The role of ensemble average differs in working memory for depth and planar information

Ke Zhang

Sun Yat-Sen University, Department of Psychology, Guangzhou, China
Shaoxing University, Center for Brain, Mind, and Education, Shaoxing, China

Jiehui Qian

Sun Yat-Sen University, Department of Psychology, Guangzhou, China

The representation of individual planar locations and features stored in working memory can be affected by the average representation. However, less is known about how the average representation affects the short-term storage of depth information. To evaluate the possible different roles of the ensemble average in working memory for planar and depth information, we used mathematical models to fit the data collected from one study on working memory for depth and 12 studies on working memory for planar information. The pattern of recalled depth was well captured by models assuming that there was a probability of reporting the average depth instead of the individual depth, compressing the recalled front-back distance of the stimulus ensemble compared to the perceived distance. However, when modeling the recalled planar information, we found that participants tended to report individual nontarget features when the target was not memorized, and the assumption of reporting average information improves the data fitting only in very few studies. These results provide evidence for our hypothesis that average depth information can be used as a substitution for individual depth information stored in working memory, but for planar visual features, the substitution of target with the average works under a constraint that the average of to-be-remembered features is readily accessible.

Introduction

Although it has long been known that humans can well perceive the depth or distance information (e.g. a 6-month-old infant can see the depth of a "visual cliff" to avoid a fatal fall; Gibson & Walk, 1960), recent studies show that our ability of memorizing depth information for only a few seconds is poor (Qian, Li, Zhang, & Lei, 2020; Qian & Zhang, 2019; Reeves & Lei, 2017). Working memory for depth (WMd) is found to be inaccurate when a single depth position is required to be remembered (with a change detection accuracy of 78%), and even worse for multiple depth positions (with an accuracy below 70% for 4 items or more; see Qian & Zhang, 2019 & Wang, Jiang, Huang, & Qian, 2021). The inadequacy in working memory seems to be unique for depth information, because working memory for multiple planar visual features is fairly accurate (Luck & Vogel, 1997) and unbiased (Wilken & Ma, 2004).

One important characteristic of WMd is the so-called contraction bias (also called compression bias) – the depth positions near the observer are later recalled to be farther and those far from the observer are recalled to be nearer, so the whole range of recalled depths is narrower than the true range (Tanaka, Yamamoto, Watanabe, & Sung-En, 2016; Zhang, Gao, & Qian, 2021). The contraction bias has been consistently reported in delayed estimation tasks, which often required participants to memorize multiple depth positions defined by binocular disparity, and then asked them to recall one of the depth positions after a brief delay (e.g. Zhang, Gao, & Qian, 2021). The bias was originated from memory not perception, because the memory displays were presented long enough (approximately 800 to 1000 ms; Tanaka et al., 2016; Zhang, Gao, & Qian, 2021) for participants to perceive the stereoscopic depth accurately. Hence, the contraction bias is likely due to the poverty of WMd, which prompts participants to use compensation strategies further causing the systematic errors. In our previous study, we used a delayed estimation task and found that the magnitude of contraction increased with memory load when the target was presented alone during the test phase (single-display condition; Zhang, Gao, & Qian, 2021), but the magnitude of contraction remained unchanged with memory load when the other nontarget memory items were presented

Citation: Zhang, K., & Qian, J. (2022). The role of ensemble average differs in working memory for depth and planar information. *Journal of Vision*, *22*(6):4, 1–16, https://doi.org/10.1167/jov.22.6.4.

https://doi.org/10.1167/jov.22.6.4 Received September 6, 2021; published May 6, 2022 ISSN 1534-7362

during the test phase (whole-display condition). We suggested that the presence of nontarget items provided consistent 2-D spatial configuration and relational depth information during memory retrieval, which alleviated the contraction bias for a larger memory load (set size approximately 3 to 6).

"Swap" error and average representation in working memory

It is still unclear why the contraction bias occurred in WMd. One possibility is that the bias could be explained by the so-called "swap" errors – mistakenly reporting a nontarget item instead of correctly reporting the target item (Bays, Catalao, & Husain, 2009). The swap error may be originated from noise in memory for the cued features of items (feature-variability account), which leads to errors in the feature by which responses were cued (e.g. location) and thus inaccurate judgments on which item to be tested (Bays, Catalao, & Husain, 2009; Schneegans & Bays, 2016). It may also reflect a binding error in working memory – falsely binding a tested feature with a spatial location even if the features and locations are themselves correctly memorized (Bays, Catalao, & Husain, 2009). Recently, researchers also suggested that the nontarget response may reflect a guessing strategy that it is better to report a memorized item, although not the tested one, than to make a random guess (Huang, 2020; Pratte, 2019). In a typical depth recall task (e.g. Zhang, Gao, & Qian, 2021), the to-be-memorized depth positions were randomly selected from a set of depth planes without replacement. When a near position was selected as the target, there was a higher chance for the nontargets to be at farther positions, and vice versa. Therefore, participants are more likely to mistakenly swap a farther nontarget with a near target, or swap a nearer nontarget with a far target, which leads to the contraction bias. In this case, the swap error reflects a biased guessing strategy induced by the experimental setting.

Another possible explanation is that the recalled individual depth position may be biased toward the average (possibly over time), which leads the recalled depth to contract toward the average. There is growing evidence indicating that working memory stores not only the information of individual items but also the ensemble information of all items, such as the average (Brady & Alvarez, 2011; Dubé & Sekuler, 2015; Huang, 2020; Lew & Vul, 2015; Utochkin & Brady, 2020). Moreover, the average representation can further affect the individual representation (Brady & Alvarez, 2011; Utochkin & Brady, 2020). Hence, researchers suggest that the bias toward the average may reflect a hierarchical encoding in working memory: individual items are not stored independently but are structured from the level of feature representations to individual objects to the level of groups or ensembles, and these levels of structure interact (Brady & Alvarez, 2011).

The average representation can affect the performance of working memory even without an explicit instruction to estimate or memorize the average (e.g. Brady & Alvarez, 2011). This phenomenon may reflect that the averaging is automatic in working memory (Dubé & Sekuler, 2015). If so, the bias toward the average should be pervasive in working memory for the visuo-spatial information. However, note that the effect of average representation may work under the constraint that the average feature value across individual items should be easily perceivable. For example, the four bars are oriented toward –30 degrees, –15 degrees, 15 degrees, and 30 degrees, and their average orientation is clearly 0 degrees; however, if they were oriented toward –90 degrees, 0 degrees, 90 degrees, and 180 degrees, their average is unoriented. Indeed, studies that investigated the average representation in working memory often chose the individual feature values based on a predefined average feature value, so that their average representation is readily accessible (e.g. Brady & Alvarez, 2011; Utochkin & Brady, 2020). Similarly, the depth planes tested in the previous studies can also be easily averaged and finding the middle of the whole depth volume may be "natural" in such experimental settings, which results in the contraction bias in WMd.

Although the behavioral evidence of the contraction bias has been consistently observed in working memory, which one of the above mechanisms contributes to the bias is yet to be explored. Recently, Huang (2020) characterized this bias in visual working memory (VWM) to distinguish these two accounts by modeling responses in delayed estimation tasks (Huang, 2020). He found that the bias is mainly caused by the swap responses, whereas directly adding a target-based bias toward the average contributes only a little to the observed data. However, behavioral evidence demonstrated by Utochkin and Brady (2020) showed that the ensemble average could substantially influence the memory of individual items even after accounting for the possibility of swap errors. We think that the divergence of the results lies in how the ensemble information plays a part in memory responses: if the target has not been correctly retained, the mean representation is then possibly recalled to serve as an informative guess, resulting in a bias toward the average; whereas if the target has been memorized with sufficient precision, the mean representation is no longer in effect. This may explain why directly adding a mean bias to the target response works under constraints. Therefore, we suggest that the contraction bias results from two separate sources of swap errors – mistakenly recalling a nontarget item or strategically using the ensemble average as a substitute.

In addition, although there is convincing evidence suggesting that the ensemble statistics plays an important role in working memory, exactly which part of memory stimuli constitutes the ensemble is still an open question. It seems that by default, the ensemble is composed of all memory items within one trial, as the aforementioned studies implied. However, researchers showed that participants can also preserve and combine visual information across trials (Akrami, Kopec, Brody, & Diamond, 2018;Chetverikov, Campana, & Kristjánsson, 2016; Crawford, Corbin, & Landy, 2019). In a recent study, Crawford et al. (2019) found that the estimate of averaging several visual stimuli in one trial was biased toward the central value of stimuli in previous trials. Therefore, we think that if ensemble average does contribute to the contraction bias, the question of whether the ensemble was defined by stimuli in the present trial, or by stimuli in all past trials, needs to be clarified.

Mathematical models for working memory

Past research on VWM has examined how mathematical models with different response components could fit to the behavioral data from working memory tasks, and thus to provide theoretical evidence for validating specific response components. For example, Zhang and Luck (2008) devised a classic model assuming that the recall performance for VWM can be characterized by a component of successfully memorizing the target and a component of random guessing if the target is absent in memory. Although the authors showed that this "target and guess" model provides relatively satisfying fits to the data, Bays, Catalao, and Husain (2009) argued that swap errors could be an important response component and the revised model showed that the swap error increases with memory load (also see Schneegans & Bays, 2016). This model is in accordance with our intuition that we are more likely to mistake one object for another when various objects are to be memorized, and it also seems to be consistent with the previous findings on WMd that the contraction bias increased with set size in the single-display condition (Zhang, Gao, & Qian, 2021). In addition, variants of models have been developed over the decades, examining the nature of the internal representations in VWM and the underlying neural mechanisms (Hardman, Vergauwe, & Ricker, 2017; Schneegans, Taylor, & Bays, 2020). For example, Hardman and colleagues (2017) developed a mathematical model of VWM that can differentiate between the continuous and categorical memory representations, and presented evidence showing that coarse categorical information is also represented in VWM. Based on the principles of stochastic sampling and neural coding, Schneegans and colleagues (2020) developed a model that establish a unified computational framework to reconcile the different accounts regarding whether the storage in VWM is discrete-slot based or continuous-resource based.

However, none of the previous modeling studies has investigated the possibility of "swapping" the target representation with the average representation. In Huang (2020), he characterized the recall estimates toward the average as an inherent bias added on to the recall estimate of the target, which was essentially different from a swap error. We think that it is important and necessary to distinguish the two types of "swapping" – swapping a nontarget or the ensemble average, because the nontargets and the ensemble average may be represented with different memory precision. In a typical recall task, it is unknown to an observer which one of the memory displays is the target until the testing stage. Hence, the variability of recall of a nontarget should in principle be the same as that of a response based on recalling the target. On the other hand, studies have shown that the representation of the ensemble average is different from that of individuals (e.g. Ariely, 2001), indicating that the variability of responses to the average could differ from that of responses to an individual item. In other words, modeling a swap error of nontargets can be different from that of the average in their variances of the underlying response distributions.

In this study, our primary purpose is to explore whether the contraction bias found in WMd can be attributed to swapping the target with the average representation or the nontarget representation. Here, we used the data from our previous study on WMd (Zhang, Gao, & Qian, 2021) and collected a new data set to perform the analysis. The behavioral data was fitted into mathematical models that consisted of one to four response components – target response (reporting the tested depth), average response (reporting the average as the target), nontarget response (swapping the target with a nontarget), and random guessing. In addition, we tested the different average responses based on an ensemble of memory items in one trial, or of all memory items in past trials. If "swapping with the average" contributes to the contraction bias, we should observe a considerably high probability of average response, and the models, including the component of average response, should outperform the models without this component.

Our secondary purpose is to evaluate and compare the contribution of average response in working memory for depth, for planar visual features, and for 2-D planar locations. If the averaging process is automatic in working memory, a tendency to report the

average should occur universally in working memory for both depth and planar locations and features. But if it is an optional strategy, it should occur only in some situations, especially when the ensemble average is easily accessible.

Methods

Empirical data

Working memory for depth

We used part of the previously published data set from Zhang et al. (2021) for analysis. The data were collected using a typical delayed estimation task, whose experimental procedure was summarized as follows (for details see Zhang, Gao, & Qian, 2021). Participants viewed an array of memory items (blue squares) presented at different stereoscopic depth positions defined by binocular disparity. They were asked to reproduce the depth of one selected item (target depth) after a brief delay by adjusting the depth position of a probe. The probe could be presented alone (single-display condition), or be presented with the other nontarget items (whole-display condition). The set size of the memory array was up to six. There were 16 participants in the original study, but seven of them lacked the detailed records for the depth positions of memory array in each trial and thus were excluded. We recruited six new participants (all women; mean age = 24.3 years), and the data sets from 15 participants with a total of 10,500 trials were used for model fitting. We received the ethical approval for this research from institutional review board (IRB) of our department. Written informed consent was obtained from each participant prior to all the experiments.

Planar visual working memory

We collected 12 published data sets of visual working memory for planar stimuli (Aagten-Murphy & Bays, 2019; Bays, 2014; Bays, Catalao, & Husain, 2009; Bays, Wu, & Husain, 2011; Gorgoraptis, Catalao, Bays, & Husain, 2011; Pratte, Park, Rademaker, & Tong, 2017; Utochkin & Brady, 2020; van den Berg, Shin, Chou, George, & Ma, 2012; see Table 1 for details), hereafter referred to as approximately E1 to E12. Three of the data sets tested working memory for planar locations (approximately E1 to E3), and the others tested working memory for planar features. In the most of these experiments, the features or locations of memory items were randomly selected: locations arranged in a circular configuration, colors on a colorful wheel, orientations form a range of 0 degrees to 180 degrees, or a range of 0 degrees to 360 degrees. In approximately E10 to 12, the authors tested how the ensemble average affected memory performances of individual orientations, and the average feature value was predetermined to aid in selecting orientations of individual memory items on each trial. Together, there were 420 subjects and 106,724 trials in total.

Mathematical modeling

We assumed that the response in a delayed estimation task could be characterized by combinations of four components – target response, average response, nontarget response, and random guessing. Figure 1 illustrates the probability distributions of each response component in the models.

A target response indicates that participants make an informative recall when they correctly memorized the target. The target component can be defined as:

$$\mathrm{p}\left(\hat{\theta}\right) = \gamma_T \varphi_{(\mu,\sigma)}\left(\hat{\theta} - \theta\right), \quad (1)$$

No.	Study	Stimulus	Set size	Display	Subjects
1	Aagten-Murphy & Bays (2019) Experiment 1	2D Location	1 2 4	Single	12
2	Aagten-Murphy & Bays (2019) Experiment 2	2D Location	4	Single	12
3	Aagten-Murphy & Bays (2019) Experiment 3	2D Location	4	Single	12
4	Bays, Catalao, & Husain (2009)	Color	1 2 4 6	Single	12
5	Bays (2014) Experiment 1	Orientation (bar)	1 2 4 8	Single	8
6	Gorgoraptis, Catalao, Bays, & Husain (2011) Experiment 2	Orientation (bar)	1–5	Single	8
7	Bays, Wu, & Husain (2011)	Orientation (bar)	1 6	Single	10
8	van den Berg, Shin, Chou, George, & Ma (2012)	Orientation (Gabor)	1–8	Single	6
9	Pratte, Park, Rademaker, & Tong (2017)	Orientation (Gabor)	1 2 3 6	Single	12
10	Utochkin & Brady (2020) Experiment 1	Orientation (triangle)	4	Single	16
11	Utochkin & Brady (2020) Experiment 2	Orientation (triangle)	4	Single	16
12	Utochkin & Brady (2020) Experiment 3	Orientation (triangle)	3	Single	296

Table 1. Summaries of the studies on visual working memory of planar stimuli reanalyzed in this study.

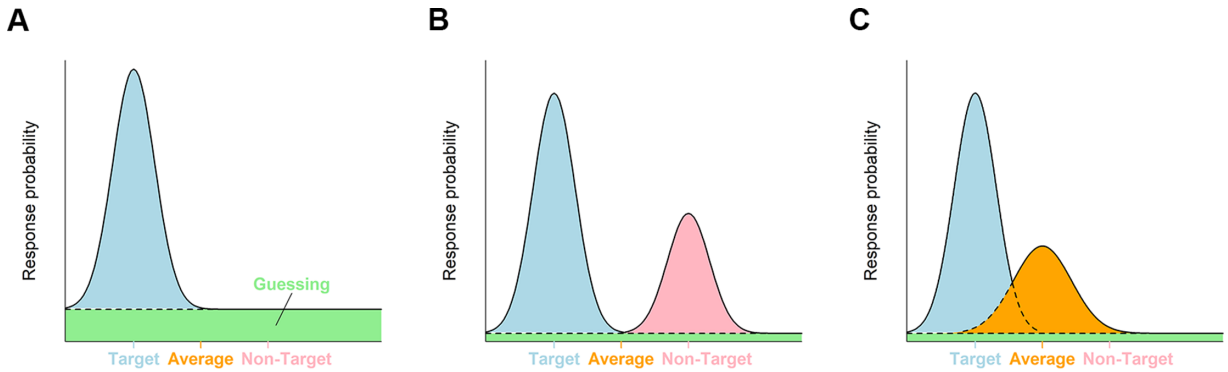

Figure 1. The probability distribution of the four response components in the model. (**A**) the distribution of the participant's response based on the target and random guessing. Variability in response for memorized target predicts a Gaussian distribution centered on the actual target feature (shown in blue), and random guessing predicts a uniform distribution (shown in green). (**B**) the distribution of response based on the target (blue), random guessing (green), and the nontarget. Variability in response for the non-target predicts a Gaussian distribution centered on the nontarget feature, with the same standard deviation as the target distribution (shown in red). (**C**) The distribution of response based on the target (blue), random guessing (green), and the average. Variability in response for the average predicts a Gaussian distribution centered on the average value, but with the different standard deviation (shown in orange).

where γ_T is the proportion of trials in which participants make a target response, θ indicates the to-be-memorized value of the target, and $\hat{\theta}$ indicates the recalled value. For WMd, the target-based recalls are assumed to be normally distributed, and thus φ represents a Gaussian distribution. For planar VWM, the responses are made on a circular analogue, and therefore φ are assumed to be a Von Mises distribution (Gaussian-like distribution for circular dimensions). The φ is characterized by its mean, μ, and its standard deviation, σ. Because studies showed that the recalled depth is not only contracted but also severely overestimated (e.g. Zhang, Gao, & Qian, 2021), for WMd, we set μ as the mean of participants' recalled depths to compensate for the overestimation. For planar VWM, we set μ as 0, because no such overestimation was found in VWM for color and orientation (Wilken & Ma, 2004). The standard deviation of the response distribution, σ, which is the variability of target-based recalls, reflects the precision of memory recalls of the target – a smaller σ indicates a more precise recall. The free parameters of this component are γ_T and σ. Because the stereoacuity for the most able observers is about 0.007 degrees (Carrillo, Baldwin, & Hess, 2020), and the least adjustable unit of depth in the task is 0.01 degrees, we restricted that σ cannot be smaller than 0.01 degrees for WMd.

An average response indicates that participants recall the average representation instead of the target. The average component can be defined as:

$$\mathrm{p}\left(\hat{\theta}\right) = \gamma_A \varphi_{(0,\sigma_A)}\left(\hat{\theta} - \bar{\theta}\right), \quad (2)$$

where γ_A is the proportion of trials in which participants make an average response, and $\bar{\theta}$ indicates the ensemble average, which was set as the mean of all the individual values. The standard deviation (variability) of responses based on the average, σ_A, reflects the precision of recalls for the average depth – a smaller σ_A indicates a more precise recall. The free parameters of this component are γ_A and σ_A. For WMd, we restricted that σ_A cannot be smaller than 0.01 degrees.

A nontarget response indicates that participants mistakenly recall a nontarget item (Bays, Catalao, & Husain, 2009). The nontarget component can be defined as:

$$\mathrm{p}\left(\hat{\theta}\right) = \gamma_N \frac{1}{m} \sum_{i}^{m} \varphi_{(\mu,\sigma_m)}\left(\hat{\theta} - \theta_i^*\right), \quad (3)$$

the to-be-memorized values of m nontarget items in one trial are donated by $\{\theta_1^*, \theta_2^*, \ldots\ldots, \theta_m^*\}$. Each nontarget has an equal probability of being mistakenly reported as an estimate of the target. Because participants could not distinguish the target item from nontarget items until the testing phase, all the items should be memorized with equal precision in principle. Hence, the nontarget response and the target response share the same parameter of variability (i.e. σ_m is equal to σ, and the mean of the response distribution, μ). The only free parameter of the nontarget component is γ_N, which is the proportion of trials in which participants make a nontarget response.

Random guessing indicates that the participants make a random guess when they fail to retain the target, so responses on these trials are drawn from a uniform distribution. The guessing component can be defined as:

$$\mathrm{p}\left(\hat{\theta}\right) = \gamma_G \frac{1}{r}, \quad (4)$$

Model	Target response	Average response	Nontarget response	Guess
T&G	√	–	–	√
TA	√	√ (Trial average)	–	√
WA	√	√ (Whole average)	–	√
TA&WA	√	√ (Trial average and whole average)	–	√
T&G&N	√	–	√	√
TA&N	√	√ (Trial average)	√	√
WA&N	√	√ (Whole average)	√	√

Table 2. Summaries of models compared in this article.

where the r is the span of the possible deviation between the recalled value and the to-be-memorized value of the target, $\hat{\theta} - \theta$. The r is 3.06 degrees for depths, 180 degrees for orientations of bars and Gabors, and 360 degrees for planar locations, colors, and orientations of the triangles used in the included experiments. γ_G is the proportion of trials in which participants make random guesses, which is equal to 1 minus the proportions of all other types of responses.

We compared seven models that differ in their assumptions of response components (Table 2).

Target and guess model

The basic model, which was developed by Zhang and Luck (2008), assumes that participants either make informative recalls based on the knowledge of the target or make noninformative random guesses. The model therefore has a target response component and a random guessing component, and is addressed as the "target and guess (T&G)" model hereafter. T&G model is defined as:

$$\mathrm{p}\left(\hat{\theta}\right) = \gamma_T \varphi_{(\mu,\sigma)}\left(\hat{\theta} - \theta\right) + \gamma_G \frac{1}{r}. \quad (5)$$

Trial-average model

This model is a variant of the T&G model, which includes a target response component, a random guessing component, and an average response component. In particular, the average response is based on the average of the to-be-memorized values in one trial (i.e. the trial average, which is termed as the TA component). It can be defined as:

$$\bar{\theta}_{TA} = \frac{\sum_j^m \theta_j}{m}, \quad (6)$$

where the θ_j is the value of the jth item among m items on a trial. For planar VWM, the $\bar{\theta}_{TA}$ is the circular mean (Fisher, 1995) of the feature values on a trial.

The TA model can be defined as:

$$\mathrm{p}\left(\hat{\theta}\right) = \gamma_T \varphi_{(\mu,\sigma)}\left(\hat{\theta} - \theta\right) + \gamma_A \varphi_{(0,\sigma_A)}\left(\hat{\theta} - \bar{\theta}_{TA}\right) + \gamma_G \frac{1}{r}. \quad (7)$$

Whole-average model

This model is otherwise identical to the TA model, except that the average response is based on the average of to-be-memorized values in the present and all previous trials (i.e. whole average, which is termed as the WA component). The whole average on the nth trial can be defined as:

$$\bar{\theta}_{WA} = \frac{\sum_i^n \sum_j^{m_i} \theta_{ij}}{\sum_i^n m_i}, \quad (8)$$

where the θ_{ij} is the value of the jth item among $\mathrm{m_i}$ items on the ith trial in a total of n previous trials. For planar VWM, the $\bar{\theta}_{WA}$ is the circular mean of all the feature values on the current trial and all previous trials. Note that replacing $\bar{\theta}_{TA}$ in Equation 7 with $\bar{\theta}_{WA}$ defines the WA model.

Target and Guess and Non-target model. This model is another variant of the T&G model, which includes a target response component, a random guessing component, and a nontarget response component. No average response component is incorporated. The model can be defined as:

$$\mathrm{p}\left(\hat{\theta}\right) = \gamma_T \varphi_{(\mu,\sigma)}\left(\hat{\theta} - \theta\right) + \gamma_N \frac{1}{m} \sum_i^m \varphi_{(\mu,\sigma_m)}\left(\hat{\theta} - \theta_i^*\right) + \gamma_G \frac{1}{r}. \quad (9)$$

The last two models incorporate an average component to the target and guess and nontarget (T&G&N) model. If the average response is based

on the trial average ($\bar{\theta}_{TA}$), the model is termed as trial-average and non-target (TA&N) model; if the average response is based on the whole average ($\bar{\theta}_{WA}$), the model is termed as whole-average and nontarget (WA&N) model. Both can be defined as:

$$\mathrm{p}\left(\hat{\theta}\right) = \gamma_T \varphi_{(\mu,\sigma)}\left(\hat{\theta}-\theta\right) + \gamma_A \varphi_{(0,\sigma_A)}\left(\hat{\theta}-\bar{\theta}\right) + \gamma_N \frac{1}{m}\sum_{i}^{m}\varphi_{(\mu,\sigma_m)}\left(\hat{\theta}-\theta_i^*\right) + \gamma_G \frac{1}{r}. \quad (10)$$

Note that $\bar{\theta}$ in Equation 10 is replaced with $\bar{\theta}_{TA}$ for the TA&N model, and is replaced with $\bar{\theta}_{WA}$ for the WA&N model.

An additional model with both a TA component and a WA component were evaluated if the models with an average response component (e.g. the TA model, the WA model, etc.) outperformed the models without (e.g. the T&G model, and the T&G&N model). This is the trial-average and whole-average (TA&WA) model, which can be defined as:

$$\mathrm{p}\left(\hat{\theta}\right) = \gamma_T \varphi_{(\mu,\sigma)}\left(\hat{\theta}-\theta\right) + \gamma_{TA} \varphi_{(0,\sigma_{TA})}\left(\hat{\theta}-\bar{\theta}_{TA}\right) + \gamma_{WA} \varphi_{(0,\sigma_{WA})}\left(\hat{\theta}-\bar{\theta}_{WA}\right) + \gamma_G \frac{1}{r}, \quad (11)$$

where γ_{TA} and γ_{WA} are the proportion of trials in which participants reported the TA and the WA, with the standard deviation of σ_{TA} and σ_{WA}, respectively. Note that, in principle, we could add a nontarget component in the TA&WA model, but this would result in too many free parameters in the model and therefore it was not evaluated.

For each model, the maximum likelihood estimates (MLEs) of the free parameters were obtained separately for each participant and each set size condition using a nonlinear optimization algorithm (Nelder & Mead, 1965). We used MATLAB and the toolbox adapted from BaysLab (http://bayslab.com) to run modeling.

Model comparison and parameter estimates analysis

Working memory for depth

We use two approaches to qualitative evaluate how well the model fits the data newly collected and those from the previous study on WMd (Zhang et al., 2021). First, we examined whether a model could simulate the contraction bias observed in WMd (i.e. the bias of overestimating the nearer depths [negative disparity] and underestimating the far depths [positive disparity]), which in the data resulted in the slope of the linear regression between the recalled depths and the true depths being smaller than one (for details, see Zhang et al., 2021). Hence, if a model could well simulate the contraction bias, the slope of the linear regression between its predicted depths and the true depths should also smaller than one. Second, we examined whether a model could simulate how the magnitude of the contraction bias varied with the set size of memory items. In the raw data, the contraction bias increased with set size in the single-display condition, but it did not vary with set size in the whole-display condition. This indicates that the probability of the possible origin(s) of contraction bias (average response or/and nontarget response) increases with set size in the single-display condition, but remains stable across set sizes in the whole-display condition. A well-fitting model should also capture this characteristic. Therefore, we conducted repeated-measures ANOVA to test whether the probability of each type of response varied with set sizes (trend analyses). In this study, p values for multiple comparisons were corrected by an adjustment of false discovery rate (FDR correction; Benjamini & Hochberg, 1995).

To quantitatively compare the model fittings, we calculated the Akaike Information Criterion (AIC; Akaike, 1974) per model fitting for each participant. Lower AIC values indicate better fittings. The best fitting model was chosen based on the AIC value.

In addition, we further examined the precision of memory representation indicated by the best fitting model. The precision of memory is reflected by the standard deviation of responses. In this study, the standard deviation of each type of response component is a parameter estimated in the corresponding model (in the unit of degree of visual angle), therefore based on the types of response component embedded in the model, we may be able to evaluate the precision of different memory responses. In this case, a paired-samples t-test (for a model including target/nontarget and average response components; e.g. Bays, Catalao, & Husain, 2009) would be conducted accordingly.

Planar visual working memory

For the studies on planar VWM, our main focus is to evaluate how pervasive the average responses are used in the task of working memory for planar information. Because the previous studies that investigated memory averaging often used different tasks compared to that investigated individual memory representation (e.g. the former usually involved a task instruction of estimating the average; Ariely, 2001), the representation precision of the memory average and individual item has not been evaluated within a single experimental paradigm. Here, we used mathematical models to differentiate the average-based responses and the individual-based (target/nontarget) responses in each study, and this allows us to evaluate the two types of memory representation without specific task instructions of estimating either an individual or the average. The same models were applied, and the AIC value per model fitting for each participant in each experiment

was calculated for comparison. The results of AIC values indicated whether incorporating an average response component in the model could improve the goodness of fit. In addition, for the best fitting model, we conduct a paired-samples *t*-test to compare the precision of memory representation indicated by the specific response component.

Results

Working memory for depth

Model fitting

Figures 2A and 2B show participants' recalled values for target depths and the estimated values predicted by the models, and Figures 2C and 2D show the 95% confidence interval of the slope of each regression line. The slope of the fitted line of raw data was 0.84 in the single-display condition and 0.90 in the whole-display condition. The contraction bias was captured by the models, including a single average component (either TA or WA), with the slopes being around 0.90 for both the single and whole displays (for the single display: TA = 0.90; WA = 0.89; TA&N = 0.91; and WA&N = 0.89; for the whole display, TA = 0.92; WA = 0.92; TA&N = 0.93; and WA&N = 0.92). However, the models without an average component predicted no such a bias, with the slopes being about 0.99 for the T&G model (single display = 0.99 and whole display = 0.98), and 0.96 for the T&G&N model (single display = 0.96 and whole display = 0.96). Note that, in principle, the T&G model

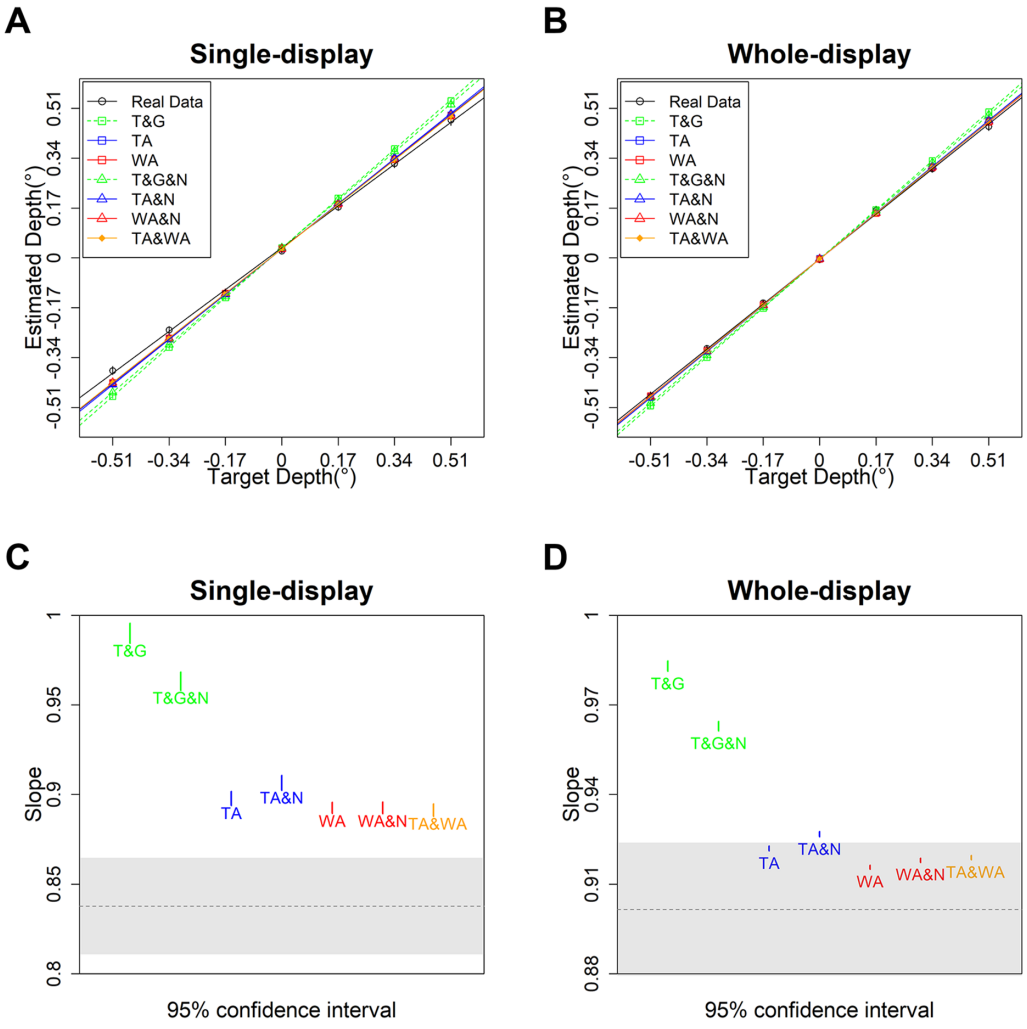

Figure 2. Model comparison of predicted performance. (**A**) and (**B**) Participants' estimates of each target depth and the corresponding predicted values of models in single- and whole-display conditions, respectively. Lines indicate the linear regression lines of the relations between depth estimates (or predicted values) and target depths. Error bars represent ± 1 SEM. (**C**) and (**D**) The 95% confidence intervals (CIs) of the slopes of the regression lines in **A** and **B**. Bars represent CIs of slopes predicted by models. The dash line represents the slope of the regression line between participants' real recalled depths and the target depths, and the gray area represents the CI of this slope.

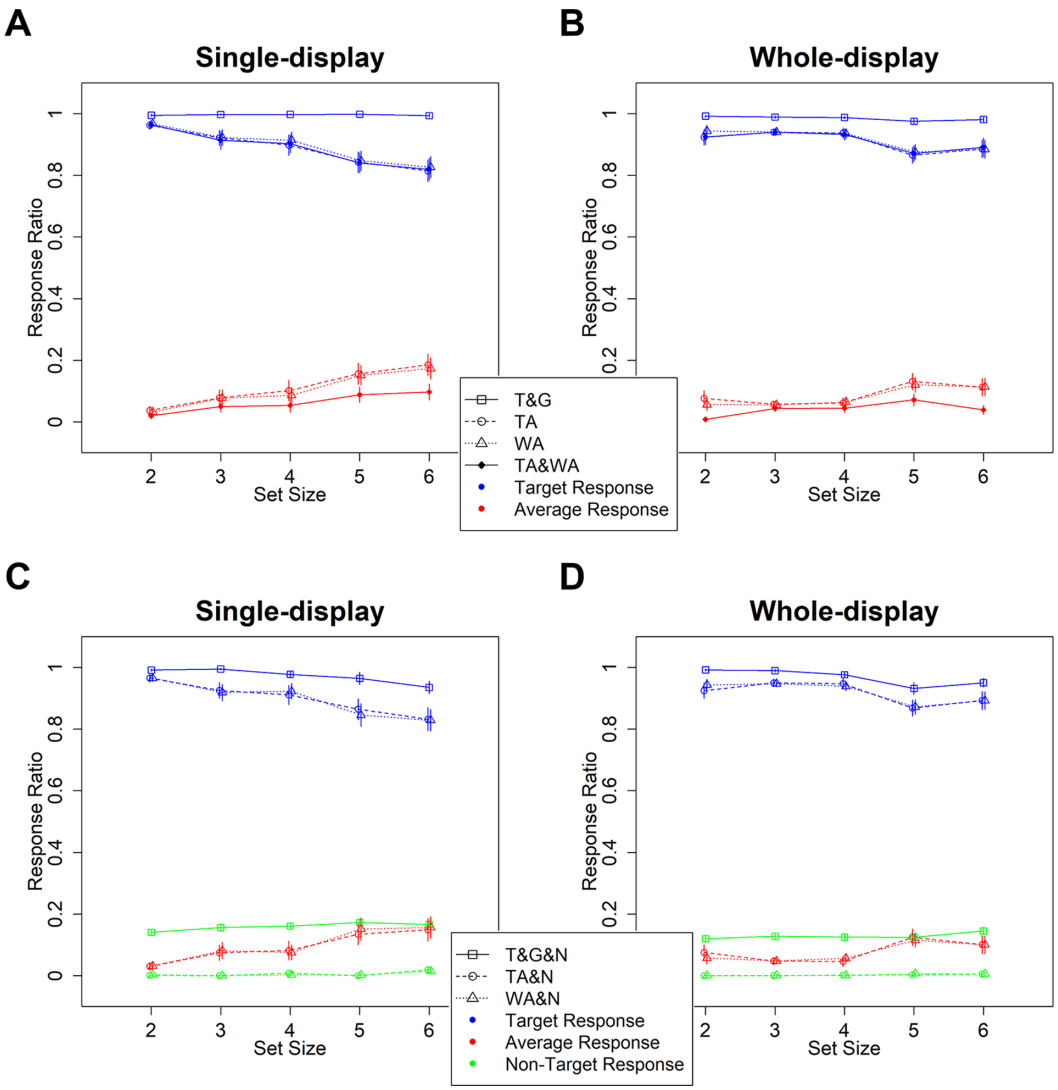

Figure 3. The estimated probability of each response component as a function of set size. (**A**) and (**B**) The response ratios of T&G, TA, WA, and TA&WA models in single- and whole-display conditions, respectively. (**C**) and (**D**) The response ratios of T&G&N, TA&N, and WA&N models in single- and whole-display conditions, respectively. Error bars represent ± 1 SEM.

cannot predict a slope other than 1.0, because the random guesses should not be expected to converge to the center of the depth volume. The experimental settings of the delayed estimation tasks for depth might induce a contraction bias in the T&G&N model (see *"Swap" error* section in *Introduction*), but the results of modeling fittings suggest that the nontarget response component cannot sufficiently explain amount of contraction bias observed in the behavioral data. Overall, adding a nontarget component did not improve the fitting for the models that already had an average component. These results showed that the average response components, not the nontarget response component, are crucial in simulating the contraction bias.

Because both the TA model and the WA model predicted the contraction bias well, we further fitted the TA&WA model. The slope is 0.89 for the single display and 0.92 for the whole display. The 95% confidence intervals (Figures 2C, 2D) showed that the TA&WA model fitted slightly better than the TA model, and slightly worse than the WA model, suggesting that the average over the previous trials might weight more than the trial average in the response.

Figure 3 shows the estimated probability of each response component as a function of set size (see Table 3 for reports on p values of ANOVAs). For the single display, the models including the average components all predict that the probability of average response increases with set size ($ps < 0.024$, although not for the TA component in the TA&WA model), which mirrors the trend that the contraction bias increased with set size in the single display. For the whole display, the models including the average components all predict

	Model	Target response	Average response	Nontarget response
Single-display	T&G	NS ($p = 0.912$)	–	–
	TA	↓ ($p = 0.002$)**	↑ ($p = 0.001$)**	–
	WA	↓ ($p = 0.004$)**	↑ ($p = 0.002$)**	–
	TA&WA	↓ ($p = 0.001$)**	TA: NS ($p = 0.118$) WA: ↑ ($p = 0.024$)*	–
	T&G&N	↓ ($p = 0.021$)*	–	↑ ($p = 0.048$)*
	TA&N	↓ ($p = 0.003$)**	↑ ($p = 0.006$)**	NS ($p = 0.135$)
	WA&N	↓ ($p = 0.001$)**	↑ ($p = 0.003$)**	NS ($p = 0.206$)
Whole-display	T&G	NS ($p = 0.183$)	–	–
	TA	NS ($p = 0.12$)	NS ($p = 0.12$)	–
	WA	NS ($p = 0.073$)	NS ($p = 0.069$)	–
	TA&WA	NS ($p = 0.152$)	TA: NS ($p = 0.543$) WA: NS ($p = 0.073$)	–
	T&G&N	↓ ($p = 0.005$)**	–	NS ($p = 0.168$)
	TA&N	NS ($p = 0.131$)	NS ($p = 0.164$)	NS ($p = 0.106$)
	WA&N	NS ($p = 0.073$)	NS ($p = 0.097$)	NS ($p = 0.105$)

Table 3. Results of trend analysis testing whether the models' estimated probability of each response component linearly increases with set size. *Note*. "↓", linear decrease; "↑", linear increase; "*", $p < 0.05$; "**", $p < 0.01$; NS, not significant.

	T&G	TA	WA	TA&WA	T&G&N	TA&N	WA&N
Single-display	−49.90 (7.70)	−50.63 (7.45)	−51.06 (7.27)	−47.73 (7.26)	−48.93 (7.66)	−48.67 (7.40)	−49.05 (7.32)
Whole-display	−75.81 (8.46)	−78.13 (8.00)	−78.15 (8.04)	−75.41 (7.92)	−75.05 (8.36)	−76.11 (7.99)	−76.37 (8.07)

Table 4. Models' AIC values of working memory for depth. *Note*. Standard errors in parentheses.

that the probability of average response does not vary with set size ($ps > 0.069$), which is also consistent with the raw data. For the three models, including the nontarget response, only the T&G&N model predicts that the probability of nontarget response increases with set size in the single display. Hence, these results showed that the models with an average response component generally performed better in capturing the characteristics of raw data.

Model comparison

The AIC value of each model in each display condition is shown in Table 4. To better demonstrate the effect of average response components on the model fitting, we subtracted AIC values of the TA, WA, and TA&WA models from that of the T&G model (see Figures 4A, 4B). For both displays (especially in the whole-display condition), the two models including a single average component (TA or WA) are better than the basic T&G model. However, the TA&WA model is worse than the T&G model. Similarly, to better demonstrate the effect of the nontarget response component on the model fitting, we subtracted the AIC values of T&G&N, TA&N, and WA&N models from the AIC values of T&G, TA, and WA (the corresponding models without a nontarget component), respectively (see Figures 4C, 4D). In each condition, for each model, adding a nontarget response component brings a larger mean AIC value, which indicates a worse fitting.

To summarize, the average response component is important for simulating the contraction biases in WMd, and the nontarget response component does not benefit, if it does not harm, the model fitting. However, including both the TA and the WA components did not improve model fitting. Therefore, the two models with a single average response component – the TA and the WA models – are selected as two candidates for the best fitting model.

Precision comparison

The above evidence shows that the TA model and the WA model are the two most competitive models, both of which include a target response component and an average response component. To further evaluate the fidelity of different types of memory representation, we compared the estimated precision of the target responses with that of the average responses in each condition. For the WA model, the mean standard deviation of target responses was significantly smaller than that of average responses, indicating that the precision of memorizing the target depth was significantly higher than the precision of responses based on the average depth in the both conditions: single-display, 0.15 vs. 0.22, $t(14) = 2.30$, $p = 0.037$; whole-display, 0.12 vs. 0.23, $t(14) = 4.17$, $p = 0.002$; and for the TA model: single-display, 0.15 vs. 0.23,

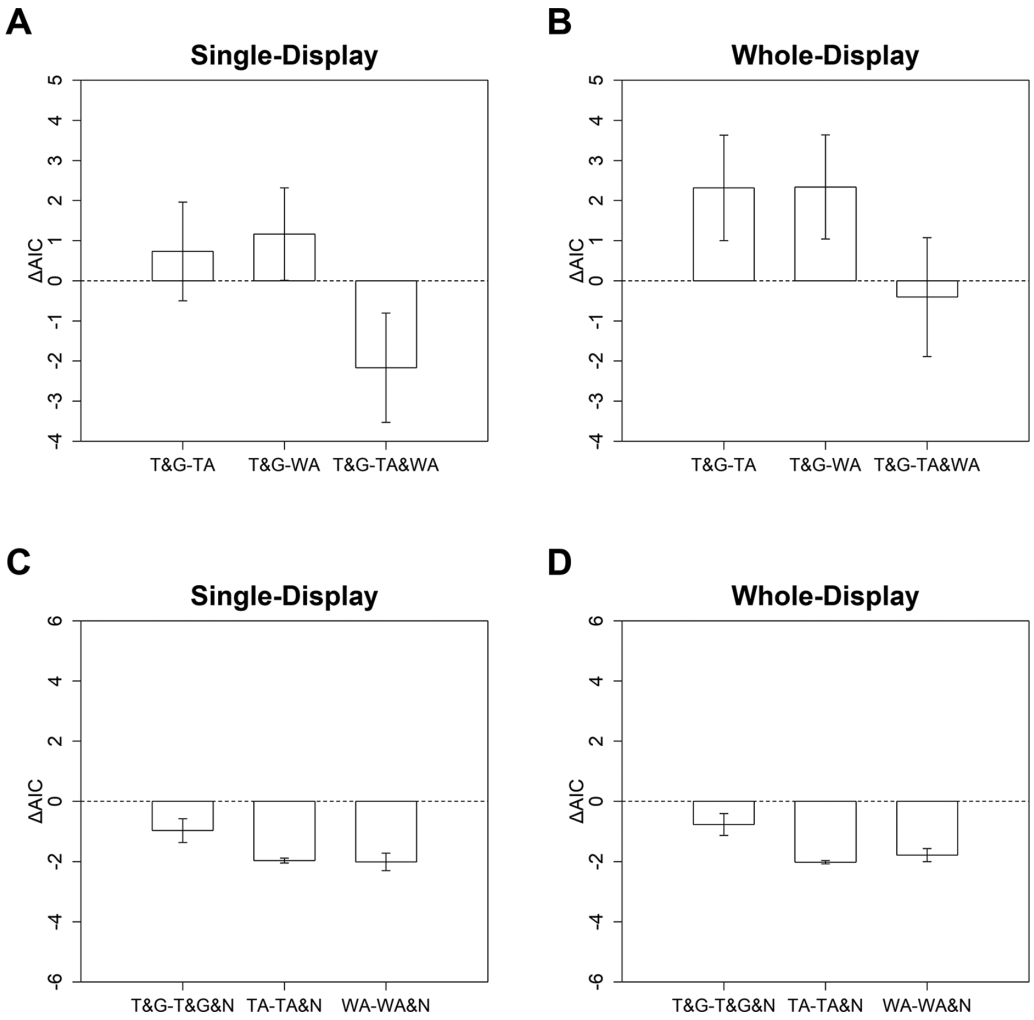

Figure 4. Models' relative AIC values. (**A**) and (**B**) The effect of the average response component on the model fitting for the single and whole displays, respectively. Bars represent the AIC value of the T&G model minus that of the TA model, the WA model, or the TA&WA model. A positive value indicated better fitting than the T&G model. (**C**) and (**D**) The effect of the nontarget response component on the model fitting for the single and whole displays, respectively. Bars represent the AIC values of models without the nontarget response component minus that of these models with such a component. A positive value indicates that the nontarget component improves the model fitting. Error bars represent ± 1 SEM.

$t(14) = 3.68$, $p = 0.003$; whole-display, 0.12 vs. 0.25, $t(14) = 4.81$, $p = 0.001$. These results show that individual memory representation is consistently more precise than the average representation.

Planar visual working memory

Model comparison

Table 5 shows the AIC values of models fitting the data of VWM for planar information. For most of the experiments, the best fitting model included a nontarget response component. Because neither the TA model nor the WA model provides the best fitting in any of the experiments, the TA&WA model was not evaluated for planar VWM.

	T&G	TA	WA	T&G&N	TA&N	WA&N
E1	−293.18	−292.06	−289.70	**−301.34**	−298.30	−296.60
E2	−185.31	−205.44	−182.29	−264.19	**−264.24**	−261.14
E3	−325.63	−336.34	−322.61	**−367.10**	−364.96	−362.02
E4	306.39	306.35	308.22	**299.32**	301.60	301.56
E5	495.34	495.45	497.59	**493.09**	495.78	495.57
E6	**185.69**	188.61		186.32	189.69	
E7	695.03	692.31		**687.39**	689.79	
E8	**781.45**	784.71		782.82	786.33	
E9	1242.97	1243.86		**1241.10**	1243.55	
E10	507.56	439.94	510.88	428.20	**422.40**	431.52
E11	840.01	784.30	842.70	777.28	**767.63**	780.45
E12	263.67	264.36	266.14	**262.19**	264.65	264.89

Table 5. Models' AIC values of visual working memory for planar stimuli. *Note*. Bold numbers are AIC values of the best model in an experiment.

Similar to what has been done for WMd, we subtracted AIC values of the TA and WA models from that of the T&G model to demonstrate the effect of average components on the model fitting (see Figure 5A). For all experiments, the WA model is no better than the T&G model, but the TA model is better than the T&G model in E2, E3, E10, and E11. Among these four experiments, E2 and E3 are two of the three experiments on working memory for planar locations. For E10 and E11, the individual feature values of memory items used in these experiments were equally spaced in a range centered at a predefined average value, which made their average representation easily accessible. In all the other experiments, individual feature values of the to-be-remembered items were randomly selected.

We further subtracted the AIC values of the T&G&N, TA&N, and WA&N models from that of their corresponding models without a nontarget response component (the T&G, TA, and WA models, respectively) to demonstrate the effect of nontarget response component (see Figure 5B). For eight of 12 experiments, adding the nontarget response component improved the model fitting for all the models tested; for two experiments (E5 and E12), adding the nontarget response component improved the fitting for the T&G and WA models but was almost equivalent for the TA model; for the other two experiments (E6 and E8), adding the nontarget response component did not affect the fitting. For the four experiments (E2, E3, E10, and E11) in which the average component improves the fitting, adding the nontarget response component greatly improved the model fitting (see Figure 5B). The best fitting model for E2, E10, and E11 is the TA&N model, and the best fitting model for E3 is the T&G&N model (see Table 5). The T&G&N model also fits best in six out of the eight other experiments. These results indicate that the nontarget response component is crucial and can be commonly found in tasks of VWM for planar information, whereas the average representation response component does not play an important role in these tasks.

Precision comparison

The precision comparison analyses can only be performed for experiments whose best fitting model includes both an individual-based response and an average response. For E2, E10, and E11, the best fitting model is TA&N. However, because the AIC values of the TA&N (–264.24) and T&G&N (–264.19) models in E2 are almost equivalent, we did not perform the precision analysis for E2. For E10 and E11, the mean standard deviation of target responses was significantly larger than that of average responses, indicating that the precision of memorizing the orientation of the target was lower than that of memorizing the average orientation: in E10, 15.23 vs. 12.56, $t(15) = 2.49$, $p = 0.025$; and in E11, 20.24 vs. 12.83, $t(15) = 3.97$, $p = 0.002$. These results seem to suggest that unlike WMd, the average representation for 2-D

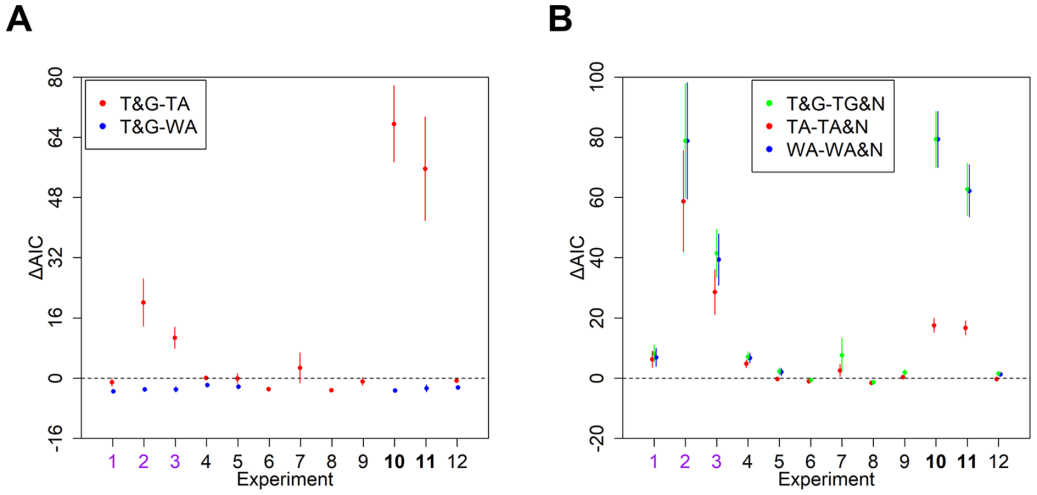

Figure 5. Relative AIC values of models fitting the data from experiments on planar VWM. (**A**) Each point represents the AIC values of the T&G model minus that of the TA (red) model or the WA (blue) model in that experiment. A positive value indicates that the model with an average component provides better fitting. (**B**) Each point represents the AIC values of the T&G model (green), the TA model (red), and the WA model (blue), minus that of the corresponding models with a non-target response component in that experiment. A positive value indicates that the model with a nontarget response component provides better fitting. Error bars represent ± 1 SEM. Along the abscissa axis, the purple numbers indicate the experiments of visual working memory for planar locations; the bold numbers indicate the experiments in which the average value of to-be-remembered features was selected in advance; the others indicate the experiments of visual working memory for planar features.

visual information, if generated, is more precise than individual representation.

Discussion

In the current study, we explore the contribution of average representation in the working memory for depth, and in the working memory for planar locations and features. We used mathematical modeling to characterize participants' responses of recalling the stored information after a brief delay. For recalls of depth information, the best-fitted model provides evidence that the average representation can be as a substitution of individual information in WMd, but swapping the target with nontargets rarely occurs in WMd. In addition, the evidence shows that the present and the past information both contribute to forming the average representation of depth. For recalls of planar features and locations, the best-fitted model indicates that, in contrast to the WMd, the representation of nontargets is commonly used as the substitution of the target, but swapping the target with the average occurs in the planar VWM only when 2-D spatial information is tested or when the average representation is easily perceived and accessed. This indicates that biasing toward the average is strategical, not inherent or automatic, in the planar VWM.

Our findings on planar VWM are consistent with previous studies. For example, (Papenmeier & Timm 2021; a replication of Brady & Alvarez, 2011) found that the memorized sizes of individual stimuli are biased toward the average only for some observers. Evidence also suggests that trial averages have little influence on recall of circular variables, such as colors (Harrison, McMaster, & Bays, 2021; van den Berg, Awh, & Ma, 2014). In addition, past research showed that the average representation is more precise than the individual representation by comparing the performances in separate memory tasks (e.g. Ariely, 2001; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001), which motivated the hypothesis that the averaging process is in parallel with the process of the individuals (Ariely, 2001; Chong & Treisman, 2005; Epstein & Emmanouil, 2021). In the current study, we used mathematical models to differentiate the average representation and individual representation in the same task, and compare the precision of the two types of representation based on the standard deviation of the corresponding distribution estimated by the best model – a larger standard deviation indicates lower precision (Zhang & Luck, 2008). We think that this method is more efficient, and feasible for the tasks not specifically designed for testing average information. For two of the experiments (E10 and E11) whose best fitting model includes the component of average-based responses, the memory precision of average representation is higher than that of the individual, which is consistent with the previous findings on the ensemble perception of planar features.

To date, less is known about how the visual system represents the average of depth information. Using mathematical modeling, we found that the models' performances were greatly improved by the assumption that participants tend to strategically report the average representation to serve as a substitution for individual depth in a WMd task. The probability of reporting the average depth increased with the memory load. Past research showed that the working memory for individual depth information is inaccurate even if only one depth position is to-be-memorized, and it gets even worse when more depths are to-be-memorized (Qian & Zhang, 2019). It is possible that the lower accuracy of memory lead to lower confidence when making recalls of individual depths, which results in reliance on other sources of information, such as the average depth representation. The average may be computed at the time of perception or at the decision stage, and is given a higher weight such that it decays less quickly than the individual depths. Therefore, participants opt for reporting the average instead of the nontargets when the target depth is not stored. In addition, the results of estimated parameters of models (standard deviations of the corresponding average or target/nontarget distributions) suggest that the memory precision of the average depth is lower than that of the individual depth, which is contrary to the findings on planar features.

The findings on WMd and VWM indicate that there may be separate mechanisms underlying the ensemble perception of spatial depth information and that of planar visual features. The higher precision for average representation in VWM suggests that the perception of ensemble information of planar feature can be independent of the perception of individual features (Chong & Treisman, 2003; Chong & Treisman, 2005; Epstein & Emmanouil, 2021). Chong and Treisman (2003) found that the thresholds for discriminating the average sizes of two displays were little affected by memory delay, whereas the thresholds for discriminating the sizes of two individual items increased with the delay. These findings support separate processing for ensemble planar information and individual planar information. However, for WMd, it is possible that the average representation of depth is calculated based on the representation of individual depth, therefore, this additional process of averaging introduces additional errors, which result in a lower precision for the average representation. Because the ensemble statistic may involve higher level of processing, it might be given a higher weight and decay more slowly, and therefore come to dominate WMd as time goes by.

We have shown that unlike in VWM, best fitting models for WMd are based on the assumption that

participants tend to strategically use average depth representation instead of nontarget representation as a substitution for the target, which may reflect that depth information stored in work memory is inherently structural. When the depth information is stored as a spatial structure or configuration in working memory, it can be easier to perceive the ensemble information, such as the average, and harder to individualize each depth location, as they are stored as a whole. Indeed, recall of depth is more accurate and less biased when a spatial configuration is provided (Qian & Zhang, 2019; Zhang et al, 2021), and manipulating the relative depth order, which requires registration of the structural depth information, significantly affects memory performance for detecting changes in depth (Qian et al., 2020). The present study, together with the past findings, provides evidence that the structural and ensemble depth plays an important role in WMd.

Our study showed that the mathematical models with an average component based on one trial and that based on all past trials performed similarly in fitting the raw data from a WMd task, although the model with the average overall past trials might have a slight advantage in model fitting (see Figure 5). To rule out the possibility that the two types of average coincided with each other on most trials for each participants, we checked the absolute difference between the average depth in one trial and the average depth in all past trials and found that the mean absolute difference across participants is about 0.1 degrees, which is much larger than the stereoacuity for normal observers (0.007 degrees; Carrillo, Baldwin, & Hess, 2020) and more than half of the relative disparity applied between the two nearest neighboring depths in the task (0.17 degrees). Therefore, the equally good performance of the two models may reflect that both the current and the past information can be served to create the average depth representation in working memory.

Studies have shown that when explicitly reporting the average of sequentially presented stimuli, the recently presented stimuli were more weighted in calculating the average (Hubert-Wallander & Boynton, 2015; Tong, Dubé, & Sekuler, 2019; also see Ebbinghaus, 1885/1974). However, the averaging of planar location shows a reverse trend that the earlier stimuli are weighted more (Hubert-Wallander & Boynton, 2015). To further examine the contribution of averaging over a single trial and all past trials, we fitted the data of WMd using a recency-prioritized weighted average model (RA model; see the Supplementary Information for details), in which the recent stimuli weighted more for computing the average (following an exponential function; Tong, Dubé, & Sekuler, 2019). We found that the RA model performed better than the TA model for the single display, but do not provide a better fitting than the WA model (the two models perform similarly), suggesting that the information from all previous trials tend to be weighted equally for computing the average depth. It is possible that the internal representation of the average depth is updated on each trial, so that information in all trials contributes equally to the formation of average depth.

The hierarchical encoding model postulates that items are stored in working memory at several levels from individual items to the whole ensemble of them (Brady & Alvarez, 2011; Utochkin & Brady, 2020). We think that the whole memory history contributes to form the ensemble (including the average), which may be at a relatively high level of representation in the framework of hierarchical encoding model. Numerous studies have found that stimulus history can influence current judgments in cognitive tasks (Akrami et al., 2018; Chetverikov, Campana, & Kristjánsson, 2016; Crawford, Corbin, & Landy, 2019). Consistently, we found that not only the currently presented depth information but also all past depth information can be useful in creating the average depth representation, indicating that the information from the memory history of stimuli is incorporated in the ensemble. As information stored in the long-term memory may benefit the working memory performance (Brady, Störmer, & Alvarez, 2016, Xie & Zhang, 2017), we think that the ensemble based on memory history of stimuli may be stored in a higher level of the hierarchy model, which also can affect individual representation.

Notably, our conclusions for WMd rely on the data from a single study and therefore need further confirmation. In addition, the current work on WMd focuses on the contribution of average representation in the cases where the target may not be stored, but we cannot rule out the possibility that the responses may be systematically biased toward the center of space, leading to the contraction bias. This center bias can be found in studies using noncircular spaces (e.g. Sims, Jacobs, & Knill, 2012; Wilken & Ma, 2004) and can be reproduced by Bayesian models that incorporate a prior reflecting the stimulus space (e.g. Salmela, Ölander, Muukkonen, & Bays, 2019). The present study did not distinguish whether the contraction bias comes from reporting the average of the tested feature values or a systematic bias to the center of the space, and future investigation is needed to clarify these two accounts.

Keywords: working memory, computational modeling, ensemble representation, depth

Acknowledgments

Supported by the Guangdong Basic and Applied Basic Research Foundation (2021A1515010840), and the Fundamental Research Funds for the Central Universities (20wkzd12). The authors have no

competing financial interests that might be perceived to influence the results and/or discussion reported in this paper.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon request.

Commercial relationships: none.
Corresponding author: Jiehui Qian.
Email: qianjh3@mail.sysu.edu.cn.
Address: Department of Psychology, Sun Yat-Sen University, Guangzhou 510006, China.

References

Aagten-Murphy, D., & Bays, P. M. (2019). Independent working memory resources for egocentric and allocentric spatial information. *PLoS Computational Biology, 15*(2), e1006563.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Akrami, A., Kopec, C. D., Brody, C. D., & Diamond, M. E. (2018). Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature, 554*(7692), 368–372.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157–162.

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience, 34*(10), 3632–3645.

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9*(10), 1–11.

Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia, 49*(6), 1622–1631.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*, 289–300.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items. *Psychological Science, 22*(3), 384–392.

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity : More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences of the United States of America, 113*(27), 7459–7464.

Carrillo, S. A., Baldwin, A. S., & Hess, R. F. (2020). Factors limiting sensitivity to binocular disparity in human vision: Evidence from a noise-masking approach. *Journal of Vision, 20*(3), 1–14.

Chetverikov, A., Campana, G., & Kristjánsson, Á. (2016). Building ensemble representations: How the shape of preceding distractor distributions affects visual search. *Cognition, 153*, 196–210.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*(4), 393–404.

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45*(7), 891–900.

Crawford, L. E., Corbin, J. C., & Landy, D. (2019). Prior experience informs ensemble encoding. *Psychonomic Bulletin and Review, 26*(3), 993–1000.

Dubé, C., & Sekuler, R. (2015). Obligatory and adaptive averaging in visual short-term memory. *Journal of Vision, 15*(4), 13.

Ebbinghaus, H. (1885/1974). *Memory: a contribution to experimental psychology*. New York, NY: Dover.

Epstein, ML., & Emmanouil, TA. (2021). Ensemble Statistics Can Be Available before Individual Item Properties: Electroencephalography Evidence Using the Oddball Paradigm. *Journal of Cognitive Neuroscience, 33*(6), 1056–1068.

Fisher, N. I. (1995). *Statistical analysis of circular data*. Cambridge, London: Cambridge University Press.

Gibson, E. J., & Walk, R. D. (1960). The "Visual Cliff." *Scientific American, 202*(4), 64–71.

Gorgoraptis, N., Catalao, R. F., Bays, P. M., & Husain, M. (2011). Dynamic updating of working memory resources for visual objects. *Journal of Neuroscience, 31*(23), 8502–8511.

Hardman, K. O., Vergauwe, E., & Ricker, T. J. (2017). Categorical working memory representations are used in delayed estimation of continuous colors. *Journal of Experimental Psychology: Human Perception and Performance, 43*(1), 30–54.

Harrison, W. J., McMaster, J. M., & Bays, P. M. (2021). Limited memory for ensemble statistics in visual change detection. *Cognition, 214*, 104763.

Huang, L. (2020). Distinguishing target biases and strategic guesses in visual working memory. *Attention, Perception, and Psychophysics, 82*(3), 1258–1270.

Hubert-Wallander, B., & Boynton, G. M. (2015). Not all summary statistics are made equal: Evidence from extracting summaries across time. *Journal of Vision, 15*(4), 5.

Lew, T. F., & Vul, E. (2015). Ensemble clustering in visual working memory biases location memories and reduces the Weber noise of relative positions. *Journal of Vision, 15*(4), 10.

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

Nelder, J. A., & Mead, R. (1965). A Simplex Method for Function Minimization. *Computer Journal, 7*(4), 308–313.

Papenmeier, F., & Timm, J. D. (2021). Do group ensemble statistics bias visual working memory for individual items? A registered replication of Brady and Alvarez (2011). *Attention, Perception, & Psychophysics, 83*(3), 1329–1336.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739.

Pratte, M. S. (2019). Swap errors in spatial working memory are guesses. *Psychonomic Bulletin & Review, 26*(3), 958–966.

Pratte, M. S., Park, Y. E., Rademaker, R. L., & Tong, F. (2017). Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 43*(1), 6.

Qian, J., Li, Z., Zhang, K., & Lei, Q. (2020). Relation matters: relative depth order is stored in working memory for depth. *Psychonomic Bulletin and Review, 27*(2), 341–349.

Qian, J., & Zhang, K. (2019). Working memory for stereoscopic depth is limited and imprecise—Evidence from a change detection task. *Psychonomic Bulletin & Review, 26*(5), 1657–1665.

Reeves, A., & Lei, Q. (2017). Short-term visual memory for location in depth: A U-shaped function of time. *Attention, Perception, and Psychophysics, 79*(7), 1917–1932.

Salmela, V. R., Ölander, K., Muukkonen, I., & Bays, P. M. (2019). Recall of facial expressions and simple orientations reveals competition for resources at multiple levels of the visual hierarchy. *Journal of Vision, 19*(3), 8.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review, 119*(4), 807.

Schneegans, S., & Bays, P. M. (2016). No fixed item limit in visuospatial working memory. *Cortex, 83*, 181–193.

Schneegans, S., Taylor, R., & Bays, P. M. (2020). Stochastic sampling provides a unifying account of visual working memory limits. *Proceedings of the National Academy of Sciences of the United States, 117*(34), 20959.

Tanaka, K., Yamamoto, K., Watanabe, K., & Sung-En, C. (2016). Memory distortion of depth of a visual stimulus for perception and action. *2016 8th International Conference on Knowledge and Smart Technology, KST 2016*, 281–286.

Tong, K., Dubé, C., & Sekuler, R. (2019). What makes a prototype a prototype? Averaging visual features in a sequence. *Attention, Perception, & Psychophysics, 81*(6), 1962–1978.

Utochkin, I. S., & Brady, T. F. (2020). Individual representations in visual working memory inherit ensemble properties. *Journal of Experimental Psychology: Human Perception and Performance, 46*(5), 458–473.

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121*(1), 124.

van den Berg, R., Shin, H., Chou, W. C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences, 109*(22), 8780–8785.

Wang, K., Jiang, Z., Huang, S., & Qian, J. (2021). Increasing perceptual separateness affects working memory for depth-re-allocation of attention from boundaries to the fixated center. *Journal of Vision, 21*(7), 8.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*(12), 1120–1135.

Xie, W., & Zhang, W. (2017). Familiarity increases the number of remembered Pokémon in visual short-term memory. *Memory & Cognition, 45*(4), 677–689.

Zhang, K., Gao, D., & Qian, J. (2021). Overestimation and contraction biases of depth information stored in working memory depend on spatial configuration. *British Journal of Psychology (London, England: 1953), 112*(1), 230–246.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233.