Title

Running after Two Hares in Visual Working Memory: Exploring Retrospective Attention to Multiple Items Using Simulation, Behavioral Outcomes, and Eye-tracking.

Taiji Ueno[1,*] and Richard J. Allen [2]

[1] Faculty of Human Sciences, Takachiho University, Japan.

[2] School of Psychology, University of Leeds

Correspondence

[*] TU is now based at Division of Psychology and Communication, School of Arts and Sciences, Tokyo Woman's Christian University, Tokyo, Japan.

E-mail: taijiueno@lab.twcu.ac.jp

Postal address: 2-6-1, Zempukuji, Suginami-ku, Tokyo Woman's Christian University, Japan.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP20H04494, JP22H01073, and JP23H01053.

Abstract

Multi-item retro-cueing effects refer to better working memory performance for multiple items when they are cued after their offset, compared to a neutral condition in which all items are cued. However, several studies have reported boundary conditions, and findings have also sometimes failed to replicate. We hypothesized that a strategy to focus on only one of the cued items could possibly yield these inconsistent patterns. In Study 1, a Monte-Carlo simulation showed that randomly selecting one of the cued items as the focus in each trial increased the chance of obtaining significant 'multi-item retro-cueing effects' on the mean accuracy over the trials, providing an incorrect conclusion if interpreted as evidence for attending all the cued items. These high rates to obtain such data fit with inconsistent patterns in the literature. To try and circumvent this situation, we conducted two new experiments (Studies 2A and 2B) where participants were explicitly instructed to fixate their gaze on all the cued positions, verified through eye-tracking (Study 2B). These produced robust multi-item retro-cueing effects regardless of previously identified boundary conditions. Notably, gazes were clearly fixated to multiple cued positions within each trial. Nevertheless, simulation revealed that our accuracy patterns could also in principle be produced by single item enhancement on each trial. The present study forms the first step to disentangle overt gaze-based allocation of attention from single-item focusing strategies, while also highlighting the need for improved methodologies to probe genuine multiplicity in working memory.

Keywords: visual working memory, retrospective attention, retro-cue, monte-carlo simulation, eye-tracking

Public Significance Statement

This study explores how attention can be directed to improve memory performance. We show via simulation that previous studies may only provide weak evidence for the ability to focus attention on more than one item in working memory when they are no longer present in the environment. We further show that people do look towards the locations of all cued items and can benefit their memory, when instructed to do so. However, questions remain about whether this really does show a genuine ability to enhance multiple items at the same time in working memory.

Introduction

Visual working memory (VWM) retains objects and operates on them over time courses of a few seconds. VWM is closely related to attention. For example, studies have demonstrated that both *pre-cueing* one of the to-be-tested items before the onset of the target array and *retro-cueing* after the offset enhance accuracy for the cued items more than a neutral condition where all items are cued (Makovski & Jiang, 2007). As Souza and Oberauer (2016) reviewed, research on retro-cueing effects allows us to investigate the nature and mechanisms involved in focusing attention to internal information in working memory (Oberauer & Hein, 2012). For example, one can investigate how information in working memory is selected, and whether capacity limitations in working memory can be mitigated against via application of retrospective attention. By increasing the number of retro-cues, one can further examine whether such functions of the focus of attention can be expanded to multiple items in working memory. Furthermore, if the focus of attention can be retrospectively allocated to multiple representations in working memory, then one can next ask whether multiple representations are focused simultaneously or sequentially (see, Ort & Olivers, 2020, for consideration of this issue in the visual search domain). In the current work, our main question examined whether retrospectively attending to multiple items enhances accuracy. As outcomes in the literature are mixed, we first explored a possible reason for these inconsistencies, and then aimed to collect potentially stronger evidence for the effect through task instructions and by tracking eye movement. As part of this approach, we inspected fixation data to clarify whether multiple items were attended simultaneously or sequentially.

Does *multi-item retro-cueing* reliably enhance accuracy for cued items? Whilst multi-item pre-cueing robustly enhances memory performance for cued items, there is still dispute regarding the effects of multi-item retro-cueing. An initial study did not find a significant multi-item retro-cueing effect (Makovski & Jiang, 2007). Subsequent studies investigated this more extensively and proposed boundary conditions for multi-item retro-cueing effects. These include (a) contextual matching between the number of cued items and the number of the probed items (Matsukura & Vecera, 2015); (b) dividing multiple cues in different hemifields (Delvenne & Holt, 2012); (c) sequentially presenting multiple retro-cues (Li & Saiki, 2014); and (d) using a feature-based cue (i.e., non-spatial cue) rather than a spatial/symbolic cue when the cues items are spatially distant from each other (Heuer & Schubö, 2016). Revealing the boundary conditions of a phenomenon is a typical and useful way to advance science because it means theories are refined. However, from a different viewpoint, boundary conditions also add complexity to theories. Thus, this is a balancing issue in that more complex theories are welcome only when these are necessary to explain the phenomenon.

However, these boundary conditions do not consistently explain the existing data. For example, there are studies that found the multi-item retro-cueing effect even without contextual matching (e.g., Heuer & Schubö, 2016), without splitting cued items into different hemifields (e.g., Heuer & Schubö, 2016; Matsukura & Vecera, 2015); and regardless of the spatial distance between cued positions (e.g., Matsukura & Vecera, 2015). Studies have also failed to find the effects of sequentially presented multi-cues on the second cued item (e.g., Van Moorselaar et al., 2015), although Souza et al (2015) replicated the effects of sequentially presented multiple cues on both the first and the second (subsequent) items. Thus, there is still disagreement on whether multi-item retro-cues enhance memory performance, particularly when using simultaneous presentation of cues (Souza & Oberauer, 2016).

It is unclear why such inconsistencies are observed regarding multi-item retro-cueing effects. Without a systematic reason that affects the presence/absence of effects in a *probabilistic* manner, it is difficult to explain the existing patterns in the literature. We hypothesized that one of the candidates for such a reason might be *a single-item focusing strategy*. Specifically, if the cued items are probed more frequently than other items (i.e., high cue validity), then focusing on only one of the multiple cued items inevitably increases the probability to achieve higher *mean accuracy over trials* than randomly focusing on one of all studied items (i.e., the neutral cue condition where no items are preferentially cued). Since it is probabilistic whether the focused single item is probed, it is also probabilistic whether each participant exhibits higher mean accuracy over trials in the multiple-cue condition than the neural cue condition (see, Ort & Olivers, 2020, for consideration of this issue in the visual search domain). As such, it is not surprising that there are mixed outcomes regarding the multi-item retro-cueing effect, as it would be inappropriate to argue for attention towards multiple items if only one of the cued items is strategically focused on in each trial.

It is therefore important to estimate the actual chance rates to obtain a statistically significant difference between mean accuracies when a single-item focusing strategy is employed. If such rates are low, then this strategy-based account cannot explain the inconsistencies. A promising approach to estimate this chance rate under a hypothetical scenario (i.e., only single item is focused) is via statistical simulation, such as the Monte-Carlo method. This was the aim of Study 1. This approach involves repeated sampling of correct/incorrect responses from binomial distributions, whose parameters are derived from published studies. The rates of detecting a significantly higher mean accuracy in the multi-item retro-cuing condition can then be calculated in scenarios where only one item was focused on rather than attending multiple items. To foreshadow the outcomes, we found that such chance rates were high when the simulation parameters were taken from the existing literature that found multi-item retro cueing effects. In other words, evidence for ‘retrospectively attending multiple items’ may be weak even if the statistically significant multi-item retro-cueing effect was obtained. In addition, it is important to note that the estimated chance rates in Study 1 were not 100%. This means that when only a single-item was strategically focused, there is also a chance to fail to obtain a statistically significant multi-item retro-cueing effect even if a genuine multi-item retro-cueing effect exists in human visual working memory.

To circumvent these situations, it is important to investigate multi-item retro-cueing under a situation where participants are actively discouraged from using a single-item focusing strategy. For this aim, we conducted two new experiments (Study 2A and Study 2B) in which participants were instructed to fixate their gazes to all the spatial positions of the cued items during the retention interval. Moreover, to collect stronger evidence, gaze was monitored in Study 2B by using an eye-tracker. Fixating gazes across all the cued positions would be a clear indicator that participants are not simply adopting a single-item focusing strategy. The rationale behind this approach is based on an active and facilitatory role of gaze position in memory after the offset of the studied items (Ferreira et al., 2008; Johansson & Johansson, 2014; Laeng et al., 2014; van Ede et al., 2019). Although Loaiza and Souza (2022) did not find a relationship between participants' spontaneous fixation duration during maintenance and recall accuracy in the fixated position, they did find a positive relationship when eye gaze to the *sequentially-presented* cued positions was explicitly instructed (see their Study 1B). Thus, this study justifies using instructed gaze fixation as an approach to attend to the multiple cued positions. We apply this technique to the simultaneously-presented multi-item reto-cueing paradigm. If retrospectively attending to multiple items is possible, then memory performance should be improved in the multi-item retro-cueing condition. Moreover, we may glean further insights on the mechanism of the multi-item retro-cueing effect (i.e., sequential vs. simultaneous processing) by comparing fixation durations across the multiple cued positions.

Study 1: Monte-Carlo Simulation

Transparency and Openness

In Study 1, we used statistical R (R Core Team, 2023) for the simulation. All the simulation codes to reproduce the simulation data are available online (DOI 10.17605/OSF.IO/DRBXT). The detailed instruction for running the simulation is also provided online. The general overview of the simulation is provided in the next section.

Aim and Overview of the Simulation

A promising approach to estimate the chance rates of obtaining statistical significance under these hypothetical scenarios is via the Monte-Carlo method of statistical simulation. This was implemented in Study 1. In each simulation trial, we can calculate the probability for a focused item to be probed by chance under the assumption that *only one* of the cued-items was randomly selected as a to-be-focused item. This probability can be precisely calculated based on number of items, number of cues, and cue validity. Then, the correctness (binary: correct or incorrect) in each trial can be simply simulated by sampling from a binomial distribution. The probability for sampling a correct case can be set to be higher when the randomly selected focused-item is probed by chance than when the focused-item is not probed by chance. The exact probabilities for these binomial distributions can be determined in an objective manner, such that the effect size of single-item retro-cueing (i.e., single retro-cue condition vs. the neutral retro-cues condition) in the simulation matches with the real human data (later in detail). Once these probabilities for the binomial distributions are determined, then the correctness of each trial can be sampled regardless of the number of cues (2 or larger). Then, after repeating this sampling procedure for the number of trials in each condition and for the number of participants, the resultant data matrix can be submitted to a conventional statistical test (e.g., *t*-test or ANOVA) to examine a multi-item retro-cueing effect. Note that this is a simulation of a scenario where only one of the cued items is focused. Thus, even if significantly higher accuracy in the multi-item retro-cueing condition than the neutral-cues condition is detected, it does not provide evidence for successful retrospective attention to multiple items. By reiterating the same procedure many times (i.e., Monte-Carlo method), we can estimate the probability of yielding statistical significance. The parameter values (i.e., number of items/cues/participants, cue validity, etc.) were taken from the existing literature that found a multi-item retro-cueing effect. We focused on the articles that reported both single-item and multi-item retro-cueing effects, as these are both necessary for our simulation (see below). Moreover, we focused on the cases where multiple cues were presented simultaneously, because sequential presentation of multiple cues appears to have a reliable effect (see, Souza & Oberauer, 2016, for a review). In addition, we also simulated the other scenarios, where cued-validity is not 100% (e.g., Li & Saiki, 2014), and where a multi-item retro-cueing effect is modulated by another factor (e.g., Heuer & Schubö, 2016). The details will be explained later.

Methods

Table 1.

Parameters from Each Human Experiments and Simulation Parameters

Parameters and variables	Articles			
	Makovski & Jiang (2007)	Li & Saiki (2014)	Matsukura & Vecera (2015)	Heuer & Schubö (2016)
Number of participants	20~23	16~18	16~32	17~23
Number of studied items	6	4	6	4[*1]
Number of retro-cues	1, 2, 3, 4, 5[*2]	1, 2	1, 3	2[*3]
Number of neutrally-cued items	6	4	6	4
Number of probes	1	1	1, 3[*4]	1
Number of trials	48~90	48[*5]	70	192
Single-cueing effect size				
Hedge's *g*	0.611	1.433[*6]	2.067	1.105[*7]
Simulation ($p_{unfocus}$)[*8]	0.600	0.600	0.600	0.600
Simulation (p_{focus})[*9]	0.639	0.726	0.738	0.649

Note. The *N* of participants/trials varied depending on the experiments within each article. Thus, we adopted the largest value for our simulation; *1 There were 8 items in the study array, but only half of them (either hemifield) was task-relevant; *2 We adopted 2-item retro-cueing condition for our simulation; *3 There was not a single-cueing condition; *4 They employed a set-probe testing procedure; *5 There were 144 trials in the double-cue condition, but there were 48 trials where each one of the cued positions were probed; *6 The only statistics for the single-cueing effect reported in Li and Saiki (2014) was the contrast between the 'withdrawal condition ' and the neural cue condition. Thus, we used these statistics in order to estimate the single-cueing effect size; *7 This value was conservatively taken from the double-cueing effect (see main text); *8 The simulation parameter was set to be 0.6. This was set arbitrarily but this absolute value is not important. What matters is the relative difference from the p_focus value; *9 These values were determined by a grid search so that the single-cueing effect size matched with human data (see main text).

Parameters from the Target Articles

The target articles that we simulated are listed in Table 1. The parameters (*N* of participants, trials, cues, targets, and probes) were taken from each article. When multiple experiments were conducted in each article, the largest value was used. The effect size index (Hedges' g) for the single-item retro-cueing effect was converted from the paired *t*-statistics (single-cue condition vs. neutral-cue condition) reported in each paper.

Simulation Dataset and Parameters from the Human Experiments

First, we specified the model to be studied using the Monte Carlo simulation. In each trial, only one of the cued items were randomly selected as a to-be-focused item. Thus, the probability to select the *i*-th item (p_{select}) simply follows a discrete uniform distribution:

$$p_{select}\ (X = i\_th\ item) = \frac{1}{N_c} \quad (i = 1, 2, \ldots, N_c) \ldots Equation\ (1)$$

, where N_c is the number of cues. In the neutral cue condition (i.e., control condition), N_c was equal to the number of the studied items.

Next, we considered the probability for each item to be probed in the test phase. In a single-probe change-detection paradigm, the probability for each item to be probed depends on the number of cues (N_c) and the cue-validity level (i.e., the degree of predictiveness of the cues). We first simulated the case of the maximum cue-validity level (i.e., 100%), and consider lower validity cases later. When cue-validity is 100%, then the probability for one (j-th item) of the cued items to be probed (p_{probe}) follows a discrete uniform distribution as follows:

$$p_{probe}\ (X = j\_th\ item) = \frac{1}{N_c} \quad (j = 1, 2, \ldots, N_c) \ldots Equation\ (2)$$

Next, when the probed item (or the lure probe in the multiple-probes situation) was coincidentally from the same position as the focused (single) item, then it is natural to expect a higher accuracy than when these were different. Then, the correctness in the k-th trial (C_k) can be simply simulated by sampling from the binominal distribution as follows:

$$C_k \sim Binomial\ (n = 1, p) \begin{cases} p = p_{focus}, when\ a\ probed\ item\ is\ taken\ from\ the\ focused\ position \\ p = p_{unfocus}, when\ a\ probed\ item\ is\ not\ taken\ from\ the\ focused\ position \end{cases}$$

$$\ldots Equation\ (3)$$

, where the probability for the binomial distribution (p_{focus}) was set to be higher than ($p_{unfocus}$). The simplicity of this equation as a model of human recognition is discussed later in the results section. The specific values for the probability parameters were determined objectively as follows: First, it is entirely reasonable to assume that participants take this single-item focusing strategy when the number of cues is one (i.e., single-item retro-cueing condition). Then, the specific *difference* between (p_{focus}) and ($p_{unfocus}$) values can be determined so that the effect size of the single-item retro-cueing effect in the Monte Carlo simulation matches that in the real human data. Taking Makovski and Jiang (2007) as an example, the single-item retro-cueing effect sizes (i.e., single cue condition vs. neutral cues condition) in their experiments were Hedges' g of 0.655, 0.470, 0.530, and 0.930 (Experiments 1a, 1b, 2a, 2b, respectively). A random-effect meta-analysis (ESCII software: Cumming, 2012) can aggregate these effect sizes to estimate a single score, Hedges' g = 0.611. Then, bearing this integrated single-cueing effect size in mind, we conducted a grid search to determine the (p_{focus}) and ($p_{unfocus}$) parameter values for simulating Makovski and Jiang (2007)'s study. More specifically, we first fixed the ($p_{unfocus}$) parameter to 0.600 whilst gradually increasing the (p_{focus}) parameter by 0.001. Then, the correctness in the k-th trial (C_k) of the single-cue condition in the simulation was sampled by the Equation (3) above for the same number of trials/participants as Makovski and Jiang (2007). This allows to compute the single-cueing effect size in the simulation under a given (p_{focus}) parameter value. Once the resultant effect size in the simulation reached Hedges' g of 0.611 (i.e., real human data), then we stopped increasing the (p_{focus}) parameter value. In case of Monte-Carlo simulation for Makovski and Jiang (2007) study, the (p_{focus}) was determined to 0.639. The same procedure was taken in determining the (p_{focus}) parameter value in simulating each study listed in Table 1.

Next, the obtained (p_{focus}) parameter values can be used in sampling the correctness value in the multi-cue trials and in the neutral-cue trials of the simulation as well (Equation 3). This is a plausible procedure because the current Monte-Carlo approach assumes the scenario where a single-item focusing strategy is taken in the multi-cue trials. By reiterating the sampling of the correctness value in the *k*-th trial (C_k) for the same number of trials/participants as the existing literature (see Table 1), we can generate the simulation data matrix under an assumed hypothetical scenario. Then, a paired *t*-test can be conducted to investigate whether a statistically significant difference is detected between the multi-item cueing condition (1 < *N* of cues < *N* of studies items) and the neutral condition (*N* of cues = *N* of studied items). If statistical significance is observed, it would be inappropriate to confidently interpret this as evidence for attending all cued items. Instead, such a result should be interpreted as supporting evidence for the argument that one can obtain significantly higher mean accuracy in the multi-cued condition than in the neutral-cues condition even when just one of the cued items is strategically focused on. Finally, all the procedures so far can be repeated multiple times in Monte-Carlo simulation (100 times in our case) so that the chance rates to observe statistical significance can be calculated in each experimental situation of the existing literature. The number of the iterations was determined by conducting a self-replication. A higher number of iterations leads to more stable estimations (i.e., less affected by sampling variations) but requires more time to conduct. The effect size of sampling variances on the outcome of the model can be evaluated empirically by re-running the simulation. We re-conducted the simulation with the number of iterations as 500, and the outcomes were very similar (e.g., the difference in the estimated chance rates was less than 5%). Thus, for the sake of efficiency we report analyses using 100 iterations.

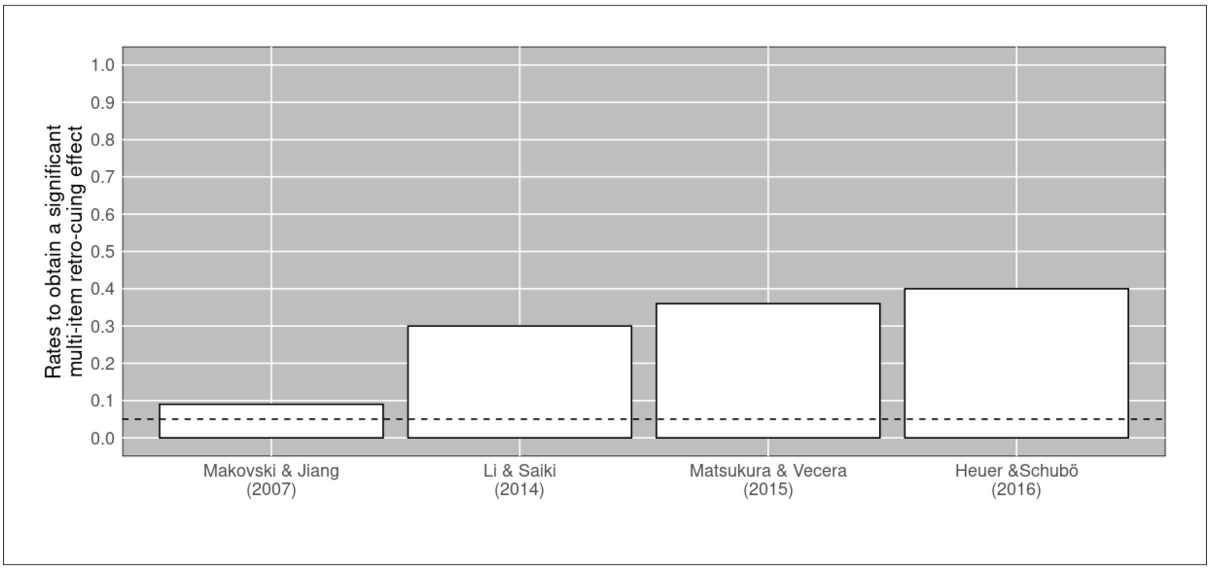
We note here that Heuer and Schubö (2016) did not include the single-item retro-cueing condition, in which case we computed the effect size of the multi-item (num of cues = 2 in this case) retro-cueing effect and used it as a conservative index for the single-item cueing effect size of the human participants. Of course, the real single-item retro-cueing effect size should be larger than the multi-item retro-cueing effect size. Thus, we obviously underestimated the size of the single-item retro-cueing effect size in this case, which is why our approach is conservative. The larger the single-item retro-cueing effect size, the higher the chance rates to detect the multi-item retro-cueing effect in the current Monte-Carlo simulation. In other words, underestimating the single-cueing effect size inevitably underestimates the chance rates to obtain statistical significance in simulation. Then, if the resultant chance rates are still found to be higher in such a conservative situation, then we can confidently argue that the rates of obtaining statistical significance in the experimental situation are much higher than our conservative estimate.

Finally, although Delvenne and Holt (2012) was one of the articles that argued to find the boundary condition for multi-item retro-cueing effect, we could not simulate their experimental situation because the detailed statistics for neither the single-item retro-cueing effect nor the multi-item retro-cueing effect were available (i.e., only *p* value was reported as $p < .05$).

Results and Discussion

Figure 1

Rates for Obtaining a Statistically Significant Difference in Accuracy between the Multi-item Retro-cueing Condition and the Neutral-cueing Condition in the Monte-Carlo Simulation.



Note. A dashed horizontal line indicates y = 0.05 (i.e., conventional alpha level)

Figure 1 shows the outcomes of the Monte-Carlo simulation. Specifically, the Y-axis is the number of the Monte-Carlo iterations that detected statistical significance between the multi-item retro-cueing condition and the neutral-cues condition where all the items are cued, divided by the total number of the simulation run (i.e., 100) in each scenario. Note that only one of the cued items was strategically focused on in this simulation. Thus, the y-axis shows the chance rates to draw an incorrect interpretation that participants were successfully attending to multiple items.

First, when simulating the initial study by Makovski and Jiang (2007), who did not find a significant multi-cueing effect, the estimated chance rates for misinterpretation were close to 5%. In contrast, when simulating the other studies, which found significant multi-cueing effects, the obtained chance rates for misinterpretation were much higher (30% ~ 40%) than the conventional alpha level (5%). This means that in the latter cases, one can find a statistically significant difference between the multi-cue and neutral-cue conditions at a chance rate of higher than 5% even if participants focused on only one of the cued items. Of course, we cannot confidently assert whether each participant in these studies covertly applied such a strategy. However, such higher chance rates are consistent with the fact that some studies obtained a statistically significant effect whilst others did not. The crucial difference between Makovski and Jiang (2007) and the other studies was the effect size of the single-item retro-cueing effect (i.e., single cue vs. neutral cues). As Table 1 shows, this effect size was much smaller in Makovski and Jiang (2007), meaning it was unlikely to generate a statistically significant multi-item retro-cueing effect even if a single-item focusing strategy was adopted in the multi-cues condition. The larger effect size in the remainder of the studies means there is more scope for a single-item focusing strategy to explain the higher accuracy observed in the multi-cue conditions.

Before simulating other scenarios, it is worthy to discuss whether the simplification of our model is justified, particularly regarding Equation (3). Specifically, we sampled the correctness value in the k-th trial (C_k) from the simple binomial distributions. One may justifiably criticize this approach in that various processes during the yes/no recognition judgment are assumed to be a black box in the current simulation. We acknowledge that our model did not specify the details of these processes in Equation (3) but argue that the output of the black box is correct. This is because we determined the probabilities of the binomial distribution (i.e., p_{focus}, and $p_{unfocus}$) based on the single-cueing effect size of the real human data in each study. In a single-cue trial, it is reasonable for a participant to use a single-item focusing strategy. Thus, as far as the outputs of the black box correctly simulate the single-cueing effect size of real participants, we can safely argue that the outcomes of the black box are close to that of the human outputs who employ a single-item focusing strategy. In a similar vein, we can justify our use of Equations 1-3 in simulating the set-probe testing method (i.e., number of probes > 1, see, Matsukura & Vecera, 2015). We acknowledge that the detailed processes should be different between a single-probe test and a set-probe test. However, once again, it is reasonable to assume that participants would naturally use a single-item focusing strategy in a single-cue trial even if a set-probe testing method is employed (Matsukura & Vecera, 2015). Then, as far as the probabilities of the binomial distributions (i.e., p_{focus}, and $p_{unfocus}$) are determined on the basis of the single-cueing effect of the participants from Matsukura and Vecera (2015), we can safely argue that the outcomes of Equation (3) reflect the real underlying (hidden) processes of single-item focusing in a set-probe test.

Lower Cue-Validity Level

Can this strategy-based account hold even if the cue-validity level is low? The simulation above assumed the case of the maximum cue-validity (100%). Therefore, the probability for one (j-th item) of the cued items to be probed (p_{probe}) was equal to one divided by the number of cues (Equation 2). However, some studies have investigated cueing effects when the cued items did not have a higher chance to be probed than the un-cued items (e.g., Experiment 3 of Li & Saiki, 2014). When simulating such a situation, one may think that Equation (2) should be changed to one divided by the number of studied items, rather than the number of cues. As a result, the simulation outcome would also change. However, we argue that the simulation outcome does not change. The reason lies in the accuracy scoring method in the multi-cued condition from the human experiments. Specifically, mean accuracy in the multi-cue condition per participant is calculated only from the cue-valid trials, where the cued item(s) are probed. In contrast, when an un-cued item is probed, the response in such a cue-invalid trial is not included when computing the mean accuracy of a participant in the multi-cued condition. This is a standard scoring method especially when contrasting accuracy in the multi-cue and neutral-cue conditions (e.g., Li & Saiki, 2014). As far as this standard scoring method is taken, Equation (2) holds without an amendment. This is because in the cue-valid trials, the chance for a single focused-item to be probed is equal to one divided by the number of cues. As we already showed in the simulation of Li and Saiki (2014) in Figure 1, the chance rate to obtain a statistically significant multi-cueing effect was much higher than 5%. In other words, a single-item focusing strategy is still an effective strategy to increase the accuracy in the multi-cues condition than the neutral-cues condition even when cue validity is low.

Interaction as a Counterargument for a Single-Item Focusing Strategy?

Finally, one can argue that boundary conditions (i.e., interactions with another factor) are incompatible with the single-item focusing strategy account. Specifically, if participants focused on only one of the cued items, then the conditions of another factor such as spatial positions/distances between cues should not have made any difference for the multi-item cueing effect (see, Heuer & Schubö, 2016, for such an explanation). However, if this strategy results in a statistically significant difference being obtained in a probabilistic manner, then it is also probabilistic to obtain a significant difference in one condition whilst not to obtain it in the other condition (i.e., the interaction). Thus, it is necessary to demonstrate actual probabilities to obtain such a discrepancy under the scenario when a single-item focusing strategy is employed.

The Monte-Carlo simulation strategy is as follows: We reiterated the abovementioned simulation of Heuer & Schubö (2016) once again, but this time the multiple-cue trials were divided (i.e., into two 'conditions') simply in terms of whether each trial was odd-numbered or even-numbered. Then, accuracy in each condition was compared to that in the neutral-cue condition. If a significant difference from the neutral-cue condition is detected in one condition whilst not in the other condition, then such a dissociation is regarded as an interaction. After repeating the same procedure 100 times, we calculated the rates for such a dissociation.

As a result, the rates for obtaining a dissociation was 0.3 (30%). We simulated the situation where a single-item focusing strategy was adopted, and the double-cue trials were randomly divided into two. As a result, there was a higher chance than 5% to obtain an interaction. Since both a focused item and a probed item were selected randomly from the cued items, it was also probabilistic either to obtain a significantly higher accuracy than the neutral-cues condition or not to obtain it.

Taken together, when interpreting a statistically significant multi-item retro-cueing effect, it is important to consider the extent to which a single-item focusing strategy can account for the outcome. The crucial factor is the effect size of the single-item retro-cueing effect. If it is too high, one needs to be cautious. Of course, there is no evidence to conclude that participants in past studies covertly applied such a strategy. It is also important to note that we are not arguing that genuine multi-item cueing is not possible. Instead, stronger support for a significant multi-item retro-cueing effect may be obtained by increasing the likelihood for attentional allocation towards multiple cued items, and capturing evidence of participants' efforts to do so.

Therefore, to collect such evidence, we conducted two new experiments with human participants by attempting to control gaze positions during the maintenance phase. Specifically, during the whole maintenance phase, participants were instructed to fixate their gaze across all the spatial positions where the cues had appeared. We know that participants can follow instructions to strategically adjust their attentional focus between items in working memory tasks (e.g. Allen & Ueno, 2018; Atkinson et al., 2022; Allen et al., 2024 for a review). The rationale behind applying such an approach in the present study is based on an active and facilitatory role of gaze position in memory after the offset of the studied items (Ferreira et al., 2008; Johansson & Johansson, 2014; Laeng et al., 2014; van Ede et al., 2019). Relatedly, Loaiza and Souza (2022) found that instructed gaze fixation to the cued positions led to higher recall accuracy of the cued (i.e., gazed) position. Evidence of direction of gaze fixation across cued positions would be a clear indicator that participants followed the instruction to attend all multi-cued items. We therefore further conducted Study 2B to replicate outcomes from Study 2A, and also to collect direct evidence for gaze positions through eyetracking. A significant multi-item retro-cueing effect in this context, along with evidence of gaze direction to multiple cued locations, would offer a stronger indication of real multi-item cueing effects than have been observed to this point.

Three points should be made regarding our gaze-based approach. First, this cannot firmly rule out the possibility that participants covertly focus attention on a single item, independent of gaze fixation. However, with evidence for gaze position, we can safely reject an account in which participants exclusively focused on only one of the multi-cued items. Namely, we can argue that at least an overt form (i.e., gaze-based) of attention is allocated to multiple representations. Second, gaze fixation at mutiple positions does not necessarily mean that the representations from those positions are activated simultaneously in each trial (see, Orts & Olivers, 2020, for raising this issue in the visual search domain). Rather, the representation for each item would be re-activated (or refreshed) sequentially as gaze is fixated to each position. This sequentiality vs. simultaneity question will be discussed further after analysis of the eyetracking data. Third, and relatedly, while gaze fixation at multiple positions clearly rejects a deliberate single-item focusing strategy, it does not necessarily mean that both cued items are activated and enhanced by gaze-based attention during the retention interval. Instead, only one of the fixated items might actually benefit from gaze-based attention. This issue will be further discussed by examining the relationship between accuracy and fixation duration.

Studies 2A and 2B: Human Experiments

Aim and Rationale

In Study 2A, we aimed to investigate the multi-item retro-cueing effect whilst explicitly instructing the participants to fixate their gaze across all the multiple cued-positions. In Study 2B, we aimed to replicate the findings of Study 2A as well as gleaning further evidence for fixating gazes across all the multiple spatial positions using eye-tracking. Our principal focus was on whether performance in the double-cue condition would be superior to that observed in neutral cue trials when a single-item focusing strategy, at least in the overt form (i.e., gazed-based attention), was explicitly discouraged. For the eye-tracking outcomes in Study 2B, we also examined whether gaze was directed to cued locations for a longer period than to uncued locations during double-cue trials.

Methods

Transparency, Openness (Data Availability, Pre-registration), and Participants

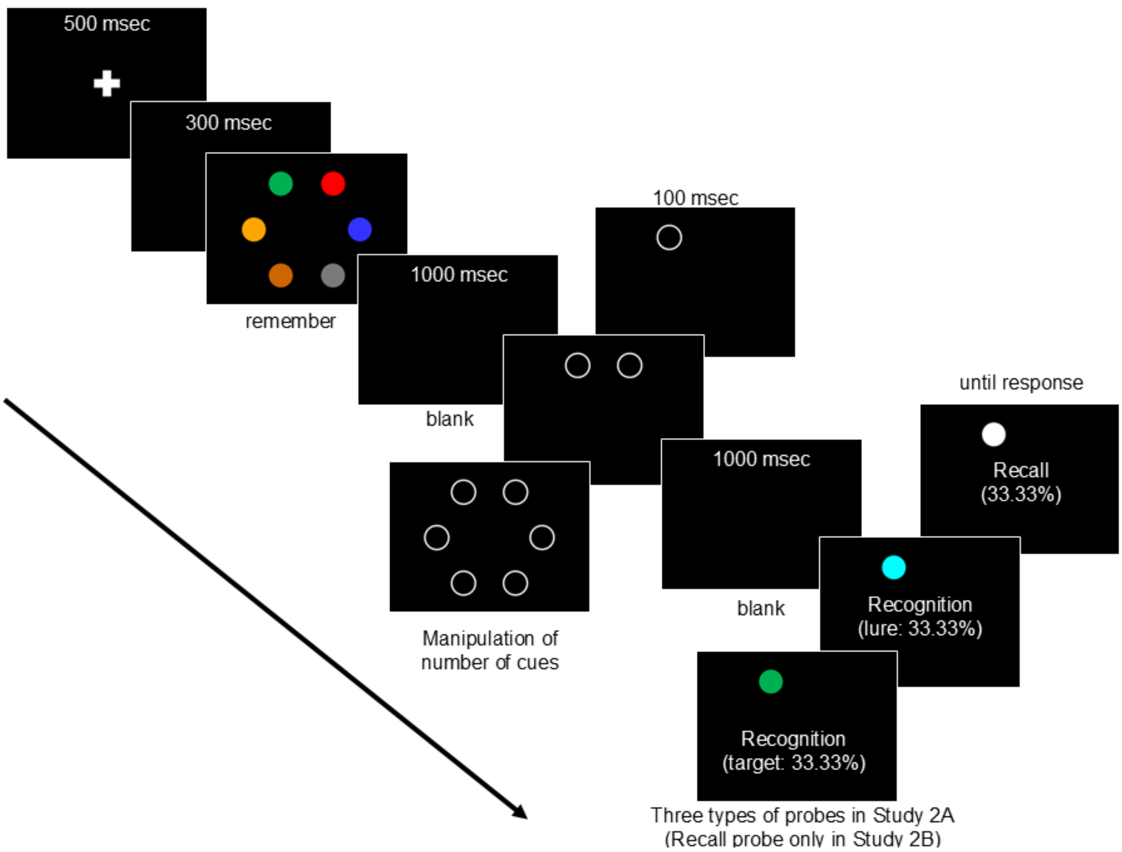
We report the sampling plan, all manipulations, and all measures in the study. All the data, analysis code, and research materials are available online (DOI 10.17605/OSF.IO/DRBXT). We used *R* (R Core Team, 2017) with *power.t.test* function to perform the power analysis. We preregistered the sampling plan and the analytic strategies of two studies beforehand (AsPredicted: https://aspredicted.org/OIA_IZA). However, for transparency, we note that the studies reported here were conducted in 2018, with subsequent delays arising due to COVID19 complications. In other words, the effect size used in the power analysis (i.e., within-group Cohen's *d* for the multi-cues vs. neutral-cues conditions) was determined based on the available literature at the time of pre-registration. Given the literature has developed since then (e.g., DiPuma et al., 2023), it is informative to conduct a sensitivity power analysis (e.g., Perugini et al., 2018) with the sample size that we collected. As a result, the minimum effect size that our study ($N = 32$) was sensitive to detect was Cohen's *d* of 0.51 (power = 0.8, alpha = 0.05). This effect size is common in the working memory literature. Thus, our original pre-registered sampling plan was appropriately powered to detect effect sizes of interest in these studies.

In each study, we continued data collection until 32 participants completed the experiment. There were 9 females and 23 males in Study 2A; 12 females and 20 males in Study 2B. The mean ages (and *SD*) were 19.15 (1.05) in Study 2A and 19.81 (1.06) in Study 2B. None of these participants were excluded due to the pre-registered criteria: i.e., participants whose mean response time (averaged over all the trials) was below 400ms; who did not complete the study; who encountered a PC problem; and whose overall mean accuracy was 2.5 SD above/below the mean in each condition. We did not pre-register any trial-level exclusion criteria. Thus, all the data were submitted to the analysis. All the participants were from Takachiho University, Japan. They took part in the 45-minute experiment and were paid (1,000 Japanese Yen) for their participation. All had normal vision and discrimination ability for the colors. The experimental protocol had been submitted to the dean of Takachiho University in advance, who was also the chair of the ethical committee and approved of the protocols. All the participants gave written informed consent.

Design, Materials, and Procedure

Figure 2

Flow of a Trial in Study 2A and Study 2B



Note. *In the recognition trials (i.e., a colored circle probe), the participants made an "old/new" judgment on the color of the target in that position via a keypress. In the recall trials (a white circle probe), the participants selected the color of the target in that position via a 9-alternative keypress (with different colored patches attached to each of 9 keys on the keyboard).*

Study 2A. This experiment followed a repeated-measures design, with the number of retro-cues (one, two, or six) as the within-participant factor. Testing was controlled using a HSP3 (Hot Soup Processor, ver.3) program (http://hsp.tv/). The materials and the flow of a trials (Figure 2) were the same as Makovski and Jiang (2007), except for the longer duration of the maintenance phase after the offset of the retro-cues, and the testing method (see later). Each trial began with a warning cross (500ms), followed by a blank screen (300ms). Then, on each memory display, six coloured circles (diameter = 1.31º) appeared for 1,000ms equidistantly on an invisible circle (diameter = 9.84º), centred around the screen center. The background color was black. The colors of the six circles were randomly selected without replacement from *green* (RGB= 0, 255, 0), *re*d (255, 0, 0), blue (0, 0, 255), *light blue* (0, 255, 255), *purple* (255, 0, 255), *yellow* (255, 255, 0), *gray* (127.5, 127.5, 127.5), *brown* (204, 102, 0), and *orange* (255, 165, 0). Before the experiment, all the participants were shown the nine coloured circles and confirmed that they were able to discriminate each color easily. After the offset of the studied items, a blank screen (1000ms) was inserted. Then, one, two, or six peripheral attentional cues appeared for 100ms in the form of open circles (diameter = 1.31º, line-color = white, filled color = black). Each represented the single-cue condition, multiple-cues condition, and neutral-cues condition, respectively. The number of trials for each of three cueing conditions was 45 in Study 2A (135 trials in total). The spatial positions of the cues were randomly selected from the six studied items' positions. The instruction was crucial in this study: Participants were informed that only the cued locations would be probed, and therefore they were instructed to focus on the items that had appeared at the retro-cued locations and to ignore the items at the un-cued locations. Thus, overall probability that a cued item would be probed was 100% (though this was obviously reduced for each individual item in the multi-cue conditions). Participants were also instructed that when two cues appeared, one of these cued positions was always probed, and therefore they should focus on *both* (this word was highlighted in red in the instruction screen) items that had appeared at the retro-cued locations. When all the six positions were cued (i.e., neutral-cues), the participants were instructed to focus on all the studied items as the cues did not predict the position of the probe. Moreover, the participants were instructed to fixate their gaze on all the cued positions during the maintenance phase (i.e., after the offset of cues and before the onset of a probe). To facilitate this gaze distribution, we made this duration longer (1,000ms) than Makovski and Jiang (2007), who set this duration at 400ms.

At the end of each trial, a single probe appeared at one of the cued position(s). Whilst Makovski and Jiang (2007) used a single-probe recognition paradigm, we added a recall probe as well, where a participant was required to recall the color in Study 2A. The response method involved selecting the color of the target item by a keypress. To aid this response, different colored patches were attached to each of 9 keys on the keyboard (“1” ~ “9”). This was done to address the issue of sensitivity. Compared to the 2-alternative choice in probe recognition, chance rate is much lower for recall from nine-possible colors. Therefore, there were three types of probes in the test phase: recall probe (i.e., a white, filled circle), a positive recognition probe (i.e., a filled circle whose color was the same as the studied item in the probed position), and a negative recognition probe (i.e., a filled circle whose color was randomly selected from one of the three, un-presented colors at that trial). These three types of probes appeared with equal probability (i.e., 33.33% for recall and 66.66% for recognition), randomly intermixed within blocks. If a colored circle appeared as a probe, then participants were required to press “s” (same) or “d” (different). If a white, filled circle appeared as a probe, then the participants were required to select the color of the target item in that position by a keypress. The probe remained on screen until a keypress response was made. Speed was not required, and accuracy was emphasized. At every trial, participants were engaged in an articulatory suppression (saying ‘da’, ‘da’, …) until a key press.

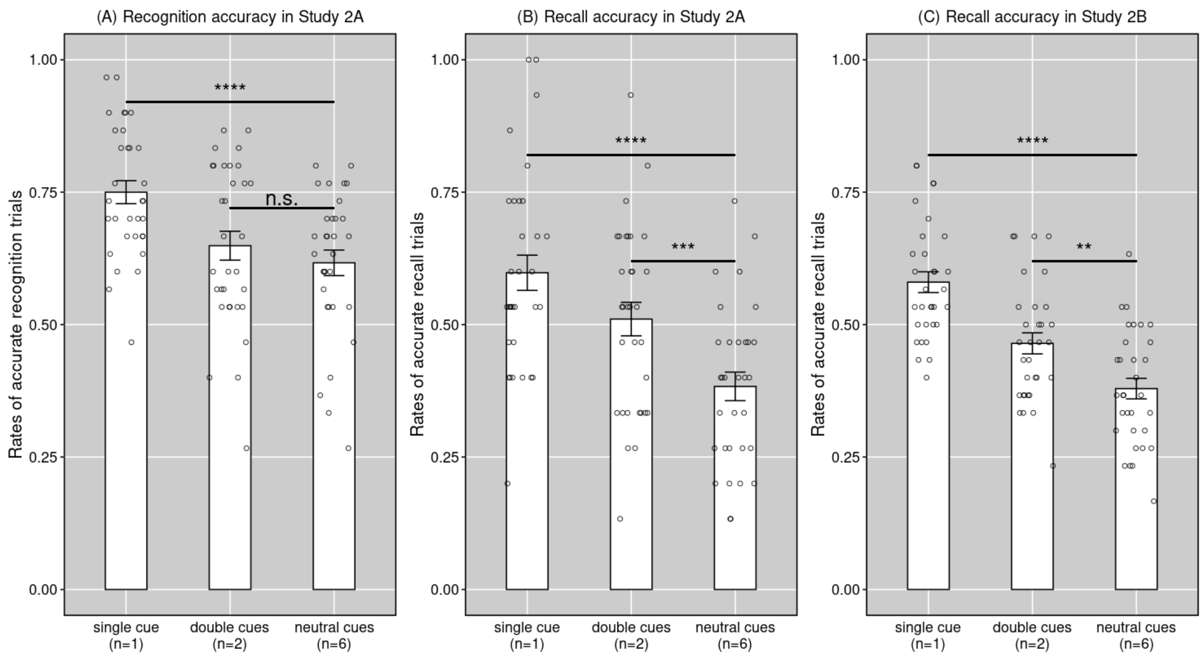
Study 2B. In Study 2B, we focused on the recall testing method to optimise sensitivity and avoid the higher chance rate of the recognition probe. Thus, all the probes were white, filled circles. The number of trials was 90 in total (i.e., 30 trials for each of the single-cue, multi-cues, and neutral-cues conditions). Moreover, in Study 2B, we measured the gaze fixations by using an eye-tracker (Tobii Pro Glasses 2) during the whole experimental session. The sampling rate was 50Hz.

Results and Discussion

Accuracy

Figure 3

Mean Accuracy and Individual Plots in Studies 2A and 2B



Note. Panel (A): Recognition trials in Study 2A. Panel (B): Recall trials in Study 2A. Panel (C) Recall trials in Study 2B. Individual dots indicate individual data. Y-axis error bars represent standard error of mean. * $p < .05$. ** $p < .01$. *** $p < .001$. **** $p < .0001$. n.s. non-significant.

Figure 3 shows the mean accuracy rates in Studies 2A and 2B (left panel: recognition trials; middle panel: recall trials in Study 2A; right panel: recall trials in Study 2B). A series of one-way repeated-measures ANOVA on each accuracy rates revealed significant main effects of cue condition on recognition accuracy in Study 2A (Panel A), F (2, 62) = 15.211, $p < .0001$, on recall accuracy in Study 2A (Panel B), F (2, 62) = 20.594, $p < .0001$, and on recall accuracy in Study 2B (Panel C), F (2, 62) = 33.481, $p < .0001$.

A pre-registered planned comparison revealed that a single-item retro-cueing effect (single-cue vs. neutral-cues) was significant on recognition accuracy in Study 2A, t (31) = 5.249, $p < .0001$, Cohen's d = 1.031, its 95% CI = [0.495, 1.567], on recall accuracy in Study 2A, t (31) = 6.722, $p < .0001$, Cohen's d = 1.258, 95% CI = [0.706, 1.810], and on recall accuracy in Study 2B, t (31) = 8.020, $p < .0001$, Cohen's d = 1.832, 95% CI = [1.229, 2.435]. Thus, a robust single-item retro-cueing effect was replicated in our study, for both experiments.

Regarding the multi-item retro-cueing effect (multi-cues vs. neutral-cues), a pre-registered planned comparison revealed that the accuracy difference was not significant on the recognition measurement in Study 2A, t (31) = 1.030, $p = .311$, *n.s.*, Cohen's d = 0.222, 95% *CI* = [-0.281, 0.726]. However, its effect was significant on the recall measurement both in Study 2A, t (31) = 3.747, p = .0007, Cohen's d = 0.768, 95% *CI* = [0.247, 1.289] and Study 2B, t (31) = 3.186, p = .0032, Cohen's d = 0.772, 95% *CI* = [0.251, 1.294]. A pre-registered internal meta-analysis (Ueno et al., 2016) across the two studies by a random-effect model (Cumming, 2012) found that the integrated effect size was Hedges' g of 0.663, 95% CI = [0.397, 0.928] on recall accuracy. Thus, whilst single-probe recognition could not detect a significant multi-item retro-cueing effect, probed recall robustly did. Our preferred interpretations regarding the differences between testing methods focus on their differing chance levels and the relative involvement of internal control. A 2-alternative (old/new) probe recognition test would be relatively noisy as the chance level is 50%. In contrast, the 9-alternative color recall has a much lower chance level (around 11%), making a guessing strategy far less effective and potentially increasing task sensitivity. Indeed, in the recognition trials, the mean accuracy minus 1*SD* was below 50% (chance level) both in the multi-cues condition (49.62%) and in the neutral-cues condition (47.53%), indicating that performance levels in the recognition trials were relatively close to floor. Moreover, relative to recognition, recall is likely to involve a more effortful, demanding, and internally driven retrieval process (e.g. Allen et al., 2018; Craik & McDowd, 1987; Craik et al., 1996), and so could be more influenced by selective attentional effects, though retro-cue benefits are of course observed on recognition measures (Souza & Oberauer, 2016, for a review). Related to this, small effects may be easier to be detected using more difficult tasks, in general.

Next, we should also note a possibility that a multi-item retro-cueing effect might be obtained only when the two cues were presented in adjacent positions. If this is the case, a more straightforward account for the effect would be chunking/grouping of the two cued items into one rather than assuming successful attention to two items. Indeed, Heuer and Schubö (2016) observed that spatial distance of the cued items can modulate the multi-item retro-cueing effect. To address this concern, we divided the double-cue trials in terms of whether the cued positions were adjacent or not. As a result, there was not a significant difference on accuracy between the adjacent trials (M = 49.47%, SD = 15.97%) and the non-adjacent trials (M = 44.44%, SD = 13.75%), t (31) = 1.476, p = .150, *n.s*. Accuracy in each case was significantly higher than that in the neutral-cue condition (ps < .05). Therefore, although there might be a chunking/grouping mechanism for the cued items in case of the adjacent positions, such an account cannot fully explain the higher accuracy in the double-cue condition than the neutral-cue condition.

Monte-Carlo Simulation on Our Accuracy Pattern

Finally, it is useful to establish whether the significant findings observed in our Studies 2A and 2B may also be explained in terms of a single-item focusing strategy as was suggested in the Study 1 simulations. Thus, we conducted the Monte-Carlo simulation with the experimental parameters derived from our Studies 2A and 2B. As a result, the chance rates were 44% to obtain a significant multi-cueing effect when only one of the cued items was strategically focused on, an estimated rate that was equivalent to those we established for existing literature. Thus, despite the instruction to attend multiple items in terms of eye movements, the accuracy patterns could in principle be caused by single-item enhancement. We return to this issue after analyzing the gaze data in the next section.

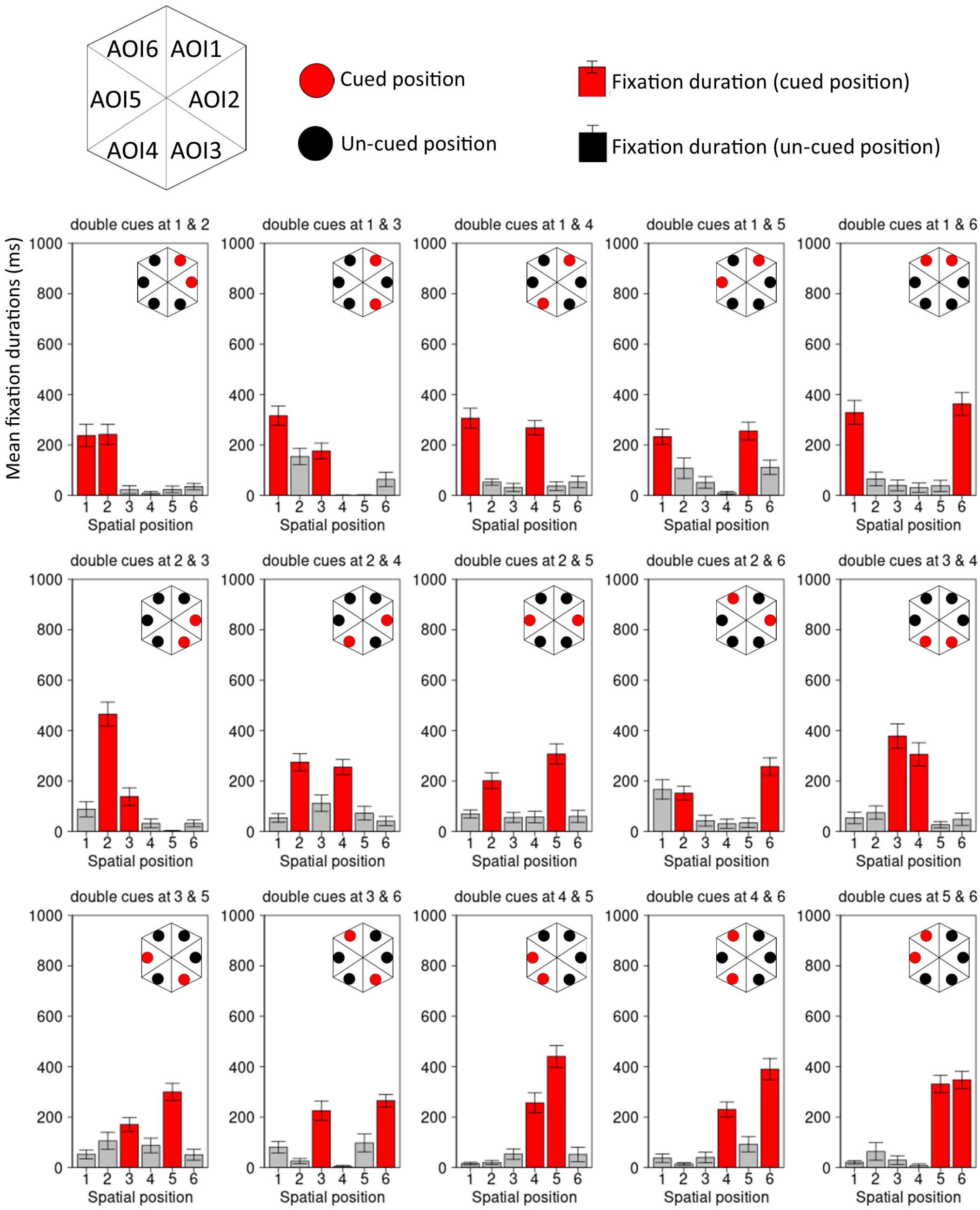
Fixation

Gaze Analysis. We used Tobii Pro Lab Analyzer software (Tobii Pro AB, Stockholm, Sweden) for eye-tracking analysis in Study 2B, following the method of Jongerius et al. (2021). Tobii Pro Glasses 2 captured the visual environment in which each participant was looking and Tobii Pro Lab Analyzer allows analysis of this gaze information. The manual analysis on this software starts with taking a snapshot of the representative video frame that captures the areas that one wants to analyze (i.e., PC screen in our case). Then, onto this snapshot, we manually drew six regions of interests, using an area-of-interest tool. In our case, we first drew an invisible hexagon on the PC screen within the snapshot, such that tangents of the six circles (i.e., six targets or six cues) were the mid-points of the six sides of the invisible hexagon. Then, this hexagon was divided into six isosceles triangles whose three apexes were the center of the invisible hexagon (i.e., center screen) and the apexes of the hexagon, respectively. The resultant six isosceles triangles were the six areas of interest and are shown in the top-row of Figure 4.

Once the areas of interests are drawn on the snapshot, then one can map eye gaze data during the time of interests (in our case, 1,000ms between the offset of the cues and the onset of a probe) onto the snapshot by using the Automatic Mapping Function. Then, various gaze information can be estimated in each time window per an area of interest. Since we aimed to visualize the distribution of gazes across the cued positions, we estimated the duration of fixations in each area of interest during the time of interest. The definition of a fixation varies depending on studies, but we used the default I-VT (Velocity-Threshold identification fixation filter) of Tobii Pro Lab Analyzer (Komogortsev et al., 2010; Salvucci & Goldberg, 2000). Specifically, the I-VT fixation classifier applies an angular velocity threshold (30 degrees/second) on each data point. Data points with angular velocity below the threshold value were classified as being part of a fixation.

Figure 4

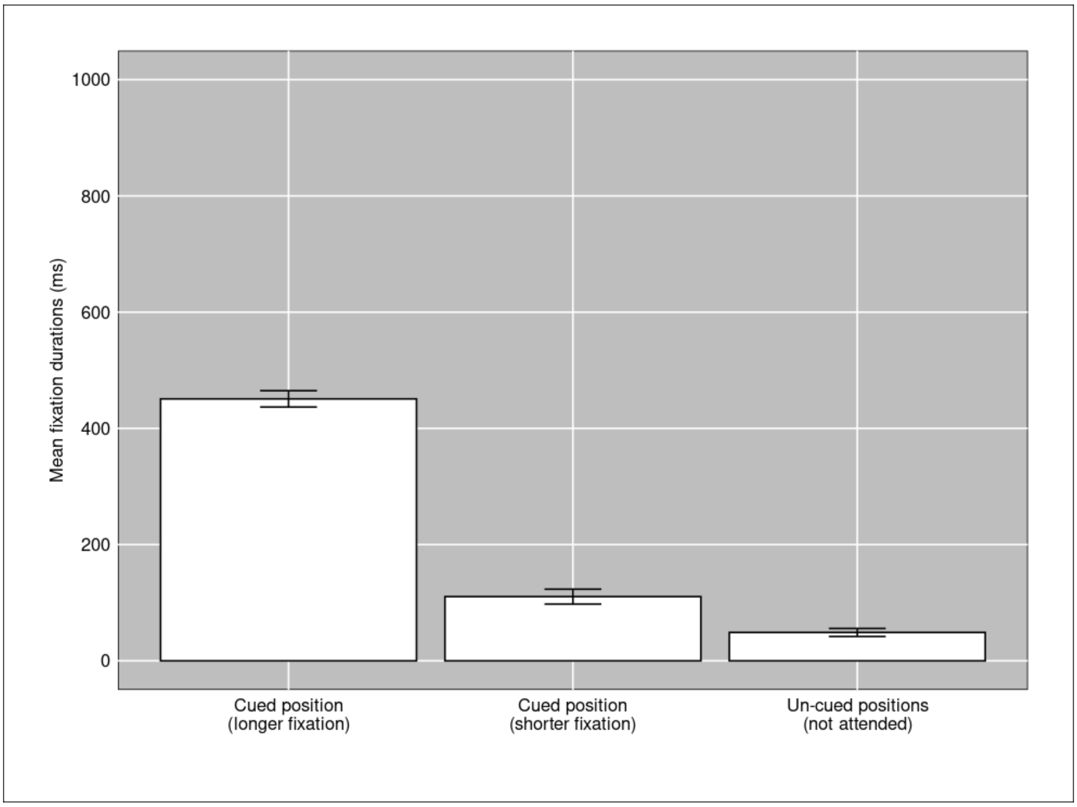
Areas of Interests (top) and Fixation Durations in Each Area in the Double Retro-Cues Condition.



Outcome. Figure 4 shows the areas of interests where fixation durations were analyzed (top row), and the duration of fixations in each area after double-cues were presented. The figure for the single-cueing condition is available online (DOI 10.17605/OSF.IO/DRBXT). The fixation data were averaged over the trials per participant and then averaged over participants. The bars and the circles are shown in red when a cue had appeared at that spatial position. This clearly shows that the mean fixation values in the multi-cued positions (red) were higher than the un-cued positions. However, to echo the focus of the Monte Carlo simulation in Study 1, this *averaged* data pattern over the trials might emerge even if a participant focused on only one of the multiple cued positions within a trial. For example, suppose positions 1 and 2 were cued, and then a participant's gaze was only fixated to position 1 in trial n and position 2 in trial $n+1$. Averaging these two trials would still result in longer fixation durations for positions 1 and 2, relative to other positions, again potentially producing artificial evidence for multi-cueing. It is important to show that each of the two cued positions received longer fixation durations than the uncued positions within a trial. Thus, we sorted the fixation data on the two cued positions within a multi-cued trial, and categorized the data as follows: (1) the longer fixation duration among the two cued positions, (2) the shorter fixation duration among the two cued positions, and (3) the fixation duration for the un-cued positions. Figure 5 shows the mean of these sorted and categorized fixation durations, averaged across trials and participants. Interestingly, the mean gaze fixation durations on the two cued positions were typically not equivalent (i.e., 450.74 ms for the longer fixation and 110.51 ms for the shorter fixation). Nevertheless, even the shorter-fixated cued position received a longer duration of gaze fixations than the uncued positions (48.83ms), t (31) = 3.280, p = .003, Cohen's d = 1.065, 95%CI = [0.526, 1.603]. Therefore, each of the cued positions received longer fixation durations than the un-cued positions, indicating participants followed the instruction to attend both cued positions. In other words, in contrast to the prior literature, we can confidently argue that participants deliberately avoided a single-item focusing strategy, at least in terms of overt (i.e., gaze-based) attention. We discuss this issue further in the general discussion in conjunction with the simulation outcomes for our accuracy data.

Figure 5

Mean Fixation durations on the two cued positions (sorted by fixation length within a trial) and those on the uncued positions.



Note. Y-axis error bars represent standard errors of means

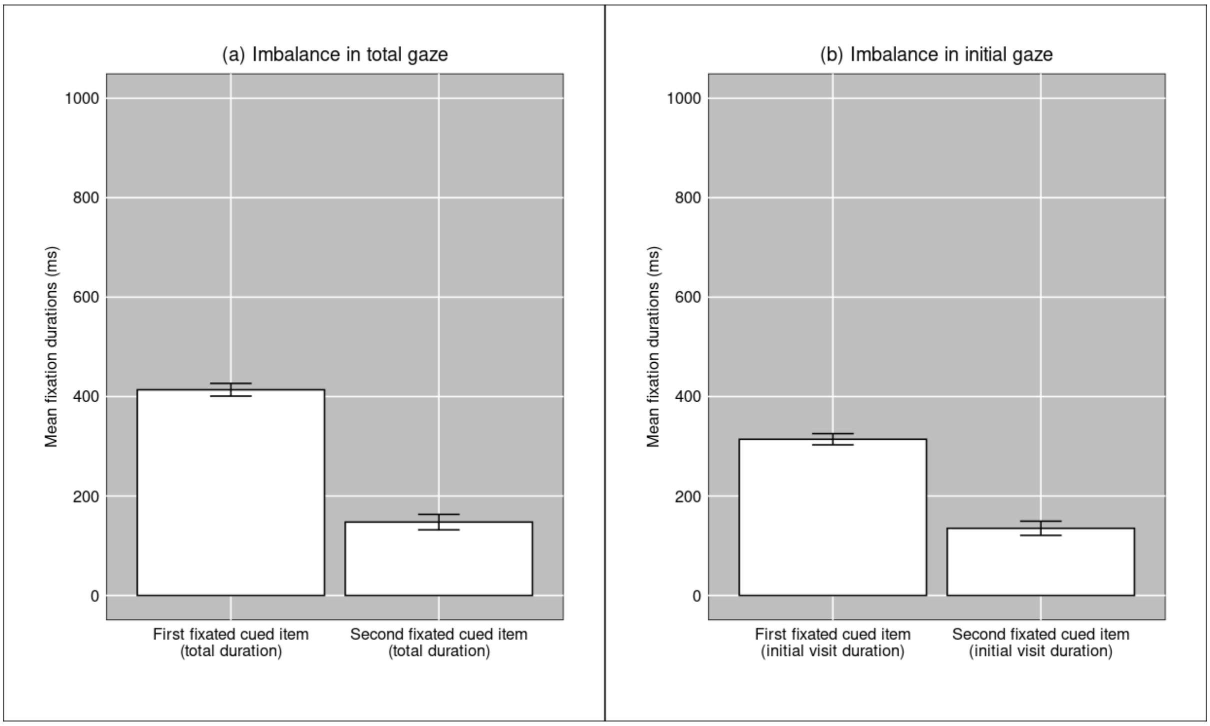
The imbalanced fixation durations for the cued items in Figure 5 indicate that although multiple cued items were being attended, rather than allocating equal attention to each cued item, a spontaneous form of prioritization may have been applied to one of the two cues. This might be part of the reason why the literature has difficulty in finding reliable multi-item retro-cueing effect, particularly when retro-cues were presented simultaneously (e.g., Makovski & Jiang, 2007; see also Delvenne & Holt, 2012; Heuer & Schubö, 2016; Matsukura & Vecera, 2015, for a significant multi-item retro-cueing effect only when a specific boundaty condition was satisfied).

Further Analyses on the Imbalanced Fixation Durations between the Cued Positions

The imbalanced fixation durations between the two cued positions motivated us to further investigate the differences between the cued positions. Thus, although not pre-registered, we conducted the following exploratory analyses. The first involves the order of gaze visits. Specifically, we investigated whether the first-fixated cued item or the second item tend to receive the longer fixation. The left panel of Figure 6 shows the total gaze fixation durations for the multi-cued spatial positions, divided in terms of whether each cued position was first-fixated or second-fixated. As a result, the first-fixated cued position received approximately 410ms of fixation durations in total (averaged over the trials and over the participants) whilst the second-fixated cued item received approximately 140ms of fixation durations. Thus, participants spent longer looking at the first-fixated position. Related to this, the right panel of Figure 6 also shows the split data in terms of the order of the gaze fixation visit, but this one shows only the fixation durations during the initial visit to each position. Extracting the fixation durations only for the initial visit time is informative as some participants may visit the same position multiple times during the retention interval. As a result, the gaze fixation duration in the initial visit was approximately 300ms on average in the first-fixated cued position whilst 135ms on average for the second-fixated cued position. In other words, the difference between the left and the right panels of Figure 6 indicates the gaze fixation duration at the second (and subsequent) visit to each position (i.e., approximately 110ms on average in the first-fixated cued position whilst only 5ms in the second-fixated cued position). Taken together, we can characterize gaze patterns as follows; the first-fixated position received a longer duration of gaze fixations (300ms) at the initial visit, after which the second one received a relatively shorter duration of gaze fixations (135ms on average), after which gaze briefly returned to the first position (110ms on average – i.e., the difference between the left and the right halves of Figure 6). Thus, overt gaze-based retrospective attention seems to be allocated to multiple positions sequentially.

Figure 6

Mean fixation durations to the two cued positions (divided in terms of the order of fixation visit within a trial)



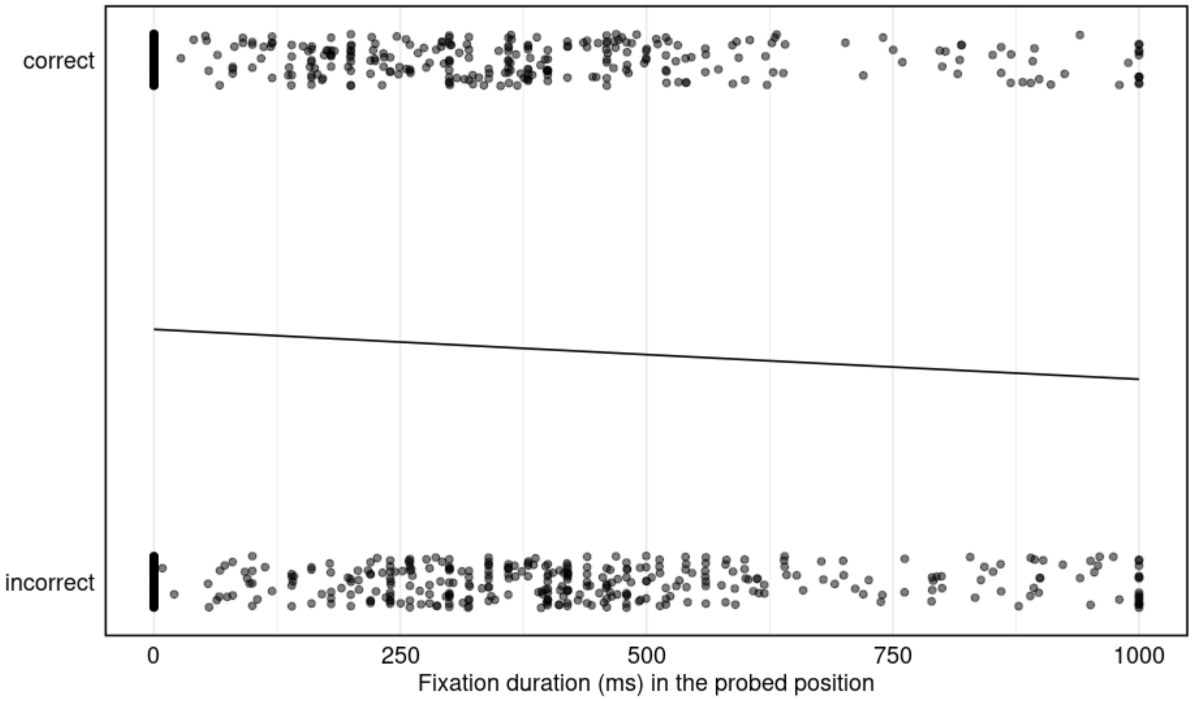
Note. Panel A (left) shows the total fixation duration in each trial; Panel B (right) shows fixation duration at the time of initial visit to each position. Y-axis error bars represent standard error of mean.

A tempting idea from these outcomes is that the first-fixated item, whose position received longer fixation durations, may show higher accuracy when it is probed. We tested this in two ways. First, we split the accuracy data in the multi-cues condition in terms of whether the probed position was first-fixated (*Mean* accuracy = 46.90%, *SD* =13.89%) or second-fixated (*Mean* = 45.47%, *SD* =14.12%). A paired *t*-test did not detect a significant difference, $t\,(31) = 0.514$, $p = .611$, $d = 0.102$, 95%*CI* = [-0.40, 0.60]. Secondly, Figure 7 plots the relationship between accuracy in each trial (a binary measure) and the total fixation duration on the probed position, including only the cued and probed position. To examine the effect of fixation duration, we conducted generalized linear mixed-effect modelling. The fixed effect was the fixation duration; the response variable was accuracy; the random variables were participants (random-intercept and random-slope) and the probed position (random-intercept). As a result, fixation duration did not significantly predict the log-odds of accuracy: coefficient = -0.0003, *z*-value = -1.438, $p = .150$. The black curve in Figure 7 represents the logistic curve (fixed-effect only) estimated by the GLMM, appearing as a straight line due to the non-significant effect of the predictor. The non-significant effect of predictor was found even when the data were split and analyzed by probed spatial position (available online: DOI 10.17605/OSF.IO/DRBXT). Taken together, our data suggests that fixation duration did not predict accuracy for cued items, and relatedly, the mean accuracy in the first-fixated position, which received longer fixation durations, was not significantly different from that in the second-fixated position.

Importantly, the equivalent accuracy rates between the first-fixated position and the second-fixated position despite their imbalanced fixation durations can help inform regarding the underlining processes in multi-cueing effect. As mentioned in the introduction of Study 2, fixating gazes at both positions demonstrates that participants made effort to avoid a single-item focusing strategy, but this does not necessarily mean that both items benefited from gaze-based attention. Instead, only one of them might benefit. However, if this were the case, it would be expected that the item in the longer-fixated position would be enhanced, which was not the case. Instead, our interpretation is that both items were enhanced by the sequentially allocation of gaze-based attention. Rather than gaze order or duration, gaze visitation towards each position determines the enhancement of memory accuracy.

Figure 7

Trial-by-trial scatterplots for the relationship between fixation duration on the probed position and its accuracy.



Note. Each point was adjusted with a small amount of random jitter (along the y-axis) to reduce overlap. The line represents the logistic regression curve (see main text).

General Discussion

Several studies have investigated the multi-item retro-cueing effect on visual working memory. However, after Makovski and Jiang (2007)'s initial study, almost all the studies have speculated on the reasons for the inconsistent effects that have observed, and some have attributed them to methodological variations between studies. We hypothesized that a single-item focusing strategy might be one reason for the inconsistencies and aimed to estimate the chance rates to obtain a statistically significant multi-item retro-cueing effect despite the use of a single-item strategy. Study 1 employed a Monte-Carlo simulation and revealed that the estimated chance rates were much higher than the conventional alpha level (5%) under the experimental settings of the existing literature. Such a high rate is consistent with the presence of mixed outcomes in the literature. The estimated chance rates represent the rate of possible misinterpretation if statistical significance is interpreted as evidence for retrospectively attending multiple items. Studies 2A and 2B investigated the multi-item retro-cueing effect in human participants, but this time a single-item focusing strategy was discouraged by instructing distribution of gaze fixations across all the cued positions. Eye-tracking in Study 2B confirmed that participants followed this instruction in each trial (see Figure 5). Thus, we collected plausible evidence that participants at least overtly attended to multiple items in each trial. Moreover, although a 2-alternative recognition test could not find a significant multi-item retro-cueing effect (indicating a relative lack of sensitivity), this effect was found in cued recall across the two studies.

Overall, our simulation indicates that prior evidence for multi-item cueing might often reflect a single item strategy. Possibly stronger evidence for a significant multi-item retro-cueing effect can be found using cued recall and the explicit direction to fixate on all cued locations. Eye-tracking confirms this, while also uncovering more complex patterns of gaze behaviour that do not directly map onto memory accuracy. Our study therefore at least somewhat reduces the possibility for a single-item focusing strategy account to explain the multi-item retro-cueing effects. However, these findings come with the caveat that simulation of our own accuracy data also indicated a relatively high chance rate for the apparent multi-cue effect on accuracy to in fact reflect a covert single item attentional focus within working memory. We will return to this possibility at the end of the discussion.

Among counterarguments or concerns regarding the explanatory power of a single-item focusing strategy, two to consider are cue validity level and possible interactions with other factors. First, is a single-item focusing strategy effective even when cue validity level is low? Our Monte-Carlo simulation demonstrated that as far as the cue-valid trials are scored, the chance rates to obtain a significant multi-item retro-cueing effect by this strategy was still high even in the lower cue validity level (e.g., Li & Saiki, 2014). Secondly, one may be tempted to exclude this strategy account if the multi-item retro-cueing effect is modulated by another factor (i.e., is involved in an interaction), such as the spatial positions/distances of the cues (e.g., Heuer & Schubö, 2016). However, we demonstrated a high chance rate for a single-item focusing strategy to result in a significant multi-item retro-cueing effect in one condition alongside a non-significant effect in the other condition (i.e., to produce an interaction). Therefore, one should be cautious about interpreting multi-item retro-cueing effects even if these effects appear to be modulated by another factor.

When developing optimal measures to detect, analyze, and interpret multi-item retro-cueing effects, what factors do we need to consider? First, task instructions should discourage the use of a single-item focusing strategy as much as possible, and one should obtain as much information as possible regarding the observance of the instruction, including gaze fixations. Another way might be to use a value-based prioritization instruction (e.g., Allen & Ueno, 2018; Atkinson et al., 2018; Hu et al., 2014; Allen et al., 2024, for a review) rather than cueing. Value-based prioritization may represent a lenient form of selective attention (Oberauer, 2019), and participants are assumed not to completely neglect the unprioritized items (Allen et al., 2024). For example, Allen and Atkinson (2021) found some evidence for retrospectively applied prioritization of the most recently encountered item in a sequence, though note that retrospective value effects may be somewhat smaller in magnitude than those of predictive retro-cueing (Hautekiet et al., 2024). Investigating possible multi-item retro-prioritization effects (i.e., two items receiving a higher reward) could be a useful future development.

A second factor to bear in mind is the size of the single-cueing effect (i.e., single-cue condition vs. neutral-cues condition). The larger this effect is, the stronger the explanatory power of a strategy-based account is (see Table 1 and Figure 1). Interestingly, the data from a recently published article (DiPuma et al., 2023) seems to be consistent with this relationship. Specifically, three experiments in DiPuma et al. (2023) could not find a significant multi-item retro-cueing effect on color/orientation memory accuracy (i.e., the absolute angular difference). However, if we look at the precision measurement (i.e., how precisely one can recall the feature of the probed stimulus at a given trial: Bays et al., 2009), then it seems to be higher in the multi-cue condition than the neutral-cue condition in Experiments 1 and 3, but such a tendency was not observed in Experiment 2. Interestingly, the effect sizes of the single-cueing effect were large in their Experiments 1 and 3 but small in Experiment 2.

We observed that overt gaze-based attention was allocated to multiple cued positions. Additionally, our exploratory analyses revealed that gaze fixation durations were imbalanced between the two cued positions and that a first-fixated position received longer fixation durations at the time of initial visit than the second-fixated item. Moreover, eye gaze appeared to return to the first-fixated position, after the second-fixating position. Thus, overt gaze-based retrospective attention is not allocated simultaneously, and is instead allocated sequentially to each of the cued items. Furthermore, fixation duration did not predict memory accuracy. Our interpretation is that gaze visit determines the cue-based enhancement of accuracy, rather than gaze duration. Once gaze is fixated to each position, refreshing processing (Souza et al., 2015) for that item starts. We note here that the non-significant fixation-accuracy relationship is inconsistent with Loaiza and Souza (2022, Study 1B), who found a positive relationship between the fixation durations and the acccuracy when eye gaze to the sequentially-presented cued positions was explicitly instructed, though they presented multi-cues sequentially whilst our study presented them simultanenously. Moreover, Loaiza and Souza used a continuous scale for reporting the color of a memorized item whilst we used a 9-alternative recall measure. These methodological differences may help explain the discrepancies in findings.

What other implications might we draw regarding possible mechanisms underlying retro-cueing effects? First, Makovski and Jiang (2007) discussed that a central bottleneck operates in consolidation of visual working memory, such that only one item at a time is processed and the consolidation process interferes with the maintenance of other items (Griffin & Nobre, 2003). We found a clear difference in accuracy between the single-cue condition and the double cue condition. Thus, there may indeed be some forms of mutual interference (or cost) to retrospectively attend multiple items (though it should be noted that this comparison is confounded by probe validity and was not a primary focus of the current work). Secondly, other studies provide more complex theories regarding working memory and attention by providing boundary conditions for the multi-item retro-cueing effects (Delvenne & Holt, 2012; Heuer & Schubö, 2016; Matsukura & Vecera, 2015), but such complexity appears to be unnecessary. Moreover, some studies have found positive effects of sequentially presented multiple retro-cues (Li & Saiki, 2014; Souza et al., 2015, but see Van Moorselaar et al., 2015). Therefore, one might be tempted to speculate that different principles apply depending on the presentation format of the cues. However, once again, such complexity is likely unnecessary. Thus, an account for single-item retro-cueing effects may be extended to multi-cueing without additional complexity. For example, Souza and Oberauer (2016) identified the following four accounts to explain a retro-cueing effect:

retrospectively attended representations are strengthened; un-attended items are removed from working memory; a head-start retrieval is provided in the accumulation of evidence for the attended items; and attended items are more protected from interference. Although our studies cannot differentiate between these accounts, we would argue that they can potentially be extended to explain multi-item retro-cueing effects without additional assumptions.

However, despite the trial-based eye-tracking evidence illustrating direction of spatial attention toward the locations of multiple cued items, simulation indicates that the overall accuracy patterns, which were aggregated across trials and participants, could in principle be produced by single item enhancement on each trial. This is because we cannot argue with 100% confidence that overt eye-movement necessarily translates into genuine multi-item effects in working memory in a way that has measurable benefits on recall performance (see also Figure 7). Thus, other forms of single-item focusing strategy rather than gaze-fixation might have contributed to aggregated accuracy pattern. For example, one could argue that participants looked at both positions because the experimenter instructed them to do so, but that directed gaze did not play much of a role in enhancing accuracy. Instead, enhancements in mean accuracy over trials in the multi-cue condition were driven by a covert single item focusing strategy. Under this account, although participants clearly observed the instruction to allocate overt gaze-based attention to multiple items, covert attentional processing within working memory did not follow on from this. Such an account is inconsistent with findings in the literature that have demonstrated an active and facilitatory role of gaze position in memory after offset of studied items (Ferreira et al., 2008; Johansson & Johansson, 2014; Laeng et al., 2014; van Ede et al., 2019, see also Loaiza & Souza, 2022, for the effect of instructional gaze), but we cannot completely reject it as a possibility. This shows that, even after adjusting task context and measurement methods to increase the probability of multi-cue implementation, it remains challenging to clearly adjudicate between single and multiple item interpretations. Thus, we see this as a call for the field to develop more sophisticated and appropriate methods to derive more confident conclusions regarding genuine multiplicity in working memory. Though not providing a perfect solution, our explicit instruction and gaze-monitoring approach forms the first step to disentangle at least an overt gaze-based allocation of attention from single-item focusing strategies. Future studies should aim to specify and de-confound other forms of single-focus strategies to isolate genuine multi-cueing effects.

Conclusions

There is mixed evidence in the literature regarding whether multiple retro-cues can facilitate working memory for each of the cued items. Our Monte-Carlo simulation indicated that a potentially misleading multi-cueing effect might emerge using a single item focus strategy, which could go some way to accounting for the inconsistent evidence to date. Two experiments then demonstrated that a multi-item cueing effect in aggregated accuracy data could indeed be observed with explicit instruction to direct gaze to all cued locations. Eye-tracking confirmed engagement with this instruction. However, simulation suggests that our observed accuracy patterns could in principle be produced by single-item enhancement on each trial. Thus, future studies should aim to develop improved methods to disentangle possible multi-item cueing from other forms of single item focusing strategies. Nevertheless, with explicit instruction and gaze monitoring, the current study forms the first step to isolate genuine retro-cueing multiplicity in working memory.

References

Allen, R. J., & Atkinson, A. L. (2021). Retrospective and Prospective prioritization in visual working memory. *PsrArXiv*. https://doi.org/10.31234/osf.io/4x8zu

Allen, R. J., Atkinson, A., & Hitch, G. J. (2024). Getting value out of working memory through strategic prioritisation; implications for storage and control. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/17470218241258102

Allen, R. J., & Ueno, T. (2018). Multiple high-reward items can be prioritized in working memory but with greater vulnerability to interference. *Attention, Perception & Psychophysics*. https://doi.org/10.3758/s13414-018-1543-6

Atkinson, A. L., Berry, E., Waterman, A. H., Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2018). Are there multiple ways to direct attention in visual working memory? *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/nyas.13634

Allen, R. J., Hitch, G. J., & Baddeley, A. D. (2018). Exploring the sentence advantage in working memory: Insights from serial recall and recognition. *Quarterly Journal of Experimental Psychology*, *71*(12), 2571-2585. https://doi.org/10.1177/1747021817746929

Atkinson, A. L., Oberauer, K., Allen, R. J., & Souza, A. S. (2022). Why does the probe value effect emerge in working memory? Examining the biased attentional refreshing account. *Psychonomic Bulletin & Review*, *29*(3), 891-900. https://doi.org/10.3758/s13423-022-02056-6

Barras, C., & Kerzel, D. (2017). Salient-but-irrelevant stimuli cause attentional capture in difficult, but attentional suppression in easy visual search. *Psychophysiology*, *54*(12). https://doi.org/10.1111/psyp.12962

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10). https://doi.org/10.1167/9.10.7

Craik, F.I.M., Govoni, R., Naveh-Benjamin, M., & Anderson, N.D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General, 125,* 159-180.

Craik, F. I., & McDowd, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 474.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Delvenne, J. F., & Holt, J. L. (2012). Splitting attention across the two visual fields in visual short-term memory. *Cognition*, *122*(2), 258–263. https://doi.org/10.1016/j.cognition.2011.10.015

DiPuma, A., Lockhart, H. A., Emrich, S. M., & Ester, E. F. (2023). Retrospective cue benefits in visual working memory are limited to a single location at a time. *Attention, Perception, and Psychophysics*, *85*(5). https://doi.org/10.3758/s13414-023-02661-0

Griffin, I. C., & Nobre, A. C. (2003). Orienting Attention to Locations in Internal Representations. *Journal of Cognitive Neuroscience*, *15*(8), 1176–1194. https://doi.org/10.1162/089892903322598139

Ferreira, F., Apel, J., & Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends in Cognitive Sciences*, *12*(11). https://doi.org/10.1016/j.tics.2008.07.007

Hautekiet, C., Langerock, N., & Vergauwe, E. (2024). Prioritization in Visual Working Memory: An Investigation of Distractor Susceptibility and Different Prioritization Modes. *Open Science Framework.*

Heuer, A., & Schubö, A. (2016). Feature-based and spatial attentional selection in visual working memory. *Memory & Cognition*, *44*(4), 621–632. https://doi.org/10.3758/s13421-015-0584-5

Hu, Y., Hitch, G. J., Baddeley, A. D., Zhang, M., & Allen, R. J. (2014). Executive and perceptual attention play different roles in visual working memory: Evidence from suffix and strategy effects. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1665–1678. https://doi.org/10.1037/a0037163

Johansson, R., & Johansson, M. (2014). Look Here, Eye Movements Play a Functional Role in Memory Retrieval. *Psychological Science*, *25*(1). https://doi.org/10.1177/0956797613498260

Jongerius, C., Callemein, T., Goedemé, T., Van Beeck, K., Romijn, J. A., Smets, E. M. A., & Hillen, M. A. (2021). Eye-tracking glasses in face-to-face interactions: Manual versus automated assessment of areas-of-interest. *Behavior Research Methods*, *53*(5). https://doi.org/10.3758/s13428-021-01544-2

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., & Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, *57*(11). https://doi.org/10.1109/TBME.2010.2057429

Laeng, B., Bloem, I. M., D’Ascenzo, S., & Tommasi, L. (2014). Scrutinizing visual images: The role of gaze in mental imagery and memory. *Cognition*, *131*(2). https://doi.org/10.1016/j.cognition.2014.01.003

Li, Q., & Saiki, J. (2014). The effects of sequential attention shifts within visual working memory. *Frontiers in Psychology*, *5*(AUG). https://doi.org/10.3389/fpsyg.2014.00965

Loaiza, V. M., & Souza, A. S. (2022). The eyes don't have it: Eye movements are unlikely to reflect refreshing in working memory. *PLoS ONE*, *17*(7 July). https://doi.org/10.1371/journal.pone.0271116

Makovski, T., & Jiang, Y. V. (2007). Distributing versus focusing attention in visual short-term memory. *Psychonomic Bulletin & Review*, *14*(6), 1072–1078. https://doi.org/10.3758/BF03193093

Matsukura, M., & Vecera, S. P. (2015). Selection of multiple cued items is possible during visual short-term memory maintenance. *Attention, Perception & Psychophysics*, *77*(5), 1625–1646. https://doi.org/10.3758/s13414-015-0836-2

Oberauer, K. (2019). Working memory and attention - A conceptual analysis and review. In *Journal of Cognition* (Vol. 2, Issue 1). https://doi.org/10.5334/joc.58

Oberauer, K., & Hein, L. (2012). Attention to Information in Working Memory. *Current Directions in Psychological Science*, *21*(3), 164–169. https://doi.org/10.1177/0963721412444727

Ort, E., & Olivers, C. N. L. (2020). The capacity of multiple-target search. *Visual Cognition*, *28*(5–8). https://doi.org/10.1080/13506285.2020.1772430

Perugini, M., Gallucci, M., & Costantini, G. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology,* 31(1): 20, 1–23, DOI: https://doi.org/10.5334/irsp.181

R Core Team. (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

R Core Team. (2023). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Eye Tracking Research and Applications Symposium 2000*. https://doi.org/10.1145/355017.355028

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, & Psychophysics*, *78*(7), 1839–1860. https://doi.org/10.3758/s13414-016-1108-5

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). https://doi.org/10.1111/nyas.12603

Ueno, T., Fastrich, G. M., & Murayama, K. (2016). Meta-analysis to integrate effect sizes within an article: Possible misuse and Type I error inflation. *Journal of Experimental Psychology: General*, *145*(5), 643–654. https://doi.org/10.1037/xge0000159

van Ede, F., Chekroud, S. R., & Nobre, A. C. (2019). Human gaze tracks attentional focusing in memorized visual space. In *Nature Human Behaviour* (Vol. 3, Issue 5). https://doi.org/10.1038/s41562-019-0549-y

Van Moorselaar, D., Olivers, C. N. L., Theeuwes, J., Lamme, V. A. F., & Sligte, I. G. (2015). Forgotten but not gone: Retro-cue costs and benefits in a double-cueing paradigm suggest multiple states in visual short-term memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, *41*(6). https://doi.org/10.1037/xlm0000124