The effects of delay duration on visual working memory for orientation

Hongsup Shin

Qijia Zou

Wei Ji Ma

We used a delayed-estimation paradigm to characterize the joint effects of set size (one, two, four, or six) and delay duration (1, 2, 3, or 6 s) on visual working memory for orientation. We conducted two experiments: one with delay durations blocked, another with delay durations interleaved. As dependent variables, we examined four model-free metrics of dispersion as well as precision estimates in four simple models. We tested for effects of delay time using analyses of variance, linear regressions, and nested model comparisons. We found significant effects of set size and delay duration on both model-free and model-based measures of dispersion. However, the effect of delay duration was much weaker than that of set size, dependent on the analysis method, and apparent in only a minority of subjects. The highest forgetting slope found in either experiment at any set size was a modest 1.14/s. As secondary results, we found a low rate of nontarget reports, and significant estimation biases towards oblique orientations (but no dependence of their magnitude on either set size or delay duration). Relative stability of working memory even at higher set sizes is consistent with earlier results for motion direction and spatial frequency. We compare with a recent study that performed a very similar experiment.

Introduction

Visual working memory can maintain multiple items for durations of up to tens of seconds. Not surprisingly, two basic variables used to characterize working memory performance are the number of items (set size) and delay duration (a third basic variable would be the complexity or number of features per item). As in recent years, set size effects have received the most scrutiny in experimental psychology (for reviews, see Luck & Vogel, 2013; Ma, Husain, & Bays, 2014); the study of delay duration effects has ended up being somewhat overshadowed.

Perhaps the first researcher to vary delay duration in a controlled experiment was Friedrich Hegelmaier (1833–1906), when he was a medical student in Tübingen (Laming & Laming, 1992). His work was published in German in 1852 and republished in English in Laming and Laming (1992). Hegelmaier measured visual working memory for the length of a line segment using a sameness judgment task, and found that his accuracy dropped only modestly, from 80% to 73%, when he increased delay duration from 3 to 60 s. More recent studies have predominantly used two-alternative forced-choice delayed-discrimination tasks. Using delay durations between 1 and 30 s, no effect of delay duration on discrimination threshold was found for spatial frequency (Greenlee, Rischewski, Mergner, & Seeger, 1993; Magnussen, Greenlee, Asplund, & Dyrnes, 1990, 1991; Magnussen, Greenlee, & Thomas, 1996; Magnussen, Idas, & Holst-Mahre, 1998; Regan, 1985), speed (Greenlee, Lang, Mergner, & Seeger, 1995; Magnussen & Greenlee, 1992), motion direction (Blake, Cepeda, & Hirsh, 1997), and motion coherence (Blake et al., 1997). Discrimination thresholds for contrast, however, seem to approximately double as delay duration increases from 1 to 10 s (Greenlee, Magnussen, & Thomas, 1991; Harvey, 1986; Lee & Harris, 1996; Magnussen et al., 1996). Working memory for orientation (Magnussen et al., 1998; Magnussen, Landro, & Johnsen, 1985; Vogels & Orban, 1986) and vernier offset (Fahle & Harris, 1992) show modest rates of decay—for example, from 77% to 71% correct when going from 1- to 10-s delay (Magnussen et al., 1998). Another paradigm that has been used to measure delay duration effects is delayed...
estimation, in which the subject matches a probe stimulus to a memorized stimulus along a continuum in the feature of interest. Blake et al. (1997) used this paradigm for motion direction to replicate their results from the two-alternative forced-choice task. Consistently, Zokaei, Gorgoraptis, Bahrami, Bays, and Husain (2011) in a sequential presentation, found a significant effect of sequence length but not of serial position on the precision of encoding as estimated from a model. Two studies employing delayed estimation for color found mild declines of performance between 1 and 26 s at a set size of one (Nilsson & Nelson, 1981) and between 4 and 10 s at a set size of three (Zhang & Luck, 2009).

Despite extensive bodies of literature on set size and delay duration effects, their intersection is sparsely populated. Blake et al. (1997) used their delayed-estimation task not only to probe the memory of a single-motion direction, but also the memory of one of \( N \) simultaneously or sequentially presented motion directions, with set size \( N \) being 3, 5, 7, or 9 (Blake et al., 1997). They found a significant effect of set size on average error, but no significant effect of delay duration (0, 10, or 30 s) and no significant interaction. In a two-item spatial-frequency delayed-discrimination task, Magnussen and Greenlee (1999) similarly found a strong effect of set size (one or two) and no effect of delay duration (1, 3, or 10 s).

The present article uses delayed estimation to investigate the joint effects of set size and delay duration on visual working memory for orientation. Since the exact shape of the error distribution in delayed estimation has in recent years become a topic of interest (Fougnie, Suchow, & Alvarez, 2012; Ma et al., 2014; van den Berg, Shin, Chou, George, & Ma, 2012; Zhang & Luck, 2008), we characterize the error distribution using a variety of metrics besides average error. In addition, we examine estimation biases (Pratte, Park, Rademaker, & Tong, 2016; Van Bergen, Ma, Pratte, & Jehee, 2015) and nontarget reports (Bays, Catalao, & Husain, 2009; van den Berg, Awh, & Ma, 2014; Zokaei et al., 2011).

The joint variation of set size and delay duration could shed light on neural mechanisms of visual working memory. A recent biophysical computational model in which working memories are represented by patterns of persistent activity (so-called “bump attractors”) that fade and merge, predicts that the effect of delay duration interacts with set size: Delay duration would affect performance at intermediate (three or four), but not at small or large set sizes (Z. Wei, Wang, & Wang, 2012). Thus, it is possible that earlier studies did not find significant effects of delay duration on orientation memory performance because they only used a set size of one. In Wei et al.’s model, the merging of multiple patterns of activity arises largely because the patterns are jointly maintained in a one-dimen-

sional population, in which neurons only have feature selectivity; metaphorically speaking, the bumps have “too little space to avoid each other.” By contrast, if visual working memories were to a large extent maintained in visual sensory areas (Harrison & Tong, 2009; Pasternak & Greenlee, 2005; Sneve, Alns, Endestad, Greenlee, & Magnussen, 2012), spatial dimensions would be available in addition to the feature dimension. Then, one might predict largely independent maintenance and no interaction between the effects of delay duration and set size.

Experiment 1 presented here was part of the PhD thesis of author H.S. in April 2015. During the review process of the present article, we became aware of a recent article that also studies the temporal decay of visual short-term memory for orientation at multiple set sizes (Pertzov, Manohar, & Husain, 2017). This convergence offers an excellent opportunity for comparison. We comment on the similarities and differences with Pertzov et al. in the Results and Discussion sections. Experiment 2 was done in response to a reviewer concern.

**Methods**

**Subjects**

All subjects gave informed consent. Treatment of subjects was in accordance with the Declaration of Helsinki.

**Experiment 1**

Five subjects participated (including one author). The data were collected at the authors’ previous institution, Baylor College of Medicine, and the study was approved by the Institutional Review Board of Baylor College of Medicine. Subjects were compensated $50 for their participation.

**Experiment 2**

Six subjects participated (all naive; two male, four female). The data were collected at New York University, and the study was approved by the Institutional Review Board of New York University. Subjects were compensated $55 for their participation.

**Apparatus and stimuli**

**Experiment 1**

All stimuli were displayed on a 19-in. Dell LCD monitor (1280 × 1024 pixels) at a viewing distance of
approximately 60 cm; in calculations of degrees of visual angle (dva), we assume a viewing distance of exactly 60 cm. The stimuli were presented on a midlevel gray background (128 on an 8-bit grayscale) with a luminance of approximately 25 cd/m².

Each stimulus was a Gabor pattern with a Gaussian envelope of standard deviation 0.2 dva and a cosine modulation with a spatial frequency of 2.86 cycles/dva, a phase of 0, and a peak luminance of approximately 110 cd/m². We did not gamma-correct the monitor but because the only variable of interest was orientation, we do not believe this mattered.

A memory array consisted of one, two, four, or six Gabor stimuli with orientations independently drawn from a uniform distribution. The center of each Gabor was located at one of six fixed positions on an imaginary circle around fixation of radius 5 dva. These positions were spaced at 60° angular intervals (within the screen, not dva), starting from the position to the right of fixation. When set size was one, two, or four, a contiguous subset of positions was used.

**Experiment 1**

Identical to Experiment 1 except for the following. All stimuli were displayed on a 24-in. ViewSonic LED monitor (3840 × 2160 pixels) at a viewing distance of 61 cm (controlled using a headrest). The stimuli were presented on a midlevel gray background (128 on an 8-bit grayscale) with a luminance of 57.9 cd/m². The Gabor stimuli had a Gaussian envelope of standard deviation 0.25 dva, a cosine modulation with a spatial frequency of 2.28 cycles/dva, and a peak luminance of approximately 240 cd/m². The imaginary circle had a radius of 6.27 dva.

**Circular statistics**

To quantify the effects of set size and delay duration on the error distribution, we consider four measures of dispersion: mean absolute error, circular standard deviation, interquartile range, and circular variance. Interquartile range is the difference between the 75th and the 25th percentiles of the sorted errors. To compute circular standard deviation and circular variance, we first multiplied all errors by 2, so that they took values between –180° and 180° (which is the standard for calculations of circular statistics). We then regarded each error as a point on a unit circle and calculated the mean resultant length R of the unit vectors corresponding to all errors. Circular standard deviation is then \( \frac{1}{2} \sqrt{-2 \log R} \) (Mardia & Jupp, 1999), where we introduced the factor \( \frac{1}{2} \) to map back to the true orientation space. Circular variance is \( 1 - R \), which is between 0 and 1.

We also consider other circular statistics: circular mean, circular skewness, and circular kurtosis. Denot-
ing errors rescaled to \([-\pi, \pi]\) by \(\epsilon_i\), where \(i\) is the trial index, the circular mean \(\bar{\epsilon}\) is the angle of the resultant vector (Mardia & Jupp, 1999). For circular skewness and circular kurtosis, we use the simple definitions by Pewsey (2004), namely skewness = \(\frac{1}{n} \sum_{i=1}^{n} \sin 2(\epsilon_i - \bar{\epsilon})\), and kurtosis = \(\frac{1}{n} \sum_{i=1}^{n} \cos 2(\epsilon_i - \bar{\epsilon})\), where \(n\) is the number of trials.

**Experiment 1: Results**

**Model-free: Population**

We measured subjects’ ability to estimate orientations stored in visual working memory as a function of both set size (one, two, four, or six) and delay duration (1, 2, 3, or 6 s). The error of an estimate on a given trial is the circular distance between the estimate and the true orientation; it takes values between \(-90^\circ\) and \(90^\circ\). For visualization (but not for any statistical analysis), we binned errors in nine equal bins. The binned error distributions across conditions are shown in Figure 1. Visually, it appears that the error distribution is strongly affected by set size but only weakly by delay duration. The effect of set size has been widely reported (e.g., van den Berg et al., 2012; Wilken & Ma, 2004) and we will not emphasize it in this article.

The measures of error dispersion are shown in Figure 2. We calculated all measures for each subject, then took means and standard errors across subjects. Visually, these plots confirm the impression from the histograms: Each measure shows a strong effect of set size and a much weaker effect of delay duration. Repeated-measures analyses of variance (ANOVAs; Table 1) show effects of set size with \(p\) values lower than \(10^{-8}\), effects of delay duration with \(p\) values of

![Figure 1](http://arvojournals.org/)

**Figure 1.** Experiment 1: Binned error histograms grouped by delay duration (A) and set size (B), averaged over subjects. Each bin was 20° wide, and we plot the histogram value in the center of the bin. Here and elsewhere, error bars represent 1 SEM across subjects.

![Figure 2](http://arvojournals.org/)

**Figure 2.** Experiment 1: Measures of error dispersion as a function of delay duration and set size. (A) Mean absolute error. (B) Circular standard deviation. (C) Interquartile range. (D) Circular variance.
0.017 and lower; the significance of the interaction depends on the measure. Averaged across trials, delay duration, and set sizes, absolute error was 19.3° ± 1.0° (mean ± standard error of the mean across subjects).

Next, we performed a linear regression of mean absolute error against delay duration, separately at each set size, and following Pertzov et al. (2017), we call the slope of this regression line the “forgetting slope.” This slope ranges from 0.42 ± 0.16 at a set size of one to 1.14 ± 0.49 at a set size of six (Figure 3, blue). None of these were significantly different from 0 after a t test with Bonferroni-Holm correction (p values = 0.062, 0.13, 0.024, and 0.082). Repeated-measures ANOVAs did not reveal a significant effect of set size on forgetting slope, F(3, 12) = 3.39, p = 0.054.

For completeness, we also examined the effect of delay duration and set size on other moments of the error distribution (Figure 4). As expected, the circular mean and circular skewness are close to zero, and all p values in a repeated-measures ANOVA are greater than 0.08. For circular kurtosis, we find a significant effect of set size, F(3, 12) = 73.5, p < 10⁻⁷, but no significant effect of delay duration, F(3, 12) = 2.42, p = 0.12, and no significant interaction, F(9, 36) = 2.08, p = 0.058.

### Table 1. Repeated-measures analysis of variance on dispersion measures in Experiment 1.

<table>
<thead>
<tr>
<th>Measure of dispersion</th>
<th>Effect of set size</th>
<th>Effect of delay duration</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute error</td>
<td>F(3, 12) = 209 p &lt; 10⁻⁹</td>
<td>F(3, 12) = 6.11 p = 0.0092</td>
<td>F(9, 36) = 2.43 p = 0.028</td>
</tr>
<tr>
<td>Circular standard deviation</td>
<td>F(3, 12) = 185 p &lt; 10⁻⁹</td>
<td>F(3, 12) = 5.78 p = 0.011</td>
<td>F(9, 36) = 2.00 p = 0.068</td>
</tr>
<tr>
<td>Interquartile range</td>
<td>F(3, 12) = 121 p &lt; 10⁻⁸</td>
<td>F(3, 12) = 5.05 p = 0.017</td>
<td>F(9, 36) = 1.59 p = 0.16</td>
</tr>
<tr>
<td>Circular variance</td>
<td>F(3, 12) = 205 p &lt; 10⁻⁹</td>
<td>F(3, 12) = 6.22 p = 0.0086</td>
<td>F(9, 36) = 2.35 p = 0.034</td>
</tr>
</tbody>
</table>

Model-free: Individual subjects

We next examined individual subjects. Visually, the histograms (Figure 5) suggest individual differences: Effects of delay duration seem to be absent in subjects S1 and S3, minimal in S5, and modest in S2 and S4; it should also be kept in mind that S2 was an author. To quantify individual-subject effects, we performed a linear regression of circular standard deviation against set size, delay duration, and their product (to model a potential interaction), for each subject individually (Table 2). Besides providing within-subjects confidence intervals (CIs), regression has the advantage over ANOVA of taking into account the ordinal nature of both set size and delay duration, albeit in a specific way (here linear). The 95% CI for the coefficient of set size contained 0 for no subject, indicating a linear effect of set size on circular standard deviation. The 95% CI for the coefficient of delay duration contained 0 for 1 subject, indicating a linear effect of delay duration on circular standard deviation for most subjects.

Model-based: Estimates of (mean) precision

So far, we have used descriptive statistics to quantify dispersion. An alternative approach is to fit a model to the error distribution and estimate the model’s precision parameter as a function of set size and delay duration. Different models have been proposed to describe the error distribution. Typically, these models have theoretical roots. In one category of models, the central concept is that of variability in encoding precision, giving rise to “variable-precision models” (Fougnie et al., 2012; Keshvari, Van den Berg, & Ma, 2013; Ma et al., 2014; van den Berg et al., 2012). In a competing tradition, the central concept is that only a subset of items are encoded (so-called “high-threshold” or “item-limit” models; Luck & Vogel, 2013; Pashler, 1988; Zhang & Luck, 2008), giving rise to a “slots-plus-resources” or a “slots-plus-averaging” model in the context of delayed estimation (Zhang & Luck, 2008).

Although the slots-plus-resources and slots-plus-averaging models as originally presented have been refuted (Keshvari et al., 2013; van den Berg et al., 2012), the item-limit concept is potentially salvageable by combining it with the notion of variable precision (van den
Berg et al., 2014). Since this debate is not the focus of the present article, we have chosen to take a theoretically neutral approach by borrowing from each theoretical tradition only the parametric form of the error distribution. Thus, we treat each model as descriptive, moving only slightly beyond the model-free statistics reported above.

In a descriptive model, it is important to avoid overfitting, as that would reduce the reliability of parameter estimates. In the present experiment, the main focus is on the dependence of precision on set size and delay duration. This inevitably introduces 16 parameters, one for each condition. Besides these 16 parameters, we choose to minimize the number of additional parameters to avoid overfitting. The models below have 0 or 1 additional parameters.

- Pure von Mises model: The error follows a von Mises distribution: \[ p(\epsilon) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos \epsilon}. \] Here, \( \kappa \) is called the concentration parameter, and \( I_0 \) is the
modified Bessel function of the first kind of order 0. We define precision \( J \) as the Fisher information, which amounts to \( J = \kappa \frac{I_1(s)}{4\pi} \) (van den Berg et al., 2012), where \( I_1 \) is the modified Bessel function of the first kind of order 1. The pure von Mises model is generally not considered a good description of the error distribution (van den Berg et al., 2014; Zhang & Luck, 2008).

- Mixture model: Zhang and Luck (2008) proposed that the error distribution is a mixture of a con Mises distribution and a uniform distribution (corresponding to guesses): 
  \[
  p(\epsilon) = \frac{1}{2\pi} + \frac{1-\lambda}{\pi I_0(k)} e^{k \cos{\epsilon}},
  \]
  where \( \lambda \in [0, 1] \) is the weight given to the uniform component. We again define precision as Fisher information. To constrain the fits of the model, we assume that \( \lambda \) is shared across all set sizes and all delay durations.

- Variable-precision model with shared scale parameter: It has been proposed that precision \( J \) itself varies across trials (and items; Fougnie et al., 2012; van den Berg et al., 2012). Such variation could be caused by a variety of factors, including stimulus differences (Bae, Olkkonen, Allred, Wilson, & Flombaum, 2008; Pratte et al., 2016), variability in decay (Fougnie et al., 2012), Poisson fluctuations in spike count (Bays, 2014), and fluctuations in attention (Cohen & Maunsell, 2010; Goris, Simoncelli, & Movshon, 2014). We have previously parametrized the variable-precision model by assuming that precision \( J \) follows a gamma distribution with mean \( \tau \) and scale parameter \( k \) and shape parameter \( \tau \) is shared across all conditions (van den Berg et al., 2014). Thus, we use Gamma \( (J; k, \frac{1}{\tau}) \) instead of Gamma \( (J; k, \frac{1}{\tau}, \tau) \). In a single condition, the two versions of the variable-precision model are equivalent; however, the different constraints across conditions make the two versions different.

The pure von Mises model has 16 parameters, one for each combination of set size and delay duration. The other three models have one parameter more, which is shared across all conditions. Our primary goal here is not to compare between these models, but to examine whether conclusions about the effect of delay duration are shared across models.

We fitted all models using maximum-likelihood estimation. Our algorithm used a combination of Bayesian optimization (Brochu, Cora, & De Freitas, 2010) and pattern search (Audet & Dennis, 2006), as implemented by Luigi Acerbi. For parameters with a range \((0, \infty)\), we fitted the logarithm of the parameter for better optimization performance. For each \((\log)\) parameter, we set a lower and upper bound. To minimize the risk of getting stuck in local maxima, we ran each optimization 10 times, with different initializations, then picked the best. Each initialization was generated by independently drawing the value of each parameter from a uniform distribution over its range. As it turned out, the algorithm almost always converged on the same maximum across runs, making it likely that local maxima were not a problem.

We find that all four models provide good fits to the data (Figures 16 through 19). The estimates of the logarithm of precision (in the pure von Mises and mixture models) and the logarithm of mean precision (in the variable-precision models) again seem to show a strong effect of set size and a weaker effect of delay duration (Figure 6). In each model, a repeated-measures ANOVA shows significant effects of both set size and delay duration on the estimated log (mean) precision, and no significant interaction (Table 3).

### Model-based: Individual subjects

Using nested model comparison, we can perform an individual-subject analog of ANOVA on (mean) precision. An ANOVA is meant to determine whether the variation in a quantity across conditions is greater than expected by chance. At the individual-subject level, we can compare a model in which the quantity is free to vary across conditions to one in which the quantity is constrained to be equal across conditions. (A fully Bayesian treatment would also marginalize over parameters [Rouder, Morey, Speckman, & Province, 2012]; we do not do that here.)

Since we are most interested in the effect of delay duration, we compare the models above against their

### Table 2. Ninety-five percent confidence intervals of the coefficients in a linear regression of circular standard deviation against set size (mean-centered), delay duration (mean-centered), and their product (interaction) in Experiment 1.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Constant</th>
<th>Set size</th>
<th>Delay duration</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[23.7, 26.0]</td>
<td>[2.91, 4.14]</td>
<td>[0.02, 1.28]</td>
<td>[−0.09, 0.57]</td>
</tr>
<tr>
<td>2</td>
<td>[20.8, 22.6]</td>
<td>[3.59, 4.55]</td>
<td>[0.92, 1.90]</td>
<td>[−0.03, 0.49]</td>
</tr>
<tr>
<td>3</td>
<td>[25.4, 27.9]</td>
<td>[4.34, 5.66]</td>
<td>[−0.70, 0.65]</td>
<td>[−0.58, 0.13]</td>
</tr>
<tr>
<td>4</td>
<td>[21.2, 24.1]</td>
<td>[3.88, 5.40]</td>
<td>[1.51, 3.06]</td>
<td>[−0.11, 0.70]</td>
</tr>
<tr>
<td>5</td>
<td>[26.9, 28.4]</td>
<td>[4.03, 4.81]</td>
<td>[0.57, 1.38]</td>
<td>[0.14, 0.56]</td>
</tr>
</tbody>
</table>

Notes: Bold = confidence interval that does not contain 0.
counterparts in which precision or mean precision depends only on set size, but not on delay duration; the models are otherwise the same. Such a “null” variant of the pure von Mises model has four parameters (one precision parameter for each set size), and the null variants of the other three models each have five parameters. Thus, the null variants have 12 parameters fewer than the full models. We compare models using the Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978). Both metrics penalize a model for having more free parameters, but the BIC penalty is steeper.

Table 4 shows the model comparison results are consistent across models, but inconsistent across subjects and metrics. Overall, on an individual-subject basis, the evidence that delay duration affects (mean) precision is mixed. This is to some extent consistent with visual inspection of the individual-subject data (Figure 5): There is no discernible effect of delay duration for subjects S1 and S3.

One concern might be that we assumed that in the three models that had an extra, nonprecision parameter, this parameter was shared across all conditions. However, we verified in a separate analysis that the conclusions did not change if we allowed that parameter to vary across set size.

### Estimation Biases

In perception (Andrews, 1965; Girshick, Landy, & Simoncelli, 2011; X.-X. Wei & Stocker, 2015) and in visual working memory (Pratte et al., 2016; Van Bergen et al., 2015), orientation biases both towards and away from the cardinal orientations have been reported. In our previous analyses of dispersion, any orientation-dependent biases would merely contribute to the dispersion. To find any potential biases, we first plot reported orientation against target orientation across all subjects, set sizes, and delay durations (Figure 7A). The data seem to deviate systematically from the diagonal, indicating biases. Indeed, we find biases away from the cardinal directions (horizontal and vertical; Figure 7B), consistent with Pratte et al. (2016), Van Bergen et al. (2015), and X.-X. Wei and Stocker (2015). Using 20° bins, a repeated-measures ANOVA reveals a significant effect of target orientation, $F(8, 32) = 8.52, p < 10^{-5}$; using 10° bins like Pratte et al. (2016), the corresponding result is $F(17, 68) = 8.26$ with $p < 10^{-10}$. In Figure 7B, we show bias for 10° bins.

Next, we break the estimation biases down by condition. Figure 7C shows an example of the raw data: error as a function of target orientation S1 for delay 1 s and a set size of one. One way to summarize these data is as the absolute circular error averaged across target orientations. The disadvantage of that metric, however, is that it is positive even if there is no true effect. Therefore, we instead define the net oblique bias as the circular mean of the weighted errors, where the weight is 1 if the target orientation is counterclockwise from the closest oblique orientation (±45°), and −1 if the target orientation is clockwise from the closest oblique orientation (Figure 7D). Note that the calculation of this metric does not involve binning. We show net oblique bias as a function of delay duration and set size in Figure

<table>
<thead>
<tr>
<th>Model</th>
<th>Effect of set size</th>
<th>Effect of delay duration</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure von Mises</td>
<td>$F(3, 12) = 140$</td>
<td>$F(3, 12) = 5.41$</td>
<td>$F(9, 36) = 1.75$</td>
</tr>
<tr>
<td>Mixture</td>
<td>$F(3, 12) = 124$</td>
<td>$F(3, 12) = 6.29$</td>
<td>$F(9, 36) = 2.99$</td>
</tr>
<tr>
<td>Variable precision with shared scale</td>
<td>$F(3, 12) = 173$</td>
<td>$F(3, 12) = 6.53$</td>
<td>$F(9, 36) = 1.96$</td>
</tr>
<tr>
<td>Variable precision with shared shape</td>
<td>$F(3, 12) = 150$</td>
<td>$F(3, 12) = 5.92$</td>
<td>$F(9, 36) = 2.03$</td>
</tr>
</tbody>
</table>

Table 3. Repeated-measures analyses of variance on the estimates of log (mean) precision in different models for Experiment 1.
A two-way repeated-measures ANOVA shows no effect of delay duration, \( F(3, 12) = 0.21, p = 0.89 \), or set size, \( F(3, 12) = 1.54, p = 0.25 \), but a significant interaction, \( F(9, 36) = 2.90, p = 0.011 \); indeed, a bit of a cross-over effect is visible. Bias as a function of target orientation is shown separately for each delay duration and set size in Figure 20.

Nontarget reports

It has previously been reported that in delayed estimation, subjects sometimes report the feature value of a stimulus other than the target (Bays et al., 2009; van den Berg et al., 2014; Zokaei et al., 2011), potentially due to a loss of binding between feature and location (Ma et al., 2014). To look for such reports, we calculated on each trial the circular distances between the subject’s response and each of the \( N - 1 \) distractor orientations on that trial. The histograms of these distances look nearly flat at all set sizes and all delay durations (Figure 8). To quantify this, we followed Bays et al. (2009) and van den Berg et al. (2014) and fitted separately for each subject, delay duration, and set size, a mixture model consisting of a von Mises distribution

![Figure 7. Experiment 1: Estimation biases. Target orientation is defined as clockwise with respect to vertical. We binned target orientation into 10° bins. (A) Scatter plot of reported versus target orientation. Data were collapsed across set sizes and delay durations. Diagonal in dashed red. (B) Bias (circular mean of error distribution) as a function of target orientation (black line with error bars). Data were collapsed across set sizes and delay durations. Individual subjects in gray. (C) Scatterplot of error against target orientation for subject S1 at delay 1 s and a set size of one. (D) Weight assigned to target orientation in the calculation of net oblique bias. (E) Net oblique bias as a function of delay duration and set size.](http://arvojournals.org/)
centered at the target orientation, a von Mises distribution at each of the distractor orientations (with equal weights), and a guessing rate (weight to a uniform component); all von Mises had the same concentration parameter. Of interest is the total weight to the von Mises distributions corresponding to the distractors. Averaged across set sizes (excluding \( N = 1 \)) and delay durations, we find this weight to be \( 0.0284 \pm 0.0060 \). Thus, the role of nontarget reports was small in this experiment. The fitted parameters of this mixture model are shown by condition in Figure 21.

![Figure 8](http://arvojournals.org/) Experiment 1: Histograms of the nontarget error for individual subjects by delay duration and set size.

![Figure 9](http://arvojournals.org/) Experiment 2: Binned error histograms grouped by delay duration (A) and set size (B), averaged over subjects.
Experiment 2: Results

In Experiment 1, delay duration was blocked. As a result, the average interstimulus interval depended on delay duration. If proactive interference (e.g., Makovski & Jiang, 2008) were to play a role in this experiment, it would then be stronger at shorter delay durations. Therefore, we conducted a control experiment in which delay durations were interleaved. We ran six subjects for 960 trials each (half the number of trials of Experiment 1).

Model-free: Population

Overall mean absolute error was 19.1° ± 1.4°, consistent with Experiment 1 (t test, p = 0.9). As in Experiment 1, the error distribution is strongly affected by set size but only weakly by delay duration (Figure 9). The measures of error dispersion confirm a strong effect of set size and a much weaker effect of delay duration (Figure 10). Repeated-measures ANOVAs (Table 5) show effects of set size with p values lower than 10−8 and effects of delay duration with p values of 0.025 and lower, except for interquartile range. No interaction were significant.

Forgetting slopes ranged from 0.25 ± 0.13 at a set size of one to 0.88 ± 0.48 at a set size of six (Figure 3, orange). None of these were significantly different from 0 after a t test with Bonferroni-Holm correction (p values = 0.12, 0.043, 0.040, and 0.13). Repeated-measures ANOVAs did not reveal a significant effect of set size on forgetting slope, F(3, 15) = 1.46, p = 0.26.

Examining secondary metrics (Figure 11), we find that the circular mean and circular skewness are close to zero, and all p values in a repeated-measures ANOVA are greater than 0.12. For circular kurtosis, we find a significant effect of set size, F(3, 15) = 99.2, p < 10−9; a significant effect of delay duration, F(3, 15) = 8.03, p = 0.0020; and no significant interaction, F(9, 45) = 0.708, p = 0.70.

Model-free: Individual subjects

For individual subjects (Figure 12), effects of delay duration seem to be modest in S3, minimal in S4, and absent in all others. A linear regression of circular standard deviation against set size, delay duration, and their product for each subject (Table 6) shows that the 95% CI for the coefficient of set size contained 0 for no subject, while the 95% CI for the coefficient of delay duration contained 0 for 4 subjects. This indicates little evidence for a linear effect of delay duration on circular standard deviation.

Model-based: Estimates of (mean) precision

For Experiment 2, all models provide good fits to the data (Figures 22 through 25). The estimates of the logarithm of precision (in the pure von Mises and mixture models) and the logarithm of mean precision measures ANOVAs did not reveal a significant effect of set size on forgetting slope, F(3, 15) = 1.46, p = 0.26.

Examining secondary metrics (Figure 11), we find that the circular mean and circular skewness are close to zero, and all p values in a repeated-measures ANOVA are greater than 0.12. For circular kurtosis, we find a significant effect of set size, F(3, 15) = 99.2, p < 10−9; a significant effect of delay duration, F(3, 15) = 8.03, p = 0.0020; and no significant interaction, F(9, 45) = 0.708, p = 0.70.

Table 5. Repeated-measures analyses of variance on dispersion measures in Experiment 2.

<table>
<thead>
<tr>
<th>Measure of dispersion</th>
<th>Effect of set size</th>
<th>Effect of delay duration</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute error</td>
<td>F(3, 15) = 84.4</td>
<td>p &lt; 10−8</td>
<td>F(3, 15) = 4.99</td>
</tr>
<tr>
<td>Circular standard deviation</td>
<td>F(3, 15) = 85.6</td>
<td>p &lt; 10−8</td>
<td>F(3, 15) = 4.18</td>
</tr>
<tr>
<td>Interquartile range</td>
<td>F(3, 15) = 68.7</td>
<td>p &lt; 10−8</td>
<td>F(3, 15) = 2.78</td>
</tr>
<tr>
<td>Circular variance</td>
<td>F(3, 15) = 73.7</td>
<td>p &lt; 10−8</td>
<td>F(3, 15) = 4.35</td>
</tr>
</tbody>
</table>

Figure 10. Experiment 2: Measures of error dispersion as a function of delay duration and set size. (A) Mean absolute error. (B) Circular standard deviation. (C) Interquartile range. (D) Circular variance.
(in the variable-precision models) again seem to show a strong effect of set size and a weak effect of delay duration (Figure 13). In each model, a repeated-measures ANOVA shows significant effects of both set size and delay duration on the estimated log (mean) precision, and no significant interaction (Table 7).

Model-based: Individual subjects

We now compare a version of each model in which (mean) precision depends on delay duration against a null variant in which it does not. For no individual subject do we find consistent evidence (between AIC and BIC) that the null model is worse (Table 8). For S3,
Table 6. Ninety-five percent confidence intervals of the coefficients in a linear regression of circular standard deviation against set size (mean-centered), delay duration (mean-centered), and their product (interaction) in Experiment 2.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Constant</th>
<th>Set size</th>
<th>Delay duration</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[25.1, 29.0]</td>
<td>[5.11, 7.15]</td>
<td>[−0.30, 1.80]</td>
<td>[−0.67, 0.42]</td>
</tr>
<tr>
<td>2</td>
<td>[25.1, 29.1]</td>
<td>[4.76, 6.88]</td>
<td>[−0.92, 1.26]</td>
<td>[−0.66, 0.48]</td>
</tr>
<tr>
<td>3</td>
<td>[24.5, 26.8]</td>
<td>[4.31, 5.50]</td>
<td>[1.65, 2.88]</td>
<td>[0.39, 1.02]</td>
</tr>
<tr>
<td>4</td>
<td>[15.5, 17.5]</td>
<td>[2.20, 3.23]</td>
<td>[0.08, 1.14]</td>
<td>[0.07, 0.63]</td>
</tr>
<tr>
<td>5</td>
<td>[20.1, 23.2]</td>
<td>[4.80, 6.39]</td>
<td>[−0.57, 1.06]</td>
<td>[−0.46, 0.39]</td>
</tr>
<tr>
<td>6</td>
<td>[25.9, 28.7]</td>
<td>[4.26, 5.68]</td>
<td>[−0.11, 1.35]</td>
<td>[−0.30, 0.46]</td>
</tr>
</tbody>
</table>

Table 7. Repeated-measures analyses of variance on the estimates of log (mean) precision in different models for Experiment 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>Effect of set size</th>
<th>Effect of delay duration</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure von Mises</td>
<td>(F(3, 15) = 95)</td>
<td>(p &lt; 10^{-9})</td>
<td></td>
</tr>
<tr>
<td>Mixture</td>
<td>(F(3, 15) = 89)</td>
<td>(p &lt; 10^{-9})</td>
<td></td>
</tr>
<tr>
<td>Variable precision with shared scale</td>
<td>(F(3, 15) = 91)</td>
<td>(p &lt; 10^{-9})</td>
<td></td>
</tr>
<tr>
<td>Variable precision with shared shape</td>
<td>(F(3, 15) = 90)</td>
<td>(p &lt; 10^{-9})</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(F(3, 15) = 4.03)</td>
<td>(p = 0.027)</td>
<td>(F(9, 45) = 0.47)</td>
</tr>
<tr>
<td></td>
<td>(F(3, 15) = 4.07)</td>
<td>(p = 0.027)</td>
<td>(F(9, 45) = 0.50)</td>
</tr>
<tr>
<td></td>
<td>(F(3, 15) = 4.00)</td>
<td>(p = 0.028)</td>
<td>(F(9, 45) = 0.47)</td>
</tr>
<tr>
<td></td>
<td>(F(3, 15) = 3.74)</td>
<td>(p = 0.035)</td>
<td>(F(9, 45) = 0.62)</td>
</tr>
</tbody>
</table>

Comparison with Pertsov et al.

Experimental differences

Our experimental design closely resembles that of Pertsov et al. (2017), but with some potentially important differences. First, we used delays of 1, 2, 3, and 6 s, whereas Pertsov et al. used delays of 0.1, 1, 2, and 3 s. Second, we used a presentation time of 100 ms, whereas Pertsov et al. used 500 ms; the latter allows for better encoding, but also for covert shifts of
attention among items in the display. Third, Pertzov et al. interleaved durations, whereas we only did that in Experiment 2. Fourth, we cued the target by location, whereas Pertzov et al. cued by color (using eight easily distinguishable colors). Fifth, the stimuli used by Pertzov were substantially larger (bars of length 2 dva) than ours (Gaussian standard deviation of 0.2 or 0.25 dva). Finally, we drew orientations independently from a uniform distribution, whereas Pertzov et al. imposed the constraint that orientations within the same trial must differ by at least $10^\circ$. We will comment on the consequences of these differences in the Discussion.

Comparison of results

We now compare the results of Pertzov et al. (2017) to those of our Experiments 1 and 2, using their chosen measure of dispersion, mean absolute error. The 0.1-s delay used by Pertzov et al. is not usually regarded as a working memory delay (Long, 1980; Phillips, 1974; Sperling, 1960). Restricting ourselves to the working memory delays of 1, 2, and 3 s that are common between our experiments and Pertzov et al., we observe that mean absolute error is very similar across all three experiments, ranging from around $10^\circ$ at a set size of one to around $30^\circ$ at a set size of six and a delay of 3 s. At the common delay durations, errors were overall slightly higher in our study (Experiment 1: $18.6^\circ \pm 1.2^\circ$; Experiment 2: $18.6^\circ \pm 1.4^\circ$) than in Pertzov et al. (about $15.1^\circ$). Our results agree with Pertzov et al. in the main effects and interaction in an ANOVA on mean absolute error, and in the main effect of delay duration on the precision parameter in a mixture model.

Our results disagree from Pertzov et al.’s (2017) in that we did not find significantly nonzero forgetting slopes, an effect of delay time on forgetting slope, or an interaction in the mixture model. Pertzov et al.’s forgetting slopes were comparable to ours at a set size of one, but higher at other set sizes (Table 9; Figure 3). Pertzov et al. did not test other model-free measures of dispersion (circular standard deviation, interquartile range, or circular variance) or other models (pure Von Mises or variable precision). They also did not perform individual-subject regressions or model comparisons. In our hands, all those analyses made the case for an effect of delay duration or an interaction between delay duration and set size much more tenuous.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Pure von Mises</th>
<th>Mixture model</th>
<th>Variable precision, shared scale par.</th>
<th>Variable precision, shared shape par.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ΔAIC</td>
<td>ΔBIC</td>
<td>ΔAIC</td>
<td>ΔBIC</td>
</tr>
<tr>
<td>S1</td>
<td>2.4</td>
<td>56.1</td>
<td>7.5</td>
<td>65.9</td>
</tr>
<tr>
<td>S2</td>
<td>4.6</td>
<td>63.0</td>
<td>6.1</td>
<td>64.5</td>
</tr>
<tr>
<td>S3</td>
<td>11.4</td>
<td>47.0</td>
<td>9.1</td>
<td>49.3</td>
</tr>
<tr>
<td>S4</td>
<td>0.9</td>
<td>59.3</td>
<td>3.2</td>
<td>61.6</td>
</tr>
<tr>
<td>S5</td>
<td>16.9</td>
<td>75.3</td>
<td>11.9</td>
<td>70.3</td>
</tr>
<tr>
<td>S6</td>
<td>13.5</td>
<td>71.9</td>
<td>13.4</td>
<td>71.8</td>
</tr>
</tbody>
</table>

Table 8. Comparison of each model against a corresponding null model in which (mean) precision only depends on set size, but not on delay duration, for Experiment 2. Notes: Negative values (bold) means that the null model is worse, suggesting an effect of delay duration. AIC = Akaike information criterion; BIC = Bayesian information criterion.
Figure 14. Experiment 2: Estimation biases. Target orientations are binned into $10^\circ$ bins. (A) Scatter plot of reported versus target orientation. (B) Bias (circular mean of error distribution) as a function of target orientation (black line with error bars). (C) Net oblique bias as a function of delay duration and set size.

Figure 15. Experiment 2: Histograms of the nontarget error for individual subjects by delay duration and set size.
What could explain these substantial discrepancies between our results and Pertzov et al.'s? First, some of our nonsignificant results might have become significant if we had run more subjects (Pertzov et al. ran 10 subjects, we five and six per experiment, respectively). However, the model-based results were obtained for individual subjects. Moreover, the effect size (as seen in the forgetting slopes) would still have been much larger in Pertzov et al. Second, all of Pertzov et al.'s analyses included a delay duration of 0.1 s. This is not considered a pure working memory delay, as sensory and iconic memory can bolster the quality of the internal representation (Long, 1980; Phillips, 1974; Sperling, 1960). Therefore, it is possible that Pertzov et al. overestimated the pure working memory forgetting slope; however, forgetting slopes estimated from only the working memory delays did not show a different pattern (Figure 3; Table 9, column 3). Third, we ran a delay duration of 6 s in both experiments. It could be that forgetting flattens off after 3 s. Our data indeed show a hint of such flattening. Fourth, Pertzov et al. cued by color rather than by location. It is possible that maintaining color-orientation bindings requires more resources than maintaining location-orientation bindings, increasing the forgetting slope.

Finally, Pertzov et al. (2017) found substantially more nontarget reports than we did. Multiple factors could play a role: First, the distance-to-size ratio of our stimuli was larger for our stimuli, thereby potentially reducing nontarget responses (Bays, 2016). Second, the colors used by Pertzov et al. to cue an item might have been less distinguishable than the locations in our experiment.

## Discussion

Using a variety of model-free and model-based metrics, we characterized how well people recall orientation from visual working memory as a function of set size and delay duration. All metrics of dispersion show a strong effect of set size, consistent with previous work. Effects of delay duration are much weaker, depend on the analysis method, and are found for only a subset of subjects. Our results are consistent with a reported lack of effect of delay duration on orientation memory at a set size of one (Magnussen et al., 1998; Magnussen et al., 1985; Vogels & Orban, 1986).

However, it should be kept in mind that our longest delay was 6 s, shorter than in previous studies (10 or 30 s). It is possible that measures of error dispersion will increase substantially beyond 6 s. We already discussed the similarities and discrepancies with Pertzov et al. (2017) above.

Our results do not provide strong support for Z. Wei et al.'s (2012) “merging attractor” model, in which unambiguous effects of delay duration are predicted. To the extent that our results can speak to neural underpinnings, they would suggest largely independent, interference-free maintenance of orientation in working memory, of the kind that would likely be provided by topographic visual areas (Harrison & Tong, 2009; Pasternak & Greenlee, 2005; Sneve et al., 2012). Based on delayed estimation at a set size of three, it has been claimed that items “die a sudden death” in visual working memory for color (Zhang & Luck, 2009). We did not find any evidence for an analogous claim for orientation. This might in part be due to differences between color and orientation, and indeed, a mild decay of working memory for color would be consistent with earlier results (Nilsson & Nelson, 1981). In addition, there might be other relevant differences in the experimental design; for example, Zhang and Luck used a longest delay of 10 s, and added an articulatory task during the delay period. However, we do not believe that their empirical findings warrant a strong qualitative conclusion of “sudden death.” They reached this conclusion by fitting a mixture of a von Mises and a uniform distribution independently at each delay duration, and then performing an ANOVA on the estimates of the weights of the uniform component. However, they did not perform the more direct AIC or BIC comparison against a null model in which the weight of the uniform component is equal across delay durations. Moreover, a reasonable fit of the mixture model does not support the process claim that people truly guess randomly; to make that claim, one would need to show that the mixture model fits better than models without truly random guessing, such as the variable-precision model. Here, we did not attempt to adjudicate between these models; instead, we tried to show that our conclusions held across models. However, Fougnie, Suchow, and Alvarez (2013) did perform such a model comparison and found that
gradual decay provides a better description for color working memory. Nevertheless, the fact that delay duration seems to affect working memory for color and contrast much more strongly than working memory for spatial frequency, orientation, or motion direction is puzzling and needs to be explained.

**Keywords:** visual working memory, orientation, precision, maintenance

## Acknowledgments

This work was supported by grant R01EY020958 from the National Institutes of Health.

Commercial relationships: none.

Corresponding Author: Wei Ji Ma.

Email: weijima@nyu.edu.

Address: Center for Neural Science and Department of Psychology, New York University, New York, New York, USA.

## References


memory for the length of a line. *Psychological Research, 54,* 233–239.


Van Bergen, R. S., Ma, W. J., Pratte, M. S., & Jehee, J.


Appendix A: Supplementary figures

Experiment 1

![Figure 16. Experiment 1: Fit of the pure von Mises model. The model has one precision parameter per condition.](http://arvojournals.org/ on 03/10/2018)
Figure 17. Experiment 1: Fit of the mixture model. The model has one precision parameter per condition and a lapse rate shared across all conditions.

Figure 18. Experiment 1: Fit of the variable-precision model with one precision parameter per condition and a scale parameter shared across all conditions.
Figure 19. Experiment 1: Fit of the variable-precision model with one precision parameter per condition and a shape parameter shared across all conditions.

Figure 20. Experiment 1: Estimation bias as a function of target orientation (binned in 20° bins), delay duration, and set size.

Figure 21. Experiment 1: Parameters fitted per condition in a descriptive model that assumes that responses follow a mixture of a von Mises distribution centered at the target orientation, a uniform distribution reflecting guesses, and von Mises distributions centered at the distractor orientations. (A) Concentration parameter of the von Mises distribution. (B) Weight to the uniform distribution. (C) Weight to nontarget responses. Note that this descriptive model has no fewer than 48 parameters (16 conditions, three parameters per condition) and the parameter values are therefore less reliable than those of the more constrained mixture model used for Figure 5, Table 3, and Table 4; this is also reflected in the large error bars.
Experiment 2

Figure 22. Experiment 2: Fit of the pure von Mises model. The model has one precision parameter per condition.

Figure 23. Experiment 2: Fit of the mixture model. The model has one precision parameter per condition and a lapse rate shared across all conditions.
Figure 24. Experiment 2: Fit of the variable-precision model with one precision parameter per condition and a scale parameter shared across all conditions.

Figure 25. Experiment 2: Fit of the variable-precision model with one precision parameter per condition and a shape parameter shared across all conditions.

Figure 26. Experiment 2: Estimation bias as a function of target orientation (binned in 20° bins), delay duration, and set size.
Figure 27. Experiment 2: Parameters fitted per condition in a descriptive model that assumes that responses follow a mixture of a von Mises distribution centered at the target orientation, a uniform distribution reflecting guesses, and von Mises distributions centered at the distractor orientations. (A) Concentration parameter of the von Mises distribution. (B) Weight to the uniform distribution. (C) Weight to nontarget responses. The same caveat as in Figure 21C applies.