Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory

Rosanne L. Rademaker Psychology Department, Vanderbilt University, Nashville, TN, USA, and Department of Cognitive Neuroscience, Maastricht University, The Netherlands

Caroline H. Tredway Psychology Department, Vanderbilt University, Nashville, TN, USA

Frank Tong Psychology Department and Vanderbilt Vision Research Center, Vanderbilt University, Nashville, TN, USA

Working memory serves as an essential workspace for the mind, allowing for the active maintenance of information to support short-term cognitive goals. Although people can readily report the contents of working memory, it is unknown whether they might have reliable metacognitive knowledge regarding the accuracy of their own memories. We investigated this question to better understand the core properties of the visual working memory system. Observers were briefly presented with displays of three or six oriented gratings, after which they were cued to report the orientation of a specific grating from memory as well as their subjective confidence in their memory. We used a mixed-model approach to obtain separate estimates of the probability of successful memory maintenance and the precision of memory for successfully remembered items. Confidence ratings strongly predicted the likelihood that the cued grating was successfully maintained, and furthermore revealed trial-to-trial variations in the visual precision of memory itself. Our findings provide novel evidence indicating that the precision of visual working memory is variable in nature. These results inform an ongoing debate regarding whether this working memory system relies on discrete slots with fixed visual resolution or on representations with variable precision, as might arise from variability in the amount of resources assigned to individual items on each trial.

Keywords: metacognition, orientation discrimination, short-term memory, conscious perception

Citation: Rademaker, R. L., Tredway, C. H., & Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *Journal of Vision*, *12*(13):21, 1–13, http://www.journalofvision.org/content/12/13/21, doi:10.1167/12.13.21.

Introduction

Visual perception is so highly efficient that people commonly report having the subjective impression of being able to take in an entire visual scene at a glance, with relatively little effort (Cohen, Alvarez, & Nakayama, 2011; Li, VanRullen, Koch, & Perona, 2002; Rousselet, Fabre-Thorpe, & Thorpe, 2002). However, studies of working memory indicate that observers can only maintain information about a limited number of objects after viewing a complex scene (Luck & Vogel, 1997; Philips, 1974). To understand the profound limitations of visual working memory, researchers have adopted psychophysical methods to better characterize the precision and capacity of this system (Magnussen & Greenlee, 1999; Regan & Beverley, 1985; Wilken & Ma, 2004). In recent years, such research has led to a bourgeoning debate regarding the fundamental properties and limits of the visual working memory system (Anderson, Vogel, & Awh, 2011; Bays, Catalao, & Husain, 2009; Bays & Husain, 2008; van den Berg, Shin, Chou, George, & Ma, 2012; Zhang & Luck, 2008).

The slot model theory proposes that visual working memory consists of three to four discrete slots, each of which can store information about a single perceptually bounded object in an all-or-none manner (Luck & Vogel, 1997). Critically, each slot is believed to support a fixed level of visual resolution. According to the slots-plus-averaging model, multiple independent slots can be used to maintain information about a single item to thereby improve the visual precision of memory for that item (Zhang & Luck, 2008). However, if the number of items exceeds the capacity limit of working memory, then each available slot will be used to store one unique item with a fixed degree of visual precision, and any remaining items will fail to be encoded into

doi: 10.1167/12.13.21 Received February 6, 2012; published December 21, 2012 ISSN 1534-7362

working memory (Anderson et al., 2011). The discrete nature of the slot model can be contrasted with resource models, which propose that visual working memory is supported by a continuous resource that can be flexibly subdivided among many items. Resource models assume that the working memory system lacks a prespecified item limit; instead, this system should be able to retain more than three to four items by flexibly trading off visual resolution for increased capacity (Bays et al., 2009; Bays & Husain, 2008; Wilken & Ma, 2004). A more recent version of the resource model, called the variable-precision model, proposes that the amount of resource assigned to each item in a display can randomly vary across items and trials (van den Berg et al., 2012). Unlike the slots-plus-averaging model, the variable-precision model assumes that the precision of visual working memory is fundamentally variable in nature and can fluctuate considerably from item to item, and from trial to trial.

Distinguishing between these models of working memory presents some empirical challenges. How might one determine whether memory precision fluctuates across trials, given that memory precision itself must be estimated by quantifying the variability of working memory performance across trials? In the present study, we addressed this issue by asking whether participants can make reliable metacognitive judgments regarding the accuracy of their own memories. If so, then it should be possible to evaluate whether ratings of subjective confidence are predictive of the precision of visual working memory.

Metacognition refers to the knowledge that one has about one's own cognitive experiences, processes, and strategies (Flavell, 1979). It provides the basis for the introspective ability to evaluate cognitive performance in the absence of direct feedback. Studies have found that people have reasonable metacognitive knowledge about the accuracy of their perceptual judgments (Kunimoto, Miller, & Pashler, 2001; Nickerson & McGoldrick, 1965; Song et al., 2011) and also their judgments regarding long-term memory (Busey, Tunnicliff, Loftus, & Loftus, 2000; Wixted, 2007; Yonelinas, 1994). Recent studies suggest that participants can also reliably evaluate the vividness of individual episodes of mental imagery (Pearson, Rademaker, & Tong, 2011; Rademaker & Pearson, 2012). However, few, if any, studies have investigated the relationship between metacognitive judgments and working memory in humans. Why might this be the case? Because the contents of working memory are consciously accessible and presumed to be maintained in an all-or-none manner (Baddeley, 2003), it would appear to be trivially easy to evaluate one's own working memory performance based simply on whether information about the target item could be reported from memory. Such a strategy would appear sufficient for evaluating working memory for distinct categorical items, such as digits, letters or words, but might prove inadequate for stimuli that vary along a continuum. For example, consider the task of maintaining a specific visual orientation in memory and later attempting to report that exact orientation by adjusting the angle of a probe stimulus. Would the participant have any metacognitive knowledge about how precisely he or she performed this task, or any insight into the precision with which that item was encoded and maintained in working memory?

We addressed this question by using the psychophysical 'method of adjustment' to obtain a continuous measure of the accuracy of working memory performance on individual trials, as has been adopted in many recent studies (Wilken & Ma, 2004; Zhang & Luck, 2008). Observers were briefly presented with displays containing three or six randomly oriented gratings at the beginning of each trial (Figure 1). After a 3-second delay period, a spatial cue appeared indicating the grating to be reported from memory. Observers first rated their confidence in their memory for the probed item on a scale from 0–5, and then adjusted the orientation of a test probe at the center of the display to indicate the remembered orientation. Monetary incentives were provided to encourage participants to respond as accurately as possible. For our analysis, we adopted the mixed-model approach to obtain separate estimates of likelihood of successful maintenance of the probed item and the precision of working memory for successfully retained items (Zhang & Luck, 2008). This allowed us to evaluate whether subjective ratings of confidence were strongly predictive of these two estimates of working memory performance across individual trials. Our results revealed that higher confidence ratings were predictive of both greater likelihood of successful memory maintenance and superior precision of memory for the probed orientation.

Methods

Procedure

Six observers (three female) between the ages of 20 and 32 participated in the main experiment. Two of these observers (CN and SJ) also participated in the additional experiment ($N = 6$, five female, ages 20–28). All observers had normal or corrected-to-normal visual acuity and received payment for participation ($10–15 per hour), with the exception of two participating authors (RR and CT). All participants provided informed written consent, and the study took place

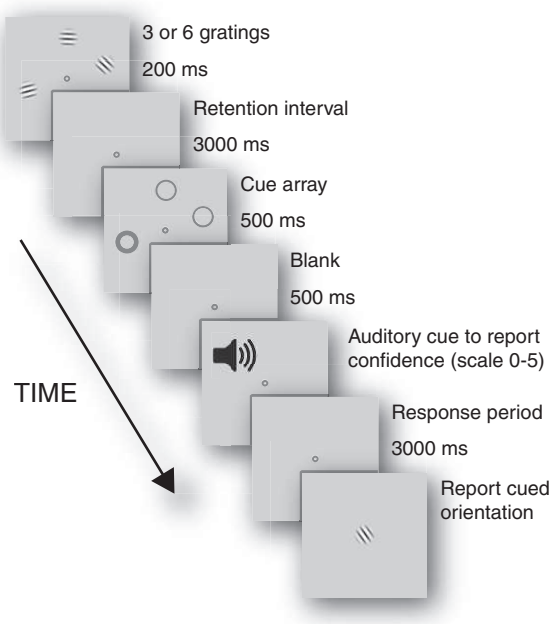

Figure 1. Experimental design. Sequence of events during a working memory trial. A display of three or six randomly oriented gratings was briefly presented, followed by a cue array that appeared 3 s later, indicating which grating to report from memory. Participants first rated their confidence in their memory for the cued item on a scale from 0 (no memory at all) to 5 (perfect memory), and then reported the orientation of the cued grating by rotating a central test grating.

under the approval of the Institutional Review Board of Vanderbilt University.

Participants viewed the stimuli in a dark room on a luminance-calibrated CRT monitor (21″ SONY Black Professional Series FLAT Trinitron) with 1152 × 870 resolution and a 75-Hz refresh rate. Stimuli were created using a Macintosh computer (Apple, CA) with MATLAB 7.5.0 (R2007b) and Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). Stimuli consisted of randomly oriented gratings (size 2°, spatial frequency 2 c/deg) presented around a central fixation point at an eccentricity of 4°. The gratings were presented at 50% contrast within a Gaussian-contrast envelope (see Figure 1) on a uniform gray background that shared the same mean luminance of 35.8 cd/m^2. Participants sat at a viewing distance of 57 cm, and used a chinrest to maintain head stability.

In the main experiment, observers were presented with randomly mixed trials of either three or six gratings to retain in working memory (Figure 1). For each grating in a visual display, the orientation was randomly and independently determined (0–180°). The location assignment for each grating allowed for a grating to be positioned anywhere at 4° eccentricity from the central fixation point, with the only constraint that each grating had to be separated from its neighbor by at least 1°. After a 3 s retention interval, spatial cues appeared for 500 ms indicating the locations of sample stimuli. A white circle of 0.10°-thickness outlined the location of the target grating to report from memory, whereas nontarget locations were indicated by thinner circles that were drawn using 0.04° outlines. After a 500 ms blank period, an auditory tone indicated that the participant should report the quality of their memory for the cued grating on a scale from 0 (no memory at all) to 5 (best possible memory) within a 3 s response window. Participants were instructed to use the full range of the rating scale to the best of their abilities. Finally, a test grating appeared at the center of the screen at an initially random orientation. Participants reported their memory for the orientation of the cued grating by rotating the test grating, using separate buttons on a keyboard for clockwise and counterclockwise rotation.

To encourage accurate performance, participants were provided with monetary incentives to respond as accurately as possible. They were told that they could earn additional payment for very accurate performance on each trial, receiving two bonus points for responses that were accurate within ±0–5° of the true orientation, and one bonus point for responses that fell between 5–10° of the true orientation. Each bonus point was worth 2 cents of additional payment. Cumulative feedback was provided regarding the number of bonus points that had been earned after every block of 20 trials. At the end of each experimental session (160 trials), the total amount of bonus points earned was displayed again, along with the associated amount in US dollars. Participants could obtain up to 50% more payment than their base payment of $10/hr, depending on the accuracy of their performance.

Each experimental session lasted between 30–45 min and consisted of 160 trials. All participants completed 10 sessions in the main experiment, resulting in a total of 1,600 trials, or 800 trials for set size 3 and an equal number of trials for set size 6.

Analysis

To separately estimate the precision of memory for successfully remembered items and the likelihood of memory failure, we adopted a mixed-model approach following the work of Zhang and Luck (Zhang & Luck, 2008, 2009). A circular Gaussian-shaped model was used to fit the distribution of orientation errors (reported orientation minus actual orientation) for each condition of interest. The model consisted of three key parameters: the mean or "center" of the Gaussian distribution, the standard deviation (*SD*) or width of the Gaussian distribution, and the extent to which the entire distribution needed to be translated along the *y*-axis to account for the frequency of

uniform responses. The mean was constrained to lie centered around a value of zero. Estimates of the other two parameters were derived using standard function fitting procedures in MATLAB ('fminsearch') to implement the simplex search method.

The mixed model assumes that the relative proportion of area under the curve corresponding to the uniform distribution reflects the probability of memory failure, whereas the standard deviation of the error distribution reflects the precision of working memory for successfully remembered items. We rely on these summary statistics throughout this paper because they provide a useful way to summarize broad trends in the data and because they may also signify distinct types of errors. However, it is important to acknowledge that the mapping between these summary statistics and underlying sources of error in the working memory system rely on an assumed model of working memory performance, and that competing models have been proposed (e.g., van den Berg et al., 2012).

Model fitting of the group-averaged data was highly robust for confidence ratings of 1 through 5. In the main experiment, R^2 values for these fits ranged from 0.73–0.99 (mean R^2 of 0.916). Estimates of memory precision for confidence level 0 were excluded from analysis, because the frequency distribution of these errors was almost flat and led to much lower R^2 values.

To obtain reliable fits of each participant's data, it was necessary to sort the confidence ratings of 1–5 into three broader cohorts of low, medium, and high confidence (Tables S1 and S2). Frequency distributions of orientation errors were calculated for each confidence level and set size using a bin width of 10°, prior to fitting the Gaussian model. More observations were required to obtain good *R*-squared measures of fit for the low confidence data, because participants exhibited reliable memory for the cued grating on a much smaller percentage of these trials. Overall, the pattern of results was robust if a sufficient number of observations were available at each confidence level. When cohorts were reassigned simply to ensure that a minimum number of 50 observations was available for analysis at each confidence level, we again observed a significant improvement in memory precision with higher ratings of confidence, $F(2, 10) = 12.1$, $p < 0.005$, and a significant decrease in the likelihood of memory failure with higher confidence, $F(2, 10) = 50.7$, $p < 0.0001$.

The measure of memory capacity K was calculated as follows:

$$K_I = (1 - P_{uniform}) \times N \quad (1)$$

where K_I is the capacity for condition I, $P_{uniform}$ is the area under the uniform distribution believed to reflect the proportion of random guessing, and N is the set size.

Results

We first evaluated whether subjective confidence ratings were predictive of objective measures of the accuracy of working memory for the cued orientation. Plots of frequency distributions indicated that participants' confidence ratings varied considerably across trials, even within each set size condition (Figure 2A). Confidence was generally higher for set size 3 than for set size 6, and in the latter condition there were frequent reports of zero confidence suggestive of a bimodal distribution. We calculated the angular difference between the reported orientation and the true orientation for all 1,600 trials performed by each participant, and plotted the *absolute* response error for each confidence rating level and set size (Figure 2B). Confidence ratings were highly predictive of this measure of working memory accuracy, $F(5, 20) = 64.32$; $p < 0.0001$. In addition, we found that absolute errors were larger at set size 6 than at set size 3 for ratings of lower confidence, as indicated by a statistically significant interaction effect between confidence and set size, $F(5, 20) = 4.21$; $p < 0.01$. As can be seen in Figure 2B, the magnitude of absolute errors steadily increased with each decrement in confidence, and almost reached the magnitude predicted by pure guessing (i.e., 45°) for ratings of zero confidence in the set size 6 condition.

Next, we plotted frequency histograms to visualize the distribution of response errors (i.e., reported orientation minus true orientation) for each confidence rating level and set size, with data pooled across all participants (Figure 3A). The distribution of errors was very different across confidence levels. When participants reported a rating of '0' indicating essentially 'no confidence,' the distribution of reported orientations was almost flat, and closely resembled a uniform distribution of random guesses. Higher ratings of confidence were associated with more frequent reports centered about the true orientation of the cued grating. Ratings of highest confidence led to the most peaked distribution with errors rarely occurring at distal orientations.

Mixed-model analysis

We adopted a mixed-model approach, which assumes that working memory performance can be described by two key parameters: the likelihood of successful memory maintenance and the precision of the memory for successfully remembered items (Zhang & Luck, 2008, 2009). This was done by fitting a circular Gaussian-shaped model to the data for each set size and each confidence level (see Methods). The extent to which the entire Gaussian curve must be translated

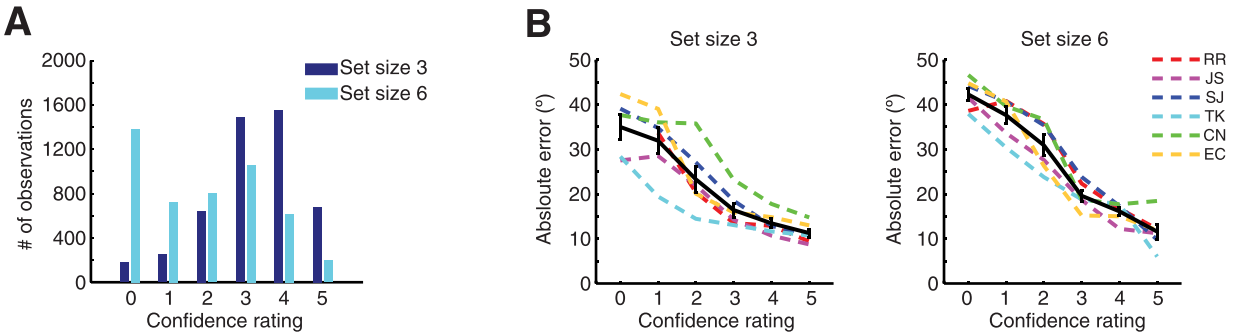

Figure 2. Confidence rating frequencies and absolute response errors. (A) Frequency distributions of confidence ratings for set sizes 3 and 6, pooled across the six participants in Experiment 1. (B) Absolute response error plotted as a function of confidence rating for set size 3 (left) and set size 6 (right). Absolute errors for each participant were calculated based on the absolute difference between reported orientation and true orientation for every trial of a given condition. Data for individual participants are plotted with dashed colored lines; group-averaged data are plotted with black solid lines and error bars showing ± 1 *SEM*.

uniformly upwards is presumed to reflect how often the probed item was forgotten (*P-uniform*), since random guesses would be expected to lead to a uniform distribution of responses. By contrast, the standard deviation (*SD*) or 'width' of the Gaussian distribution provides an estimate of the precision of memory (with smaller *SD* indicating greater precision), since the amount of internal noise with which an item is stored should be independent of the proportion of guess responses. (For the purposes of this study, we adopted these measures to evaluate the data according to the predictions of the slots-plus-averaging model, though it should be noted other researchers have argued that these two statistical measures may not necessarily reflect the proportion of guessing responses or the precision of memory (van den Berg et al., 2012).

When observers reported poor confidence in their memory, we observed very high estimates for the probability of memory failure for the cued item. Ratings of zero confidence were accompanied by a

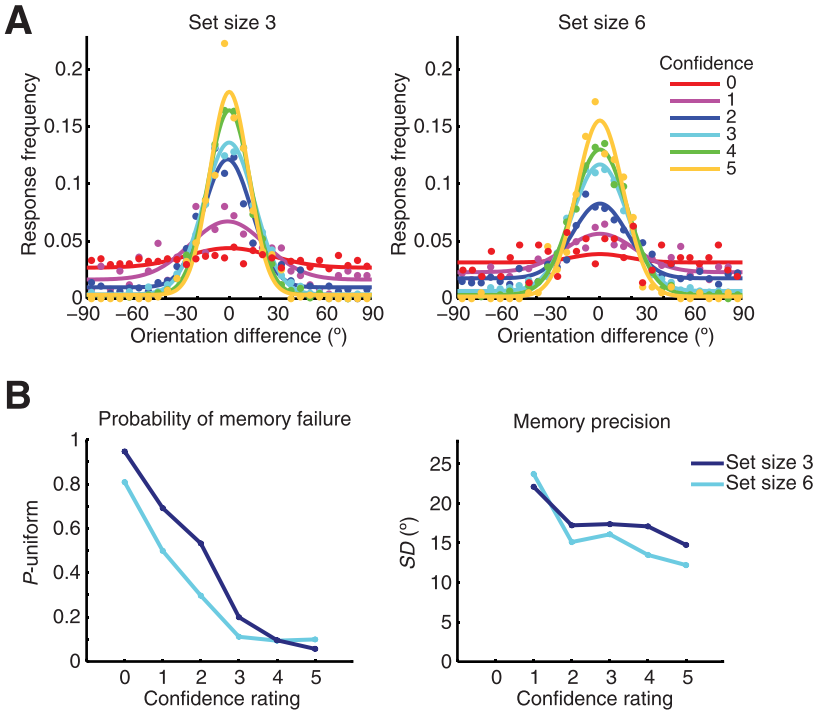

Figure 3. Mixed-model analysis of distribution of orientation errors for data pooled across participants. (A) Distribution of orientation differences between reported orientation and true orientation (centered at 0°), plotted by confidence rating for set size 3 (left) and set size 6 (right). Data points indicate frequency distributions using a bin width of 6°; curves show the best-fitting circular Gaussian function (centered around 0°) based on the mixed model analysis. (B) Parameter estimates of the probability of memory failure (left) and precision of memory for successfully remembered items (right) based on model fits of the data, collapsed across all participants in Experiment 1.

failure of memory on the vast majority of trials, whereas ratings at the highest confidence level led to memory failure on less than 10% of trials (Figure 3B, left panel). These results imply that observers can make accurate metacognitive judgments regarding whether they have successfully retained information about an item in working memory.

Might participants also have reliable metacognitive knowledge regarding the precision of their working memory for a visual item? Analysis of the pooled data across participants suggested that the visual precision of memory tended to improve as a function of self-reported confidence (Figure 3B, right panel). These variations in memory precision could also be seen in the Gaussian-shaped distribution of deviation errors for the orientation judgments, which appeared narrower and more precise for ratings of high confidence as compared to low confidence (Figure 3A).

We analyzed the data of individual participants to determine whether trial-by-trial variability in working memory performance could be predicted by individual confidence ratings. This analysis was needed to ensure that the trends found in the pooled group data could not be explained by individual differences, as could arise if individuals with superior memory precision also reported generally greater confidence in their memory. Moreover, this analysis allowed us to measure the consistency of these effects across participants, making standard statistical testing possible. Each participant's confidence ratings from 1–5 were subdivided into categories of low, medium, and high confidence, to obtain a sufficient number of observations for reliable estimation of the memory parameters at each set size for that individual. (Confidence ratings of 0 were excluded from this analysis, since they led to almost flat distributions that could not be well-fitted by the Gaussian-shaped model.)

We found that failure of memory for the probed item was much more likely to occur on trials in which observers reported low rather than high confidence in their memory (Figure 4A). This effect was confirmed by a within-subjects analysis of variance, $F(2, 10) = 56.3$, $p < 0.0001$. These findings imply that participants had very reliable metacognitive knowledge regarding whether they had successfully retained information about the probed item, though it is interesting to note that reports of subjective confidence were graded and probabilistic rather than all or none (Figure 2A). This apparent lack of memory for the probed item, based on estimates from the mixed-model analysis, could potentially arise from both failures of encoding due to the limited capacity of visual working memory (Zhang & Luck, 2008), as well as failures of active maintenance, especially for longer delays that considerably exceed 4 s in duration (Zhang & Luck, 2009). We found that the estimated probability of memory failure was far greater for set size 6 than for set size 3 (see Figure 5A), which is consistent with the predictions of the slot model regarding strict item limits for encoding. On a subset of trials, it is possible that successfully encoded items failed to be actively maintained throughout the delay period, though it is worth noting that our retention interval was not unusually long (~4 s) and our participants had an economic incentive to perform as well as possible.

Of particular interest, it can further be seen that each of the participants exhibited a general trend of better memory precision on high as compared to low confidence trials (Figure 4B). Statistical analyses confirmed this general trend, indicating that greater confidence was predictive of more precise reports of the cued orientation, $F(2, 10) = 18.3$, $p < 0.0005$. This proved true even for large display sizes of six gratings, $F(2, 10) = 7.224$; $p < 0.05$, which exceed the proposed three- to four-item limit of visual working memory (Luck & Vogel, 1997). (Among the participants tested here, estimates of working memory capacity (K) ranged from 2.0 to 3.87 items.) These results indicate that observers have reliable metacognitive knowledge of the precision of their working memory, which can vary across trials.

Is it possible that subjective confidence might have varied across trials because of spurious factors, such as momentary blinking or lapses in attention, irrespective of the demands of the memory task itself? Given that participants were economically rewarded for accurate performance and could choose when to advance to the next trial, we would expect such lapses to be very rare. Moreover, such an account cannot explain why confidence ratings differed greatly between set size conditions. With a mixed-trial design, participants could not predict the set size in advance of viewing the briefly flashed display (200 ms duration), and yet reports of zero confidence were rare at set size 3 while they were the most frequent response at set size 6 (Figure 2A). Thus, confidence ratings were strongly determined by the demands of the working memory task. An analysis of memory performance for each set size, pooled across all confidence ratings, further indicated that the probability of memory failure was far greater at the larger set size ($T = 9.78$; $p < 0.0005$; Figure 5A. In addition, working memory performance was significantly more precise at set size 3 than at set size 6 ($T = 3.98$; $p < 0.05$; Figure 5B).

Effects of training across sessions

We also considered the possibility that working memory performance and subjective confidence ratings might have generally improved over the course of the experiment, due to repeated training at the task. Such

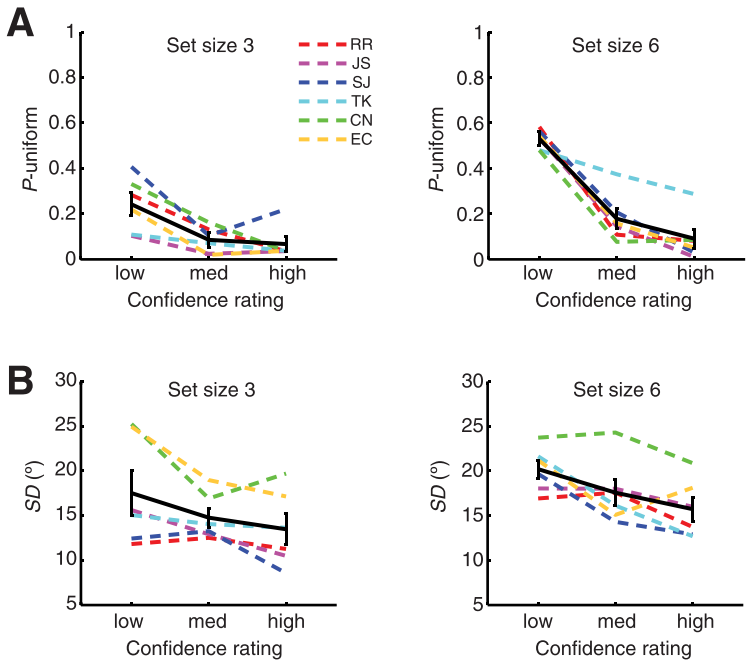

Figure 4. Probability of memory failure and memory precision across different levels of subjective confidence. (A) Estimated probability of memory failure (*P-uniform*) for trials rated with low, medium, or high subjective confidence. Data for individual participants are plotted with dashed colored lines; group-averaged data are plotted with black solid lines and error bars showing ± 1 *SEM*. (B) Precision of memory (*SD*) for orientations reported with low, medium or high confidence. Higher levels of confidence were predictive of smaller values of *SD*, indicating superior memory precision.

day-to-day improvements could potentially account for the apparent relationship between subjective confidence and objective performance, rather than trial-to-trial variability. To address this issue, we calculated estimates of the probability of memory failure and the precision of memory for each of the 10 experimental sessions (Figure 6A, B), pooling across set sizes 3 and 6 to ensure that there was sufficient data to obtain highly reliable fits ($R^2 > 0.75$ in all cases). Our analyses failed to reveal evidence of a change in the likelihood of memory failure, $F(9, 45) = 1.787$; $p = 0.097$, or a change in the precision of visual memory, $F(9, 45) = 1.61$; $p = 0.14$, across sessions. Moreover, we found that confidence ratings remained stable across sessions, $F(9, 45) = 0.732$; $p = 0.678$; Figure 6C, and were consistently higher for set size 3 than for set size 6, $F(1, 45) = 61.3$; $p < 0.0001$. These findings indicate that learning effects across sessions cannot account for the strong relationship we observe between subjective confidence and objective memory performance.

Analysis of nontarget responses

To what extent might spatial confusions between the cued grating (i.e., target) and other nontarget gratings

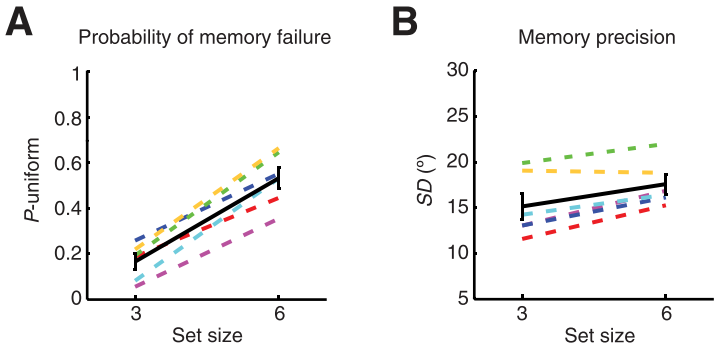

Figure 5. Comparison of working memory performance for set sizes 3 and 6. (A) Estimated probability of memory failure (*P-uniform*) for set sizes 3 and 6, for data pooled across all confidence levels. Individual data plotted with dashed colored lines; group average in solid black. (B) Estimated precision of memory (*SD*) for set sizes 3 and 6.

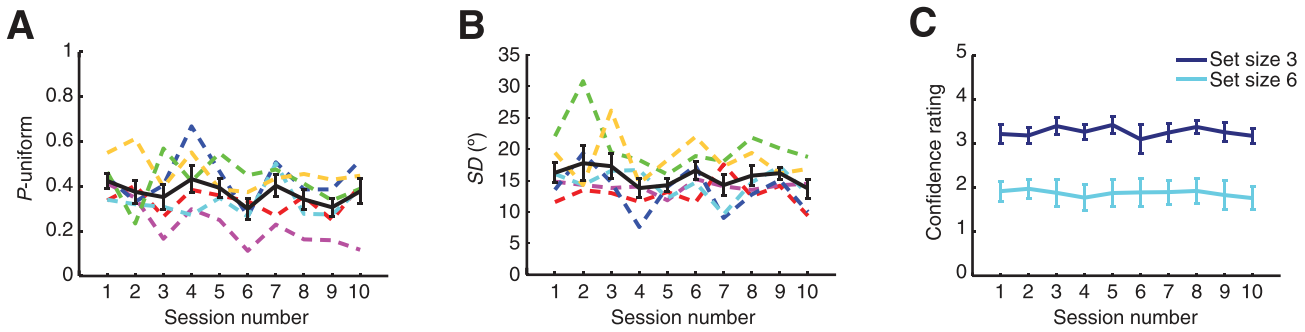

Figure 6. Effects of training on working memory performance and confidence ratings. (A) Probability of memory failure for each of the 10 test sessions in Experiment 1. Data were pooled across set sizes 3 and 6 to obtain reliable parameter estimates for all individual participants ($R^2 > 0.75$ in all cases). (B) Estimates of memory precision plotted by session. (C) Confidence ratings plotted by experimental session for set sizes 3 and 6. Statistical analyses indicated that there were no reliable changes in memory performance or confidence ratings across sessions.

have contributed to the pattern of results found in this study? Is it possible that on low confidence trials, participants were more likely to exhibit such confusions due to the misbinding of object and location information, and might this have affected our estimates of memory precision? According to Bays et al. (2009), a significant proportion of apparent "guessing" responses in working memory tasks can instead be attributed to the report of a nontarget stimulus. To address this issue, we applied their maximum likelihood fitting procedure to our data for low, medium, and high confidence trials, and obtained separate estimates of the proportion of random guess responses, nontarget responses, and the precision of memory for the reported item.

As can be seen in Figure 7, we observed the same overall pattern of results after incorporating nontarget responses into our analysis. The estimated proportion of uniform guessing responses was significantly greater at low confidence than at high confidence, $F(2, 10) = 17.68$, $p < 0.001$, and also larger for set size 6 than for set size 3, $F(2, 10) = 31.3$, $p < 0.0001$. Guessing responses occurred most frequently on low confidence trials for set size 6, comprising about 40% of those trials. In comparison, nontarget responses occurred much less frequently. We did observe significantly more nontarget responses on trials with lower confidence, $F(2, 10) = 4.99$, $p < 0.05$, suggesting that confusions between target and nontarget items are more prevalent when participants report low confidence. However, even on these low confidence trials, the proportion of nontarget responses did not exceed 3.4% and 11.9% for set sizes 3 and 6, respectively.

To what extent did these nontarget responses impact our estimates of memory precision?

With nontargets incorporated into our analysis, we again found that higher ratings of subjective confidence were predictive of more precise memory for orientation, $F(2, 10) = 21.6$, $p < 0.0005$. Thus, greater confidence is associated with decreased likelihood of guessing, decreased confusions between target and

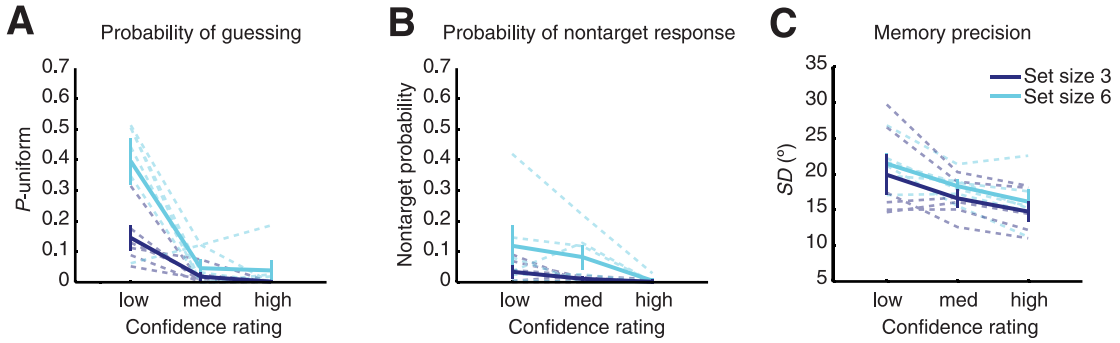

Figure 7. Parameter estimates with nontarget responses included in mixed model analysis. (A) Estimated probability of random guessing (*P-uniform*) for different confidence levels, using a mixed model analysis that includes possible nontarget responses. (B) Estimated probability of nontarget responses. Such responses were quite rare overall, but did occur more frequently on trials involving lower self-reported confidence and larger set size. Note that target responses (not shown) included all remaining trials that were not classified as nontarget responses or random guessing. (C) Estimated precision of memory (*SD*) for this analysis. Individual data shown with dotted lines, group data shown with solid lines and error bars indicating ± 1 *SEM*. Dark blue, set size 3; light blue, set size 6.

nontargets, and superior memory precision for successfully maintained targets.

Additional control experiment

We conducted a separate control experiment to ensure that our main results did not depend on the fact that subjective confidence ratings were made prior to reporting the orientation of the cued grating. In Experiment 2, participants first reported the orientation of the cued grating, and then reported their confidence in their memory (Figure 8A). Observers were first asked if they had any memory at all for the cued grating, and if they responded yes, were then asked to rate how well they remembered the cued grating on a scale of 1 (not well) to 5 (very well). The stimuli and experimental conditions were otherwise the same as those of the main experiment, with participants performing a total of 1,600 trials over a series of 10 testing sessions. Participants did not receive additional pay for accurate performance in this experiment; nevertheless, the same pattern of results was found in the pooled data across participants (Figure 8B, C). We analyzed the data of individual participants, after binning trials according to low, medium and high confidence ratings (Table S2). Our analyses indicated that both the likelihood of successful memory maintenance, $F(2, 10) = 28.2$, $p < 0.0001$, and the precision of visual memory, $F(2, 10) = 6.1$; $p < 0.05$, were better at high than low confidence (Figure 9).

We directly compared the results of this second experiment with data from the first experiment, by performing a combined ANOVA with experiment type as a between-groups factor. We found no evidence of a difference in overall probability of memory failure or memory precision across experiments ($F < 1$ in both cases). More important, we found that the combined data across the two experiments led to highly significant effects of confidence for probability of memory failure, $F(2, 20) = 76.7$; $p < 0.0001$, and memory precision, $F(2, 20) = 18.1$; $p < 0.0001$, with no evidence of an interaction between these effects of confidence and experiment type ($F < 1$ in both cases). From these results, we can conclude that the relationship between subjective confidence and working memory performance is highly reliable, and unaffected by whether the confidence rating or the memory report is performed first.

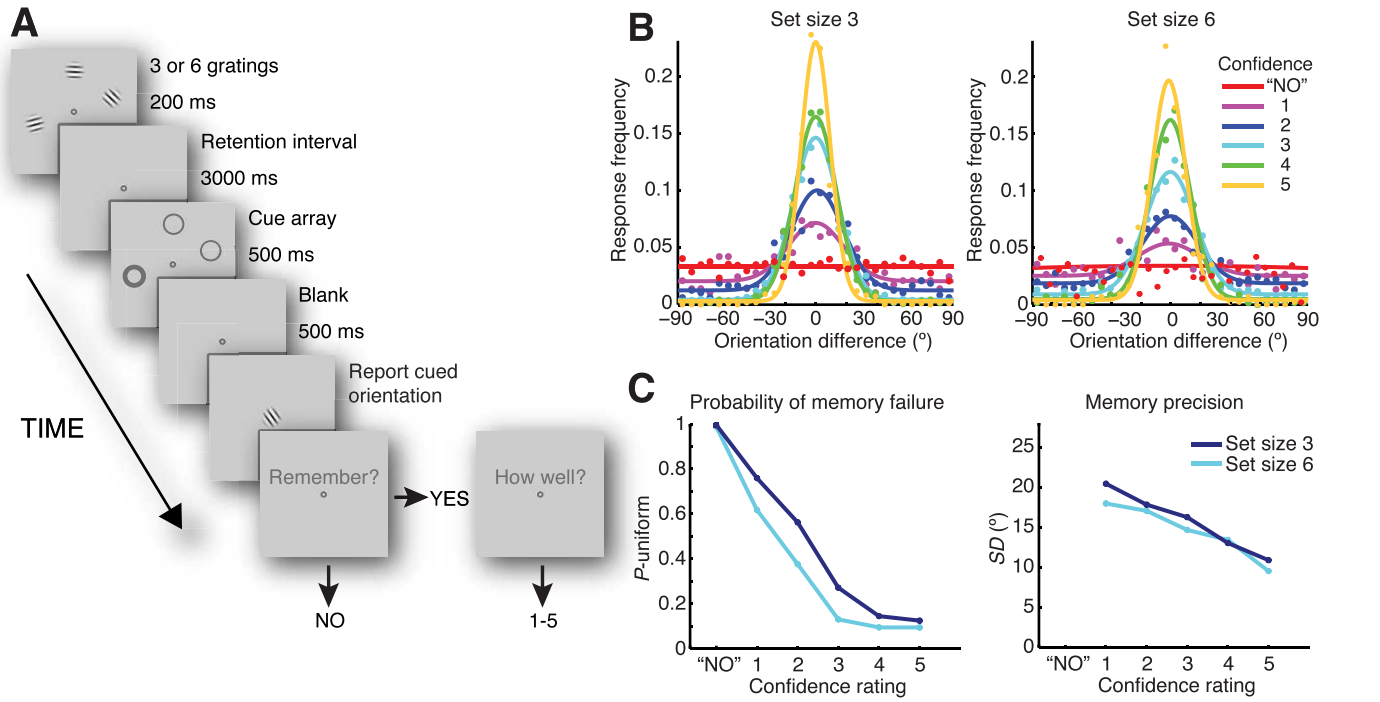

Figure 8. Experimental design and pooled group results for Experiment 2. (A) Design for Experiment 2. After presentation of the sample display, participants first reported the orientation of the cued grating and then rated the accuracy of their memory. They were cued ("Remember?") to make a binary decision indicating whether they had any memory of the cued grating. For 'yes' responses, they were then asked to rate how well they remembered the grating's orientation, from 1 (not well at all) to 5 (very well). (B) Pooled group data showing the distribution of orientation errors, plotted by confidence rating separately for set sizes 3 and 6. Data points indicate frequency distributions using a bin width of 6°; curves show the best-fitting Gaussian function (centered around 0°) based on the mixed model analysis. (C) Parameter estimates of the probability of memory failure (left) and the precision of memory for successfully remembered items (right) based on model fits of the pooled group data.

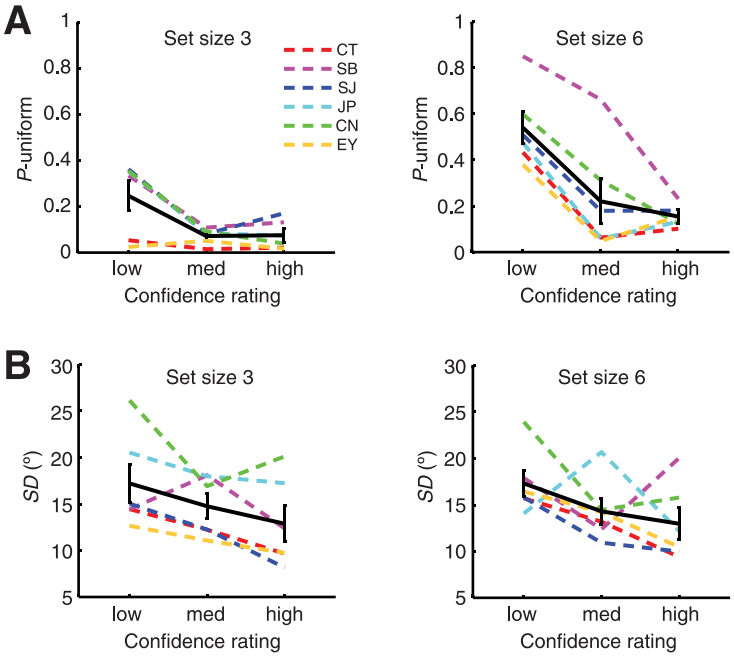

Figure 9. Memory performance across different confidence levels for Experiment 2. (A) Estimated probability of memory failure (*P-uniform*) for trials rated with low, medium, or high subjective confidence in Experiment 2. (B) Estimates of memory precision (*SD*). Working memory performance was significantly more precise on trials with higher confidence ratings, $F(2, 10) = 6.1$; $p < 0.05$. Dashed colored lines show individual data; solid black lines show group-averaged data.

Discussion

Our study provides novel evidence that people have accurate metacognitive knowledge regarding the reliability and precision of their working memory for specific items. Through the use of subjective confidence ratings, we found that the visual precision of working memory can indeed fluctuate from trial to trial. These findings demonstrate the variable nature of the visual working memory system, and may help inform current theories and models.

According to the slots-plus-averaging model, working memory consists of three to four discrete slots, each of which can store a single object with a fixed degree of visual resolution (Zhang & Luck, 2008). If the number of display items exceeds the capacity of working memory, then each available slot will be used to retain a unique item with fixed precision, and any residual items must necessarily be discarded. Current versions of this model do not address or account for the possibility of variability in memory precision, for conditions in which the number of items exceed working memory capacity. Indeed, several studies by Zhang and Luck have emphasized the fact that memory precision appears fixed across a variety of experimental manipulations. For example, changes in display duration (110 ms vs. 340 ms) were found to affect the probability of successful memory maintenance but did not affect memory precision for successfully maintained items (Zhang & Luck, 2008). Similarly, experimental manipulations of retention duration, spatial attentional cuing (comparing neutral and invalid conditions), and strategic instructions to emphasize memory capacity or precision, were found to have no effect on memory precision (Zhang & Luck, 2008, 2009, 2011). Some researchers have reported modest decrements in memory precision for increases in set size beyond three to four items or for backward masked stimuli that are displayed for less than 100 ms (Bays et al., 2009; Bays, Gorgoraptis, Wee, Marshall, & Husain, 2011), but these effects required direct manipulations of the visual display. By contrast, here we find that the resolution of visual working memory can indeed fluctuate from trial to trial due to strictly *endogenous* factors, in the absence of experimental manipulations of the stimuli or task.

However, our results also run contrary to the predictions of standard resource models (Bays & Husain, 2008; Wilken & Ma, 2004). These models assume that it should be possible to retain more than three to four items by trading off visual resolution for increased capacity, but do not address the possibility that memory precision might vary across items or trials when set size is held constant. Another problem arises from the fact that most resource models rely on a signal detection framework to account for changes in performance across set size, which does not allow for item limits or uniform guessing responses. Such a framework would have difficulty accounting for the heavy tails that were observed in the

distribution of errors on low confidence trials (e.g., Figure 3A). Based on our mixed-model analysis, we found that estimates of memory precision for set size 6 were only fractionally worse on low as compared to high confidence trials (*SD* of 20.2° vs. 15.7°, respectively), and yet the likelihood of memory failure was almost six times greater on low than high confidence trials (53% vs. 9%, respectively) due to the heavy tails in the error distribution. Signal detection-based models would require a much greater loss of memory precision (i.e., much larger *SD*) to account for the heavy tails on low confidence trials.

How then might our findings be reconciled with current models of working memory? One possibility is that the visual working memory system can maintain a limited number of individual objects (Zhang & Luck, 2008) or features (Fougnie & Alvarez, 2011; Fougnie, Asplund, & Marois, 2010), as predicted by slots-plus-averaging, but that the precision of each stored visual representation can vary to some extent. Based on our estimates from the mixed-model analysis, average working memory precision (*SD*) varied over a modest range of about 15°–20° between conditions of low and high confidence. Modest variations in memory precision could partly arise from limitations of perceptual encoding, perhaps due to variations in visual sensitivity across the visual field (Westheimer, 2003) or other such factors. Fluctuations in memory precision might also reflect the degree of internal noise present in the nervous system (Pasternak & Greenlee, 2005; Wilken & Ma, 2004), as greater noise during the encoding or maintenance period would lead to less precise memory performance. Finally, memory precision might depend on the amount of attention directed to each item in the visual display during the encoding phase, as focal attention has been shown to improve the spatial resolution of visual processing in both perceptual and short-term memory tasks (Bays et al., 2011; Yeshurun & Carrasco, 1998; Zhang & Luck, 2008). However, caution is required here if one assumes that attentional resources can be distributed in a continuous manner across items in the visual field, as the slots-plus-averaging model is distinguished by its core assumption that central resources are allocated in discretized units.

An alternative account of our findings would be a continuous resource model that allows for variability in memory precision as part of its core framework. Recently, van den Berg et al. (2012) proposed a variable-precision model of working memory, in which the amount of encoding resources assigned to each item in a display can randomly vary across items and trials. Variable precision in this model is based on the assumption that the distribution of errors at a given set size should reflect the sum of a continuous set of circular Gaussian (or von Mises) distributions of varying precision, ranging from extremely coarse to very fine. This model assumes that the working memory system has no discrete limits; instead, what appears to be a uniform random-guessing component in the distribution of errors is attributed to extremely coarse memory representations on a subset of trials. It would be interesting for future studies to evaluate whether the variable-precision resource model might provide a good fit to our data across variations in self-reported confidence, and to compare this with a modified slots-plus averaging model with discrete item limits that allows for modest variations in memory precision.

Our study further demonstrates that people can make remarkably accurate metacognitive judgments regarding the accuracy of their own memories. Confidence ratings were highly predictive of whether or not an item was successfully retained, with failures of memory rarely occurring during reports of high confidence. These findings are consistent with the proposal that the contents of working memory appear to be immediately accessible to consciousness (Baars & Franklin, 2003; Baddeley, 2000). Our results suggest that this is likely to be true, as participants were able to introspect whether they had a memory representation for what was previously presented at the cued location. However, current theories of working memory would not necessarily anticipate that confidence ratings should be predictive of the precision of visual working memory. Such metacognitive judgments of memory precision are much less straightforward to implement. Participants did not have the opportunity to directly compare their memories to the previously seen items, nor did they receive feedback regarding how accurately they had performed on the task for any specific trial. This implies that observers can assess the quality or clarity of their internal visual representations to some degree, and evaluate the degree of internal noise with which an item is stored. This ability to evaluate the quality of one's internal visual representations may be related to recent neuroimaging studies that have revealed the presence of item-specific activity in early visual areas during working memory maintenance (Harrison & Tong, 2009; Serences, Ester, Vogel, & Awh, 2009). Our findings add to a growing literature on the cognitive and neural bases of metacognitive judgments (Fleming, Weil, Nagy, Dolan, & Rees, 2010; Kiani & Shadlen, 2009; Song et al., 2011), as well as recent work on the reliability of metacognitive judgments pertaining to mental imagery (Pearson et al., 2011).

The present study illustrates how research on human metacognition can provide insights into the nature of cognitive systems. An interesting question for future research would be to explore whether reliable metacognitive performance might be integral to the effective functioning of the working memory system. If a system's contents must be continually updated while operating near its capacity limits, then such a system might greatly

benefit from having a mechanism to evaluate the accuracy of its own contents. More generally, the ability to evaluate whether or not a perception or memory is accurate could prove helpful for making informed decisions in a variety of settings, especially those in which one must act or bet on the reliability of one's own judgments (Hampton, 2001; Kiani & Shadlen, 2009; Persaud, McLeod, & Cowey, 2007).

Conclusions

The visual precision of working memory for individual items can fluctuate from one trial to the next. Observers were able to make accurate metacognitive judgments about the content and also the precision of their working memory for a specific item. This latter type of judgment is considerably more subtle and complex, as participants did not have the opportunity to directly compare their memory to the item itself, which was previously seen several seconds ago; nor did they receive immediate feedback regarding their performance. We found that visual working memory is fundamentally variable in nature. Such variability has not been considered in most previous models of working memory, although variable precision could be incorporated into models that assume limited discrete slots or continuous resources. The present findings inform an ongoing debate regarding the defining functional properties of the visual working memory system.

Acknowledgments

The authors would like to thank M. Pratte for helpful discussions, and E. Counterman and C. Neely for technical assistance. This research was supported by the NEI grant R01 EY017082 and NSF grant BCS-1228526 to FT. It was also facilitated by support from center grant P30-EY008126 to the Vanderbilt Vision Research Center, directed by Dr. Schall.

Commercial relationships: none.
Corresponding author: Rosanne L. Rademaker.
Email: rosanne.rademaker@maastrichtuniversity.nl.
Address: Department of Cognitive Neuroscience, Maastricht University, The Netherlands.

References

Anderson, D. E., Vogel, E. K., & Awh, E. (2011). Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. *Journal of Neuroscience, 31*(3), 1128–1138.

Baars, B. J., & Franklin, S. (2003). How conscious experience and working memory interact. *Trends in Cognitive Sciences, 7*(4), 166–172.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417–423.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4*(10), 829–839.

Bays, P., Catalao, R., & 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, http://www.journalofvision.org/content/9/10/7, doi:10.1167/9.10.7. [PubMed] [Article]

Bays, P., Gorgoraptis, N., Wee, N., Marshall, L., & Husain, M. (2011). Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision, 11*(10):6, 1–15, http://www.journalofvision.org/content/11/10/6, doi:10.1167/11.10.6. [PubMed] [Article]

Bays, P., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321*(5890), 851.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*(1), 26–48.

Cohen, M. A., Alvarez, G. A., & Nakayama, K. (2011). Natural-scene perception requires attention. *Psychological Science, 22*(9), 1165–1172.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906–911.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science, 329*(5998), 1541–1543.

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of Vision, 11*(12):3, 1–12, http://www.journalofvision.org/content/11/12/3, doi:10.1167/11.12.3. [PubMed] [Article]

Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory? *Journal of Vision, 10*(12):27, 1–11, http://

www.journalofvision.org/content/10/12/27, doi:10.1167/10.12.27. [PubMed] [Article]

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America, 98*(9), 5359–5362.

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

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324*(5928), 759–764.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition, 10*(3), 294–340.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America, 99*(14), 9596–9601.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279–281.

Magnussen, S., & Greenlee, M. W. (1999). The psychophysics of perceptual memory. *Psychological Research, 62*(2-3), 81–92.

Nickerson, R. S., & McGoldrick, C. C., Jr. (1965). Confidence Ratings and Level of Performance on a Judgmental Task. *Perceptual & Motor Skills, 20,* 311–316.

Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience, 6*(2), 97–107.

Pearson, J., Rademaker, R. L., & Tong, F. (2011). Evaluating the mind's eye: The metacognition of visual imagery. *Psychological Science, 22*(12), 1535–1542.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience, 10*(2), 257–261.

Philips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics, 16*(2), 283–290.

Rademaker, R. L., & Pearson, J. (2012). Training visual imagery: Improvements of metacognition, but not imagery strength. *Frontiers in Psychology, 3,* 224.

Regan, D., & Beverley, K. I. (1985). Postadaptation orientation discrimination. *Journal of the Optical Society of America A, 2*(2), 147–155.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience, 5*(7), 629–630.

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science, 20*(2), 207–214.

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness & Cognition, 20*(4), 1787–1792.

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 of the United States of America, 109*(22), 8780–8785.

Westheimer, G. (2003). Meridional anisotropy in visual processing: Implications for the neural site of the oblique effect. *Vision Research, 43*(22), 2281–2289.

Wilken, P., & Ma, W. (2004). A detection theory account of change detection. *Journal of Vision, 4*(12):11, 1120–1135, http://www.journalofvision.org/content/4/12/11, doi:10.1167/4.12.11. [PubMed] [Article]

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152–176.

Yeshurun, Y., & Carrasco, M. (1998). Attention improves or impairs visual performance by enhancing spatial resolution. *Nature, 396*(6706), 72–75.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1341–1354.

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

Zhang, W., & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. *Psychological Science, 20*(4), 423–428.

Zhang, W., & Luck, S. J. (2011). The number and quality of representations in working memory. *Psychological Science, 22*(11), 1434–1441.