Psychological Research (2011) 75:259–271
DOI 10.1007/s00426-010-0304-5

ORIGINAL ARTICLE

Integrating visual mental images and visual percepts: new evidence for depictive representations

Katie J. S. Lewis · Grégoire Borst · Stephen M. Kosslyn

Received: 6 April 2010 / Accepted: 2 August 2010 / Published online: 24 August 2010

Abstract In two experiments, we used a temporal integration task to investigate visual mental images based on information in short-term memory or generated from information stored in long-term memory (LTM). We specifically asked whether the two sorts of images rely on depictive representations. If mental images rely on depictive representations, then it should be possible to combine mental images and visual percepts into a single representation that preserves the spatial layout of the display. To demonstrate this, participants were asked to generate mental images and then combine them with visual percepts of grids that were partially filled with different numbers of dots. Participants were asked to determine which cell remained empty when the two grids were combined. We contrasted predictions of propositional or verbal description theories with those of depictive theories, and report findings that support the claim that mental images—based on either short-term or LTM—depict information.

Introduction

Introspectively, visual mental images appear akin to "mental pictures," which has led many to liken visual mental imagery to "seeing with the mind's eye". However, some researchers argue that the picture-like qualities of visual mental imagery, so evident to introspection, are in fact epiphenomenal—they play no role in information processing (e.g., Pylyshyn, 1973; for a review, see Kosslyn, Thompson, & Ganis, 2006). According to this view, all mental activity relies on symbolic, propositional (descriptive) representations, and the experience of imagery is like the heat thrown off by a light bulb while one reads—of no functional consequence. In contrast, others have argued that the depictive properties of visual mental images are functional, in that they convey information originally encoded during perception. According to this view, if the representations that underlie imagery are depictive, they use space in a representational medium to represent space in the world, such that each part of the representation corresponds to a part of the represented object and the distances among the representations of parts mirrors the actual distances among the parts themselves (see Kosslyn et al., 2006).

A wealth of behavioral research has been conducted in an effort to resolve the debate about whether visual mental images rely, at least in part, on depictive representations. This behavioral data has indicated that participants are able to scan (e.g., Kosslyn, Ball, & Reiser, 1978), rotate (Shepard & Metzler, 1971), and inspect (Kosslyn, 1975) objects in visual mental images, suggesting that their representations possess depictive qualities. However, propositional theorists have proposed alternative explanations for such data. In many cases, the accounts rely on the notion that participants drew on general knowledge of physical laws to produce the results, and such knowledge is presumed to be represented propositionally (Pylyshyn, 2002, 2003).

The fact that the behavioral data have failed to resolve the issue thus far does not imply that such data in principle cannot bear on the issue. In this article, we consider an alternative method, which relies on a simple logic: First,

K. J. S. Lewis
Department of Psychology, University of Bath, Bath, UK

G. Borst (✉) · S. M. Kosslyn
Department of Psychology, Harvard University,
836 William James Hall, 33 Kirkland Street,
Cambridge, MA 02138, USA
e-mail: borst@wjh.harvard.edu

we know that perceptual displays do depict, by definition, and we know that the initial representations of visual information in the brain clearly depict information (Sereno, Dale, Reppas, Kwong, Belliveau, Brady et al., 1995). Second, if visual mental images rely on these same representations, they would also depict. Third, one way to find out whether images do in fact rely on the same depictive representations used in perception is to determine whether participants can combine information in visual mental images with information in perceptual displays, producing a single representation that preserves the spatial layout of the display.

Researchers have developed methods that can be used to investigate how visual mental images and visual percepts can be combined. However, until recently, such methods have been used solely to study temporal integration of percepts (e.g., Di Lollo, 1980; Eriksen & Collins, 1967; Loftus & Irwin, 1998). In these studies, participants are required to integrate two stimuli, presented sequentially, and then to use the resulting stimulus to solve a task. These studies generally reported optimum levels of performance for interstimulus intervals (ISIs) shorter than 50 ms, and sharp decreases in accuracy when ISIs exceeded this delay (see Brockmole, Wang, & Irwin, 2002). This finding implies that temporal integration of percepts is limited to a small time frame, given that successful integration relies on the visible persistence of the stimuli presented.

The studies reported in this article followed directly on the heels of those reported by Brockmole et al. (2002). They reasoned that increasing the ISI of two sequentially presented displays beyond 100 ms should allow participants to form a mental image of the first stimulus, which could be integrated with a subsequent perceptual stimulus; if so, then performance should improve when ISIs are increased beyond 100 ms. Brockmole et al. adapted a paradigm from Di Lollo (1980), and asked participants mentally to combine pairs of patterns that were presented sequentially. Each pattern included seven or eight dots, with each dot occupying a cell of a 4×4 grid. The task was to determine which cell in the grid remained unoccupied by a dot when the patterns were combined. Both patterns were presented for 33 ms, with ISIs ranging from 500 to 5,000 ms (0 and 100 ms served as baseline conditions; 0 ms being percept–percept integration and 100 ms used as delays in which percept–percept integration typically does not occur). In the percept–percept integration condition (ISI of 0 ms); participants were accurate on a mean of 79% of the trials, indicating that they were able to integrate two perceptual images. Consistent with previous research on the temporal characteristics of perceptual integration, increasing the ISI from 0 to 100 ms led to a rapid decrease in accuracy (21%). However, the researchers found that when the ISI exceeded 100 ms, accuracy actually increased—with an asymptote of 68% at 1,500 ms. Brockmole et al. inferred that increasing the ISI allowed participants to generate a visual mental image of the first pattern, which was subsequently combined with the second. In addition, the fact that accuracy increased to roughly the same level as percept–percept integration led Brockmole et al. to conclude that these mental image representations can be processed the same way as perceptual representations. On the face of things, these data suggest that visual mental images rely on depictive representations.

However, some researchers have argued that participants perform better with longer ISIs (greater than 1,300 ms) because they compare representations of the two patterns within visual short-term memory (VSTM), and not because they generate mental images and integrate these images with percepts (Jiang, 2004; Jiang & Kumar, 2004; Jiang, Kumar, & Vickery, 2005). According to this view, to isolate the empty cell, participants attend to the empty cells of the first pattern, store in VSTM a representation of the negative (i.e., empty) space, and then compare this representation to the pattern of dots in the second array. To evaluate the two possible methods of performing the task, Hollingworth, Hyun and Zhang (2005) manipulated the complexity of the pattern in the first array; either the dot patterns were simpler than the negative space patterns or vice versa. Participants were better at locating the empty cell when the negative space patterns were simple (as compare to the simple dot array). These results suggest that participants remember the negative space patterns instead of the array patterns. By doing so, they turn the empty cell localization task into a change detection task, in which the representations of the two arrays are not integrated within a single representation.

These findings raise the spectre that Brockmole et al.'s results may not demonstrate successful image-percept integration. However, we note that in another integration task, Brockmole, Irwin and Wang (2003) provided evidence that participants focus their attention on the first dot array rather than the empty cells in the second dot array. These findings support the claim that, in this type of task, participants locate the empty cell by maintaining a visual representation of the first array and combining it with the second array.

In the present article, we present results from two variations of the Brockmole et al. paradigm in which participants were explicitly asked to form a mental image of the first pattern and to integrate this image into a visually presented second pattern. In one condition, participants generated the mental images of the first stimulus based on information stored in short-term memory (STM); in the other condition, the first stimulus was generated entirely from long-term memory (LTM). By comparing the effects

of identical manipulations in the two conditions, we are able to assess the characteristics of visual mental images created on the basis of a STM representation versus a LTM representation, which is of interest in its own right.

Experiment 1

The goal of Experiment 1 was to determine whether visual mental images can be successfully combined with visual percepts. To test this hypothesis, we used the integration task initially created by Di Lollo (1980) and modified by Brockmole et al. (2002). In this paradigm, pairs of 4 × 4 grids are presented sequentially at varying ISIs. In each grid, some of the cells are filled with dots whereas others remain empty. When each pair of grids is mentally combined, all cells but one are occupied by a dot. On each trial, participants attempt mentally to integrate each pair of grids to determine which cell is left empty. We modified this paradigm by asking participants to create a mental image of the first grid and combine this image with a second grid displayed briefly on a computer screen.

In our experiment, participants generated their mental images on the basis of information stored either in STM (i.e., STM condition) or in long-term memory (i.e., LTM condition). In the STM condition, the first grid was presented on the screen for 5 s. After a 5 s delay, participants generated a visual image of this grid before the brief presentation of the second grid. In the LTM condition, prior to performing the task, participants memorized three grids, each containing a different number of dots, and learned to associate each grid with a letter. On each trial, participants heard the name of a letter and generated an image of the grid corresponding to that letter. Then, the second grid was presented, as in the STM condition.

In addition to the response times, we also recorded the time the participants required to generate the images. We can derive three contrasting predictions for the effects of the number of dots on these generation times: (1) The depictive theory makes a clear prediction for the image generation times: Previous studies have demonstrated that each part of an imaged object is generated sequentially (e.g., Kosslyn, 1980, 1994; Kosslyn, Cave, Provost, & Von Gierke, 1988; Kosslyn, Reiser, Farah, & Fliegal, 1983), and hence objects with more parts take more time to visualize. Therefore, if participants generated depictive mental images, then we expect generation times to increase as participants visualized grids containing more filled cells. (2) As the number of dots increases, the complexity of negative space decreases—and thus if the participants have memorized the empty space, then they should require less time to generate images of patterns with greater numbers of dots. (3) Finally, if verbal descriptions are used, we expect no effect of the number of dots on the generation time, given that no representation needs to be created. In fact, we would expect participants to take as little time as possible in the generation phase, in order to respond before the information has faded from memory. If so, then the intercept of the function relating stimulus complexity to generation time should be lower in the STM condition than in the LTM condition.

If visual mental images and visual percepts both rely on a depictive format, then it should be possible to combine these representations into a single depictive representation. Therefore, we expect participants to be able to determine accurately which cell is not filled when the two grids are integrated (i.e., they will perform above chance level in the integration task). On the other hand, if mental images do not rely on depictive representations, then it should not be possible to combine them easily with percepts. In this case, participants should perform at chance level in the integration task.

However, one could argue that participants store the locations of dots in a description, using a propositional format, and then use this description to search the cells of the grid. If so, then we can make two predictions: First, participants should be less accurate and require more time to retrieve a stored representation of Pattern 1 when more dots are included in the grids, if a list of filled locations in Pattern 1 is retrieved from memory and used to search the cells of Pattern 2. Second, as more dots are included in Pattern 1, participants should be less accurate and require more time in the STM condition than in the LTM condition: the descriptions would need to be formulated on the fly in the STM condition, and it should be increasingly difficult to formulate and store accurate descriptions of the more complex patterns in the STM condition. Thus, in order to test the propositional theory, we manipulated the number of filled cells in the grids participants visualized.

If participants identify the empty cell by integrating mental images with visual percepts, then two predictions follow: First, the number of dots in the first array should affect performance only if it exceeds the capacity of spatial mental images. Second, participants' performance (accuracy and speed) should be similar in the LTM and the STM conditions, given that the source of generated mental images should not influence how they are subsequently processed (Kosslyn, 1994).

Method

Participants

Thirty-one volunteers from Harvard University and the local community participated in the experiment (23 females and 8 males, mean age 24 years and 4 months).

Twenty-eight participants were right-handed and three were left-handed. Participants were recruited via the Harvard Study Pool website and were compensated $10 or course credit. All participants provided written consent and reported normal or corrected-to-normal vision. All participants (including those reported in Experiment 2) were tested in compliance with national and international rules and regulations governing the ethical treatment of human research participants. The study was approved by the Harvard University Faculty of Arts and Sciences Committee on the Use of Human Subjects.

Materials

Stimuli consisted of 4 × 4 square grids. Each grid consisted of light blue grid lines and a gray background. Each grid measured 15.9 cm, subtending approximately 18° of visual angle. The black dots used to fill cells within the grid each measured 3.3 cm (3.8° visual angle) and each cell in the grid measured 3.96 cm (4.5° visual angle). Stimuli were created in pairs so that the first grid (i.e., Pattern 1) contained two, four, or eight dots whereas the second grid (i.e., Pattern 2) contained, respectively, 13, 11, or 7 dots. When combined, the two patterns of dots filled all but one cell. The configurations of dot patterns and the location of the unoccupied cell were randomly generated to produce 24 pairs of dot patterns for the main trials, and 18 pairs for the practice trials. When creating these patterns, we were constrained by the fact that the LTM condition utilized only one two-dot pattern, one four-dot pattern, and one eight-dot pattern across all trials. This meant that, for the pattern containing eight dots, there were only eight possible locations for the unoccupied cell. Thus, for each level of image complexity (two, four, or eight dots in Pattern 1), we created eight Pattern 2 stimuli. Stimuli were presented using E-prime version 2.0 software on an IBM ThinkVision L171 monitor (17 in.) with a resolution of 1,280 × 1,024 pixels and a refresh rate of 75 Hz.

Procedure

Participants were tested individually, sitting approximately 50 cm from a computer monitor. Participants performed both the STM condition and the LTM condition of the integration task. The order of two conditions was counterbalanced across participants.

In the STM condition, participants generated a visual-mental image of Pattern 1 from information stored in STM, and combined this image with the visual percept of Pattern 2. On each trial, participants first studied Pattern 1 for 5 s. Then, participants kept their gaze focused on a fixation point for 5 s, which should have allowed any afterimage of the pattern of dots to fade before participants started to visualize the dots. In addition, participants were explicitly instructed not to visualize the dots before the end of the 5 s. During this interval, we discouraged participants from visualizing the dots by presenting a blank gray background; this background did not contain the grid, and thus made it more difficult to visualize the dots. Following this, an empty grid was presented in the centre of the screen along with a beep sound. At this point, participants were instructed to form a mental image of Pattern 1. Once they had formed a visual mental image of the pattern, participants pressed the spacebar. We measured the time between the onset of the empty grid and this response, which allowed us to record the time taken to generate the image. Immediately after participants pressed the spacebar, Pattern 2 was presented for 33 ms, followed by the presentation of the second empty grid and a mouse cursor (positioned in the centre of the grid on each trial). Participants were instructed to combine their mental image of Pattern 1 with Pattern 2, presented on the screen, to locate the cell that was not filled in either grid. Participants indicated their judgments by moving the mouse to shift the cursor, and then clicking on the appropriate cell in the empty grid. The presentation of the second empty grid started a timer that was stopped when participants clicked the mouse button. We recorded response times (RTs) and whether participants selected the correct cell.

In the LTM condition, participants integrated a visual mental image of Pattern 1 generated on the basis of information stored in LTM, and combined this image with a visual percept of Pattern 2. Before beginning the computer task, participants memorized three dot-patterns (i.e., Pattern 1), named by the letters A (containing two dots), B (containing four dots), or C (containing eight dots). Participants studied each pattern for 20 s and then drew the pattern from memory into an empty grid. This procedure was repeated five times for each pattern, with a review of all patterns twice before proceeding to the main task (at this point, all participants could reproduce the patterns accurately). Participants were told that they would later be cued to retrieve each of the patterns from memory and were reminded throughout the learning phase to associate each letter with its corresponding pattern.

On each trial in the LTM condition, an empty grid first appeared on the screen and the sound of the name of a letter cue—A, B, or C—was produced by the computer; the cue indicated that the participants should form a visual mental image of the respective Pattern 1. Participants pressed the spacebar as soon as they had visualized a vivid image of Pattern 1. As in the STM condition, we recorded the time taken to generate the mental image. Pressing the spacebar resulted in the immediate presentation of Pattern 2 for 33 ms. As in the STM task, participants were told to combine the mental image of Pattern 1 with Pattern 2,

thereby allowing them to determine which cell in the grid remained unoccupied by either pattern. They indicated their responses by using the mouse cursor to click on the correct cell in the grid. As in the STM task, RTs and accuracy were recorded.

Participants performed two blocks of 24 trials in each condition for a total of 96 trials. In each block, a pair of patterns was presented once. Thus, in the STM condition, each of the 24 Pattern 1 grids was presented twice—which did not provide much opportunity for the participants to store these patterns in LTM. This was not an issue in the LTM condition, and thus each of the three Pattern 1 grids was presented 16 times. The order of the trials was random, except that the number of dots in the first pattern (two, four, or eight) could not occur more than three times in a row. In each condition, participants first performed two sets of practice trials: first, a block of six practice trials where participants visualized three-dot patterns presented on a 3×3 grid. Second, participants performed a block of 18 practice trials with dot patterns presented on a 4×4 grid. In these practice trials, participants visualized patterns containing two, four, or eight dots. In all practice trials, but not in the actual test trials, participants received feedback on their accuracy.

At the end of the experiment, the participants completed a debriefing questionnaire to ensure that they did not infer the purpose of the experiment and that they had followed the instructions.

Results

We began by asking whether participants rely on visual mental images of Pattern 1 to perform the integration task. We conducted two analyses of variance (ANOVAs) on the generation times (the time taken for participants to visualize Pattern 1) to determine the effect of the number of dots participants visualized in Pattern 1 and of the image source (STM vs. LTM) on these times. The differences in the procedure and materials used in the STM and LTM conditions preclude our directly comparing these two conditions. However, our comparison of the relevant underlying processes in the LTM and the STM conditions rested on the effects of the pattern complexity in each of the two conditions, and this manipulation was the same in both conditions. Thus, although we report the overall comparisons of the two conditions in the following analyses, it is difficult to interpret the main effect of image source per se. Then, we turned to the analyses of participants' accuracy in the two memory conditions to determine whether mental images can be integrated with visual percepts. Finally, we analyzed the decision RTs (how long it took participants to locate the unoccupied cell in the grid). For each of the analyses, we report the effect size of the ANOVA (partial eta squared). Preliminary analyses revealed no effect of the gender of the participants or of the order of the tasks on the dependent variables; thus we pooled the data over these variables.

Image generation times

Preliminary analyses indicated no differences between the image generation time on correct and incorrect trials (i.e., trials on which participants subsequently made a correct or incorrect response), respectively $t(27) = 1.32$, $p = 0.2$ in the LTM condition, and $t < 1$ in the STM condition, and hence we combined the data from both types of trials in subsequent analyses. However, before performing the analysis we excluded outliers, which were defined as participants for whom the average generation time was more than 3 SD from the group mean. Based on previously published data on the time to generate visual mental images, it was clear that these participants were not pressing the spacebar after generating the image—and thus, retaining the data from these participants for this analysis was not appropriate. We removed 3 participants out of 31 from the following analyses. However, we note that including these participants in these analyses did not alter the pattern of results reported below.

First, we conducted a 2 (image source, i.e., STM versus LTM) $\times$ 3 (complexity, i.e., number of filled cells) ANOVA, which did not reveal a two-way interaction, $F < 1$, but—as shown in Fig. 1—did reveal that participants took more time to generate the image from LTM, $F(1, 27) = 19.80$, $p < 0.001$, $\eta_p^2 = 0.42$. Post hoc pairwise comparisons showed that this was the case for all three levels of complexity, ps < 0.001. Image generation times were also affected by the number of filled cells in the visualized pattern (i.e., Pattern 1), $F(2, 54) = 18.15$, $p < 0.001$, $\eta_p^2 = 0.40$. In the LTM condition, participants required different amounts of time to generate images of Pattern 1 for the three levels of complexity, $F(2, 54) = 10.19$, $p < 0.01$, $\eta_p^2 = 0.27$. Post hoc pairwise comparisons were significant between two versus four-dot images (2,722 vs. 3,219 ms, $p < 0.001$), but not between four versus eight-dot images (3,219 vs. 3,442 ms, $p = 0.24$).

In the STM condition, participants required different amounts of time to generate images of Pattern 1 for the three levels of complexity, $F(2, 54) = 10.42$, $p < 0.001$, $\eta_p^2 = 0.28$. Specifically, participants took less time to generate images containing two dots than images containing four dots (1,975 vs. 2,458 ms, $p < 0.01$), but took comparable amounts of time for four-dot and eight-dot patterns (2,458 vs. 2,610 ms, $p = 0.24$). Critically, they were not faster when more dots were included, as predicted by the negative space hypothesis.

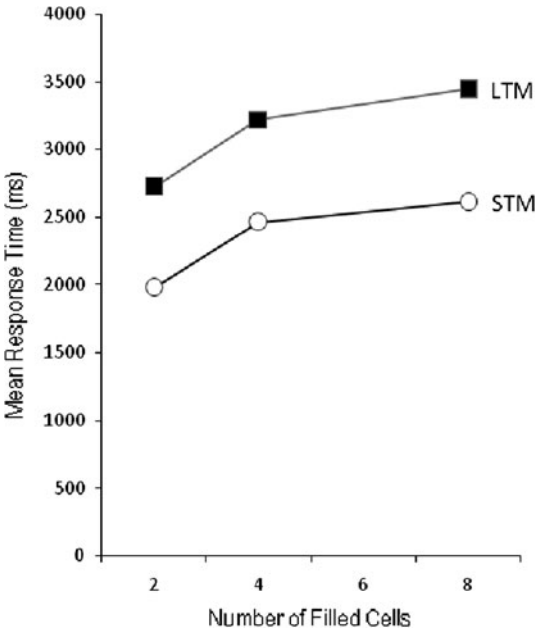

Fig. 1 Mean image generation times (ms) for Pattern 1 in Experiment 1 when the image was generated from long-term memory (*LTM*) or short-term memory (*STM*)

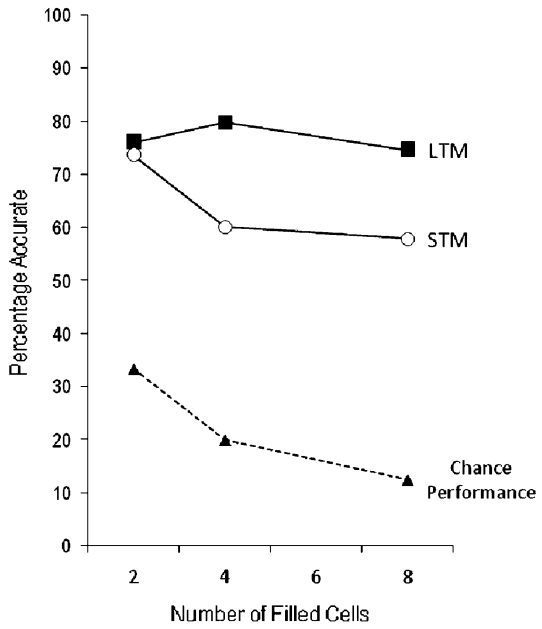

Fig. 2 Mean percentage accuracy in the integration task of Experiment 1 when the mental image was generated from long-term memory (*LTM*) or short-term memory (*STM*). "Chance Performance" represents the percentage accuracy expected at the most conservative criterion for a chance level of performance

Accuracy rates

Next, we carried out *t* tests to determine whether participants' accuracy exceeded what would be expected by chance alone. For these analyses, we adopted the most conservative estimate of chance level by assuming that participants could retain a perfect residual image of Pattern 2. If so, then participants would have a probability of selecting the correct cell by chance on 1 trial out of 3 (the number of empty cells in Pattern 2), 1 trial out of 5, or 1 trial out of 8, for each of the three levels of complexity. Thus, for the three levels of complexity, we set the chance level of accuracy, respectively, at 33.33, 20 or 12.5%. As shown in Fig. 2, in the LTM condition, participants were substantially more accurate than predicted by chance when Pattern 1 contained two dots ($M = 76\%$), $t(30) = 12.1$, $p < 0.001$, four dots ($M = 79.8\%$), $t(30) = 20.27$, $p < 0.001$, or eight dots ($M = 74.6\%$), $t(30) = 16.81$, $p < 0.001$. The same was true for the three levels of the STM condition, with accuracy greater than that predicted by chance when participants visualized patterns containing two dots ($M = 73.6\%$), $t(30) = 12.39$, $p < 0.001$, four dots ($M = 60.1\%$), $t(30) = 10.94$, $p < 0.001$, and eight dots ($M = 57.9\%$), $t(30) = 16.06$, $p < 0.001$ (see Fig. 2).

A 2 (image source, i.e., STM or LTM) × 3 (complexity, i.e., number of filled cells) repeated measures ANOVA revealed that image complexity affected accuracy differently for STM and LTM images, as witnessed by an interaction between the two factors, $F(2, 60) = 10.15$, $p < 0.001$, $\eta_p^2 = 0.25$. Although image complexity affected accuracy in the STM condition, $F(2, 60) = 18.12$, $p < 0.001$, $\eta_p^2 = 0.38$, it did not affect accuracy in the LTM condition, $F(2, 60) = 1.98$, $p = 0.15$, $\eta_p^2 = 0.06$. Post hoc pairwise comparisons between means using Tukey HSD test revealed that in the STM condition participants were more accurate when Pattern 1 contained two dots versus four dots ($p < 0.0001$), but not when Pattern 1 contained four dots versus eight dots ($p = 0.97$). However, as revealed by the lack of an interaction between image source and image complexity (when restricted to accuracy for four- and eight-dot Patterns), we also found that the difference in accuracy between the four- and eight-dot patterns was the same in the STM and LTM conditions, $F < 1$. Lastly, participants were more accurate in general in the LTM condition than in the STM condition, $F(1, 30) = 43.57$, $p < 0.001$, $\eta_p^2 = 0.59$. However, this difference occurred only when Pattern 1 contained four or eight dots ($p < 0.001$ in both cases), and not when it contained two dots ($p = 0.96$).

Decision response times

Finally, we analyzed how much time the participants required to locate the empty cell in the grid when Pattern 1 and Pattern 2 were combined. To limit measurement error, we only included in the following analyses participants with more than six correct responses, leading us to exclude 6 out of 31 participants. Again, we found that including

these participants did not alter the pattern of results reported below. We again conducted a two-way repeated measures ANOVA, with image source and image complexity as the within-participants factors.

As shown in Fig. 3, RTs were affected by image complexity, $F(2, 48) = 12.65$, $p < 0.001$, $\eta_p^2 = 0.35$, but not by the image source, $F < 1$. Critically, contradicting the prediction of the propositional theory, the two-way interaction did not even approach significance, $F(2, 48) = 1.06$, $p = 0.35$. In both the LTM and STM conditions, participants differed in the amount of time they took to locate the missing cell for the different levels of complexity, $F(2, 48) = 6.01$, $p < 0.01$, $\eta_p^2 = 0.20$ for the LTM condition, and $F(2, 48) = 7.57$, $p < 0.01$, $\eta_p^2 = 0.24$ for the STM condition. Post hoc pairwise comparisons revealed that in the LTM condition the participants required comparable amounts of time to evaluate the four-dot patterns (1,392 ms) versus the two-dot patterns ($M = 1{,}258$, $p = 0.47$) as well as for the four-dot patterns versus the eight-dot patterns ($M = 1{,}512$, $p = 0.60$). In the STM condition, participants were faster to locate the empty cell when Pattern 1 contained two dots ($M = 1{,}195$ ms) than when it contained four dots ($M = 1{,}457$ ms, $p < 0.025$), but required comparable amounts of time for the four-dot and eight-dot patterns ($M = 1{,}440$ ms, $p = 0.99$).

Discussion

The results suggest that people can in fact integrate mental images with percepts. The simple fact that participants could perform better than what would be expected by chance indicates that they could in fact integrate visual mental images and visual percepts. Moreover, if they had encoded negative space, they should have generated the images faster (or prepared for the upcoming stimulus faster) when more dots were present—but there was no hint of such a trend.

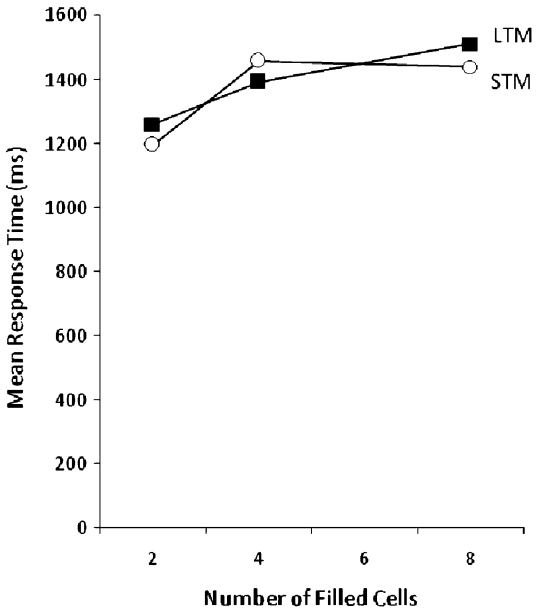

Fig. 3 Mean decision response times (ms) in Experiment 1 for locating the missing cell when Pattern 2 was combined with an image generated from long-term memory (*LTM*) or short-term memory (*STM*)

However, these findings do not guarantee that depictive images of Pattern 1 were used to perform the task. In principle, a description of the cells filled in Pattern 1 could have been used to perform the task. This theory generated three predictions: First, as predicted, participants were less accurate and required more time to retrieve Pattern 1 when more dots were included in the grid—but, counter to prediction, this was only true in the STM condition, not in the LTM condition. Second, this theory also led us to expect that the disparity in accuracy and decision response times between the STM and LTM conditions should be larger when more dots are included in the first pattern, because it should be increasingly difficult to formulate and store accurate descriptions of the more complex patterns in the STM condition. Instead, we found that the decrement in performance was the same for four versus eight-dot patterns. Also counter to prediction, the increase in decision times for more complex patterns was comparable for the STM and LTM conditions.

Although the propositional theory made predictions that were not confirmed, some of its predictions did bear fruit. However, we note that these results are easily accommodated within a depictive theory. Because of constraints on image maintenance (see Kosslyn, 1980, 1994), as the number of parts contained in each mental image increases, the images should become more difficult to maintain. This should make the integration process more difficult for images containing more parts, and be reflected by longer decision RTs and lower accuracy for more complex images.

We also investigated whether mental images generated from STM and those generated from LTM are equally precise in representing the spatial layout of a stimulus. When mental images were created from information stored in LTM, participants required more time to generate the mental images of the first grid, but were more accurate when locating the empty cell. Taken alone, these results might suggest a speed-accuracy trade-off. However, for two-dot patterns, participants were equally accurate in both STM and LTM conditions, although they generated the image more quickly in the STM condition. We note, however, that in the STM condition participants could have begun to generate the visual mental images of Pattern 1 before they were cued to generate their images. This may explain the differences in the generation times observed between the STM and LTM conditions.

Nevertheless, the fact that complexity affected RTs indicates that the participants did not fully generate the images prior to the cue in the STM condition.

In contrast, the time participants required to determine the location of the missing cell was not affected by whether the image was generated from LTM or STM. This is to be expected if image representations are processed equivalently once generated (Kosslyn, 1994)—but, as noted earlier—is inconsistent with the idea that a list of locations, and not a mental image, was used to perform the task. This result is also consistent with previous research on image-percept integration (Brockmole et al., 2002), which indicates that once the composite image (of percept and image) is created, each part of the overall representation is processed equivalently.

Finally, we must consider some unexpected trends in the data. We expected a smaller increase in image generation times from two- to four-dot patterns than from four- to eight-dot grids, if in fact an image of each dot needed to be generated individually. However, participants did not take significantly longer to generate images containing eight dots than images containing four dots. This finding may reflect the way we designed the task. In particular, because the ratio of empty versus filled cells decreased as the number of dots in each grid increased, there was a greater chance that the dots grouped into simple clusters as more cells were filled with dots. If this occurred, the similarity in generation times between four- and eight-dot patterns might reflect the fact that these patterns contained comparable numbers of perceptual units. Indeed, previous research has identified the number of perceptual units to be the primary factor in influencing image generation times (Kosslyn, et al., 1983). This could also account for the corresponding trend in RTs and accuracy rates.

We therefore conducted an additional experiment to follow-up the present findings, and to examine further the properties of image-percept integration.

Experiment 2

We demonstrated in Experiment 1 that participants could integrate information in visual mental images with visual percepts. However, the effect of pattern complexity did not exhibit the trend we predicted. This finding might indicate that participants organized the patterns into higher-order perceptual units, and hence the number of dots did not correspond to the number of perceptual units; if so, then it is possible that the participants organized Pattern 1 in different ways in the STM and LTM conditions, and this was responsible for the observed differences between the conditions.

In Experiment 2, we designed the Pattern 1 grids to ensure that participants would organize the dots into a particular number of perceptual units, and designed the grids so that the number of perceptual units increased with the number of dots. In addition, to ensure that the dot patterns did not become more crowded as the number of dots increased, we kept the ratio of filled versus empty cells constant across the three levels of complexity. If participants perform the task by constructing visual mental images of the Pattern 1 grids and integrating them with visual percepts of the second grids, then participants should require more time to generate the image when the first grid contains more perceptual units.

Method

Participants

A total of 32 volunteers from Harvard University and the local community participated in the experiment (16 females, 16 males). Their ages ranged from 18 to 35 years (mean = 23 years and 1 month). Twenty-nine were right-handed, and three were left-handed. Participants were recruited via the Harvard Study Pool website and were compensated $10 or credit for a course requirement. Prior to taking part in the study, all participants provided written consent and reported normal or corrected-to-normal vision. No participants had taken part in Experiment 1.

Materials

We designed three sets of eight patterns of dots for a total of 24 patterns to be visualized in the STM condition. To ensure that the participants created mental images of increasing complexity, perceptual units needed to increase with the number of dots in each pattern. Thus, each set contained two, four, or eight dots organized respectively in one, two, or three perceptual units (see Fig. 4). Perceptual units were defined by the Gestalt laws of similarity, proximity and continuity. Thus, two or more adjacent dots formed one perceptual unit. In order to keep the ratio of filled to empty cells as similar as possible between the three sets of dot patterns, two-dot patterns were presented on 4×4 grids, four-dot patterns were presented on 5×6 grids, and eight-dot patterns were presented on 8×8 grids. Although the number of cells increased as the number of dots increased in the grids, we ensured that all grids subtended the same visual angle ($18° \times 18°$) by adjusting the size of the cells. We designed three Pattern 1 grids with respectively one, two, and three perceptual units for the LTM condition. The visually presented grids had the same dimensions, only differing in the number of cells filled with dots (13, 25 and 55 dots respectively).

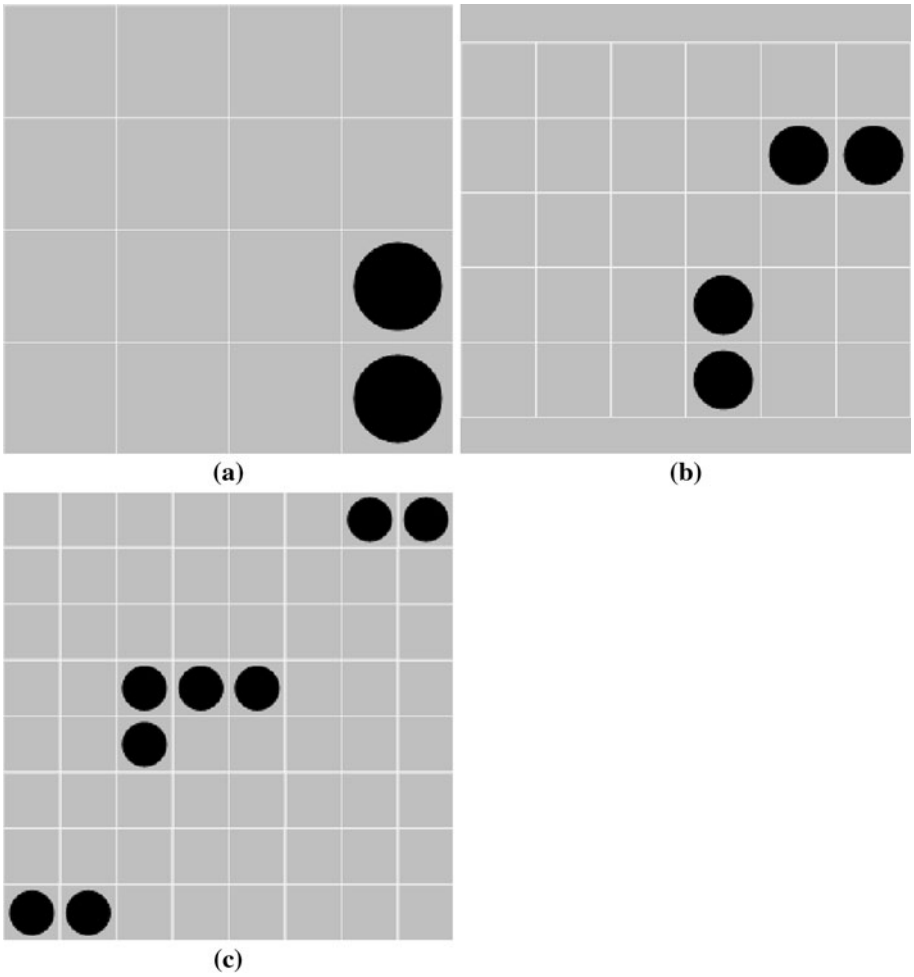

Fig. 4 Examples of **a** 2 dots/1 perceptual unit array, **b** 4 dots/2 perceptual units array and **c** 8 dots/3 perceptual units array

Procedure

On each trial in the STM condition, the procedure was the same as that for Experiment 1, with three exceptions: First, the Pattern 1 grid was now presented for 3, 6 or 9 s depending of number of perceptual units; we thus provided participants with enough time to encode more information. Second, the Pattern 2 grid was presented for 190 ms instead of 33 ms, to allow participants enough time to encode all the information in this grid; we limited the exposure to 190 ms in order to discourage eye movements. Third, given that the experiment was more difficult than Experiment 1, before each condition participants performed 24 practice trials with the same type of grids used in the experimental trials (4×4, 5×6, and 8×8). These patterns were not used in the experimental trials.

On each trial in the LTM condition, the cue for Pattern 1 was presented as in Experiment 1, and the second and third modifications in the STM condition, just noted, were also adopted here. In addition, the learning phase was modified: Participants still saw each pattern for 20 s, followed by drawing the pattern from memory, but now they repeated this procedure until they could draw each pattern correctly twice in a row. This technique ensured that the participants learned all patterns to the same level before they began the experimental trials.

Results

As in Experiment 1, in each condition we analyzed participants' accuracy for each level of perceptual units in order to determine whether participants could integrate a mental image and a percept into a single representation, We then analyzed the effect of the image source (STM vs. LTM) and of the complexity of the image (i.e., the number of perceptual units) on the accuracy rates, generation times, and decision RTs. These analyses allowed us to (1) determine whether image complexity had similar effects for representations created from information stored in LTM and STM, and (2) assess whether participants actually used mental images. If the data from all ANOVAs did not meet Mauchly's assumption of sphericity, degrees of freedom were corrected using the Greenhouse-Geisser epsilon.

Accuracy rates

As in Experiment 1, we adopted the most conservative estimate of chance level, with chance levels of performance set at 33.33, 20, or 11.11% for, respectively, 13, 25, and 55 cells filled with dots in the second grid. As shown in Fig. 5, in the LTM condition, participants performed significantly above chance level for two-dot/one perceptual unit grids, $M = 94\%$, $t(31) = 34.83$, $p < 0.001$, for four-dot/two perceptual units grids, $M = 72.7\%$, $t(31) = 17.91$, $p < 0.001$, and for eight-dot/three perceptual units grids, $M = 52.2\%$, $t(31) = 11.13$, $p < 0.001$. The same was true in the STM condition, with grids containing two dots, $M = 95.7\%$, $t(31) = 58.51$, $p < 0.001$, four dots, $M = 81.1\%$, $t(31) = 25.38$, $p < 0.001$, and eight dots, $M = 34.8\%$, $t(31) = 7.58$, $p < 0.001$.

A two-way 2 (image source, i.e., STM versus LTM) × 3 (complexity, i.e., number of perceptual units) repeated measures ANOVA revealed that participants were no more accurate in the STM condition than in the LTM condition, respectively 70.5 vs. 72.9%, $F(1, 31) = 1.9$, $p = 0.18$, but that they became less accurate when the number of perceptual units in the Pattern 1 increased, $F(1.46, 45.12) = 229.47$, $p < 0.001$, $\eta_p^2 = 0.88$. Two separate one-way ANOVAs revealed that the number of perceptual units affected the performance in the STM condition, $F(1.67, 51.68) = 302.92$, $p < 0.001$, $\eta_p^2 = 0.91$, and the LTM condition, $F(1.64, 50.75) = 89.33$, $p < 0.001$, $\eta_p^2 = 0.74$. In both conditions, post hoc pairwise comparisons computed with a Tukey HSD test revealed that participants were more accurate for grids containing one perceptual unit than for grids containing two perceptual units, and were more accurate for grids containing two perceptual units than for grids containing three perceptual units, all ps < 0.001. In addition, there was an interaction between image source and the number of perceptual units, $F(2, 62) = 38.82$, $p < 0.001$, $\eta_p^2 = 0.56$. Post hoc comparisons revealed that participants were comparably accurate in the STM and LTM conditions when Pattern 1 had one perceptual unit ($p = 0.96$), but were less accurate in the LTM than in the STM condition when Pattern 1 had two perceptual units ($p < 0.005$); the reverse was true when Pattern 1 had three perceptual units ($p < 0.001$).

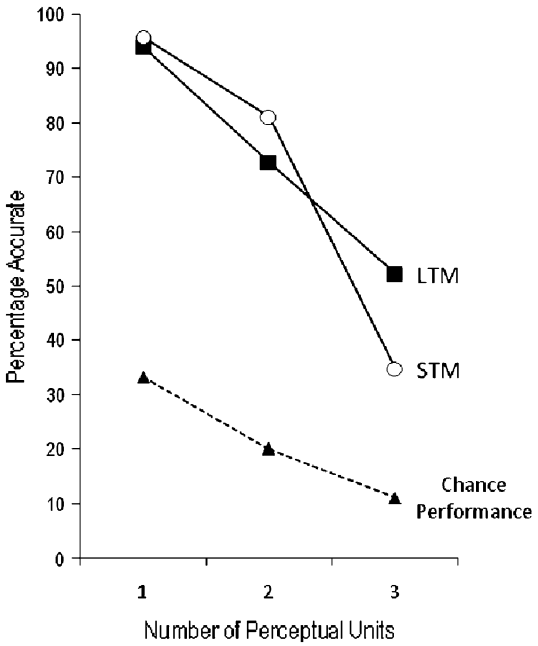

Fig. 5 Mean percentage accuracy in the integration task of Experiment 2 when the image was generated from long-term memory (*LTM*) or short-term memory (*STM*). Chance performance represents the percentage accuracy expected at the most conservative criterion for a chance level of performance

Image generation times

We analyzed image generation data from both correct and incorrect trials (i.e., trials on which participants subsequently made a correct or incorrect response), and defined outliers using the same criterion as Experiment 1. Consequently, we removed data from 4 out of 32 participants from the following analyses; we note that follow-up analyses revealed that, when included, these participants did not alter the pattern of results.

First, we conducted a 2 (image source, i.e., STM versus LTM) × 3 (complexity, i.e., number of perceptual units) repeated measures ANOVA, which revealed that both image source and the number of perceptual units affected the image generation times (see Fig. 6): participants required more time in the LTM condition than in the STM condition, $F(1, 27) = 7.47$, $p < 0.05$, $\eta_p^2 = 0.22$, and required more time for different numbers of perceptual units, $F(1.21, 32.6) = 60.67$, $p < 0.001$, $\eta_p^2 = 0.69$; participants required less time to generate the image when the grid contained one perceptual unit (for the LTM and STM conditions, respectively, $M = 1{,}567$ and $M = 1{,}696$ ms) than when it contained two perceptual units ($M = 3{,}017$ and $M = 2{,}413$ ms), and when it contained two perceptual units than when it contained three perceptual units ($M = 4{,}408$ and $M = 3{,}014$ ms), ps < 0.01 for all pairwise comparisons using post hoc Tukey HSD test in each condition. The effect of perceptual units was significant in both conditions, with $F(1.33, 35.83) = 65.16$, $p < 0.001$, $\eta_p^2 = 0.71$ for the LTM condition, and $F(1.27, 34.28) = 16.37$, $p < 0.001$, $\eta_p^2 = 0.38$ for the STM condition. Finally, there was a significant interaction between the number of perceptual units and the image source, $F(1.27, 34.32) = 13.27$, $p < 0.001$, $\eta_p^2 = 0.33$, which indicated that participants required a smaller increment of time for each additional perceptual unit in the STM

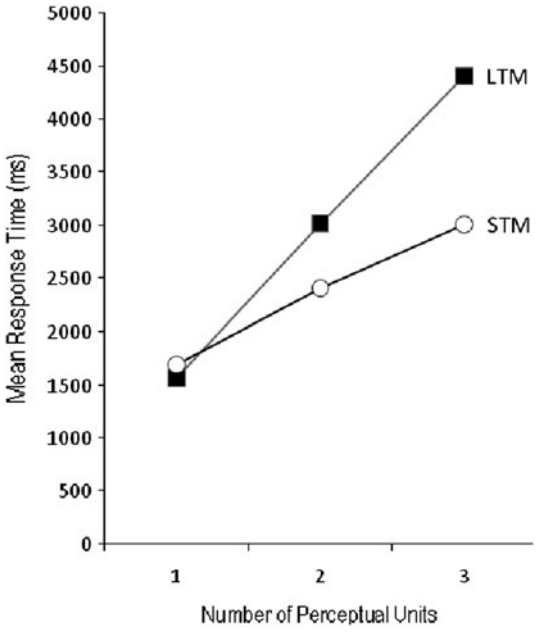

Fig. 6 Mean image generation times (ms) for Pattern 1 in Experiment 2 when the image was generated from long-term memory (*LTM*) or short-term memory (*STM*)

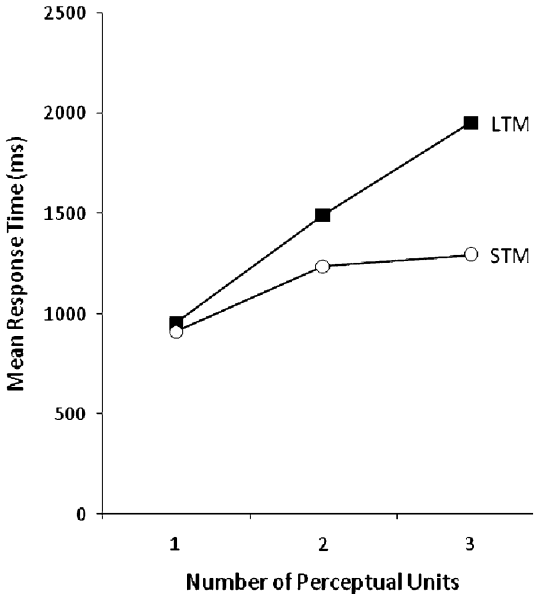

Fig. 7 Mean decision response times (ms) in Experiment 2 for locating the missing cell when Pattern 2 was combined with an image generated from long-term memory (*LTM*) or short-term memory (*STM*)

condition than in the LTM condition; this effect indicates that at least some different processes were drawn upon in the two conditions.

Decision response times

We restricted our analyses of decision RTs to participants who provided at least four correct responses (out of 16) for each level of perceptual units of Pattern 1. Eight out of 32 participants were removed from the following analyses. However, we note that including these participants in the analyses did not alter the pattern of results from those reported below.

As shown in Fig. 7, RTs were affected by image source, $F(1, 23) = 12.60$, $p < 0.005$, $\eta_p^2 = 0.35$, and the number of perceptual units, $F(2, 46) = 45.61$, $p < 0.001$, $\eta_p^2 = 0.67$. The effect of the number of perceptual units on the RTs differed in the two conditions, as indicated by an interaction between source and the number of perceptual units, $F(2, 46) = 12.01$, $p < 0.001$, $\eta_p^2 = 0.34$.

One-way ANOVAs on data from the LTM and STM conditions revealed that the number of perceptual units affected RTs in both conditions, with $F(2, 46) = 31.05$, $p < 0.001$, $\eta_p^2 = 0.57$ for LTM and $F(2, 46) = 31.33$, $p < 0.001$, $\eta_p^2 = 0.58$ for STM. In the LTM condition, participants required 951, 1,493 and 1,951 ms for one, two, and three perceptual units in Pattern 1, respectively (all ps < 0.001 as revealed by Tukey HSD tests). Pairwise comparisons in the STM condition revealed a different pattern of results, with longer RTs for images containing two perceptual units than one perceptual unit (1,236 vs. 908 ms, $p < 0.001$), but no difference between three and two perceptual units ($M = 1{,}291$ ms, $p = 0.99$).

Discussion

Consistent with the results from Experiment 1, we found that participants were able to perform the integration task above a chance level of accuracy. This was true even for the most complex grids, comprising 64 cells. These results provide good evidence that participants can integrate even highly complex visual mental images with visual percepts. In addition, we found that when we removed the confound of perceptual grouping (by maintaining a constant ratio of filled to empty cells and by ensuring that the number of perceptual units increased systematically as the number of dots increased), participants required more time to generate images of more complex patterns. Given that image generation involves a sequential retrieval of encoded parts, the increase in generation time with increasing number of perceptual units supports the inference that participants used mental images of the Pattern 1 grids and integrated these with the visual percepts of the Pattern 2 grids.

However, in the STM condition, we found that participants took less time and were less accurate when locating the empty cell for grids containing three perceptual units, which suggest a speed-accuracy tradeoff. This result may

have occurred because complex visual stimuli fade from STM at a faster rate than simpler visual stimuli (Philips, 1974), and hence participants may have rushed to respond before the image faded from STM, which thereby led them to respond quickly and commit more errors.

Finally, our results suggest that the source of the image affects the efficiency of image-percept integration. We found that when the image of Pattern 1 contained two perceptual units, images generated from STM yielded more accurate image-percept integration than those generated from LTM. However, when the Pattern 1 grid had more than two perceptual units, mental images generated from STM were less precise than mental images generated on the basis of information stored in LTM. For reasons outlined earlier, in these instances participants' performance may have been compromised by the limits of visual STM. Another explanation for this effect may be that participants needed to memorize finer metric information about the locations of each perceptual unit as the number of perceptual units increased, and this was more difficult to encode in STM.

General discussion

The results of both experiments reported here clearly demonstrate that participants' mental images preserve structural information of the pattern they represent, and can be integrated with percepts to create a single composite representation. These two characteristics imply that visual mental images rely on a depictive format to represent information. Moreover, we found evidence against the idea that participants described the grid of Pattern 1 and used this description to search for the empty cell in Pattern 2. These results are consistent with those from previous studies indicating that mental images preserve metric information of the scene they represent (e.g., Anderson & Helstrup, 1993; Finke & Slayton, 1988; Kosslyn, et al., 1978; Thompson, Kosslyn, Hoffman, & Van der Kooij, 2008).

To confirm that mental images are combined with visual percepts, it was necessary to demonstrate that participants generated mental images of the Pattern 1 grids. In order to garner such evidence, we relied on one hallmark of imagery: Images of more complex patterns require more time to generate (for reviews, see Kosslyn, 1980, 1994; Kosslyn et al., 2006). Thus, we manipulated the number of dots in the grid participants visualized. We found this trend in Experiment 1, but only when the complexity of the first grid was increased from two to four dots. In Experiment 2, after we controlled for perceptual grouping of the dots, we found that participants required more time to generate mental images with more perceptual units. Taken together, these results provide good evidence that participants did generate mental images of the first grid.

In addition, our paradigm allowed us to compare mental images that were generated on the basis of information stored in LTM to images generated on the basis of information stored in STM. The results indicate that when mental images are based on a just-seen grid (STM), and these images contain one or two perceptual units (such as those used in Experiment 2), they tend to be more accurately and more efficiently integrated with visual percepts than are mental images based on information stored in LTM; this finding is consistent with studies of image scanning (see Borst & Kosslyn, 2008). However, when participants visualized images containing more than two perceptual units, we found that this effect reversed. In these instances, the capacity of visual STM may have limited the precision of participants' images.

This new evidence that images rely on depictive representations has clear implications for the debate about the nature of mental representations (see Kosslyn, et al., 2006). In particular, our results bear on theories that posit an exclusively propositional account of mental imagery (Pylyshyn, 1973, 2002, 2003). The present results are difficult to explain if all stored information relies on a descriptive, symbolic mode of representation. Rather, these results are consistent with previous evidence that both the visual system and imagery system rely, at least in part, on representations in a depictive format (Craver-Lemley, Arterberry, & Reeves, 1999; Kosslyn & Thompson, 2003).

Acknowledgments This research was supported by Grant R01 MH060734 from the National Institute of Mental Health. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Institute of Mental Health. We wish to thank Ned Block for inspiring this research by bringing the earlier work, and the possible problems with it, to our attention.

References

Anderson, R. E., & Helstrup, T. (1993). Visual discovery in mind and on paper. *Memory and Cognition, 21*, 283–293.

Borst, G., & Kosslyn, S. M. (2008). Visual mental imagery and visual perception: Structural equivalence revealed by scanning processes. *Memory and Cognition, 36*, 849–862.

Brockmole, J. R., Irwin, D. E., & Wang, R. F. (2003). The locus of spatial attention during the temporal integration of visual memories and percepts. *Psychonomic Bulletin & Review, 10*, 510–515.

Brockmole, J. R., Wang, R. F., & Irwin, D. E. (2002). Temporal integration between visual images and visual percepts. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 315–334.

Craver-Lemley, C., Arterberry, M. E., & Reeves, A. (1999). "Illusory" illusory conjunctions: The conjoining of visual and imagined stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1036–1049.

Di Lollo, V. (1980). Temporal integration in visual memory. *Journal of Experimental Psychology: General, 109*, 75–97.

Eriksen, C. W., & Collins, J. F. (1967). Some temporal characteristics of visual pattern perception. *Journal of Experimental Psychology, 74*, 476–484.

Finke, R. A., & Slayton, K. (1988). Explorations of creative visual synthesis in mental imagery. *Memory and Cognition, 16*, 252–257.

Hollingworth, A., Hyun, J., & Zhang, W. (2005). The role of visual short-term memory in empty cell localization. *Perception & Psychophysics, 67*, 1332–1343.

Jiang, Y. (2004). Time window from visual images to visual short-term memory: Consolidation or integration? *Experimental Psychology, 51*, 45–51.

Jiang, Y., & Kumar, A. (2004). Visual short-term memory for two sequential arrays: One integrated representation or two separate representations? *Psychonomic Bulletin & Review, 11*, 495–500.

Jiang, Y., Kumar, A., & Vickery, T. J. (2005). Integrating visual arrays in visual-short term memory. *Experimental Psychology, 52*, 39–46.

Kosslyn, S. M. (1975). Information representation in visual images. *Cognitive Psychology, 7*, 341–370.

Kosslyn, S. M. (1980). *Image and mind*. Cambridge: Harvard University Press.

Kosslyn, S. M. (1994). *Image and brain*. Cambridge: MIT Press.

Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 47–60.

Kosslyn, S. M., Cave, C. B., Provost, D., & Von Gierke, S. (1988). Sequential processes in image generation. *Cognitive Psychology, 20*, 319–343.

Kosslyn, S. M., Reiser, M. J., Farah, M. J., & Fliegal, S. L. (1983). Generating visual images: Units and relations. *Journal of Experimental Psychology: General, 112*, 278–303.

Kosslyn, S. M., & Thompson, W. L. (2003). When is early visual cortex activated during visual mental imagery? *Psychological Bulletin, 129*, 723–746.

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. New York: Oxford University Press.

Loftus, G. R., & Irwin, D. E. (1998). On the relations among different measures of visible and informational persistence. *Cognitive Psychology, 35*, 135–199.

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

Pylyshyn, Z. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin, 80*, 1–24.

Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioural and Brain Sciences, 25*, 157–238.

Pylyshyn, Z. (2003). Return of the mental image: Are there pictures in the brain? *Trends in Cognitive Sciences, 7*, 113–118.

Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, B. R., et al. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science, 268*, 889–893.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 701–703.

Thompson, W. L., Kosslyn, S. M., Hoffman, M. S., & Van der Kooij, K. (2008). Inspecting visual mental images: Can people "see" implicit properties as easily in imagery and perception? *Memory & Cognition, 36*, 1024–1032.