid.

Freek van Ede, **Melissa Lê-Hoa Võ**, **William Ngiam**, **Heleen Slagter**, **Jeremy Wolfe**, thank you for being so encouraging and supportive of the work of some PhD candidate who you have (or in Jeremy's case *had*) no official connection to. I am very appreciative of that.

Folmer, ik heb het ontzettend getroffen met jou als huisgenoot voor vier jaar. Je hebt nu de reden in je handen dat ik 's avonds thuis vaak geen energie meer had om te kletsen of te koken. Zorg goed voor de plantjes!

Floris (*Flo*), **Inigo** (*Ini*), **Roderic** (*Ro*), **Timo** (*Timochan*), **Zoril** (*Zorro*), daar is Dr. Axle dan toch echt. Wat ooit begon als groepsopdracht bij Psychofarmacologie is uitgegroeid tot een al bijna tien jaar lange vriendschap, en het is jullie betrouwbaarheid als vrienden die me steun en afleiding gaven wanneer ik het even helemaal zat was. Iemand 20:00 Rocket League?

Hub (*Hubanjo*), **Jeroen** (*JK*/*J-nerd*), **Leon** (*Broeder Ham*/*Kobe*), **Ruben** (*Ruupie*/*Abel*), **Sander** (*San*), **Wouter** (*Drief*/*Juan Carlos*), heel veel liefde voor de Bonkers sinds 2006. Pour la révolution!

Bedankt aan mijn lieve familie. **De Engeltjes**, met wie ik afgelopen kerst niet heb kunnen vieren omdat ik dit proefschrift zat te typen. It was a long fahrt but I made it! **Ashley**, **Jaleesa** en **Thomas** (en schoonfamilie en kids), die ik veel te weinig heb gezien, maar waarvan ik weet dat ze er altijd voor me zijn. **Oma Joke**, die altijd zó graag net als ik psycholoog hadden willen worden; en **Jaap**, die genoeg kon lullen over binnenvaart, hijskranen en aquariums om z'n eigen proefschrift te vullen. **Opa en oma Poezen**, die hier te weinig van hebben kunnen meemaken. **Mama** en **Rob**, **Papa** en **Jitske**, het is natuurlijk lastig om 31 jaar in een dankwoord te vatten, dus dan maar zo; bedankt en ik hou van jullie.

Curriculum Vitae

Alex Jan Hoogerbrugge was born on March 21st 1994 in Capelle aan den IJssel.

Alex completed primary school in 2006, and in 2013 he obtained his VWO diploma at IJsselcollege. After a gap year (during which he worked at a bank, in hospitality, and as a volunteer), Alex started a degree in Psychology in 2014. In 2017, Alex succesfully completed the CreateDAV summer school at York University, Canada. He obtained his Bachelor's degree in Psychology from Utrecht University in 2018, after completing his thesis under the supervision of dr. Krista Overvliet (*Gestalt Principles in Haptics: Similarity Cues in Temperature Perception and Its Rivalry with Proximity Cues*; which was nominated for the Peter G. Swanborn award for outstanding theses).

Alex subsequently obtained his Master's degree *cum laude* in Artificial Intelligence from Utrecht University in 2020, with a minor in Applied Data Science and a diploma from the Interdisciplinary Graduate Honours Programme. During this period, he worked as a research assistant and wrote his thesis under the supervision of prof. dr. Stefan Van der Stigchel and dr. Tanja C. W. Nijboer (*Reliability of Visual Access: Modeling the trade-off between internal storage and external sampling in a Visual Working Memory task*).

Alex started as a PhD candidate in 2021, under the supervision of prof. dr. Stefan Van der Stigchel, dr. Tanja C. W. Nijboer. Dr. Christoph Strauch joined the supervision team a year later. During his PhD candidacy, Alex managed two eye tracking labs, established and co-organised the Diversity Event at the 2023 NVP Winter Conference, was interviewed for BNR Nieuwsradio, and was involved in three public engagement demonstrations (NEMO Science Museum, Betweter Festival, Lowlands Festival).

From January 2025 until March 2025, Alex worked as a postdoctoral Research Associate under the supervision of dr. Christoph Strauch; he first-authored two manuscripts during that time. Since May 1st 2025, Alex works at the University of Manchester as a postdoctoral Research Associate under the supervision of dr. Johan Hulleman and prof. dr. Jeremy Wolfe.

Bibliography

Peer-reviewed articles

As first author

Hoogerbrugge, A. J., Strauch, C., Nijboer, T. C. W., & Van der Stigchel, S. (2024). Persistent resampling of external information despite 25 repetitions of the same visual search templates. *Attention, Perception, & Psychophysics, 86*(1). doi.org/10.3758/s13414-024-02953-z

Hoogerbrugge, A. J., Strauch, C., Böing, S., Nijboer, T. C. W., & Van der Stigchel, S. (2024). Just-in-Time Encoding Into Visual Working Memory Is Contingent Upon Constant Availability of External Information. *Journal of Cognition, 7*(1). doi.org/10.5334/joc.364

Hoogerbrugge, A. J.*, Strauch, C.*, Baer, G., Hooge, I. T. C., Nijboer, T. C. W., Stuit, S. M., & Van der Stigchel, S. (2023). Saliency models perform best for women's and young adults' fixations. *Communications Psychology, 1*(1). doi.org/10.1038/s44271-023-00035-8 * Authors contributed equally and share first-authorship.

Hoogerbrugge, A. J., Strauch, C., Nijboer, T. C. W., & Van der Stigchel, S. (2023). Don't hide the instruction manual: A dynamic trade-off between using internal and external templates during visual search. *Journal of Vision, 23*(7), 14. doi.org/10.1167/jov.23.7.14

Hoogerbrugge, A. J., Strauch, C., Oláh, Z. A., Dalmaijer, E. S., Nijboer, T. C. W., & Van der Stigchel, S. (2022). Seeing the Forrest through the trees: Oculomotor metrics are linked to heart rate. *PLOS ONE, 17*(8). doi.org/10.1371/journal.pone.0272349

As co-author

Strauch, C.*, **Hoogerbrugge, A. J.**, & Ten Brink, A. F.* (2024). Gaze data of 4243 participants shows link between leftward and superior attention biases and age. *Experimental Brain Research, 242*(1). doi.org/10.1007/s00221-024-06823-w * Authors contributed equally.

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). doi.org/10.3390/jcm12113630

Brouwer, V. H. E. W., Stuit, S., **Hoogerbrugge, A. J.**, Ten Brink, A. F., Gosselt, I. K., Van der Stigchel, S., & Nijboer, T. C. W. (2022). Applying machine learning to dissociate between stroke patients and healthy controls using eye movement features obtained from a virtual reality task. *Heliyon, 8*(4). doi.org/10.1016/j.heliyon.2022.e09207

Articles in preparation & preprints

Hoogerbrugge, A. J., Strauch, C., Hoevers, N., Olivers, C. N. L., Nijboer, T. C. W., & Van der Stigchel, S. (Under review). Multi-target visual search flexibly switches between concurrent and sequential search modes.

Hoogerbrugge, A. J., Sahakian, A., Brouwer, R., Klauss, G., Strauch, C., Nijboer, T. C. W., & Van der Stigchel, S. (2024). Three unpublished, publicly available visual search datasets with 151 participants. *OSF Preprints*. doi.org/10.31219/osf.io/kgpq4

Conference contributions

2024	**Perception day. Utrecht, NL**	Talk
	Concurrent multi-target search is possible, but sequential search is sometimes preferred	
2024	**The 4th International Conference on Working Memory (ICWM). Leeds, UK.**	Poster
	Persistent resampling of external information despite twenty-five repetitions of the same search templates	
2024	**Vision Sciences Society (VSS). St. Pete Beach, FL, USA.**	Poster
	Persistent resampling of external information despite twenty-five repetitions of the same search templates	
2023	**NVP Winter Conference on Brain & Cognition (NVP). Egmond aan Zee, NL.**	Poster
	Persistent resampling of external information despite twenty-five repetitions of the same search templates	
2023	**Vision Sciences Society (VSS). St. Pete Beach, FL, USA.**	Talk
	Don't hide the instruction manual: A dynamic trade-off between using internal and external templates during visual search	
2023	**Perception day. Utrecht, NL.**	Talk
	Don't hide the instruction manual: A dynamic trade-off between using internal and external templates during visual search	
2022	**European Conference on Visual Perception (ECVP). Nijmegen, NL.**	Poster
	Seeing the Forrest through the trees: Oculomotor metrics are linked to heart rate	
2022	**European Conference on Eye Movements (ECEM). Leicester, UK.**	Poster
	Seeing the Forrest through the trees: Oculomotor metrics are linked to heart rate	
2022	**Vision Sciences Society (VSS). St. Pete Beach, FL, USA. Online.**	Poster
	Seeing the Forrest through the trees: Oculomotor metrics are linked to heart rate	
2022	**NVP Winter Conference on Brain & Cognition (NVP). Egmond aan Zee, NL.**	Poster
	Reliability of Visual Access: A trade-off between internal storage in visual working memory and external sampling	
2021	**Vision Sciences Society (VSS). St. Pete Beach, FL, USA. Online.**	Poster
	Reliability of Visual Access: Modeling the trade-off between internal storage in visual working memory and external sampling	

2021 **Human-Centered Artificial Intelligence Seminar. Utrecht, NL. Online.** Talk

Using Machine Learning to distinguish stroke patients from healthy controls, based on eye movement data in a Virtual Reality setting