Attention flexibly trades off across points in time
Denison, R.N., Heeger, D.J., Carrasco, M.

SUPPLEMENTARY INFORMATION

Here we report statistical procedures, reaction time (RT) results for Experiments 1 and 2, all ANOVA results and pairwise tests not reported in the main manuscript, and model comparison results for Experiment 3 (estimation task).

General methods

Statistics

Repeated measures ANOVAs were used to test for main effects and interactions across experimental conditions, and two-tailed paired t-tests were used for pairwise comparisons. In Experiment 1, we tested the effects of probed target interval (T1/T2), target contrast, and cue validity in a 3-way ANOVA. Because target contrast had no effect, all pairwise tests were for data averaged across contrasts.

For RT in Experiment 2 and the precision measure in Experiment 3, Mauchly's sphericity test was violated for parametric ANOVA and pairwise differences were not normally distributed for some comparisons (assessed using Q-Q plots, Shapiro-Wilk normality tests), so we used non-parametric statistics. An Aligned Rank Transform (ART) (Wobbrock, Findlater, Gergle, & Higgins, 2011) was applied to the data prior to the ANOVA, a non-parametric method suitable for repeated measures factorial designs. Pairwise comparisons were tested using a two-sided paired Wilcoxon signed-rank test.

For each target, we performed planned pairwise comparisons between valid and neutral and neutral and invalid conditions to assess benefits and costs separately from the valid vs. invalid cueing effect (Keppel & Wickens, 2004). The conclusions remained the same after correcting for these three comparisons per target using Holm's method (Holm, 1979).

All pairwise differences between validity conditions were confirmed using permutation tests. Group average differences between two conditions were computed after shuffling the validity condition labels across trials for each observer. This procedure was repeated 1,000 times to generate a null distribution corresponding to the expected group difference if cue validity had no effect. The actual group difference was compared to both sides of the null distribution to determine the 2-tailed p-value. All significant pairwise differences ($p<0.05$) reported in this manuscript were also found to be significant with the permutation tests. Some additional comparisons were significant in the permutation tests but not in the reported tests. The permutation tests may have been more liberal because they tested for fixed as opposed to random effects.

We calculated benefit and cost indices to allow comparisons across experiments with different performance measures. The benefit/cost indices were calculated from the group condition means (see Methods in main text), because the ratios were not stable when computed separately for each individual observer. To generate error bars on these values, we resampled trials (with replacement) from each observer and condition and recomputed the group means and benefit/cost indices. This procedure was repeated 1,000 times to generate a bootstrapped distribution for each index. The 16th and 84th percentiles of that distribution gave 68% confidence intervals.

Experiment 1

Results

We tested the effects of probed target interval (T1/T2), target contrast, and cue validity in a 3-way ANOVA. Performance accuracy did not significantly depend on whether the probed target was in the first or second interval (F(1,9)=2.76, p=0.13, η_G^2 = 0.10) and was only marginally better at the higher contrast (F(1,9)=4.95, p=0.053, η_G^2 = 0.021). Performance improvements with higher cue validity did not depend on the interval or contrast of the probed target (no interactions: F(2,18)<1.7, p>0.1).

We found a similar pattern of cueing effects for RT as for accuracy, with fastest RT on valid trials (1127 ms), slowest on invalid trials (1377 ms), and intermediate RT on neutral trials (1271 ms) (**Figure 1e,f**), reflecting an 18% decrease in RT from invalid to valid trials. A 3-way ANOVA confirmed faster RTs with higher cue validity (main effect: F(2,18)=12.82, p=0.0003, η_G^2 = 0.015). There were significant or marginally significant differences for all comparisons among valid, neutral, and invalid trials (valid vs. invalid: t(9)=4.19, p=0.002, d=1.32; valid vs. neutral: t(9)=3.63, p=0.006, d=1.15; neutral vs. invalid: t(9)=2.24, p=0.052, d=0.71). Collapsing across targets (**Figure 1f**), valid performance was higher than invalid (valid vs. invalid: t(9)=4.19, p=0.0023, d=1.32), and there were significant benefits (valid vs. neutral: t(9)=3.63, p=0.0055, d=1.15) and marginally significant costs (neutral vs. invalid: t(9)=2.24, p=0.052, d=0.71). There was no evidence of a dependence on contrast (F(1,9)<1) or interaction between contrast and the other factors (all F<1.3, all p>0.3). RTs were faster for T2 than T1 (F(1,9)=7.20, p=0.025, η_G^2 = 0.005), but performance improvements with higher cue validity did not depend on the interval of the probed target (no interaction: F(2,18)<1). These results not only confirm that the observed accuracy effects were not driven by speed-accuracy tradeoffs, but also show that the differences in accuracy were paralleled by differences in response speed.

Experiment 2

Results

We tested the effects of probed target interval (T1/T2/T3) and cue validity in a 2-way ANOVA. Performance accuracy differed for the three targets (F(2,18)=9.75, p=0.0014, η_G^2 = 0.30), with highest accuracy for T3 (83%), intermediate for T1 (75%), and lowest for T2 (70%). Collapsing across targets (**Figure 1c**), valid performance was higher than invalid (valid vs. invalid: t(9)=3.79, p=0.0043, d=1.20), and there were significant costs (neutral vs. invalid: t(9)=2.84, p=0.020, d=0.90) but not benefits (valid vs. neutral: t(9)=1.67, p=0.13, d=0.53).

RT was not expected to be strongly affected by cue validity because of the forced 1500 ms response delay. Nevertheless, RT was fastest on valid trials (432 ms), slowest on invalid trials (504 ms), and intermediate on neutral trials (466 ms) (**Figure 2d**), reflecting a 14% decrease in RT from invalid to valid trials. A 2-way ANOVA confirmed faster RTs with higher cue validity (main effect: F(2,18)=6.58, p=0.0072, η_P^2 = 0.42). RT was similar for all probed targets (no main effect: F(2,18)=1.35, p=0.28, η_P^2 = 0.13), and performance improvements with higher cue validity did not depend on the interval of the probed target (no interaction: F(4,36)=0.77, p=0.55, η_P^2 = 0.079). Collapsing across targets (**Figure 2e**), valid performance was higher than invalid (valid vs. invalid: t(9)=2.50, p=0.034, d=0.79), and there were significant costs (neutral vs. invalid: t(9)=2.50, p=0.034, d=0.79) but not benefits (valid vs. neutral: t(9)=1.76, p=0.11, d=0.56).

Experiment 3

Methods

The estimation task yielded a distribution of errors (differences between actual target orientation and reported target orientation) for each observer in each condition (valid/neutral/invalid x T1/T2). We first determined what model of the errors was most appropriate for the data using a stepwise series of model comparisons (**Table 1**), testing four parameters used to describe estimation data (Bays, Catalao, & Husain, 2009; Suchow, Brady, Fougnie, & Alvarez, 2013; Zhang & Luck, 2008). The comparisons were performed for each observer using data from all conditions combined (640 trials). Model comparison metrics Akaike information criterion (AIC) and Bayesian information criterion (BIC) were computed for each pair of models (van den Berg, Awh, & Ma, 2014). A model was selected if the AIC or BIC values were significantly in its favor, as assessed by a Wilcoxon signed-rank test across observers (a non-parametric test was used because the AIC and BIC difference values were not normally distributed). The selected model was then used for the next comparison. If model performance was indistinguishable for the two models (no statistical difference across observers), then the model with fewer parameters was selected.

Four different models for the probability (*p*) of a given estimation error were compared using this procedure (**Table 1**). Model 1 (Equation 1) was a von Mises distribution (see Equation 5) centered on the probed target orientation (mean of 0 error). The standard deviation of the von Mises (σ) reflected the precision of the target representation.

$$p = VM(0,\sigma) \tag{1}$$

Model 2 (Equation 2) was a standard mixture model (Zhang & Luck, 2008) consisting of a von Mises distribution centered on the postcued target orientation and a uniform distribution. The height of the uniform distribution (*g*) gave the guess rate, or the probability of not representing the target (to a level above noise), together with lapses.

$$p = (1-g)VM(0,\sigma) + g/180 \tag{2}$$

Model 3 (Equation 3) was the standard mixture model, but with an additional parameter for the mean of the distribution (μ), which modeled an observer's overall bias toward CW/CCW responses.

$$p = (1-g)VM(\mu,\sigma) + g/180 \tag{3}$$

Model 4 (Equation 4) was the standard mixture model plus a swapping distribution (Bays et al., 2009). The swapping distribution was a second von Mises distribution centered on the non-probed target orientation (mean error θ = non-target orientation – target orientation) with standard deviation equal to that of the target-centered distribution. The swap parameter (β) specified the probability of a swap.

$$p = (1-g-\beta)VM(0,\sigma) + \beta VM(\theta,\sigma) + g/180 \tag{4}$$

The von Mises probability density function is defined as

$$VM(\mu,\sigma) = p(x;\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(x-\mu)}, \tag{5}$$

where *x* is orientation, κ is the von Mises concentration parameter corresponding to the standard deviation σ of a circular normal distribution, and I_0 is the modified Bessel function of the first kind of order 0.

Models were fit to the error distributions for each observer in each condition using an MCMC sampling procedure to find the maximum a posteriori (MAP) parameter estimates (**Figure 3b**). These parameter estimates were then subjected to statistical tests to compare experimental conditions. All model fitting and comparison was performed using MemToolbox (Suchow et al., 2013).

Results

We first determined which of our hypothesized sources of error were consistent with the orientation estimation error distributions by performing a series of model comparisons (**Table 1**). The best model was a mixture model (von Mises + uniform), parameterized by the von Mises standard deviation (σ) and the uniform guess rate (g). This model outperformed a model that also included a swapping parameter (AIC: Z=2.04, p=0.042, r=0.58; BIC: Z=3.06, p=0.0005, r=0.88).

	Model A	Model B	AIC difference	BIC difference
1	Precision only (σ)	**Mixture (σ,g)**	55.1 (44.4)***	48.8 (44.4)***
2	**Mixture (σ,g)**	Mixture + bias (σ,g,μ)	7.1 (16.5)	1 (16.5)
3	**Mixture (σ,g)**	Mixture + swap (σ,g,β)	-1.1 (1.4)*	-7.4 (1.4)***

Table 1. Stepwise model comparisons combining all trials from each observer to determine the most appropriate model parameters for fitting each condition separately. AIC and BIC columns show mean difference scores (Model A–Model B) across observers, with SD in parentheses. Negative difference scores favor Model A. The model selected from each comparison is in bold (see Methods). * p<0.05, *** p<0.001.

We fit the mixture model to the data from each experimental condition. Pooled across T1 and T2, we found the highest precision (lowest standard deviation) for valid (12.0°), lowest precision for invalid (14.2°), and intermediate precision for neutral trials (13.0°). So overall precision increased 18% from invalid to valid trials. Precision for T2 was higher than for T1 (main effect of target interval: F(1,11)=57.75, p<0.0001, η_P^2=0.84), and cueing improved precision more for T1 than for T2 (interaction between cue validity and target interval: F(2,22)=3.62, p=0.044, η_P^2=0.25). T1 precision was indistinguishable for neutral and invalid trials (Z=0.24, p=0.85, r=0.07). For T2 individually, cue validity did not alter precision reliably (all Z<1.6, p>0.1), but we note that it was higher for valid than invalid. Collapsing across targets (**Figure 3d**), valid precision was higher than invalid (valid vs. invalid: Z=2.28, p=0.023), and there were significant benefits (valid vs. neutral: Z=2.82, p=0.0047) but not costs (neutral vs. invalid: Z=1.10, p = 0.27). There was no evidence for a difference between T1 and T2 guess rates (F(1,11)=2.39, p=0.15) or in how cue validity affected guess rate for the two targets (no interaction: F(2,22)=1.03, p=0.37).

REFERENCES

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), 7.1-11. doi: 10.1167/9.10.7

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*.

Keppel, G., & Wickens, T. D. (2004). *Design and Analysis: A Researcher's Handbook* (4 ed.). New Jersey: Pearson Prentice Hall.

Suchow, J. W., Brady, T. F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision, 13*(10), 1-8. doi: 10.1167/13.10.9.doi

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121*(1), 124-149. doi: 10.1037/a0035234

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011, May 7-12, 2011). *The Aligned Rank transform for nonparametric factorial analyses using only ANOVA procedures.* Paper presented at the ACM Conference on Human Factors in Computing Systems (CHI '11), Vancouver, British Columbia.

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