Behavioural and neural correlates of visual working memory reveal metacognitive aspects of mental imagery

Kathryn E. Bates[1], Marie L. Smith[2], Emily K. Farran[3]*, Maro G. Machizawa[4]*

1. Department of Psychology, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, UK
2. Department of Psychological Sciences, Birkbeck, University of London, UK
3. Department of Psychology, University of Surrey, UK
4. Center for Brain, Main & KANSEI Sciences Research, Hiroshima University, Hiroshima, Japan

*joint senior authors

The authors declare no competing interests.

Corresponse should be directed to Dr Kathryn Bates: kathryn.2.bates@kcl.ac.uk

Abstract

Mental imagery (MI) is the ability to generate visual phenomena in the absence of sensory input. MI is often likened to visual working memory (VWM): the ability to maintain and manipulate visual representations. How MI is recruited during VWM is yet to be established. In a modified orientation change-discrimination task, we examined how behavioural (proportion correct) and neural (contralateral delay activity; CDA) correlates of precision and capacity map onto subjective ratings of vividness and number of items in MI within a VWM task. During the maintenance period, seventeen participants estimated the vividness of their MI or the number of items held in MI while they were instructed to focus on either precision or capacity of their representation and to retain stimuli at varying set sizes (1, 2 and 4). Vividness and number ratings varied over set sizes; however, subjective ratings and behavioral performance correlated only for vividness rating at set size 1. While CDA responded to set-size as was expected, CDA did not reflect subjective reports on high and low vividness and on non-divergent (reported the probed number of items in mind) or divergent (reported number of items diverged from probed) rating trials. Participants were more accurate in low set sizes compared to higher set sizes and in fine (15º) orientation changes compared to coarse (35º) orientation changes. We failed to find evidence for a relationship between the subjective sensory experience of precision and capacity of MI and the precision and capacity of VWM.

Introduction

Our ability to generate perceptual phenomena in mind allows us to contemplate the future and remember the past, whilst navigating through the present. Mental imagery (MI) is defined as the ability to generate visual mental images in mind in the absence of sensory input (Kosslyn, 1980). MI is consistently likened to visual working memory (VWM; Tong, 2013); the ability to maintain and manipulate visual information in mind (Baddeley, 2003; Baddeley & Andrade, 2000; Cowan, 2001; Logie, 1995). However, the evidence does not yet warrant this conclusion. Previous research has suggested some people appear to recruit MI strategies in VWM tasks, while others do not (Bates & Farran, 2021; Keogh & Pearson, 2014). In the context of Aphantasia, individuals report no sensory experience of MI while holding typical abilities in VWM (Pounder et al., 2022; Jacobs et al., 2018), which further suggests a distinction between these seemingly unified sub-processes**.** Empirical studies are currently limited to directly comparing the behavioural and neural substrates of MI to VWM. To explain the relations between MI and VWM, direct evidence is required to examine how MI is recruited within a VWM task.

Delineating the relationship between MI and VWM

The investigation of how MI and VWM are related is limited, and the suggestion that they are similar functions is largely based on the parallels between the definitions of MI and VWM and the evidence for overlapping functional activation underpinning the two abilities (Lorenc et al., 2015; Miller & D'Esposito, 2005; Pearson, 2019; Sreenivasan et al., 2014; Spagna et al., 2021). Much like in the MI neuroimaging literature (Spagna et al., 2021), there is evidence for a functional role of the frontal regions in VWM, namely the lateral prefrontal cortex (Miller & D'Esposito, 2005; Sreenivasan et al., 2014), but there is also evidence that the visual cortex plays an important role (Albers et al., 2013; Serences, 2016). This has led to the argument that conflicting findings regarding the importance of either frontal or visual regions in VWM are likely dependent on individual differences in the recruitment of visual strategies in VWM (Linke et al., 2011; Pearson & Keogh, 2019). For example, some studies show visual representations in VWM are decoded in early visual areas (V1-V3; Albers et al., 2013), and others demonstrate the importance of top-down connectivity between high level regions, such as the lateral prefrontal cortex and the visual cortex (Sreenivasan et al., 2014). It is therefore speculated that not all individuals approach *visual* memory tasks in the same manner; however, research is yet to test how individuals use different visual strategies – namely, MI – within a VWM task.

There is, however, evidence for shared visual representations between MI and VWM. Findings have shown that oriented gratings held in mind in VWM can be decoded using multi-voxel pattern analysis (MVPA) in visual areas V1-V4 (Harrison & Tong, 2009). This has then been extended to show that a classifier trained on early visual area (V1-V3) activation in VWM trials reliably decoded activation in MI trials and vice versa (Albers et al., 2013). Based on this evidence, we might conclude that MI and VWM are therefore not distinguishable (Tong, 2013). However, behavioural evidence does not entirely align with this suggestion. Behavioural studies adopting a sensory strength measure of MI, which measures the extent to which perception is altered following an imagery period in a binocular rivalry paradogm (Pearson et al., 2008). Results from this task have implied that the recruitment of visual strategies in VWM is dependent on MI strength. When visual noise is presented during the delay period of a VWM task, it negatively impacts performance, which is taken to suggest it disrupts the visual information from being held in mind (Baddeley & Andrade, 2000). This interpretation is supported by the finding that MI is also disrupted when background luminance is modulated (Pearson et al., 2008). In turn, it has been shown that VWM performance was significantly poorer in the modulated background luminance condition but only in those that scored highly on the MI sensory strength measure. It was therefore interpreted that only “good imagers” recruit visual strategies in VWM (Keogh & Pearson, 2011, 2014). In addition, our group has recently found no significant associations between visual and transformation components of MI and maintenance and manipulation measures of VWM (ANONYMISED FOR REVIEW), further adding to ambiguity around the types of strategies individuals recruit in VWM tasks.

This is not the only study to imply individual differences in the recruitment of MI strategies for VWM. A 2020 study that examined the effects of training a visualisation strategy for a set of VWM tasks in adults found that in the control group (no strategies trained) only 4% reported visualisation (e.g., “I visualised the numbers”) and no participants in the control group reported a self-generated imagery strategy (e.g., “I tried to associate each digit with some image in my mind.”) (Forsberg et al., 2020). Instead, self-generated strategies included rehearsal (“I repeated the list of letters in my mind”), grouping (“I remembered the digits in groups”) and other (“I made up a song…”). Examining the extent to which individual differences in MI impact VWM would further elucidate the role of visual strategies/MI in supporting memory. Research thus far has been restricted to comparisons between absolute performance on MI measures and on VWM measures. To fully elucidate how MI supports VWM, it is necessary to investigate how within-task individual differences in the precision of visual representations and the sensory experience of MI impacts VWM performance.

Measuring MI within a VWM task

The visual precision and capacity of VWM maintenance have been documented using the study of event-related potentials; namely, contralateral delay activity (CDA). In their seminal paper, Vogel and Machizawa (2004) found that CDA is modulated as a function of the number of items held in mind up to 4 items. The finding that CDA can index VWM capacity has since been replicated (see Luria et al., 2016 for review), and there is evidence for individual differences in that greater CDA amplitude is denoted in individuals with good VWM compared to those with poorer VWM (Adam et al., 2018).

The visual precision of VWM representations held in mind can also be indexed by CDA amplitudes. Researchers applied an orientation-discrimination paradigm to not only discriminate between CDA amplitudes associated with increasing set size but also those associated with coarse (45°) and fine (15°) orientation discriminations. Here it was found that at smaller set sizes, there was greater CDA amplitude in fine orientation discriminations compared to coarse. Thus, it was interpreted that at lower capacities, individuals exert wilful control over the visual precision of their representations and that the CDA amplitude can reflect both the precision and capacity of maintained representations (Machizawa et al., 2012). This evidence has been extended to show that CDA is modulated by instruction. When participants were instructed to focus on precision, CDA was associated with grey matter volume in the left lateral occipital area, whereas when instructed to focus on capacity, CDA was associated with grey matter volume in the right intra-parietal sulcus (Machizawa et al., 2020). These findings support a threshold model of VWM and demonstrate the importance of accounting for both the visual precision and number of items.

Notably, there is overlap between the how visual representations are described in the parallel VWM and MI literatures. Specifically, what might be described in the VWM literature as visual precision of representations, would ultimately be described as the visual vividness or quality of mental images in MI literature. We might therefore assume that at smaller set sizes, neural correlates of precision, i.e., CDA, reflects the visual quality of visual images held in mind during VWM, and otherwise CDA reflects the capacity of visual items held in mind. However, this has not been measured alongside the reported subjective, sensory experience of MI. For simplicity, we will continue with the term visual precision when referring to instruction to attend to the precision of the representation, vividness when referring to the subjective vividness rating of representations, and capacity when referring to instructions to attend to capacity of the representations and number of items when referring to the subjective rating of number of items held in mind.

The most common approach to MI research is to measure the sensory experience of MI using subjective ratings. This is not surprising given that MI is an inherently private and variable sensory experience. In the quest to establish evidence to suggest that visually

depictive representations are recruited during MI, research has examined the relationship between the subjective, sensory experience of MI and selective neural activation of visual areas. For example, studies have adopted trial-by-trial vividness ratings. A significant positive association between the behavioural MI sensory strength score and trial-by-trial subjective vividness ratings (1 = almost no imagery, 2 = some weak imagery, 3 = moderate imagery, 4 = strong imagery almost like perception) has been evidenced (Pearson et al., 2011). This was interpreted to suggest that individuals have good insight into their MI. More recently, it has been shown that the overlap between brain regions activated during MI and during visual perception is positively associated with trial-by-trial subjective vividness (1 = not vivid at all to 4 = very vivid; Dijkstra et al., 2017). With respect to confidence, Williams et al. (2022) manipulated instruction to demonstrate confidence in responses reflects memory strength in a VWM task. However, whether individuals have good insight into the precision and capacity of their representations during a VWM task is yet to be tested.

Taken together, based on the behavioural and neural findings we might assume that the subjective sensory experience of the vividness of MI maps onto the precision of visual representations. However, this has not been directly assessed with respect to VWM because these processes have been examined in parallel literatures. While evidence in the VWM literature suggests that the number of items and the precision of items held in mind during the delay period in VWM can be quantified by CDA, the extent to which this reflects the subjective sensory experience of the number of items and precision of items in MI is yet to be addressed. Therefore, adapting a VWM paradigm to include trial-by-trial subjective vividness ratings and capacity ratings (number of items in mind) presents a novel opportunity to address the current gap in the literature in understanding how individual differences in MI impact VWM.

The current study

The current study was designed to directly examine how MI is recruited in a VWM task in the form of two aims. The first aim is to characterise how behavioural and neural correlates VWM are modulated by expectations of instruction (precision/capacity) and type of subjective ratings (vividness/number). For clarity, precision is adopted from the VWM literature (such as Machizawa et al., 2012; Zhang & Luck, 2008) and forms the dependent variable of proportion correct in an orientation-discrimination task where stimuli are presented at varying levels of precision (fine precision/15° orientation change and coarse precision/45° orientation change). The term vividness is adopted from the MI literature (e.g., Pearson et al., 2011; Marks, 1973), and refers to subjective ratings of how vivid participants deem the representation they held in mind during each orientation-discrimination trial. The second aim is to establish the metacognitive link between the *subjective* sensory experience

of MI and *behavioural* and *neural* correlates of VWM (CDA). Our hypotheses are outlined at the end of the methods section.

Materials and methods

Participants

Participants were recruited from the SONA database at ANOYMISED FOR REVIEW and the surrounding community of ANONYMISED FOR REVIEW. All participants gave written informed consent and had the option of the receiving £25 to participate or the equivalent course credit. Ethical approval was provided by the University Ethics Committee. Participants had normal or corrected-to-normal vision and each participant completed the Ishihara 38 Plates CVD Test (https://www.color-blindness.com/ishihara-38-plates-cvd-test/) to check for red-green colour deficiencies and were required to score “none” to participate. A total of 23 individuals were recruited for the final experiment. Prior to artefact rejection, two participants were excluded due to incomplete datasets because of technical errors and three more participants were excluded as they did not respond in any of the trial-by-trial subjective ratings and thus did not produce any behavioural ratings data. One more participant was excluded following artefact rejection due to there being less than 75% of the total trials remaining. A total of 17 participants are included in the reported results (age: M = 26.00, SD = 4.39, 10 female). Power is outlined in the next section alongside trial numbers.

Materials and procedure

A classic orientation discrimination VWM paradigm developed by Machizawa and colleagues (2012, 2020) was adapted to include within-trial subjective ratings of MI (see Figure 1 for schematic of trial sequence and outline of blocks). Participants were instructed to memorise an array of bars, hold the orientation of bars in mind, rate either the vividness or capacity of their MI and subsequently determine whether the highlighted bar in the probe array had been rotated clockwise or counter-clockwise. A fixation point was presented in the centre of the screen throughout the trial and participants were required to maintain their gaze at the fixation point. First, participants were cued to memorise either the bars presented to the left or right side of the screen. Second, the sample display was presented which consisted of two, four or eight bars (one, two or four bars to be remembered and presented to each hemifield, respectively). Participants were instructed to maintain fixation at the central fixation point and hold the bars in mind as accurately as possible during the subsequent delay. Following the delay, a tone rating cue was presented to cue participants to rate either the vividness of the representation held in mind or number of items they had in mind, depending on the block. The tone was generated in Cogent 2000 and comprised a

250Hz sine wave lasting 200ms, which was played from a speaker placed behind the participant's chair. Finally, a probe array was presented which was the same as the sample array except that the highlighted bar/item had been rotated. Fine (15° orientation change) and coarse (45° orientation change) trials were randomised within each block, as were clockwise and counter-clockwise orientations. Participants were required to respond as to whether the highlighted bar had been rotated right (clockwise) or left (counter-clockwise).

FIGURE 1 HERE

Figure 1: **A)** Trial sequence. Inter-trial interval ranged between 500-700ms. For each trial, an arrow cue was presented for 200ms to indicate which side of the screen should be attended to. This was followed by a 300-500ms interval before the sample array was presented for 200ms. The sample array consisted of 1, 2 or 4 bars on each side of the screen (set size 2 pictured) and either red (precision-focused instruction block) or green (capacity -focused instruction block, pictured) bars. This was followed by a 1300ms delay period whereby participants had to hold the image in mind. After the delay, a tone rating cue was played and participants provided either a vividness or capacity rating, depending on the block. Subsequently, a probe array was presented until the participant responded (or 2500ms) whereby all stimuli except the target stimulus were presented in black. Participants were required to judge whether the target was rotated clockwise (pictured) or counter-clockwise compared to the memorised sample array. **B)** Schematic representation of experiment procedure. Order of blocks was counterbalanced per participant.

A total of four blocks of 96 trials (384 total trials) were presented with two breaks within each block and an additional break between blocks to reduce fatigue and boredom (see Figure 1B). Blocks were differentiated by instruction and rating type. In the precision-focused instruction block, participants were asked to focus on to holding a visually precise image in mind and in the capacity-focused instruction, participants were required to focus on holding as many items in mind as required (i.e., they should try and hold all four items in mind in the 4-item condition). There were two rating types: vividness and number. In vividness rating blocks, participants were required to rate the vividness of the representation held in mind on a scale of 1-4 in line with previous paradigms: 1 = almost no image, 2 = weak image, 3 = moderate image, 4 = strong image/almost like perception (as in Pearson et al., 2011). In capacity rating blocks, participants were required to rate the number of items they held in mind (see Figure 1B for the procedure). The order of blocks was counterbalanced per participant. Set size (1 item, 2 items, 4 items), precision (fine, coarse) and attended side (left, right) were randomised within each block resulting in eight trials per

condition. A study conducted simulations to estimate how many participants and how many trials are required for different levels of power in CDA analyses (Ngiam et al., 2021). It was suggested that 30-50 trials were required per condition to detect the presence of CDA and up to 400 trials per condition with 25 participants could be needed to detect differences between set size conditions in CDA with 80% power. The task with 384 trials already takes just under an hour to complete, therefore adding more trials would distort the quality of the data. Moreover, robust CDA effects have been established in previous studies with ~20 subjects and ~80 trials per condition (Machizawa et al., 2012, 2020).

To familiarise participants with the task, they completed a precision-focused block and capacity-focused block (with either vividness or capacity ratings, counterbalanced) with 24 trials per block as practise. The practise blocks were repeated if participants scored < 65% percentage correct. A confidence rating was included at the end of each experimental block where participants were asked to rate their confidence in their behavioural performance of that block. While the subjective rating is purposefully placed before the probe array in the trial sequence to reduce the confound of confidence, a weak correlation between subjective ratings and confidence was expected. To test this, participants were presented with a blank grey screen at the end of the block with “confidence?” in the centre and they were required to answer according to a standard 5-point Likert scale: 1 = not confident at all, 2 = slightly confident, 3 = somewhat confident, 4 = fairly confident, 5 = completely confident (4 confidence trials in total).

EEG recording

EEG data was continuously recorded offline at 1,000Hz sampling rate using a fitted cap (EASYCAP) with 64 Ag-AgCl passive electrodes according to the international 10-20 system using a BrainVision BrainAmp amplifier. No online filters were applied during the recording. The cap included two horizontal EOG channels mounted in the cap at FT9 and FT10 locations. A vertical EOG channel was placed directly underneath the right eye to monitor blinks and saccades. Electrical impedance was kept below 5 kΩ. During the recording, FCz acted as the reference electrode and AFz as the ground electrode.

Pre-processing of EEG data and CDA extraction

After the recording, the continuous data was pre-processed offline in MATLAB (2016b) using the MATLAB toolbox EEGLAB (version 2019.1.; Delorme & Makeig, 2004). Data were filtered offline with an 8th-order Butterworth bandpass filter at 0.05–30Hz and resampled at 500Hz. Data were then epoched to –200 to 1400ms around the sample array onset and baseline corrected (–200–0ms). Blinks during the sample array onset (0–200ms)

were first detected using a moving window peak to peak detection algorithm with a window size of 200ms, a step of 10ms and a threshold of 50μV, trials with blinks during the sample array onset were then rejected (M ± SD*:* 23 ± 18, *range* = 4 to 61).
Next, an algorithm to detect square waves in the bipolar HEOG channel was applied with the threshold criteria set to +/– 18μV. A bipolar HEOG channel was derived (right horizontal EOG channel subtracted from left horizontal EOG channel) to observe the magnitude of left and right saccades, respectively. Mean amplitudes between 300-500ms following cue-onset were calculated for each visual angle (2º: *M* = 10.88, *SD* = 2.43; 4º: *M* = 22.82, *SD* = 6.07; 6º: *M* = 36.91, *SD* = 10.04; 8º: *M* = 48.65, *SD* = 12.21; 10º: *M* = 59.94, *SD* = 14.09). A repeated measures ANOVA of amplitude with a within-subject factor of visual angle (2º, 4º, 6º, 8º, 10º), which revealed increasing amplitude with visual angle ($F(4,16) = 73.82$, $p < .001$, $\eta_p^2 = .94$). Post hoc comparisons showed no overlap across visual angles (*p*s < .001). Based on the mean amplitudes, a simple formula can be applied to estimate the degree of horizontal eye movement: y = x / 6; where x = bipolar HEOG channel amplitude and y = degrees the eyes moved. The stimuli in the main experiment were presented between 2.5º–6.5º visual angle and the formula indicates that 2º saccades would be characterised as a mean bipolar HEOG channel amplitude of +/- 12μV and 3º would be characterised as mean bipolar HEOG channel amplitude of +/- 18μV. To avoid overcorrection of data, 18μV was chosen as the final value to detect savvades in the main experiment trials.

While this was effective in detecting saccades, the algorithm also detected +/– 18μV square waves that were too quick to be saccades (i.e., 50ms) (mean number of trials detected = 86, *SD* = 64, *range* = 7 to 200). Therefore, the trials flagged by the algorithm were checked by eye to determine whether the square waves detected were in fact saccades, i.e., the square wave spanned ~200ms (mean number of trials detected = 32, *SD* = 34, *range* = 7 to 118). As can be seen from the range, if all trials with saccades were removed, this would result in more participants being excluded due to insufficient data. Research has shown that applying independent component analysis (ICA) to remove saccade and blink components does not distort data for CDA analyses and is therefore an efficient method to retain data (Drisdelle et al., 2017). ICA was therefore conducted using the SOBI algorithm in EEGLAB and components were observed using ICLabel (Pion-Tonachini et al., 2019). Saccade and blink components were detected with the aid of ICLabel, which labels components according to the pattern of activity (e.g., eye component, muscle components etc.). An average of 2 (*SD* = 1, *range* = 1 to 5) components that were deemed either blink or saccade components were removed.

The blink and saccade algorithms were re-applied to the ICA corrected data and any remaining trials with saccades exceeding the 18μV threshold and blinks exceeding the 50*μV*

threshold were rejected (4 ± 2, *range* = 0 to 8). Finally, extreme values of +/-75*μV* and abnormal trends of linear drift over the entire epoch time-window (50*μV*, *r* = .80) were detected and rejected. The average number of remaining trials including all conditions for the CDA analyses following all artefact rejection was 335 trials (*SD* = 30, *range* = 278 to 364). The number of trials remaining following artefact rejection was similar across all conditions (mean = 6.98, *SD* = .16, *range* = 6.65-7.25). The cleaned data was then computationally re-referenced to bilateral mastoid electrodes (T9 and T10), in line with previous literature conducting CDA analyses (Machizawa et al., 2012). Channels rejected due to noise by the EEGLAB automated criteria were interpolated (1 ± 1, *range* = 0 to 2). Finally, as in convention, the average CDA waveform was obtained from posterior parietal and temporal-occipital channels (namely, P5/6, P7/8, PO3/4, PO7/8, and O1/2); and CDA amplitude was computed from 400-1400ms after sample onset for each condition.

Data analysis

Tests of normality revealed some variables were not normally distributed, however parametric analyses were applied given that ANOVA is robust to violations of assumptions of normality (Blanca et al., 2017). All within-subject post hoc comparisons are reported with Bonferroni corrections. Where assumptions of sphericity were violated, Greenhouse-Geisser estimates are reported. Reaction times (RTs) for subjective ratings and behavioural responses to the probe array that were less than 250ms or equal to 2500ms (no response) were excluded in analyses as inappropriate responses.

Hypotheses and aims

With reference to the first aim, behavioural outcomes are firstly expected to replicate previous findings (Vogel & Machizawa, 2004; Machizawa et al., 2012). Specifically, accuracy (proportion correct) was predicted to be greater in coarse vs. fine orientation-discrimination trials and greater in lower (1 and 2 items) vs. higher (4 items) set sizes. We also predicted an interaction between instruction (precision-focused vs. capacity-focused) and precision (fine vs. coarse) in that there would be greater accuracy in coarse vs. fine precision in precision-focused (try to maintain a highly precise representation) trials only but not in capacity-focused trials. With regards to the focus on attention, instruction was also expected to modulate subjective ratings in that greater vividness ratings were expected in precision-focused blocks compared to capacity-focused (try to maintain as many items as required) blocks and greater capacity ratings are expected in capacity-focused blocks compared to precision-focused blocks.

Measuring EEG during the behavioural VWM task allows for the unique opportunity to directly measure the visual precision and capacity of items held in mind during the delay

period (via CDA). We further expected instruction to modulate the usage of memory resource indexed by the CDA (Machizawa et al., 2020). As CDA is measured during the delay period and before the behavioural response, if CDA is modulated by instruction, this would demonstrate that individuals could flexibly control the precision and capacity of their visual representations at will (as implied in previous evidence: Zhang & Luck, 2008; Machizawa et al., 2020). If this is the case, differences in CDA were expected between precision-focused trials compared to capacity-focused trials at low set size but not between fine and coarse precision trials, this is because participants *expect* and can prepare for either precision- or capacity-focused responses, but actual difficulty (fine and coarse trials) was not cued in this experiment, therefore they cannot prepare for this. In sum, this will extend previous findings by examining how *instruction* modulates VWM consumption as indexed by CDA amplitude and how it interacts with the number of items. Finally, it was assumed that the established CDA set size effect would be replicated here in that CDA would increase as a function of set size up to 4 items (e.g., Vogel & Machizawa, 2004; Vogel, McCollough & Machizawa, 2005; Machizawa et al., 2012).

Next, we tested the relationship between confidence ratings and vividness and number ratings. We expected a significant positive association between confidence and vividness ratings, and a significant negative association between confidence and non-divergent ratings (when participants reported holding the correct number of items in mind). As previous evidence has demonstrated a positive association between confidence and strength of memory (Rademaker et al., 2012), we expect to find such a relationship with the vividness and number of items reported in this VWM task.

Finally, we examined how the sensory experience of MI was associated with behavioural and neural correlates of VWM. Evidence for significantly greater accuracy in trials rated as high vividness compared to low vividness was predicted and significantly greater accuracy in non-divergent number ratings (e.g., rated 2 items in mind when required to remember 2 items) compared to divergent number ratings (e.g., rated 2 items in mind when required to remember 4 items) was also predicted. With regard to neural correlates, it was predicted that CDA amplitudes would be significantly larger in high vividness trials compared to low vividness trials, and this effect would likely be greater in precision-focused trials at smaller set sizes. It was also predicted that CDA amplitudes would be significantly larger in larger set sizes in trials with non-divergent number ratings compared to trials with divergent number ratings. Together, this would support the assumption that individuals have good insight into their visual representations in both MI and VWM. Moreover, it will demonstrate that CDA not only maps the visual precision and/or capacity of representations but also the subjective sensory experience of MI within VWM. This would therefore provide a novel method for measuring the role of MI in VWM.

Results

Characterising the visual precision and capacity of VWM maintenance as indexed by proportion correct, subjective MI ratings and CDA

Accuracy (proportion correct)

Overall accuracy (as measured by proportion correct) was .71 (*SD* = .09), which is comparable to previous reports with a similar version of this task (Machizawa, et al., 2012). Descriptive statistics of proportion correct for all conditions are reported in Figure 2.

FIGURE 2 HERE

Figure 2: Mean (and +/– SE) accuracy (proportion correct) per condition, e.g. “Right, Fine, Precision, Vividness” refers to trials for the right attended, fine precision (15 orientation), precision-focused instruction, vividness rating

A repeated measures 4-way (3x2x2x2) ANOVA was conducted with proportion correct as the dependent variable and within-subject factors of set size (1 item, 2 items, 4 items), precision (fine, coarse), instruction (capacity-focused, precision-focused), rating (vividness, capacity), and attended side (left, right). Firstly, as was expected, accuracy significantly varied with set size ($F(2,32) = 82.59$; $p < .001$; $\eta_p^2 = .84$), Bonferroni corrected post hoc comparisons revealed a significant decrease in proportion correct between all comparisons (all *ps* < .001). Also in line with previous findings, accuracy significantly varied with required precision ($F(1,16) = 10.31$; $p = .005$; $\eta_p^2 = .39$), such that there was greater proportion correct in coarse precision (45° orientation-change) trials compared to fine precision (15° orientation-change) trials. There was no main effect of instruction ($F(1,16) = 2.58$; $p = .128$; $\eta_p^2 = .39$, $BF_{10} = .82$), rating ($F < 1$, $BF_{10} = .33$) or attended side ($F(1,16) = 2.58$, $p = .128$, $\eta_p^2 = .39$, $BF_{10} = .81$) nor an interaction between precision and instruction ($F(1,16) = 3.31$, $p = .088$, $\eta_p^2 = .17$).

There was a significant 3-way interaction between attended side, set size and instruction ($F(2,32) = 4.31$, $p = .022$, $\eta_p^2 = .21$). Follow up ANOVAs for each set size were conducted to explore this interaction. There was a significant interaction between attended side and instruction only in the 4-item condition ($F(1,16) = 7.22$; $p = .016$; $\eta_p^2 = .31$) (1-item condition attended side x instruction interaction: $F < 1$; 2-item condition attended side x instruction interaction: $F(1,16) = 1.88$; $p = .189$; $\eta_p^2 = .11$). Follow up t tests revealed an effect of attended side; significantly greater proportion correct in the *right* attended trials compared to the left attend trials for the *capacity*-focused condition ($t(16) = 3.01$; $p = .030$, $d = .36$), but not in the precision-focused condition ($t(16) = .58$; $p = 1.00$; $d = .07$). The 3-way interaction between attended side, precision and rating was not significant ($F(1,16) = 3.97$; $p = .064$; $\eta_p^2 = .19$), and the 4-way interaction between attended side, rating, instruction and rating was also not significant ($F(3,32) = 2.98$; $p = .065$; $\eta_p^2 = .16$).

Trial-by-trial subjective ratings on vividness and number

Next, separate ANOVAs were conducted on vividness ratings and capacity ratings, respectively. The within-subject factors were set size (1 item, 2 items, 4 items), precision (fine, coarse), instruction (capacity-focused, precision-focused) and attended side (left, right). Descriptive statistics of vividness ratings are presented in Figure 3.

FIGURE 3 HERE

Figure 3: Mean and SE of vividness (top) and number (bottom) ratings per condition. The y axis labels indicate condition: e.g., Right, Fine, Precision indicates right-attended, fine precision (15º orientation) and precision-focused instruction condition.

The vividness ratings ANOVA revealed a significant main effect of set size ($F(2,32) = 3.58$; $p = .040$; $\eta_p^2 = .18$), where post hoc comparisons showed marginally significantly higher vividness ratings when participants were required to remember 1 item compared to when they remembered 4 items ($p = .055$) (all other *ps* > .05). There was no main effect of precision ($F < 1$, $BF_{10} = .33$), attended side ($F < 1$, $BF_{10} = .32$) or instruction ($F(1,16) = 3.71$; $p = .072$; $\eta_p^2 = .18$, $BF_{10} = .97$). There were no significant interactions (all $F < 1.06$; n.s.).

An equivalent ANOVA was conducted on capacity ratings with the same within-subject factors as the vividness ratings ANOVA. Contrary to the vividness rating, number rating monotonically varied as a function of set size ($F(2,32) = 32.04$, $p < .001$, $\eta_p^2 = .67$), where capacity ratings increased with each increase in number of items (2 items > 1 item: $p < .001$, 4 items > 2 items: $p = .004$, 4 items < 1 item: $p < .001$). There were no main effects of precision ($F < 1$, $BF_{10} = .45$), instruction ($F < 1$, $BF_{10} = .46$) or attended side ($F < 1$, $BF_{10} = .59$) and there were no significant interactions (all *F*s < 1, n.s.).

Contralateral delay activity (CDA)

To examine how CDA was modulated by condition, an ANOVA was conducted on grand-averaged CDA and within-subject factors of set size (1 item, 2 items, 4 items), precision (fine, coarse), instruction (capacity-focused, precision-focused), rating (vividness, capacity) and attended side (left, right). Mean and standard error of grand-averaged CDA for all conditions are presented in Figure 4.

FIGURE 4 HERE

Figure 4: Mean and SE of grand-averaged CDA per condition. The y axis labels indicate condition: e.g., Right, Fine, Precision indicates right-attended, fine precision (15º orientation) and precision-focused instruction condition.

In line with previous reports, CDA significantly increased as set size ($F(2,32) = 14.06$; $p < .001$; $\eta_p^2 = .47$). Post hoc comparisons revealed significantly greater CDA between 1 item ($M = -.99$, $SD = .60$) and 2 items ($M = -1.31$, $SD = .59$) ($p = .020$) as well as 1 item and 4 items ($M = -1.58$, $SD = .85$) ($p < .001$), and the difference between 2 items and 4 items was not significant ($p = .07$). Given our sample size, we computed a power calculation to confirm this effect. We found the effect is powered to .91 with just 8 participants (calculation: $f^2 = .94$, p = .001, power = .80, number of groups = 1 number of measurements = 3). There was no main effect of precision ($F < 1$, $BF_{10} = .33$), instruction ($F < 1$, $BF_{10} = .33$), rating ($F < 1$, $BF_{10} = .35$) or attended side ($F < 1$, $BF_{10} = .37$). There was a significant 3-way interaction between precision, instruction and attended side ($F(1,16) = 6.01$; $p = .026$; $\eta_p^2 = .27$) and a significant 4-way interaction between instruction, attended side, set size and rating ($F(2,16) = 4.06$; $p = .027$; $\eta_p^2 = .20$).

Follow up ANOVAs on precision-focused and capacity-focused blocks, respectively, were conducted to explore the significant 3-way interaction. There was a significant interaction between attended side and precision in the capacity-focused condition only ($F(1,16) = 5.85$; $p = .028$; $\eta_p^2 = .27$) (precision-focused condition attended side x precision interaction: $F < 1$). T tests of the effect of precision for each attended side revealed significantly greater (more negative) CDA in coarse trials compared to fine trials in the *right* attend condition ($t(16) = 2.74$; $p = .015$; $d = .66$) but there was no difference between fine and coarse in the *left* attend condition ($t(16) = 1.23$; $p = .238$; $d = .29$).

With regard to the 4-way interaction, there was a significant interaction between attended side, set size and rating in the capacity-focused trials only ($F(2,32) = 3.35$; $p = .048$; $\eta_p^2 = .17$) (precision-focused condition attended side x set size x rating interaction: $F(1,32) = 1.41$; $p = .259$; $\eta_p^2 = .08$). Follow up ANOVAs for each set size for capacity-focused trials revealed a significant interaction between rating and attended side in the 2-item condition only ($F(1,16) = 4.50$; $p = .050$; $\eta_p^2 = .08$) (1-item condition rating x attended side interaction: $F(1,16) = 2.39$; $p = .141$; $\eta_p^2 = .14$; 4-item condition rating x attended side interaction: $F(1,16) = 2.93$; $p = .107$; $\eta_p^2 = .16$). While the means point towards greater CDA amplitude in left ($M = -1.65$, $SD = 1.93$) compared right ($M = -1.04$, $SD = .04$) attend trials in the capacity ratings, this was not significant ($t(16) = .97$; $p = .345$; $d = .24$). There was also no significant difference between left attend ($M = -1.16$, $SD = 1.35$) and right attend trials ($M = -1.28$, $SD = 1.59$) in the vividness ratings condition ($t(16) = .183$; $p = .857$; $d = .04$). There were no other significant interactions (set size x rating: $F(2,32) = 2.65$, $p = .086$, $\eta_p^2 = .14$, all

other *Fs* < 1). Grand averaged ipsilateral, contralateral and CDA waveforms for each set size are presented in Figure 5A, and waveforms per block presented in Figure 5B.

FIGURE 5 HERE

Figure 5: A) Grand-averaged waveforms for the 1 item trials (left), 2 items (centre) and 4 items (right). Sample onset is at 0-200msec and vertical dotted line at 400ms added for reference (CDA amplitude calculated as mean amplitude between 400ms and 1400ms after sample onset). *B)* CDA waveform per condition, CDA was calculated from 350ms to 1400ms. Note: n.s. stands for not significant

The relationship between subjective trial-by-trial MI ratings and VWM maintenance

The first set of analyses were conducted to investigate how behavioural and neural correlates of VWM were modulated by expectations of instruction (precision/capacity) and subjective ratings of items held in mind during the maintenance period (vividness/number). The next set of analyses were conducted to address aim 2: to examine the metacognitive link between subjective ratings of MI and behavioural and neural indices of VWM maintenance.

Behavioral contrast between low vs. high vividness trials and non-divergent vs. divergent capacity trials

To investigate whether individual's subjective ratings reflected VWM accuracy, two paired sample t tests were conducted to examine the difference in proportion correct between trials rated with high vividness and low vividness and non-divergent and divergent capacity ratings, respectively. High vividness ratings were trials where the participant rated either 3 (moderate image) or 4 (strong image/almost like perception) and low vividness ratings were trials where the participant rated either 1 (almost no image) or 2 (weak image). Non-divergent capacity ratings were trials where participants rating did not diverge from the number of items they were required to hold in mind (e.g., required to hold 4 items in mind, reported holding 4 items in mind, score for trial = 0) and divergent capacity ratings were trials where participant diverged from number of items they were required to hold in mind (e.g., required to hold 4 items in mind, reported holding 2 items in mind, score for trial = 2). Firstly, there was no significant difference between proportion correct in high vividness trials (M = .73, SD = .15) and low vividness trials (M = .67, SD = .12) ($t(16) = 1.51$; p = .152; d = .37). For the capacity ratings analysis, one participant was excluded because none of their trials were divergent, and another participant was excluded as none of their trials were non-divergent. There was a significant difference between proportion correct in non-divergent ratings (M = .77, SD = .07) and divergent ratings (M = .68, SD = .19) ($t(14) = 2.21$; p = .040; d = .57), which showed greater accuracy in non-divergent trials compared to divergent trials, see Figure 6.

FIGURE 6 HERE

Figure 6: Mean accuracy (proportion correct) for non-divergent and divergent number rating trials (left) and high and low vividness rating trials (right). Note: n.s. stands for not significant

CDA in high vs. low vividness ratings and non-divergent vs. divergent capacity ratings

To examine CDA between rating type at each set size and instruction, an ANOVA was planned with grand-averaged CDA as the dependent variable and rating (high vividness, low vividness, non-divergent capacity, divergent capacity), set size (1 item, 2 items, 4 items), instruction (precision-focused, capacity-focused) and attended side (left, right) as the within-subject factors. However, as the conditions were based on participant responses, there was at least one condition per participant where there were no responses (e.g., some participants did not rate any 4 item trials as high vividness). Therefore, an ANOVA was conducted for vividness ratings and capacity ratings collapsed across all conditions except vividness (number of high vividness responses: *M* = 91, *SD* = 43, *range* = 34 to 151; number of low vividness responses: *M* = 68, *SD* = 46, *range* = 4 to 140) and capacity (number of non-divergent responses: *M* = 115, *SD* = 35, *range* = 65 to 174; number of divergent responses: *M* = 50, *SD* = 35, *range* = 0 to 110) respectively. Despite the imbalance of trial numbers in accuracy variables, high and low vividness (W = .982, p = .971) and non-divergent and divergent (W = .910, p = .137) were normally distributed, and therefore the assumptions for correlations are met. The vividness rating ANOVA included a within-subject factor of vividness (high, low), which revealed no main effect of vividness ($F(1,16) = 1.38$; $p = .258$; $\eta_p^2 = .08$) on grand-averaged CDA. The capacity ratings ANOVA included within-subject factors of divergence (non-divergent, divergent). Similarly, to the vividness ANOVA, there was no main effect of divergence ($F < 1$).

FIGURE 7 HERE

Figure 7: Mean grand-averaged CDA in high and low vividness trials, and in divergent and non-divergent number rating trials. Note: n.s. stands for not significant

Relationship between proportion correct and subjective MI ratings as a function of set size

Spearman's correlations were conducted to examine the relationship between proportion and rating at each set size. As the analyses above indicate only an effect of set size in ratings and proportion correct; precision, instruction and attended side were collapsed across to retain power in the following analyses. Individual differences in *vividness* ratings were significantly and positively associated with proportion correct only in 1 item trials (r_s = .578; p = .015), however there were no significant correlations in 2 item trials (r_s = .143; p = .585) or 4 item trials (r_s = .010; p = .974).

For the capacity ratings analysis, the divergence score was included. Capacity divergence was not associated with proportion correct in 1 item trials (r_s = -.427; p = .088), 2 item trials (r_s = -.369, p = .144) or 4 item trials (r_s = -.327, p = .200). Taken together, the findings suggest participants have relatively poor insight into the visual precision (vividness rating) of representations held in VWM and the number of visual items (capacity rating) in representations held in VWM, except for visual precision (vividness) at the smallest set size (1 item).

Relationship between CDA and subjective MI ratings as a function of set size

To assess the relationship between CDA and subjective MI ratings, separate correlations were conducted for vividness ratings and capacity ratings. For vividness ratings, the CDA dependent variable was computed as the difference between grand-averaged CDA for 1-item trials and 2-items trials per participant, given that vividness is expected to be more prominent in smaller set sizes. The vividness ratings dependent variable consisted of the mean vividness ratings for 2-item trials per participant. This is based on the logic that if vividness ratings map onto the number of items in mind as indexed by CDA, there should be a positive association between vividness ratings in 2-item trials and the difference in CDA between 1- and 2-item trials, i.e., the greater the set size effect in CDA, the higher the vividness rating. Thus, we were motivated to assess how ratings were related to CDA modulation effect of VWM (e.g., Machizawa et al., 2021). Moreover, a difference calculation of CDA rather than individual ERP allows us to control for non-neural influence on the signal,

e.g., participant's skull thickness or scalp condition. There was no relationship between the difference between CDA in 1-item and 2-item trials and vividness ratings in 2-item trials (r_s = -.314; p = .220).

For capacity ratings, the CDA dependent variable was computed as the difference between grand-averaged CDA for 1-item trials and 4-item trials. The capacity ratings dependent variable consisted of the mean capacity rating for 4-item trials. As above, this is based on the logic that if capacity ratings map onto the number of items held in mind as indexed by CDA, there should be a positive association between capacity ratings in 4-item trials and the difference between CDA between 1-item and 4-item trials, i.e., the greater the set size effect in CDA, the more items the participant reports holding in mind. However, there was no relationship between the difference between CDA in 1-item and 4-item trials and capacity ratings in 4-item trials (r_s = .302; p = .239).

Relationship between subjective MI ratings and confidence ratings

A Pearson's correlation was conducted between mean confidence ratings for vividness rating blocks (M = 3.53, SD = .62) and mean vividness ratings (M = 1.58, SD = .26). This revealed a strong positive correlation between confidence ratings and vividness ratings (r = .508; p = .037), which suggests the higher participants rated vividness, the greater the confidence participants had in their VWM accuracy. The equivalent Pearson's correlation was conducted between mean confidence ratings for capacity blocks (M = 3.47, SD = .91) and mean capacity divergence score (M = .58, SD = .40). A mean divergence score was calculated based on divergent and non-divergent responses. Non-divergent responses were scored 0 and were trials where the participant rated that they had all items in the array clearly in mind (e.g., they were required to remember 4 items and rated 4). Divergent responses were scores where the rating diverged from the number of items the participant was required to remember (e.g., required to remember 4 items, reported remembering 2 items, divergence score for trial = 2, while divergence score for a perfect report = 0). This showed a strong negative correlation between confidence ratings and divergence score (r = – .737; p < .001), suggesting the lower the divergence between the number of to-be-remembered items and the number of items in mind, the greater the confidence participants had in their VWM performance.

FIGURE 8 HERE

Figure 8: Scatter plots for vividness rating as a function of confidence (left panel) and for divergent score and as a function of confidence rating (right panel)

Discussion

The overarching goal of this study was to investigate how MI is recruited in VWM. The first aim was to characterise how instruction (precision-focused vs. capacity-focused) and the type of subjective rating (vividness vs. number) modulated the neural (CDA) and behavioural (accuracy) correlates of VWM. The second aim was to examine the relationship between the subjective sensory experience of MI (vividness and number) and the behavioural and neural correlates of VWM. We failed to find evidence that instruction, type of rating (vividness or number) or precision (fine vs. coarse orientation) modulated CDA or proportion correct. Previous findings regarding set size were replicated; poorer proportion correct with increasing set size, and greater (more negative) CDA amplitude with increasing set size. We found no evidence for a relationship between MI and the visual precision and capacity of representations held in VWM. This may have implications for theory on the role of consciousness in VWM and for future methodology applied to understand individual differences in VWM. The findings are discussed in turn below.

The interaction between subjective ratings of MI and the behavioural and neural correlates of VWM maintenance

Previous findings were replicated in that proportion correct was greater in smaller set sizes compared to larger set sizes (Machizawa et al., 2012, 2020). Contrary to expectations, proportion correct was not modulated by the cued conditions of instruction and type of subjective rating. Previous evidence suggests that individuals exert willful control over the precision of visual representations, as instructed, and this in turn influences their performance (Machizawa et al., 2012, 2020; Zhang & Luck, 2008). However, this effect was not found in the context of conditions instructing participants to consider the visual precision of their representations, i.e., the precision-focused instruction instructs participants to hold a precise visual image in mind and the capacity-focused instruction instructs them to hold the correct number of visual items in mind (capacity). Thus, this calls into question the role of consciousness in MI compared to VWM. Vividness ratings were found to be higher at smaller set sizes and capacity ratings increased with increasing set size, yet there were no effects of instruction on vividness or capacity ratings. This suggests that the type of instruction did not modulate individuals subjective experience of the number of items held in mind, which is perhaps not surprising given that the ratings are subjective in nature. New evidence has suggested that when encouraged to use imagery, those with high imagery

vividness perform better on a VWM task compared to those with low imagery vividness (Slinn et al., 2023). Future research with powered sample sizes and number of trials should test the nuances of instruction, as in our study, and how this might differentially modulate accuracy in VWM dependent on imagery vividness group.

The effects of attended side in proportion correct and CDA amplitude are notable. Proportion correct was greater in right attended trials compared to left attended in the capacity-focused condition and 4-item trials only. Comparatively, greater CDA amplitudes were indexed in coarse compared to fine trials in the right attended but not left attended trials in the capacity-focused blocks. While laterality differences were not initially hypothesised, the suggestion of hemispheric differences in proportion correct is consistent with recent findings in a similar paradigm. Namely, Machizawa et al. (2020) report that behavioural performance and CDA amplitudes in their precision-focused instruction condition (fine trials only) were associated with the grey matter volume in the right parietal cortex whereas behavioural performance and CDA amplitudes in their capacity-focused condition (coarse trials only) were associated with grey matter volume in the left lateral occipital cortex. The findings presented here, that are specific to the largest set size when participants were required to rate the number of items in mind (capacity rating) and were following a capacity-focused instruction, support the indication of left hemispheric specialisation of VWM capacity.

The finding of greater amplitude in coarse trials compared to fine in the right attend trials only is perhaps not entirely surprising as it is partially in line with an association between coarse (capacity-focused) performance and left lateral occipital volume in Machizawa et al.'s (2020) study, although in their study coarse precision was cued. Therefore, the finding that there is a difference between coarse and fine trials is unexpected in that participants were not cued for the precision (fine, coarse) modulation in the current study. Given that the 3- and 4-way interactions include individual conditions with limited number of trials per condition, it is not possible to make general conclusions regarding hemispheric differences in CDA based on these findings and further research is warranted.

Distinction between subjective MI ratings and the visual contents of VWM

We failed to find evidence for a significant relationship between subjective MI ratings and vividness and capacity of VWM, except that vividness was significantly correlated with proportion correct in 1-item trials, but not in 2- or 4-item trials. This result may be explained by to *willful* control of our VWM resources at low set-sizes (Zhang & Luck 2008; Machizawa, Goh & Driver 2011). Evaluation of whether willful control of our resources and awareness on perceived resolution of our mental imagery is also constrained by VWM capacity should be examined in the future studies. Proportion correct was higher in non-divergent compared to

divergent capacity ratings, indicating that individuals have some insight into the number of items held in mind. There were no significant differences between CDA amplitude between high vividness rating trials and low vividness rating trials and no significant association between the CDA set size effect and vividness ratings. The important term here is "subjective". We can make conclusions about individuals' subjective insight into their mental imagery during this task, rather than their explicit ability in mental imagery. Therefore, we would not rule out a functional relationship between VWM and mental imagery or VWM and CDA based on this evidence.

In Pearson & Keogh's (2019) review, they argued that individual differences in the neural correlates of VWM may be dependent on the types of strategies recruited in VWM, i.e., imagery strategies vs. propositional strategies akin to general thought, and that measuring strategies recruited in VWM tasks might explain these individual differences. The study presented here directly addresses this proposition by measuring trial-by-trial subjective ratings of MI within a VWM task. However, we failed to find evidence for a relationship between self-reported subjective ratings/MI strategies and the precision and capacity at which visual information is held in mind. Firstly, propositional/verbal strategies are unlikely in this task given the very short stimuli presentations (200ms) and delay period (1400ms). Moreover, the modulations in proportion correct and CDA amplitude depending on instruction demonstrate that individuals have flexible control over the precision and capacity at which visual information is held in mind, as discussed in detail above. In the few studies that have investigated the relationship between behavioural outcomes in VWM and MI, some have found an association (Keogh & Pearson, 2011; 2014), whereas others have not (Bates & Farran, 2021). For example, findings show that MI sensory strength was positively associated with VWM capacity at set size 3 (Keogh & Pearson, 2014) and only VWM performance in those with high MI sensory strength was disrupted by background luminance manipulations (Keogh & Pearson, 2011; 2014). While in these studies it was argued that individuals with stronger MI recruit MI strategies in VWM, the findings presented in this study call into question whether assessing subjective strategies in VWM is akin to behavioural and neural indices of the visual precision and number of visual items maintained in VWM.

General considerations and limitations

It is important to consider potential methodological constraints. Previous evidence has suggested a relationship between saccades and MI in that participants tend to make similar gaze patterns when imagining a previously viewed stimulus as they do when viewing a stimulus, known as the "looking at nothing" effect (Brandt & Stark, 1997; Johansson & Johansson, 2014). Given the nature of EEG data, trials with saccades present artefacts which must be removed prior to analysis. While ICA was conducted to retain as many trials

as possible and the number of trials retained per participant was > 75% in this study, this is an important consideration given that high MI trials may have been rejected due to saccade artefacts. That being said, a 2021 study examining gaze patterns during MI found that gaze patterns during MI trials were not associated with vividness of MI as measured by the VVIQ (Gurtner et al., 2021). Therefore, it appears that it is unlikely that rejection of saccade trials would have influenced results examining the link between MI and VWM in this study. Future research examining gaze patterns alongside subjective ratings of MI within a VWM task would further elucidate this relationship.

It is also notable that participants appear to rarely rate at either end of the rating scales. For example, individuals rarely report having 4 items in mind in the number ratings. The fact that the vividness rating scale ranged from 1–4 and the capacity rating scale ranged from 0–4 could have been confusing for the participants. Thus, another study with only number report might be able to eliminate such confusion. However, the vividness rating scale was chosen as so in line with previous studies (Pearson et al., 2011; Dijkstra et al., 2017). This is the first study to adopt a capacity rating scale and it appears individuals are reluctant to rate at either end of the scale. One previous study has used a continuous scale (i.e., visual analog scale) for rating vividness using a sliding bar (Dijkstra et al., 2020), however responses broadly fell into the 1-4 category ratings and were therefore binned as such. Further research is required to test whether ratings are distorted by the Likert-scale. The findings regarding the link between MI and VWM are somewhat limited due infrequent responses. For example, some participants did not rate any 4-item trials as low vividness, therefore it was not possible to test the relationship between individual differences in ratings and CDA for each set size or instruction, for instance. Moreover, our vividness rating did not capture the strength or contrast of mental images, which are another important facet of imagery vividness (Riley & Davies, 2023). Future studies sampling participants based on low vividness, high vividness, non-divergent and divergent ratings, as well as including more detailed assessments of vividness, would be useful to further examine individual differences.

In addition, it is important to recognise the limited sample size and its statistical power. Twenty-three participants were recruited, which is in line with previous studies demonstrating robust CDA effects in precision and capacity in VWM (Machizawa et al., 2012, 2020). However, due to exclusion, only 17 participants remained in the final sample. We subsequently conducted Bayes Factor analyses of the main effects with null findings but some of the outcomes were inconclusive, therefore further replication is needed. The instruction condition was added after piloting to reduce difficulty and to replace the expected condition or precision (fine, coarse). While this allowed us to investigate questions regarding expectation, it rendered a 5-factorial design with low power. Findings should therefore be interpreted with caution and future replications should consider a simplified design. Small

sample sizes are a common issue in neuroimaging studies (Button, et al., 2013) given the resource and time constraints associated with this research. Recently, it was suggested that only 30-50 trials are sufficient to detect the *presence* of CDA but for differences between set sizes 2 and 4, up to 400 trials per condition could be required (Ngiam et al., 2021). While this is informative, up to 400 trials per condition is practically very difficult as this would lead to lengthy experiments and therefore participant fatigue and boredom, which would rather induce a distortion of the data. It is important to strike a balance in methodological design and to take sample size and trial numbers into account when drawing conclusions on analyses of CDA. To note, we had relatively sufficient and *feasible* number of trials for simple main effect comparisons (i.e., approximately 90 to 120 trials per set-size, collapsing across the other factors).

Conclusion

Ultimately, this study provides a much-needed account of the interaction between subjective ratings of MI and the behavioural and neural correlates of VWM. Contrary to hypotheses, participants appear to have poor insight into both of the visual precision and capacity of representations held in VWM. Rather than providing a novel method for measuring the role of MI in VWM using subjective ratings, we failed to find evidence for a relationship between the subjective sensory experience of MI and the visual precision and capacity of VWM. As our reports were mostly on averaged scores, future investigation on trial-by-trial approach may reveal momentary association or dissociation of our MI and VWM relations. This has methodological implications for examining how individual differences in MI support VWM and contributes to the theoretical interpretations of the role of consciousness in VWM.

Data availability statement

The participants of this study did not give written consent for their data to be shared publicly, therefore the data is not available.

Author contribution

Kathryn E. Bates: Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualisation, Funding Acquisition; **Marie L. Smith**: Software, Resources, Writing – Review & Editing; **Emily K. Farran**: Conceptualization, Methodology, Supervision, Writing – Review & Editing, Funding Acquisition; **Maro G. Machizawa**: Conceptualization, Methodology, Writing – Review & Editing, Supervision

Acknowledgements

This research was supported by a 1+3 ESRC PhD Studentship (1788622) awarded to KEB, and the JST COI grants (JPMJCE1311; JPMJCA2208) and JST Moonshot Goal 9 (JPMJMS2296) plus Hiroshima University Grant-in-Aid Basic Research to MGM.

References

Adam, K. C. S., Robison, M. K., & Vogel, E. K. (2018). Contralateral delay activity tracks fluctuations in working memory performance. *Journal of Cognitive Neuroscience*, *30*(9), 1229–1240. https://doi.org/10.1162/jocn_a_01233

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & De Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, *23*(15), 1427–1431. https://doi.org/10.1016/j.cub.2013.05.065

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews: Neuroscience*, *4*(October), 829–839. https://doi.org/10.1038/nrn1201

Baddeley, A. D., & Andrade, J. (2000). Working Memory and the Vividness of Imagery. *Journal of Experimental Psychology: General*, *129*(1), 126–145. https://doi.org/10.1037/0096-3445.129.1.126

Bates, K. E., & Farran, E. K. (2021). Mental imagery and visual working memory abilities appear to be unrelated in childhood: Evidence for individual differences in strategy use. *Cognitive Development*, *60*, 101120.

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, *29*(4), 552–557. https://doi.org/10.7334/psicothema2016.383

Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, *9*(1), 27–38. https://doi.org/10.1162/jocn.1997.9.1.27

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14(5), 365–376. https://doi.org/10.1038/nrn3475

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Cui, X., Jeter, C. B., Yang, D., Montague, P. R., & Eagleman, D. M. (2007). Vividness of mental imagery: Individual variability can be measured objectively. *Vision Research*, *47*(4), 474–478. https://doi.org/10.1016/j.visres.2006.11.013

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dijkstra, N., Ambrogioni, L., Vidaurre, D., & van Gerven, M. A. J. (2020). Neural dynamics of perceptual inference and its reversal during imagery. *ELife*, *9*, 1–19. https://doi.org/10.7554/eLife.53588

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. J. (2017). Vividness of Visual Imagery Depends on the Neural Overlap with Perception in Visual Areas. *The Journal of Neuroscience*, *37*(5), 1367–1373. https://doi.org/10.1523/jneurosci.3022-16.2016

Drisdelle, B. L., Aubin, S., & Jolicoeur, P. (2017). Dealing with ocular artifacts on lateralised ERPs in studies of visual-spatial attention and memory: ICA correction versus epoch rejection. *Psychophysiology*, *54*(1), 83–99. https://doi.org/10.1111/psyp.12675

Forsberg, A., Fellman, D., Laine, M., Johnson, W., & Logie, R. H. (2020). Strategy mediation in working memory training in younger and older adults. *Quarterly Journal of Experimental Psychology*, *73*(8), 1206–1226. https://doi.org/10.1177/1747021820915107

Gurtner, L. M., Hartmann, M., & Mast, F. W. (2021). Eye movements during visual imagery and perception show spatial correspondence but have unique temporal signatures. *Cognition*, *210*, 104597. https://doi.org/10.1016/j.cognition.2021.104597

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635. https://doi.org/10.1038/nature07832

Jacobs, C., Schwarzkopf, D. S., & Silvanto, J. (2018). Visual working memory performance in aphantasia. *Cortex*, *105*, 61-73. https://doi.org/10.1016/j.cortex.2017.10.014

Johansson, R., & Johansson, M. (2014). Look Here, Eye Movements Play a Functional Role in Memory Retrieval. *Psychological Science*, *25*(1), 236–242. https://doi.org/10.1177/0956797613498260

Keogh, R., & Pearson, J. (2011). Mental Imagery and Visual Working Memory. *PLOS ONE*, *6*(12), e29221. https://doi.org/10.1371/journal.pone.0029221

Keogh, R., & Pearson, J. (2014). The sensory strength of voluntary visual imagery predicts visual working memory capacity. *Journal of Vision*, *14:7*(12), 1–13. https://doi.org/10.1167/14.12.7.doi

Kosslyn, S. M. (1980). *Image and Mind*. Havard University Press.

Lee, S. H., Kravitz, D. J., & Baker, C. I. (2012). Disentangling visual imagery and perception of real-world objects. *NeuroImage*, *59*(4), 4064–4073. https://doi.org/10.1016/j.neuroimage.2011.10.055

Linke, A. C., Vicente-Grabovetsky, A., Mitchell, D. J., & Cusack, R. (2011). Encoding strategy accounts for individual differences in change detection measures of VSTM. *Neuropsychologia*, *49*(6), 1476–1486. https://doi.org/10.1016/j.neuropsychologia.2010.11.034

Logie, R. H. (1995). *Visuo-spatial working memory*. L. Erlbaum Associates.

Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience and Biobehavioral Reviews*, *62*, 100–108. https://doi.org/10.1016/j.neubiorev.2016.01.003

Machizawa, M. G., Driver, J., & Watanabe, T. (2020). Gray Matter Volume in Different Cortical Structures Dissociably Relates to Individual Differences in Capacity and Precision of Visual Working Memory. *Cerebral Cortex*, *30*(9), 4759–4770. https://doi.org/10.1093/cercor/bhaa046

Machizawa, M. G., Goh, C. C. W., & Driver, J. (2012). Human Visual Short-Term Memory Precision Can Be Varied at Will When the Number of Retained Items Is Low. *Psychological Science*, *23*(6), 554–559. https://doi.org/10.1177/0956797611431988

Marks, D. (1973). Visual imagery differences in the recall of pictures. *British Journal of Psychology*, *64*(1), 17–24. https://doi.org/10.1111/j.2044-8295.1973.tb01322.x

Marks, D. F. (1995). New directions for mental imagery research. *Journal of Mental Imagery*, *19*(3–4), 153–167.

Miller, B. T., & D'Esposito, M. (2005). Searching for 'the top' in top-down control. *Neuron*, *48*(4), 535–538. https://doi.org/10.1016/j.neuron.2005.11.002

Ngiam, W. X. Q., Adam, K. C. S., Quirk, C., Vogel, E. K., & Awh, E. (2021). Estimating the statistical power to detect set size effects in contralateral delay activity. *Psychophysiology*, *58*(5), e13791. https://doi.org/10.1111/psyp.13791

Pearson, J. (2019). The human imagination: the cognitive neuroscience of visual mental imagery. Nature Reviews Neuroscience, 20(10), 624–634. https://doi.org/10.1038/s41583-019-0202-9

Pearson, J., Clifford, C., & Tong, F. (2008). The Functional Impact of Mental Imagery on Conscious Perception. *Current Biology*, *18*(13), 982–986. https://doi.org/10.1016/j.cub.2008.05.048

Pearson, J., & Keogh, R. (2019). Redefining Visual Working Memory: A Cognitive-Strategy, Brain-Region Approach. *Current Directions in Psychological Science*, *28*(3), 266–273. https://doi.org/10.1177/0963721419835210

Pearson, J., Rademaker, R., & Tong, F. (2011). Evaluating the Mind's Eye: The Metacognition of Visual Imagery. *Psychological Science*, *22*(12), 1535–1542.

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026

Pounder, Z., Jacob, J., Evans, S., Loveday, C., Eardley, A. F., & Silvanto, J. (2022). Only minimal differences between individuals with congenital aphantasia and those with

typical imagery on neuropsychological tasks that involve imagery. *Cortex*, *148*, 180–192. https://doi.org/10.1016/j.cortex.2021.12.010

Rademaker, R. L., & Pearson, J. (2012). Training visual imagery: Improvements of metacognition, but not imagery strength. *Frontiers in psychology*, *3*, 224. https://doi.org/10.3389/fpsyg.2012.00224

Riley, S. N., & Davies, J. (2023). Vividness as the similarity between generated imagery and an internal model. *Brain and Cognition*, *169*, 105988.

Serences, J. T. (2016). Neural mechanisms of information storage in visual short-term memory. *Vision Research*, *128*, 53–67. https://doi.org/10.1016/j.visres.2016.09.010

Slinn, C., Nikodemova, Z., Rosinski, A., & Dijkstra, N. (2023, February 1). Vividness of visual imagery predicts performance on a visual working memory task when an imagery strategy is encouraged. https://doi.org/10.31234/osf.io/34wsv

Spagna, A., Hajhajate, D., Liu, J., & Bartolomeo, P. (2021). Visual mental imagery engages the left fusiform gyrus, but not the early visual cortex: A meta-analysis of neuroimaging evidence. *Neuroscience & Biobehavioral Reviews*, *122*, 201–217. https://doi.org/10.1016/j.neubiorev.2020.12.029

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89. https://doi.org/10.1016/j.tics.2013.12.001

Tong, F. (2013). Imagery and visual working memory: One and the same? *Trends in Cognitive Sciences*, *17*(10), 489–490. https://doi.org/10.1016/j.tics.2013.08.005

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

Williams, J. R., Robinson, M. M., Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2022). You cannot "count" how many items people remember in visual working memory: The importance of signal detection–based measures for understanding change detection performance. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(12), 1390. https://doi.org/10.1037/xhp0001055

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. https://doi.org/10.1038/nature06860