Atten Percept Psychophys
DOI 10.3758/s13414-013-0601-3

Working memory for relations among objects

Pamela E. Clevenger · John E. Hummel

Abstract Across many areas of study in cognition, the capacity of working memory (WM) is widely agreed to be roughly three to five items: three to five *objects* (i.e., bound collections of object features) in the literature on visual WM or three to five *role bindings* (i.e., objects in specific relational roles) in the literature on memory and reasoning. Three experiments investigated the capacity of observers' WM for the spatial relations among objects in a visual display, and the results suggest that the "items" in WM are neither simply objects nor simply role bindings. The results of Experiment 1 are most consistent with a model that treats an "item" in visual WM as an object, along with the roles of all its relations to one other object. Experiment 2 compared observers' WM for object size with their memory for relative size and provided evidence that observers compute and store objects' relations per se (rather than just absolute size) in WM. Experiment 3 tested and confirmed several more nuanced predictions of the model supported by Experiment 1. Together, these findings suggest that objects are stored in visual WM in pairs (along with all the relations between the objects in a pair) and that, from the perspective of WM, a given object in one pair is not the same "item" as that same object in a different pair.

Keywords Visual relations · Spatial relations · Visual working memory · Objects · Role bindings

Working memory (WM) is the cognitive resource responsible for the active maintenance and manipulation of information. The capacity of WM, which is sharply limited, determines how much information one can maintain and manipulate—that is, how much one can perceive or think about—in parallel. As such, the capacity of WM is a key bottleneck in perception (Simons & Ambinder, 2005; Simons & Chabris, 1999), attention (Treisman, 1998; Treisman & Gelade, 1980; Wolfe, 1994), memory (Baddeley & Hitch, 1975; Cowan, 2001), reasoning (Halford, Baker, McCredden, & Bain, 2005; Hummel & Holyoak, 1997; Morrison, Holyoak, & Truong, 2001), and virtually everything else we do. Individual differences in the capacity of WM have also been linked to various performance measures, such as fluid intelligence (e.g., Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Gray, Chabris, & Braver, 2003; Heitz, Unsworth, & Engle, 2005; Kane & Engle, 2002).

A large body of converging evidence from the literatures on both visual cognition and "higher" cognition suggests that, for most people and across a variety of tasks, the capacity of WM is roughly three to five items. This is not to say that there is necessarily only a single system for WM, but only that the various systems have strikingly similar capacity limits (Avons, Wright, & Pammer, 1994; Baddeley, Thomson, & Buchanan, 1975; Cowan, Wood, Nugent, & Treisman, 1997; Halford et al., 2005; Hitch, Burgess, Towse, & Culpin, 1996; Levy, 1971; Longoni, Richardson, & Aiello, 1993; Luck & Vogel, 1997; Murray, 1968; Peterson & Johnson, 1971; Song & Jiang, 2006; Woodman & Vogel, 2008; for reviews, see Baddeley, 2003; Cowan, 2001). More recent evidence suggests that visual WM may also be characterized in terms of the amount and/or quality of information stored about the items in WM, rather than strictly in terms of the number of those items (e.g., Alvarez & Cavanagh, 2004; Alvarez & Oliva, 2009; Oliva & Torralba, 2006). The work presented here does not speak to the distinction between these discrete (i.e., item-based) and continuous (i.e., information-/quality-based) models of WM. Instead, we aim to elucidate the currency and capacity of WM for the spatial relations among objects in a display. This capacity has important implications for our ability to interpret spatial layouts, to reason about objects

P. E. Clevenger (✉) · J. E. Hummel
Department of Psychology, University of Illinois at Urbana-Champaign, Urbana, IL, USA
e-mail: glosson2@illinois.edu

Published online: 31 December 2013

arrayed in scenes, to interpret graphs, and to make basic judgments about individual pairs of objects (e.g., whether one can fit on top of or inside another) and even for our ability to visually recognize individual, multipart objects (e.g., Biederman, 1987; Hummel, 2001; Hummel & Biederman, 1992). More important for our present purposes, understanding the capacity of WM for the spatial relations among objects can lend insight into the currency of WM more generally: What *are* those "items" of which we can hold three to five in WM?

According to a great deal of literature on visual cognition, the "items" that occupy slots in WM (e.g., Luck & Vogel, 1997) or otherwise consume the finite resources of WM (e.g., Alvarez & Cavanaugh, 2004) are discussed as *objects*—that is, bound collections of object features. For example, using a change detection paradigm, Luck and Vogel (1997), Experiment 3 varied both the number of objects in the array (two, four, or six) and the complexity of those objects (i.e., whether the objects were defined by a single feature or a conjunction of two features). They found that subjects' accuracy in detecting a change between two displays decreased with the number of objects in the displays, but not with the number of features defining each object. They concluded that the capacity of WM is about four objects, regardless of the number of features per object. (More recently, Vogel, Woodman, & Luck, 2001, have revised their estimate down to 2.8 items.) The basic idea is that once an object occupies a slot in WM, additional features on that object can all fit within the same slot with little or no additional cost.

In the literature on memory and reasoning, the capacity of WM is discussed most often in terms of *role–filler bindings* (see Cowan, 2001; Halford et al., 2005; Hummel & Holyoak, 2003). A role–filler binding (or, equivalently, *variable–value binding*) is a binding of one relational role (or variable) to its argument (filler). For example, the proposition *loves* (John, Mary) consists of two such bindings: John bound to *lover* and Mary to *beloved*. (The proposition *loves* [Mary, John] specifies the same roles and arguments, but the bindings are reversed.) The expression 2 = 6/3 consists of three such bindings: 2 to *result*, 6 to *numerator*, and 3 to *denominator*. The consensus view in this literature is that the capacity of WM is roughly 4 ± 1 such bindings (see Cowan, 2001; Morrison, 2005).

For example, Halford et al. (2005) had subjects view graphical representations of two-, three-, and four-way interactions, which entail two, three, and four variable–value bindings, respectively. The subjects' task was to indicate whether "greater" or "smaller" would correctly complete the final sentence of a verbal description of each graph. For example, in the three-way problem shown in Fig. 1, the verbal description was the following: "People prefer fresh cakes to frozen cakes. The difference depends on the flavor (chocolate vs. carrot) and the type (iced vs. plain). The difference between fresh and frozen increases from chocolate cakes to carrot cakes. This increase is (greater/smaller) for iced cakes than for plain cakes" (Halford et al., 2005, p. 71). Halford et al. (2005) found that subjects' error rates increased with the order of the interaction. In Experiment 2, when the researchers increased the number of variables to five, subjects performed at chance. In broad agreement with Luck and Vogel (1997), these researchers concluded that the capacity of WM is roughly four variable–value bindings. Numerous other experiments have converged on a similar estimate for the capacity of WM (see Cowan, 2001, for a review; but see Vogel et al., 2001).

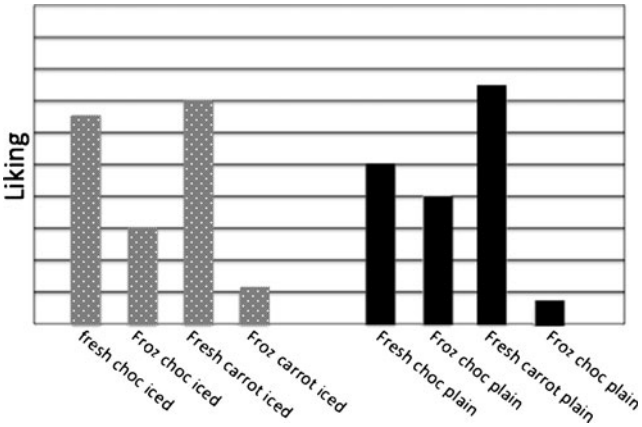

Fig. 1 An example of the stimuli from Halford, Baker, McCredden, and Bain's (2005) experiments on working memory for variables. Subjects were presented with a graph as shown here with the following verbal description: "People prefer fresh cakes to frozen cakes. The difference depends on the flavor (chocolate vs. carrot) and the type (iced vs. plain). The difference between fresh and frozen increases from chocolate cakes to carrot cakes. This increase is (greater/smaller) for iced cakes than for plain cakes." Each subject's task was to indicate whether "greater" or "smaller" was the correct answer. (Here, the correct answer is "smaller")

Although there is comparatively broad agreement that the capacity of WM is roughly three to five items (but see Alvarez & Cavanagh, 2004; Alvarez & Oliva, 2009; Oliva & Torralba, 2006; Vogel et al., 2001), what remains less clear is precisely what these "items" are. As was noted previously, the literature on visual WM tends to discuss items as objects (i.e., bound collections of object features), whereas the cognitive literature tends to define the items as variable–value (role–filler) bindings. That these two conceptions are not necessarily consistent becomes clear when one asks about the capacity of visual WM for the spatial relations among objects in a display. It is known that perceiving spatial relations requires visual attention (Logan, 1994) and, therefore, consumes WM resources.

Consider, for example, a four-object display and imagine that an observer is tasked with remembering three spatial relations among the objects in the display (e.g., for any given pair of objects, which is *larger* than the other, which is *above* the other, and which is *right of* the other; see Fig. 2). According to the items-as-objects account, the WM load imposed by this task is simply the number of objects in the display. With a capacity of 4 ± 1 objects, it should be easy to perceive all the relations among the four objects in Fig. 2. But

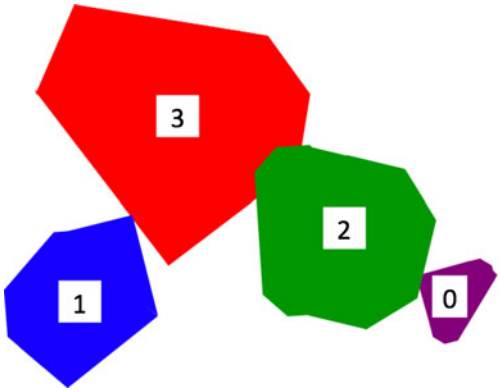

Fig. 2 A display of four objects (0–3). Encoding three relations (e.g., *larger*, *above*, and *left/right*) between each pair of objects would entail encoding 36 role bindings: *larger* (2, 1), *below* (1, 2) . . . *larger* (2, 0) . . . *left of* (3, 2) . . . and so forth.

if the items that occupy WM are role–filler bindings, as suggested by the items-as-role-bindings account, then the load imposed by three relations among each of four objects should scale as $r(n^2 - n)$, where r is the number of relations in the vocabulary (e.g., three in the present example) and n is the number of objects to be related. Three relations among four objects comes to 36 role bindings, vastly exceeding the 4 ± 1 capacity of WM. These accounts thus make very different predictions about our ability to encode in WM the spatial relations among objects in a display.

Both the items-as-objects and the items-as-role-bindings accounts have important limitations as accounts of the representation of the spatial relations among visual objects. According to the simplest version of the items-as-objects account, a relational role is just another feature on an object. For example, this approach would represent the fact that object 2 in Fig. 2 is larger than object 1 by including *larger* among the features of 2 and *smaller* among those of 1 (see, e.g., Hummel & Biederman, 1992; for a related proposal, see Franconeri, Scimeca, Roth, Helseth, & Kahn, 2012). This approach works well as long as there are only two objects in the display, but it fails catastrophically with three or more objects. For example, object 2 is larger than both 0 and 1 (and would therefore have the feature *larger* bound to its representation) and smaller than object 3 (and would therefore also have *smaller* bound to its representation). The resulting representation would specify that 2 is both larger and smaller than something, but it would fail to specify what, specifically, it is larger or smaller than. (Hummel & Biederman, 1992, observed that this property of their JIM model of object recognition constituted a novel prediction—one that was soon falsified by Logan & Compton, 1996.) This prediction of the strict items-as-objects account of visual WM seems absurd on its face, but it is at least logically possible that it is empirically true of human visual WM (modulo; Logan & Compton, 1996); and, more important, it really is a prediction of the account.

Role–filler based representations of the kind postulated by the items-as-bindings account (including both traditional symbolic representations [e.g., Anderson, Matessa, & Lebiere, 1997; Falkenhainer, Forbus, & Gentner, 1989] and symbolic-connectionist representations [e.g., Doumas, Hummel, & Sandhofer, 2008; Hummel & Holyoak, 1997, 2003]) do not suffer from this kind of ambiguity, because they represent relations as full-blown propositions: The propositions *larger* (1, 0), *larger* (2, 1), and *larger* (3, 2) explicitly specify which objects are larger than which. However, the disadvantage of this kind of representation is that every role binding requires its own slot in WM. For example, representing that 2 is both larger than and above 1 requires the propositions *larger* (2, 1) and *above* (2, 1): a total of four role bindings—four slots in WM—to represent just two relations between two objects.

A hybrid account of the currency of WM for spatial relations

An alternative way to represent that 2 is both above and larger than 1 would be to "stack" pairs of relational roles onto pairs of objects, in much the same way as the items-as-objects approach "stacks" visual features into representations of complete objects. For example, rather than explicitly coding the separate propositions *larger* (2, 1) and *above* (2, 1), encode instead the mixed relation *above-and-larger* (2, 1). The resulting "stacked" representation would consume two, rather than four, slots in WM but would leave the question of what the object was larger than or what it was above unambiguous. This kind of parallel encoding of multiple relations among a single pair of objects is also broadly consistent with the massively parallel processing of early and intermediate vision (see Hummel & Biederman, 1992): As long as one is calculating and encoding the fact that 2 is larger than 1, one may as well bring the fact that 2 is also above 1 along for the ride.

According to this hybrid account, the load imposed by representing r relations among n objects would be simply $(n^2 - n)$. Under this approach, each object is encoded in relation to one other object—eliminating the ambiguity of the items-as-objects approach—but with all the roles describing all the relations between two objects "stacked" into the two WM "slots" occupied by those objects. Thus, WM capacity for the relations among objects would be limited by the number of pairs of objects to be stored, regardless of the number of relations between those pairs.

Experiment 1 was designed to test which of these three accounts best characterizes the capacity of human WM for the spatial relations among objects. On each trial, the observer viewed a display of two, three, or four random polygons differing in color, size, and location in the display (see Fig. 3), followed by a pattern mask. After the mask, a query

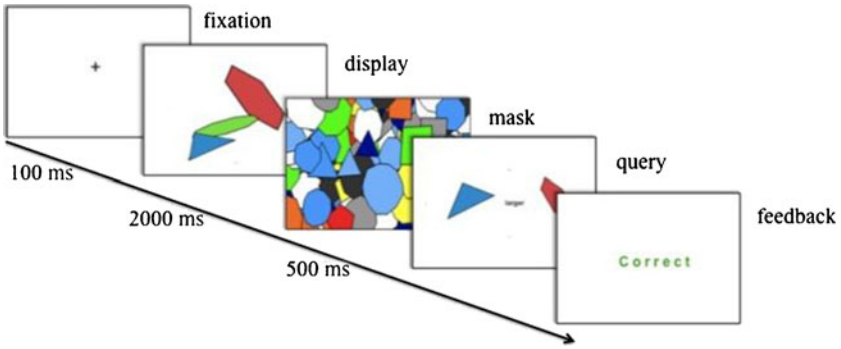

Fig. 3 An example of a trial from Experiment 1. A fixation cross was presented for 500 ms, followed by a two-, three-, or four-object display for 2,000 ms. A pattern mask appeared for 500 ms, followed by a query, which remained on the screen until the subject responded with a keypress. The response was followed by accuracy feedback

appeared consisting of two objects from the display, side-by-side and of the same size, with a relation word ("larger," "above," or "right-of") between them. The observer's task was to indicate which of the two objects had been larger than the other, above the other, or to the right of the other in the previous display. On any given trial, some subjects were always asked about the same relation, others were asked about either of two relations, and still others were asked about any of three relations.

Each of the accounts reviewed above makes a unique set of predictions about how performance on this task should vary as a function of the number of relations to be remembered (varied between subjects) and of the number of objects in the original display (varied within subjects). The probability, $p(\mathrm{c})$, of responding correctly to any given query is $p(\mathrm{r})$, the probability that the queried relation will be remembered (i.e., in WM), plus the probability of guessing correctly (.5 in the case of our two-alternative forced choice task) times the probability that the item will not have been remembered:

$$p(\mathrm{c}) = p(\mathrm{r}) + 0.5 * [1 - p(\mathrm{r})]. \quad (1)$$

The probability, $p(\mathrm{r})$, of remembering any given relation from the original display is the probability that the queried relation will get into a WM "slot", $p(\mathrm{s})$, times $p(\mathrm{r}|\mathrm{s})$, the probability that it will remain in that slot until queried:

$$p(\mathrm{r}) = p(\mathrm{s})p\left(\mathrm{r}\middle|\mathrm{s}\right). \quad (2)$$

The probability of getting into a slot, $p(\mathrm{s})$, is simply the number of WM slots, k, divided by the load, l (i.e., the number of items vying to occupy those slots), truncated above 1:

$$p(\mathrm{s}) = \left\lfloor \frac{k}{l} \right\rfloor^{1}. \quad (3)$$

These equations describe the predictions of all three models of WM capacity described above. The models differ only in how they predict that load, l in Eq. 3, will scale with the number of relations to be remembered and the number of objects entering those relations. Recall that the items-as-objects model predicts that load, l, is simply n, the number of objects; the items-as-bindings model predicts that load is equal to the number of role bindings: $l = r(n^2 - n)$; and the hybrid model predicts that it is simply the number of unique object pairs, where ordering within the pair matters [i.e., $r(a, b) \neq r(b, a)$]: $l = n^2 - n$. Figure 4 summarizes the predicted loads, l, and probabilities $p(\mathrm{c})$ of responding correctly for each model at each of three values of r and n.

Experiment 1

Experiment 1 was designed as a preliminary test of these predictions. Observers performed a task identical to that illustrated in Fig. 3.

Method

Subjects

Sixty-three University of Illinois undergraduate students earned course credit for participating in the experiment. Each was randomly assigned to one of three conditions: one, two, or three relations to remember on every trial. In the one- and two-relation conditions, assignment of which relations were to be remembered was counterbalanced.

Stimuli

Subjects viewed displays on a 24-inch Apple iMac with 1920 x 1200 resolution. Displays consisted of two, three, or four irregular polygons, which were organized to be categorically distinguishable from one another in terms of the spatial relations among them. Each polygon could be described as below or above, left or right of, and larger or smaller than any other polygon on the screen. Polygons were semirandomly colored (subject to the constraint that no two polygons be similarly colored) in order to facilitate subjects' memory for them.

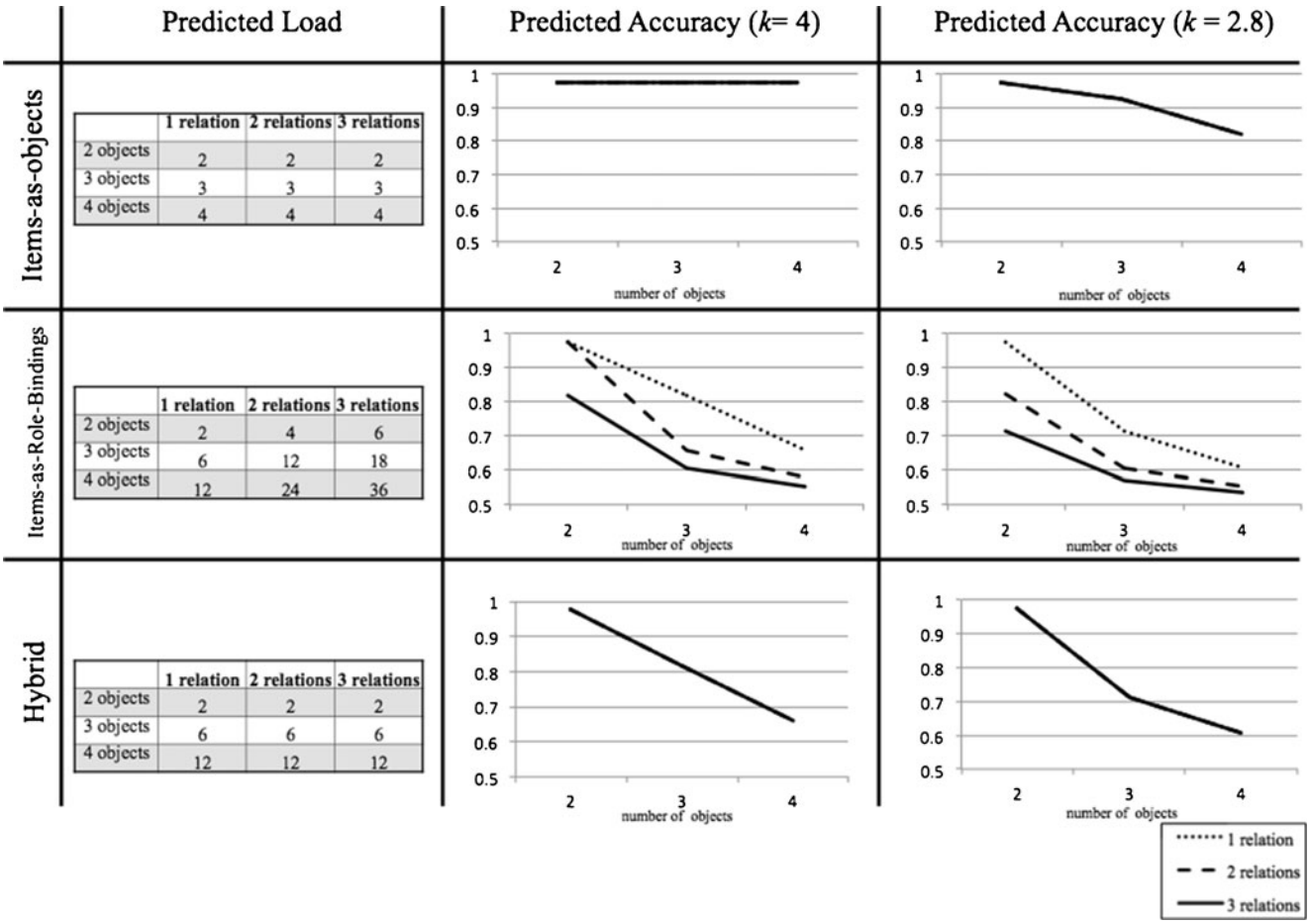

Fig. 4 Load and accuracy predictions of the items-as-objects, items-as-bindings, and hybrid accounts. The items-as-objects account predicts that performance will vary with the number of objects but not the number of relations, at least for a working memory capacity, *k*, of <4. The items-as-bindings account predicts that accuracy will vary with the number of pairs of objects times the number of relations. The hybrid account predicts that accuracy will vary with the number of pairs of objects, regardless of the number of relations

The query images displayed a pair of objects chosen semirandomly from the original display (i.e., subject to the constraint that equal numbers of objects were chosen that either had or had not been touching in the prior display). The objects appeared on the screen side-by-side and normalized to be the same size but otherwise maintained their original shape, color, and orientation. The word "larger," "above," or "right-of" was displayed between the objects in blue and served (in the two- and three-relation conditions) to tell the subject which relation to report.

Design and procedure

Each trial was structured as follows: After the 500-ms presentation of a fixation cross, a two-, three-, or four-object display appeared on the screen for 2,000 ms, followed by a pattern mask of randomly generated and randomly colored polygons. Subjects were then shown the query display consisting of two objects and a relation word. The subjects' task was to indicate, with a keypress, which object had stood in that relation to the other in the original display. The number of objects in the original display (two, three, or four) varied within subjects, but the number of relations to be queried (one, two, or three) varied between subjects.

In the one-relation condition, each subject was informed of the relation to which he or she was to pay attention and was asked about that same relation on every trial of the experiment. Assignment of relations to subjects was counterbalanced. Subjects in the two-relation condition could be queried about either of two relations on any given trial (counterbalanced). Subjects in the three-relation condition could be queried about any of the three relations on any given trial. In the two- and three-relation conditions, each of the possible relations was queried equally often in a random order.

For each subject, a session consisted of 15 practice and 70 actual trials. Trials were presented in a random order.

Results

Figure 5 shows the mean proportion of correct responses in all nine conditions.

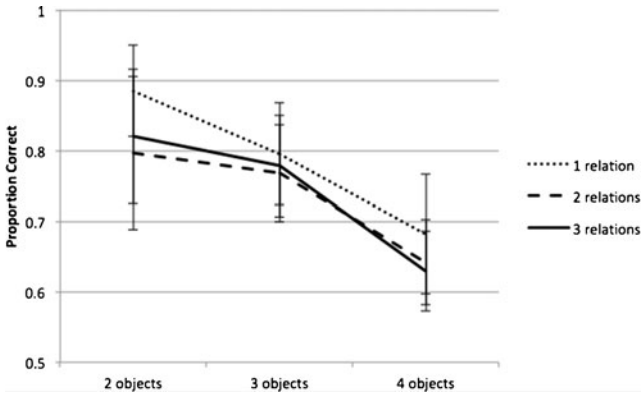

Fig. 5 Mean proportion of correct responses by condition in Experiment 1. Error bars depict the standard error of the mean

The results reveal a main effect of the number of objects, $F(2) = 22.81$, $p < .0001$, but no effect of number of relations that could be queried, $F(2) = 2.07$, $p = .129$, and no interaction, $F(4) = 0.31$, $p = .871$. There was no effect of type of relation queried in either the two-relation, $F(2) = 1.51$ $p = .230$, or the three-relation $F(2) = 0.62$, $p = .541$, condition.

We fitted these accuracy data against the predictions of all three models with values of k (WM capacity) from 2.0 to 7.1 in increments of 0.1. The best-fitting items-as-objects model had a k of 2.7 (consistent with the estimate of Vogel et al., 2001) and accounted for 89.9 % of the variance in subjects' accuracy. With $k = 3.0$, this model accounts for 83.0 % of the variance, and with $k = 4.0$ and $k = 5.0$, it accounts for 0 % of the variance. (With $k \geq 4.0$, this model's performance is at ceiling in all conditions.) The best-fitting items-as-bindings model had a k of 5.2 and accounted for 71.2 % of the variance. With $k = 3.0$, 4.0, and 5.0, this model accounts for 69.0 %, 67.5 %, and 71.0 % of the variance, respectively. Finally, the best-fitting hybrid model had $k = 5.0$ and accounted for 90.0 % of the variance. With $k = 3.0$, 4.0, and 5.0, this model accounts for 74.2 %, 84.8 %, and 90.0 % of the variance, respectively.

In terms of the proportion of subjects' error variance accounted for, all three models provide good fits to the results of Experiment 1. However, for sufficiently large values of k, all three models fail to account for any of the variance by predicting ceiling effects in all conditions. For the items-as-bindings model, this value is $k = 36$ (i.e., a person who could simultaneously hold 36 role bindings in WM would perform equally well in all conditions). For the hybrid model, this value is $k = 12$. That is, both models predict nonzero effects of our experimental manipulations within the normally accepted range of three to five items for WM capacity. By contrast, the items-as-objects model predicts no effects of our manipulations for any value of k equal to or above 4.0.

We also calculated root mean squared difference (RMSD) between the various model predictions and the observed accuracy data. With one degree of freedom (fixing WM capacity at 4, allowing only $p(r|s)$ to vary between 0.00 and 1.00 in increments of .05), the best-fitting hybrid model gives an RMSD of 0.045 (at $p(r|s) = .75$), and the best-fitting items-as-bindings model gives an RMSD of 0.102 (at $p(r|s) = .85$). (Recall that smaller values of RMSD indicate better fits.) Since the items-as-objects model accounted for none of the variance in the data with $k = 4$, we did not compute RMSD for that model.

With two degrees of freedom (again allowing $p(r|s)$ to vary between 0.00 and 1.00 in increments of .05 and allowing WM capacity to vary from 2.0 to 7.1 in increments of 0.1), the best-fitting items-as-objects model gives an RMSD of 0.043 at $k = 2.3$ and $p(r|s) = .65$. The best-fitting items-as-bindings model gives an RMSD of 0.072 at $k = 7.1$ and $p(r|s) = .70$. And the best-fitting hybrid model gives an RMSD of 0.030 at $k = 5.3$ and at $p(r|s) = .65$. As is visible in Fig. 6, the range of

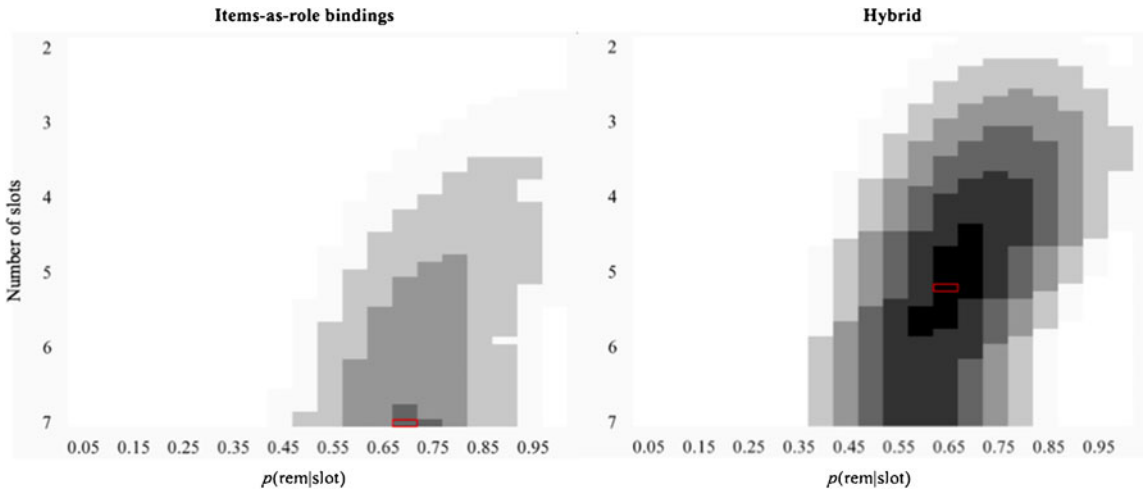

Fig. 6 Root mean squared difference (RMSD) plots for the items-as-objects, items-as-bindings, and hybrid model fits to the data from Experiment 1, varying k (working memory capacity) and $p(r|s)$. Darker colors in these plots indicate better fits. White indicates RMSD $\geq$ 0.13, and successively darker shades of gray indicate steps of 0.01, down to RMSD < 0.02, which is plotted as black. The best fit (lowest value of RMSD) in each plot is highlighted in a red box

parameter values over which the hybrid model gives small values of RMSD is wider than the corresponding range for either the items-as-objects or the items-as-bindings model.

It is interesting to note that the best-fitting hybrid model (in the two degrees of freedom (DF) case) has a WM capacity closer to the three-to-five range than does the best-fitting items-as-bindings model, although even the best-fitting hybrid model appears to have a WM capacity that is roughly a point too high relative to the limit of three to five. As was elaborated in the context of Experiment 3, this inflated value may reflect subjects' systematically answering correctly on a subset of the trials on which they had not encoded the queried pair into WM. Moreover, the best-fitting items-as-bindings model lies at the extreme of the range of WM capacities tested, suggesting that the fit could likely be improved further by assuming an even more unrealistic WM capacity. For the hybrid and items-as-objects models, by contrast, RMSD was nonmonotonic in the range of WM capacities tested, meaning that further increases in WM capacity would not improve the models' fits.

Experiment 1 conclusions

According to both proportion of variance accounted for and RMSD, the hybrid model provides the best account of the data from Experiment 1: Accuracy decreased with the square of the number of objects, but not with the number of relations between them. This finding suggests that relational roles can be "stacked" such that encoding a pair of objects in WM entails encoding all the relations between that pair at no additional cost (much as additional features of one object can be encoded at no additional cost). This result suggests that the visual system may encode only two pairs of objects in WM at a time but that it computes all the relations between the objects in each pair in parallel. More specifically, in a WM task for visual relations, an item in WM appears to be one member of a pair with a stack of relational roles relating it to the other member of the same pair.

One potential objection to Experiment 1 is that we do not know that subjects were actually computing the spatial relations between the objects and storing them in WM during encoding (but see Franconeri et al., 2012; Jung & Hummel, 2009; Roth & Franconeri, 2012; Saiki & Hummel, 1996, 1998; Tomlinson & Love, 2006). Perhaps, instead, subjects were simply memorizing the metric details of the display (e.g., in an image-like format) and computing the relevant relations only at the time of query. In this case, our task would not be a test of WM for relations among objects, but only a test of WM for objects. This interpretation of the findings of Experiment 1 is challenged by the fact that this "compute the relations after the fact" account predicts the same performance as the items-as-objects model. This "after the fact" interpretation is also challenged by the results of Experiments 2 and 3. Experiment 3 was designed to test very specific predictions of the hybrid model about accuracy as a function of the relation between the pairs a subject encodes and those on which he or she is queried. The items-as-objects model accounts for none of the variance in subjects' accuracy in this experiment, even if the capacity of WM is assumed to be less than four. Experiment 2 was deigned to explicitly compare subjects' memory for relative size with their memory for absolute size and provides evidence that observers do, indeed, encode relative size in WM.

Experiment 2

In order to directly test the hypothesis that observers perform the task in Experiment 1 by remembering the objects' absolute sizes and locations and compute their relative sizes and locations only at the time of query, Experiment 2 was designed to compare subjects' memory for absolute size with their memory for relative size. (In this experiment, as in Experiment 3, we used relative size as a proxy for all the relations investigated in Experiment 1. Recall that Experiment 1 showed no reliable effects of the number of relations to be remembered on a trial.)

On each trial of Experiment 2, the observer saw either a single random polygon or two copies of a polygon, side-by-side, differing slightly in size (Fig. 7). After a brief exposure, this display was replaced by a pattern mask, an interstimulus interval (ISI; a blank screen), and a test depicting either the same polygon (if the previous display had shown a single polygon) or the same pair (if the previous display had shown two). If the displays had shown single polygons, the observer's task was to say whether the second polygon was smaller or larger than the first (i.e., an absolute size comparison task). If the displays had shown pairs of polygons, the task was to say whether the size *difference* between the polygons in the second pair was smaller or larger than the size difference between the polygons in the first (i.e., a relative size comparison task). As will be detailed shortly, size differences on the absolute size comparison task (i.e., one polygon with one polygon) were numerically equated with size-difference differences on the relative size comparison task (i.e., two polygons with two polygons).

To the extent that observers encode only absolute size in memory and compute relative size only at the time of test, performance on the absolute size judgment should exceed performance on the relative size judgment at all exposure durations and ISIs. In this case, the relative size judgment task would require the subject to encode and compare four sizes (and compare the results of two size comparisons), whereas the absolute size task only requires them to encode and compare two. But to the extent that observers can encode relative size explicitly, performance on the absolute and relative size judgment tasks might diverge in other ways.

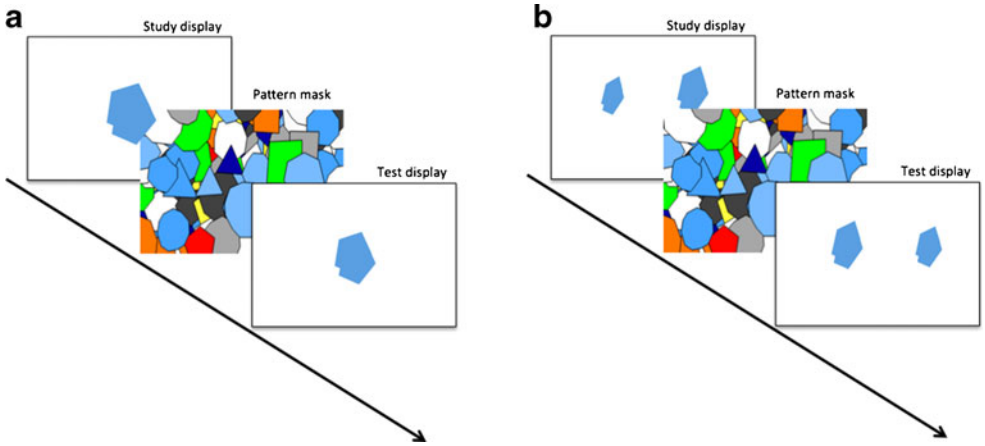

Fig. 7 **a** On absolute size trials, subjects must indicate whether the first (study) or second (test) polygon was larger. **b** On relative size trials, subjects must indicate whether the first or the second size *difference* was larger

One possible divergence concerns the effect of exposure duration. To the extent that relative size takes longer to compute and encode than absolute size (e.g., because it is based on estimates of absolute size), judgments of absolute size might be more accurate than judgments of relative size at short exposure durations. (Or, more generally, any advantage for absolute size over relative size might be greater at shorter, rather than longer, exposure durations, or any advantage for relative size over absolute size might be smaller at shorter than at longer exposure durations.)

A second possible divergence concerns the effect of delay between study and test. Relative size is a better measure of the distal stimulus than is absolute retinal size (e.g., the latter, but not the former, changes with distance from the viewer), so people may be biased to encode the relative sizes of objects in memory, rather than their absolute sizes. To the extent that this bias holds, judgments of relative size may be more robust to longer delays than are judgments of absolute size.

A third possible divergence is simply in overall accuracy. To the extent that numerical differences in absolute and relative size can be equated (as we have done in Experiment 2b), differences in observers' ability to detect one versus the other must reflect differences in how they encode the two properties in memory and compare them at test.

Method

Subjects

Sixteen University of Illinois undergraduate students earned course credit for participating in this experiment. All comparisons were within subjects.

Stimuli

Subjects viewed displays on a 24-inch Apple iMac with 1920 x 1200 resolution. The experiment was created and run in Python and Pygame. The study and test stimuli on all trials consisted of blue randomly generated polygons. We equated the absolute and relative size trials by using a fixed set of ratios ([1.04, 1.0816, 1.1.248, 1.1698]) to relate study items to test items on both kinds of trials.

We generated each absolute size stimulus by creating a randomly generated irregular convex *base* polygon, b_1, whose size (i.e., area on the computer screen) was randomly chosen from a square distribution in the range 8,000–15,000 pixels. We next made a *reference* polygon, r_1, by multiplying the area of b_1 by a scaling factor, s_1:

$$r_1 = s_1 b_1, \tag{4}$$

$s_1 \in (1.04, 1.0816, 1.1.248, 1.1698)$. Other than the difference in size, b_1 and r_1 were identical. Half the trials presented b_1 first, and the other presented r_1 first; the observer's task was to decide whether the first or the second polygon had been larger.

We generated the relative size trials as follows. First, we created a pair, p_1, of polygons, b_1 and r_1, as described previously. We next created a second pair of polygons, p_2, identical to the polygons in p_1 in shape but not size. Specifically, b_2 (the "first" member of p_2), like b_1, took a random size in the range 8,000–15,000 pixels. Polygon r_2 (the "second" member of p_2) took as its size the same scaling factor, s_1, that related r_1 to b_1 times an additional scaling factor, s_2, in the same range:

$$r_2 = s_2 s_1 b_2, \tag{5}$$

$s_2 \in (1.04, 1.0816, 1.1.248, 1.1698)$. As a result, exactly the same set of ratios relating b_1 to r_1 on absolute size trials related *pairs* polygons, p_1 and p_2, on relative size trials (i.e., s_1 and s_2 were chosen from the same set of values). However, none of the absolute sizes, b_1, b_2, r_1, or r_2, were the same on any given trial, making it impossible to perform the relative size judgment on the basis of the absolute sizes of b_1 and b_2 (i.e., even though the size difference in p_2 was always larger than the size

difference in p_1, b_1 was just as likely to be larger as smaller than b_2). However, by virtue of the way p_1 and p_2 were constructed, there was a small statistical tendency for r_2 to be the largest object on any given trial. (Specifically, r_2 will be the largest object on any trial on which b_1 is less than 1.082 larger than b_2, which is slightly more than half the trials. Experiment 2b was designed to explicitly control the appearance of the largest/smallest polygons in the p_1/p_2 pairs.) On half the trials, p_1 was presented first, and on the other half, p_2 was presented first. The subject's task was to determine whether the size difference relating the first pair was larger or smaller than the size difference relating the second pair.

Design and procedure

The experiment consisted of 8 practice trials, followed by 200 trials on which accuracy data were collected. During the experiment, absolute and relative trials were randomly intermixed.

Absolute size trials consisted of one irregular polygon (b_1 or r_1) displayed on the screen for 34, 68, 136, 273, or 544 ms. This display was followed by a pattern mask (ISI), which stayed on the screen for 200, 500, 1,000, or 4,000 ms, followed by the test display, which remained on the screen until the subject responded. The subjects' task was to indicate whether the first or the second polygon had been larger, using a keypress (see Fig. 7a).

Relative size trials consisted of one pair of polygons (p_1 or p_2) displayed on the screen for 34, 68, 136, 273, or 544 ms, followed by a pattern mask (ISI) for 200, 500, 1,000, or 4,000 ms. The test display (p_2 or p_1) depicted the same two pair of polygons, but with a slightly larger or smaller size difference between the members of the pair. The subjects' task was to indicate whether the first or the second size difference had been larger. The second display remained on the screen until the subject responded with a keypress (see Fig. 7b).

In this experiment, exposure duration, ISI, and the ratios, s_1 and s_2, did not vary orthogonally. Instead, in one block, the ratios varied while exposure duration and ISI were held constant, both at 200 ms. In the other block, the ratios were held constant at 1.1248 while exposure duration and ISI varied within subjects. The order of blocks was counterbalanced. Each block consisted of 200 trials.

Results

Unsurprisingly, as the ratios s_1 and s_2 increased, so did accuracy on both absolute size and relative size trials, $F = 65.639$, $p < .001$. More to our present interest, Fig. 8 shows accuracy

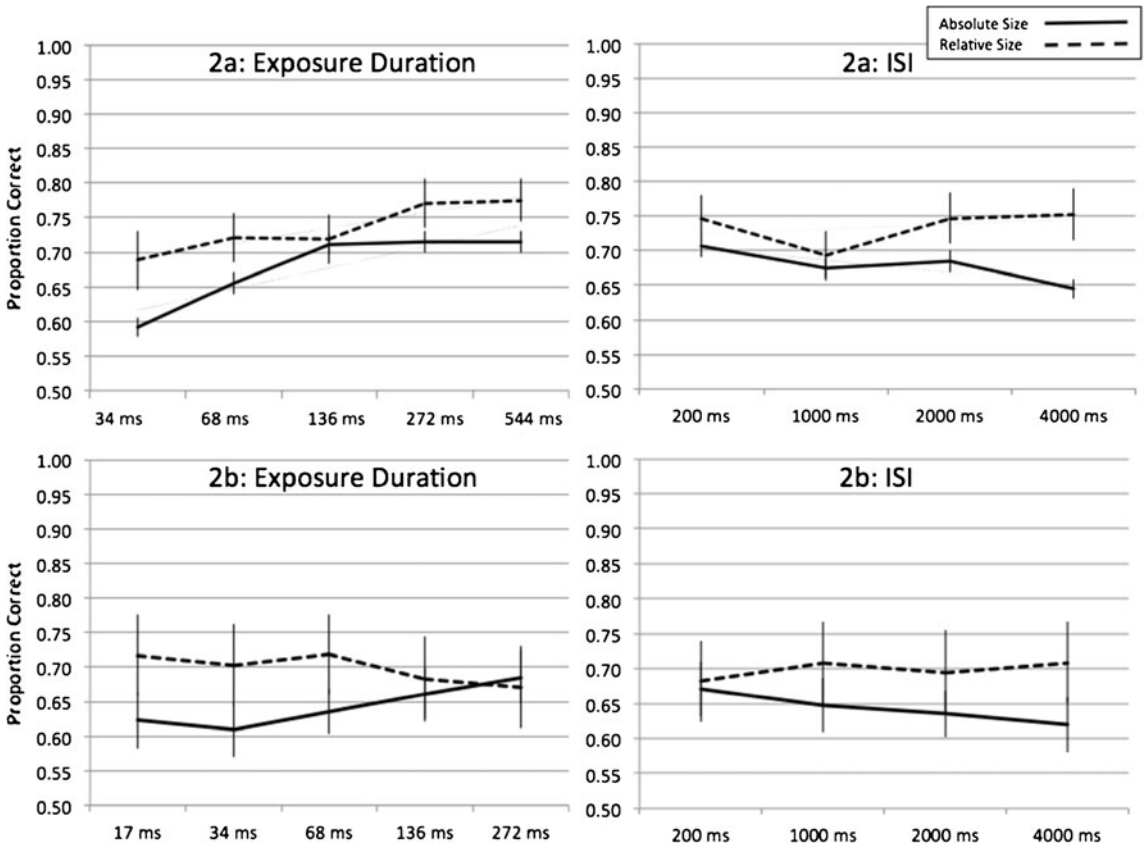

Fig. 8 Subjects' accuracy in Experiment 2a and in 2b as a function of exposure duration and ISI. Experiment 2b is a replication of Experiment 2a with more careful control

as a function of exposure duration (averaged over ISI, s_1, and s_2) and ISI (averaged over exposure duration, s_1 and s_2).

As the ISI between initial display and test increased, accuracy decreased on absolute size trials but remained relatively stable on relative size trials across all the delays tested. In the relative size condition, there was no reliable difference between performance with the shortest (200 ms) and longest (4,000 ms) ISIs, $t(16) = 0.86$, $p = .40$, but there was a reliable difference in performance between the shortest and the longest ISIs in the absolute size condition, $t(16) = 2.68$, $p = .01$. As is evident in Fig. 8, exposure duration had no reliable effect on performance in the relative size condition, $t(16) = 1.71$ $p = .11$, but longer exposure durations did facilitate performance, relative to shorter exposure durations, in the absolute size condition, $t(16) = 5.89$ $p < .001$.

Experiment 2b

Experiment 2b was a direct replication of the block of Experiment 2a that varied exposure duration and ISI, except that the stimuli were designed to more precisely control the largest and smallest values of both the smaller (b_1 and b_2) and larger (r_1 and r_2) members of p_1 and p_2 in the relative size condition. Specifically, the absolute sizes of the individual polygons on relative size trials were staggered as illustrated in Fig. 9, ensuring that b_2 was the smallest polygon on 60 % of the trials, with b_1 the smallest on the other 40 %, and r_2 was the largest polygon on 60 % of the trials, with r_1 the largest on the other 40 %. As such, responding only to the largest or the smallest polygon in any pair of pairs would ensure 60 % correct performance, and responding to both the largest and smallest would ensure correct performance on 20 % of the trials (see Fig. 9).

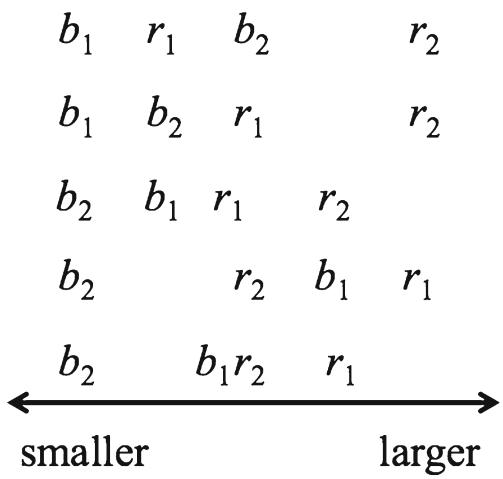

Fig. 9 Illustration of the size staggering used in Experiment 2b. Each row represents the sizes of the polygons presented on one kind of relative size trial, with larger polygons depicted to the right in the figure. The absolute sizes of the smallest (b_1 and b_2) and largest (r_1 and r_2) members of polygon pairs (p_1 and p_2) were controlled so that b_1 was the smallest polygon on 40 % of the trials (rows 1 and 2) and b_2 the smallest on the other 60 % (rows 3–5), and r_1 was the largest polygon presented on 40 % of the trials (rows 4 and 5) and r_2 the largest on the other 60 % (rows 1–3)

This staggering was accomplished by first setting the size of b_1 to a value between 8,000 and 15,000 pixels and r_1 to a size value based on b_1 and s_1, as described above (Eq. 4). (In this experiment, s_1 was randomized to a value in the set [1.04, 1.0816, 1.1.248, 1.1698] on each trial, and s_2 varied within subjects as in Experiment 2a.) To construct a [b_1, r_1, b_2, r_2] trial (Fig. 9, row 1), we set b_2 to the size of r_1 plus a random number of pixels between 0 and 1,000 and then set r_2 to b_2 times the larger relation size, s_1s_2 (Eq. 5). To construct a [b_1, b_2, r_1, r_2] trial (Fig. 9, row 2), we set b_2 to a value halfway between b_1 and r_1 and then set $r_2 = b_2s_1s_2$ (Eq. 5). To construct a [b_2, b_1, r_1, r_2] trial (Fig. 9, row 3), we set r_1 to b_1s_1, b_2, and r_2 to r_1. We then iteratively made b_2 smaller and r_2 larger until the ratio (r_2/ b_2)/(r_1/ b_1) = s_2. To construct a [b_2, r_2, b_1, r_1] trial (i.e., Fig. 9, row 4), we set r_2 to b_1 minus a random number of pixels between 0 and 1,000 and then set b_2 so that relation r_2/b_2 was equal to s_2 times the relation r_1/b_1. That is, $b_2 = (s_2r_1)/(r_2r_1)$ so that $r_2 = b_2s_1s_2$ (Eq. 5). To construct a [b_2, b_1, r_2, r_1] trial (i.e., Fig. 9, row 5), we set the size of r_2 to a value halfway between b_1 and r_1 and then set $b_2 = (s_2r_1)/(r_2r_1)$ so that $r_2 = b_2s_1s_2$ (Eq. 5) Fig. 10.

In all other respects, Experiment 2b was identical to the block of Experiment 2a that varied ISI and exposure duration, except that (1) as was noted previously, we randomized s_1 on a trial-by-trial basis and (2) the shortest exposure duration we tested was 17 ms (i.e., one screen refresh) rather than 34 ms.

Results

Just as in Experiment 2a, accuracy decreased with increasing ISI for absolute size trials but remained relatively stable for relative size trials. In the relative size condition, there was no reliable difference between performance in the shortest (200 ms) and longest (4,000 ms) ISIs, $t(20) = 1.39$, $p = .17$, but there was a reliable difference in performance between the shortest and the longest ISIs in the absolute size condition, $t(20) = 3.09$, $p = .005$.

Also, although the shapes of the lines appear different for Experiments 2a and 2b for exposure duration, the reliable pattern remains the same between the two experiments. That is, exposure duration had no reliable effect on performance in the relative size condition, $t(20) = 1.59$ $p = .12$, but longer exposure durations did facilitate performance relative to shorter exposure durations in the absolute size condition, $t(20) = 3.22$ $p = .004$.

Experiment 2 Conclusion

Experiment 2 demonstrated qualitative differences between observers' memory for absolute and relative size. Most strikingly, subjects' memory for relative size was overall more accurate than their memory for absolute size (i.e., around

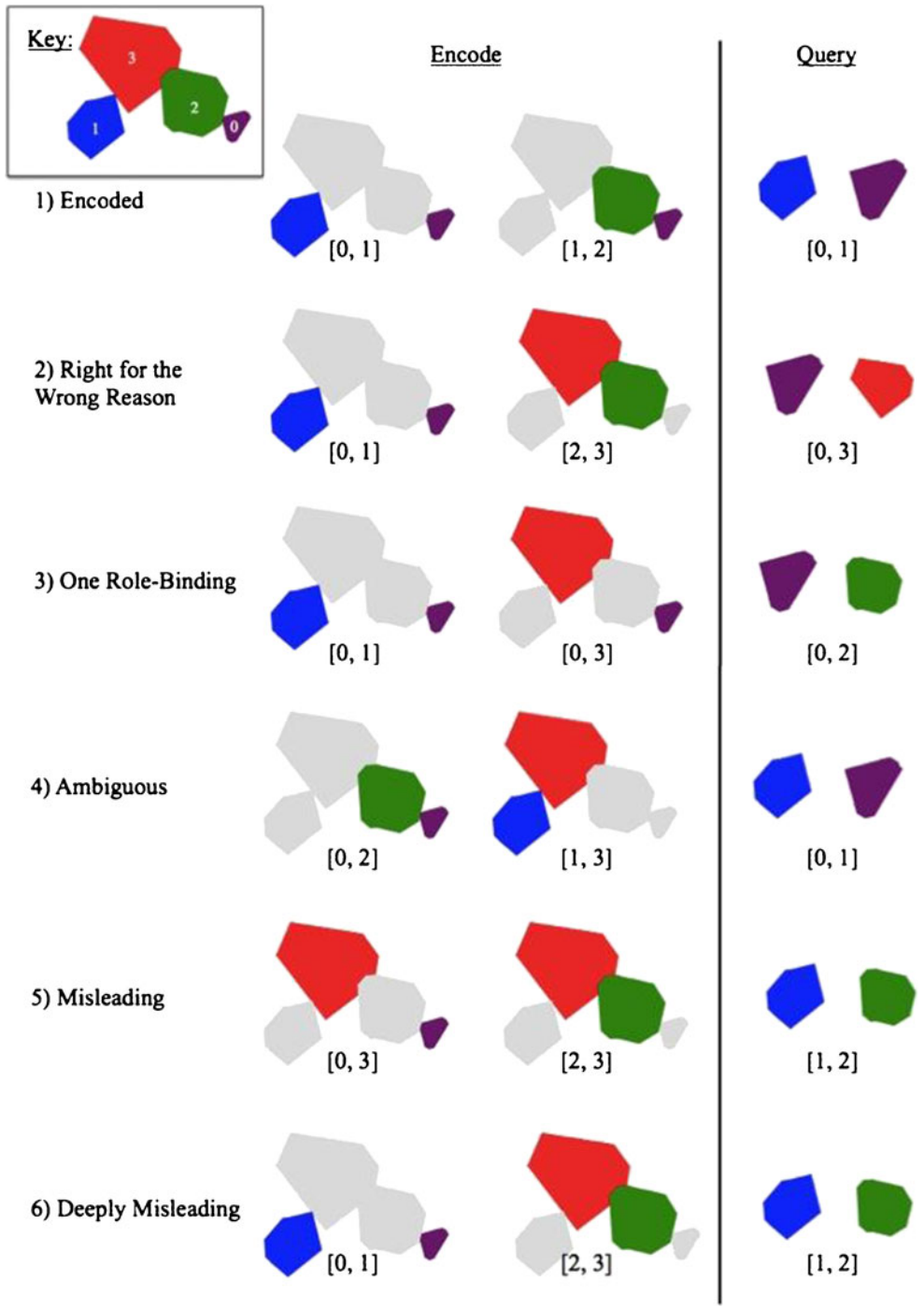

Fig. 10 Schematic illustration of the six categories of encoding-query sets. Colored objects in the encoded pairs were presented first and last on the encoding trial. See the text for details

75 % vs. 70 % accuracy on our tasks, respectively). Memory for relative size also lasted longer than memory for absolute size, although over the durations we tested, this effect was modest. Interestingly, memory for relative size was superior to memory for absolute size at all the exposure durations we tested. This result suggests that relative size can be computed and encoded very rapidly, at least under the presentation conditions to which we exposed our observers.

Most important for our present purposes, these results suggest that observers can and do compute and encode relative size as a property of two objects in its own right: It is not the case that people encode absolute size and compute relative size only later, as required by the task in which they are engaged. Of course, this result does not imply in any strong sense that the subjects in Experiment 1 were encoding our displays in terms of the relations among the depicted objects,

but only that they were at least capable of doing so and that, to the extent that Experiment 1 required our subjects to make "fine" (for some definition of "fine") discriminations between the objects in terms of size or location, it would have been in their best interest to do so.

Experiment 3 and computational model

On the basis of mathematical instantiations of the items-as-objects, items-as-bindings, and hybrid models of WM for spatial relations, the results of Experiment 1 provided support for the hybrid account. This model accounted for the largest proportion of the variance in subjects' accuracy on the object–relation memory task and provided the closest fit to the subjects' data in terms of RMSD.

Experiment 3 was designed to replicate and extend the results of Experiment 1. Whereas Experiment 1 was designed to test the broad performance predictions of abstract mathematical instantiations of each of the three models, Experiment 3 was designed to test more nuanced predictions of the hybrid model itself. To the extent that the hybrid model provides an accurate account of the manner in which we encode the spatial relations among objects into visual WM, it ought to be able to predict not only overall accuracy, but also specific patterns of accuracy as a function of the relations between the objects and relations the observer encodes into WM and those on which he or she is subsequently queried.

In order to test these more nuanced predictions, we first constructed a process version of the hybrid model—that is, a version of the model that actually performs the observer's task—and observed the model's accuracy as a function of (1) which object pairs it had encoded into WM on a given trial and (2) the pair on which it was queried on that trial. We tested the model on all possible combinations of encodings and queries. We next briefly summarize the process model and its predictions, followed by Experiment 3, which was designed to test those predictions. The model is described in detail in the Appendix.

Computational model

The fundamental tenet of the hybrid model is that two pairs of objects can be held in WM, along with all the relations between the objects within each pair. Accordingly, the process version of the model encodes pairs of objects in memory along with all the relations (*larger*, *above*, and *right-of*) between the members of each pair. We simulated each trial of Experiment 1 in two phases, an *encoding* phase and a *test* phase. During encoding, the model stores two pairs of objects in memory in terms of their shapes and the spatial relations within each pair. Queried with a pair of objects during test, the model compares the queried objects with the pairs it has stored in memory and attempts to activate spatial relations on the basis of the match between the queried pair and the stored pairs. The relations so activated serve as the model's "memory" for—more accurately, estimate of—the likely relation between the objects in the query. On the basis of this estimate, the model generates a response (e.g., "object 1 was larger than object 0"), which is compared with the correct relation in order to determine the accuracy of the model's response.

In the model's memory, displays such as those used in Experiment 1 are encoded at three hierarchical levels: (1) as objects bound to specific relational roles (e.g., object 0 bound to *smaller*, *right-of*, and *above*); (2) pairs of objects in specific collection of relations to one another (e.g., *smaller*, *right-of*, *above* [object 0, object 1]); and (3) collections of two such pairs. During the test phase, the model's response to a queried pair is based on the match between the objects in that pair and the relations and relational roles encoded in the model's memory.

Simulations and predictions

All our simulations used four-object displays, since such displays afford the richest set of potential pairs for encoding and test that can be compared with the conditions of Experiment 1. In a four-object display, there are six [$6 = (4^2 - 4)/2$] unique pairs of objects to use as queries and 15 [$15 = (6^2 - 6)/2$] unique pairs of pairs to use for encoding, for a total of 90 possible trial types (where a trial is defined by the pair of pairs encoded and the pair queried; see Table 1). These predictions fall into six categories.

1. *Encoded* is when the queried pair is one of the ones the model encoded. For example, if the model encodes pair [0, 1] and pair [1, 2] and is then queried about [0, 1], it is likely to report the correct relation between 0 and 1 because it had encoded it.
2. *Right for the wrong reason* is when the model can get the correct answer without having encoded the queried pair. For example, if the model encodes [0, 1] and [2, 3] and is queried on [0, 3], it should answer correctly that 0 was smaller than 3 because 0 is encoded (in the first level of the hierarchy) as *smaller than* something and 3 is encoded as *larger than* something.
3. *One role binding* ("1RB" in Table 1) is when the model encodes only one of the role bindings on which it is queried—for example, encoding [0, 1] and [0, 3] and then being queried on [0, 2]. Even though the model doesn't know anything about object 2, it knows that 0 is smaller than something, and so it may have an opportunity to guess the correct answer.
4. *Ambiguous* occurs either (a) when both queried objects are encoded in the same role or (b) when only one of the encoded objects is queried and that object was encoded in both roles. An example of (a) is when the model encodes

Table 1 For 15 possible pairs of parts and six possible queries, there are six categories of responses from the model, which are based on the relationship between pairs of objects that were encoded and the pair queried

	possible query pairs					
	[0, 1]	[0, 2]	[0, 3]	[1, 2]	[1, 3]	[2, 3]
possible encoding pairs						
[0, 1] [0, 2]	encoded	encoded	1RB	ambig.	misleading	misleading
[0, 1] [0, 3]	encoded	1RB	encoded	misleading	ambig.	1RB
[0, 1] [1, 2]	encoded	RWR	1RB	encoded	ambig.	misleading
[0, 1] [1, 3]	encoded	1RB	RWR	ambig.	encoded	1RB
[0, 1] [2, 3]	encoded	ambig.	RWR	**DEEP misl.**	ambig.	encoded
[0, 2] [0, 3]	1RB	encoded	encoded	1RB	1RB	ambig.
[0, 2] [1, 2]	ambig.	encoded	1RB	encoded	1RB	misleading
[0, 2] [1, 3]	ambig.	encoded	RWR	RWR	encoded	ambig.
[0, 2] [2, 3]	1RB	encoded	RWR	ambig.	1RB	encoded
[0, 3] [1, 2]	ambig.	RWR	encoded	encoded	RWR	ambig.
[0, 3] [1, 3]	ambig.	1RB	encoded	1RB	encoded	1RB
[0, 3] [2, 3]	1RB	ambig.	encoded	misleading	1RB	encoded
[1, 2] [1, 3]	misleading	1RB	1RB	encoded	encoded	ambig.
[1, 2] [2, 3]	misleading	ambig.	1RB	encoded	RWR	encoded
[1, 3] [2, 3]	misleading	misleading	1RB	ambig.	encoded	encoded

[0, 2] and [1, 3] and is queried on [0, 1]; then it knows that 0 is smaller than something and 1 is smaller than something, leaving it unclear which of these two had been smaller than the other. An example of (b) is when the model encodes [0, 1] and [1, 2] and is queried on [1, 3]. In this case, it knows nothing about 3, and the two bindings it knows about 1 contradict one another.

5. *Misleading* is when one encoded role binding points in the wrong direction. For example, if the model encodes [0, 3] and [2, 3] and is queried on [1, 2], it has encoded 2 as smaller than something even though 2 was, in fact, larger than 1 (a relevant fact the model failed to encode).
6. *Deeply misleading* occurs when the model encodes [0, 1] and [2, 3] and then is queried on [1, 2]. It has encoded that 1 was larger than something and that 2 was smaller, so it is very unlikely to answer correctly that 1 had actually been smaller than 2.

Figure 11a illustrates the proportion of the model's correct responses over 100 runs of all 90 possible encoding × query pairings. In the cases of *encoded* and *right for the wrong reason*, the model performs near ceiling. In the cases of *one role binding*, *ambiguous*, and *misleading*, the model performs near chance. And in the case of *deeply misleading*, the model nearly always answers incorrectly.

Figure 11b depicts predicted accuracy in Experiment 3 if subjects encode the objects' absolute sizes at encoding and compute their relative sizes only at the time of query. These predictions are based on a WM capacity, k, of 4.0 and $p(\mathrm{r}|\mathrm{s}) = 1.0$, but the ordinal predictions remain exactly the same (i.e., with all values scaled toward 0.5) if k is assumed to be less than 4.0 or $p(\mathrm{r}|\mathrm{s})$ is assumed to be less than 1.0.

Experiment 3

The simulation results in Fig. 11a constitute detailed predictions about subjects' performance with four-object displays as a function of which pairs they happen to encode on a given trial and the pair on which they are queried. Experiment 3 was designed to test these predictions. Every trial of Experiment 3 presented subjects with a four-object display to encode and queried them about which of two objects had been larger in that display (recall that the effect of number of relations was not reliable in Experiment 1, so we held that variable constant in Experiment 3).

Although it is straightforward to experimentally manipulate the pair on which a subject is queried on any given trial, it is more challenging to manipulate the pairs they happen to encode. To this end, Experiment 3 manipulated the timing of the presentation of the objects in the encoding displays (Fig. 12). On each trial, two of the four to-be-encoded objects were presented first for 100 ms, followed by the four-object display as a whole (1,000 ms), followed by two of those objects for 100 ms. Our intuition was that presenting pairs of objects in isolation before and after the display as a whole would bias subjects to encode the relations between the objects in those pairs. Accordingly, the objects presented first and last on any given trial were chosen to correspond to the rows of Table 1. To the extent that the process model provides

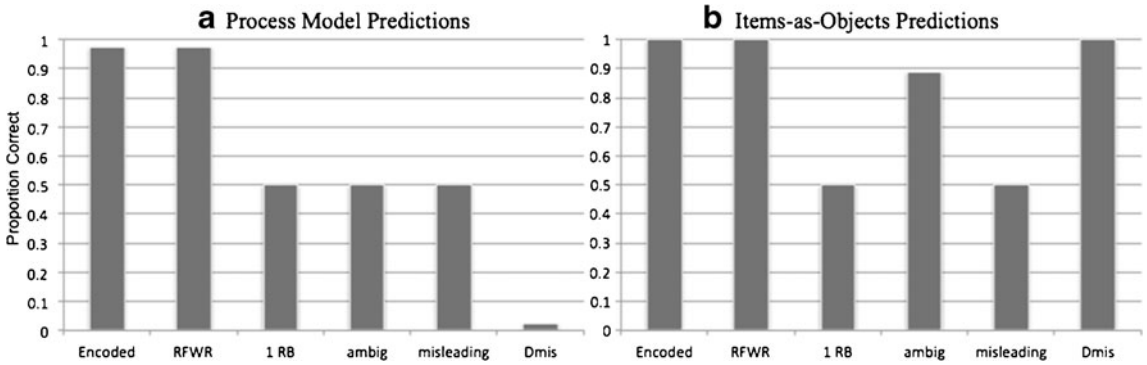

Fig. 11 **a** Results of 100 runs of the process version of the hybrid model. **b** Predictions of the encode size only (i.e., items as objects) model in the same six conditions

an accurate account of the manner in which people perform our task, subjects' performance as a function of the encoding and query manipulations ought to conform to the model predictions presented in Fig. 11a.

Method

Subjects

Forty-eight University of Illinois undergraduate students earned class credit for participating in the experiment.

Stimuli

The stimuli were like those of Experiment 1, except that all stimuli presented four objects.

Design and procedure

The procedure was similar to that of Experiment 1. Subjects viewed objects on a 24-inch Apple iMac with 1920 x 1200 resolution. First a fixation cross was presented; then the display, pattern mask, and, finally, a query were presented, followed by accuracy feedback. In this experiment, however, we were interested in comparing the six categories of accuracy predictions from the model with the performance of human subjects. For this reason, only four-part objects were used, and only one relation (larger/smaller) was queried.

In order to compare pairs encoded by the subjects on a given trial with the query, the displays appeared on the screen in a way designed to encourage subjects to encode particular pairs of objects into WM (see Fig. 12). Specifically, one pair of objects was presented for 100 ms, then the whole display was presented for 1,000 ms, and then the second pair remained alone on the screen for 100 ms. All of the 15 possible part pairings (rows in Table 1) were presented equally often.

A session consisted of 15 practice and 90 actual trials. Trials were constructed to match the conditions depicted in Table 1.

Results

Figure 13 shows subjects' accuracy in each of the six categories of conditions. The data closely match the hybrid model predictions (r^2 = .88). In contrast to the model, the subjects showed a strong recency effect, in that they had a better memory for the second pair of objects presented during encoding than for the first pair, $t(47) = 4.17$, $p < .001$. Considering only the pairs subjects encoded second during encoding, r^2 increases to .93. By contrast, the items-as-objects model (Fig. 11b) accounts for only .003 (0.3 %) of the variance in subjects' accuracy.

The model predicted that memory for *encoded* and *right for the wrong reasons* would be better than memory for *one role binding*, *ambiguous*, and *misleading*. The subjects' data exhibited the same pattern, $t(47) = 2.44$, $p = .01$. However,

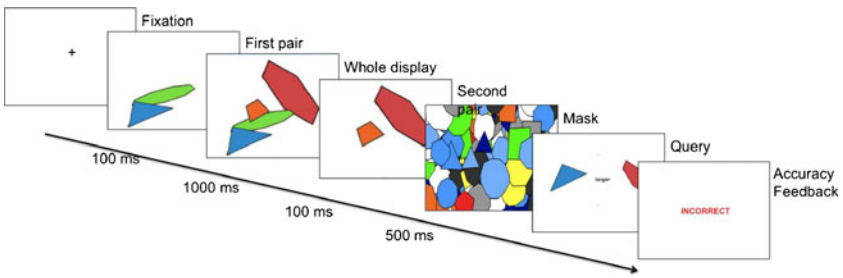

Fig. 12 Sequence of events on a trial in Experiment 3

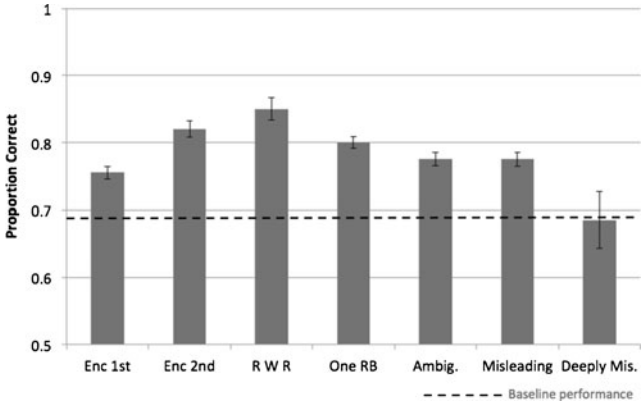

Fig. 13 Accuracy as a function of category of response in Experiment 3. Baseline performance is based on the four-object, one-relation condition from Experiment 1

counter to the model's prediction, performance in *one role binding*, *ambiguous*, and *misleading* was not reliably better than performance in *deeply misleading*, $t(47) = 1.65$, $p = .10$.

The baseline value depicted in Fig. 11 corresponds to accuracy in the four-object, one-relation condition from Experiment 1. In Experiment 3, performance in the *deeply misleading* condition was numerically worse than baseline performance, although this difference was not statistically reliable, $t(20) = 1.29$, $p = .208$.

Experiment 3 conclusions

When recency effects are excluded from the subjects' data, the process model accounts for 93 % of the variance in subjects' accuracy in Experiment 3. As was predicted, accuracy is higher in *encoded* and *right for the wrong reasons* than in *one role binding*, *ambiguous*, *misleading*, and *deeply misleading*. The only case in which performance fell numerically below baseline was *deeply misleading*. Along with the results of Experiment 1, these results are consistent with the hybrid model's prediction that we can hold two pairs of objects in WM, along with all the relations between the objects within each pair. These results also stand in stark contrast to the predictions of the items-as-objects model or any model that assumes that subjects hold absolute size in WM and compute relative size only at the time of query.

One notable difference between the model's performance and that of the human subjects is that the model's accuracy varies between zero and one, whereas subjects' accuracy is bounded between chance (50 %) and one. Although the reason for this difference is not completely clear, one likely explanation is that the model, unlike the human subject, cannot guess when the evidence it gets from the feedback from memory is weak: The model simply compares activation accumulated in units representing *larger* and *smaller* and responds when a threshold difference is reached. It is not sensitive to the strength of the evidence that led to that difference. The human observer, by contrast, may have some sense of the strength of the evidence he or she brings to bear on his or her decision and is more likely to guess when that evidence is weak. A more important difference between the model and our human observers is that the model could encode only those pairs we told it to on any given simulation. Our experimental manipulation, by contrast, serves merely to bias subjects to encode some object pairs over others. To the extent that this biasing effect is imperfect, subjects can be expected to encode pairs other than those we intended on any given trial.

Another interesting property of the simulation results and the subjects' data is that the model predicted slightly better performance in the *right-for-the-wrong-reason* condition than in the *encoded* condition, and a trend toward a similar pattern is visible in the human data (although the difference is not statistically reliable). In the model, this difference derives from the fact that a *right-for-the-wrong-reason* response is driven by a larger number of pairs of objects in memory than is an *encoded* response. Specifically, *right-for-the-wrong-reason* activates two objects in two separate pairs in memory (one for each pair in which each object participated; recall that *right-for-the-wrong-reason* trials present, at test, objects the subject/model has encoded in relations consistent with the right answer, just not in the same pair). By contrast, *encoded* activates two objects in memory but only one pair. Due to the nonlinearity of the activation function of the units composing the model (see the Appendix), this two-and-two versus two-and-one difference is sufficient to generate more evidence for the right answer in the case of *right-for-the-wrong-reason* than in the case of *encoded*. It is tempting to wonder whether a similar effect was operating in the visual systems of the human subjects.

The fact that the process model can respond correctly to some queries it had not encoded into memory (e.g., in the case of *right for the wrong reason* and *one-role binding*) has important implications for our estimates of the capacity of WM on tasks requiring memory for the spatial relations among objects. Recall that the RMSD fits of the data from Experiment 1 to the predictions of (the mathematical version of) the hybrid model yielded an estimated WM capacity of 5.3, which is above the normally accepted range of 3–5. We speculate that the model's (and, by hypothesis, human's) ability to exceed the 4 ± 1 capacity limit reflects the role of retrieval-based heuristics, such as those used by the model, that make it possible to make intelligent guesses.

To this end, we used RMSD to fit the simulation results of the process model against the mathematical version of the hybrid model (in a manner precisely analogous to that in which we fit the human data; i.e., for the purposes of this analysis, we treated the process model as a human subject). For this analysis, we assumed that the model would perform perfectly in the two-object condition of Experiment 1. The model's WM capacity is set to exactly 4, and its $p(r|s) = 1.0$, so

there is no reason why it should make any errors in the two-object case. In the three-object condition of Experiment 1, the process model is expected to perform at a mean accuracy of 0.89: On seven of the nine possible encoding/query pairs, the model will have encoded the queried pair [giving it a $p(c) = 1.0$], and on the remaining two, it will have encoded one role binding [giving it a $p(c) = .5$], for a total of 8/9 = .89 overall proportion correct in that condition. And for the four-object condition of Experiment 1 (which is equivalent to Experiment 2), the model achieved an overall accuracy of 0.66. With these numbers, the best RMSD fit of the process model simulations to the mathematical version of the hybrid model yields RMSD = 0.021at WM capacity 4.5 and $p(r|s) = 1.0$ (see Fig. 14). That is, the model, like our human subjects, best fits a WM capacity slightly greater than 4.0. And, crucially, it did so in spite of the fact that we explicitly built it to have a WM capacity of *exactly* 4.0. We take this result as strong suggestive evidence that our human subjects' seemingly exaggerated WM capacity in Experiment 1 may reflect retrieval/relation-matching strategies similar to those used by the process model.

General discussion

The literature on WM, from the study of both visual WM (e.g., Luck & Vogel, 1997) and WM in memory and reasoning (e.g., Baddeley & Hitch, 1975; Halford et al., 2005) suggests that, for most people and across many tasks, people can hold and manipulate about three to five "items" in WM at a time. However, this apparent agreement about the capacity of WM belies an implicit disagreement about the currency of WM: In the vision literature, the "items" occupying WM are often assumed to be objects (i.e., bound collections of object features), whereas in the literature on memory and reasoning, these "items" are assumed to be role–filler (or, equivalently, variable–value) bindings. The importance of this implicit disagreement becomes apparent when one poses the question, What is the capacity of visual WM for the spatial relations among objects in a visual display?

According to the traditional visual account, which holds that WM load scales with the number of objects to be remembered, people ought to have no difficulty remembering the spatial relations among up to four (or at least 2.8; Vogel et al., 2001) objects. But according to the traditional account from memory and reasoning, which holds that load scales with the number of role bindings to be encoded, remembering, say, three relations among just four objects ought to impose a catastrophically large load of 36 role bindings. According to the hybrid account proposed here, WM load should scale with the square of the number of to-be-encoded objects (as predicted by the items-as-bindings account) but should be unaffected by the number of relations to be encoded by the members of each pair (as predicted by the items-as-objects account from the vision literature).

Experiment 1 evaluated these three accounts as mathematical models that predict performance on an object relation memory task as a function of the number of objects to be remembered and the number of relations to be remembered among those objects. The results supported the hybrid model, which accounted for 90 % of the variance in subjects' accuracy on this task. Experiment 2 compared observers' memory for objects' relative sizes with their memory for objects' absolute sizes and found that the two kinds of memory are differentially sensitive to the time between encoding and test, an effect that suggests that observers can encode relative size in memory explicitly, as a visual property in its own right (i.e., it is not something we simply compute after the fact by storing objects' absolute sizes in memory). Experiment 3 tested the predictions of a process version of the hybrid model. Rather than simply predicting accuracy as a function of the number of

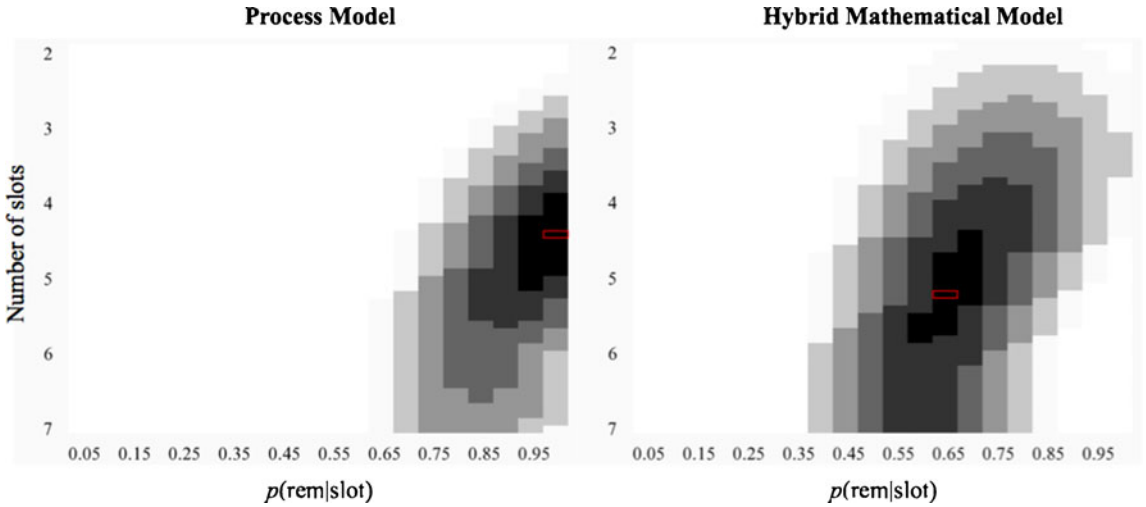

Fig. 14 Left: RMSD fits of the process model simulation results against the predictions of the mathematical version of the hybrid model. Right: For comparison, RMSD fits of human data (Experiment 1) to the same mathematical model

objects and relations to be remembered (as tested in Experiment 1), this model makes detailed predictions about subjects' performance as a function of which object pairs they encode during study and which they are tested on at query. Ignoring recency effects in subjects' performance (which the model, with its perfect memory, does not show), this model accounts for 93 % of the variance in subjects' accuracy on an object relation memory task.

Together, the results of Experiment 2 suggest that observers can represent visual relations (at least relative size) explicitly in WM, and those of Experiments 1 and 3 support the hypothesis that visual WM can hold roughly two pairs of objects along with all the spatial relations between the members of each pair. These results have a number of counterintuitive implications for our understanding of the capacity and currency of visual WM.

First, they suggest that conceptualizing the currency of WM simply as "objects" or as "role-bindings" is too simple. Visual WM, at least, appears to be more intelligent than that, encoding pairs of objects in pairs of "slots," but "stacking" all the relations between the objects into the same "slots" in WM. This approach avoids both the ambiguity of the purely object-based account (e.g., making it possible not only to know that object 1 is larger than something, but also to know what, specifically, it is larger than) and the large WM resources demanded by the pure items-as-bindings account (e.g., 36 "slots" to hold just three relations among four objects). In this sense, the hybrid account seems like an intelligent compromise between these two traditional accounts of the currency of WM.

A second counterintuitive implication of the hybrid model is that an "object," as defined in the stimulus or by the experimenter, is not the same thing as an "object" as defined in terms of the resource limitations of WM: In the stimulus (and probably in the mind of the experimenter), object 1 in Fig. 2 is simply object 1. But if an observer encodes, for example, *larger* (1, 0) and *larger* (2, 1) into his or her WM, then object 1 in the context of its relation to object 0 occupies a different slot in WM—is effectively a different item in the currency of WM—than object 1 in the context of its relation to object 2. In this respect (as in many others in experimental psychology), the mind of the subject may not respect the definitions assumed in the mind of the experimenter.

A third implication of the findings reported here—in particular, those of Experiment 3—is that performance on a WM task is not necessarily a straightforward function of the relation between the query and the contents of WM. As is illustrated by subjects' high level of accuracy in the *right for the wrong reason* condition of Experiment 3, a person may perform well on a WM query even if he or she never actually encoded the relation embodied in the query. In this context, it is important to note that subjects' high accuracy in this condition cannot be attributed to simple, after-the-fact strategies such as reasoning by transitive inference (e.g., "I know that object 2 was larger than object 1 and 1 was larger than 0, so I can infer that 2 was larger than 0"). This interpretation is inconsistent with the fact that encoding [0, 1] and [2, 3] helps subjects to correctly answer that 3 was larger than 0. But since the pairs [0, 1] and [2, 3] share no arguments, they do not afford transitive reasoning. It appears that, rather than using a rational, deliberative process such as transitive inference, our subjects were basing their 3 versus 0 judgment (and related *right for the wrong reason* judgments) on a process more akin to the retrieval-based heuristic embodied in the hybrid model. As is illustrated by our subjects' performance in Experiment 3, such heuristics may provide a useful (albeit fallible; recall the deeply misleading case) basis for making (mostly accurate) relational judgments even in the face of a sharply limited WM capacity.

A related implication of the results of Experiment 3 concerns the difference between the subjects' (and the model's) performance in the *right for the wrong reason* and *one role binding* conditions. In *right for the wrong reason*, the model/subject encodes two role bindings that both point in the direction of the right answer (e.g., encoding both that 0 is smaller and that 2 is larger and so answering correctly that 2 was larger than 0). In *one role binding*, only one of these bindings gets encoded. Although, in principle, the one binding could be used to make an educated guess at the right answer, neither the model nor the human subjects appear to use this information, as evidenced both by the superior performance in *right for the wrong reason*, relative to *one role binding*, and by the equivalence of *one role binding* to, for example, *misleading* (where one queried binding is again encoded but points to the wrong answer). Together, these results suggest that the human observer bases his or her judgments on consistent *conjunctions* of role bindings (e.g., 0 is smaller and 2 is larger), rather than on single bindings in isolation (e.g., 2 is larger). This result is inconsistent with any account that simply treats relational roles as object features (e.g., Hummel & Biederman's, 1992, JIM model and the items-as-objects account from the visual WM literature) or otherwise asymmetrically codes relations as properties of a single object (e.g., Roth & Franconeri, 2012).

Visual processing may be especially well-suited to relational "stacking" as embodied in the hybrid model proposed here. Although visual perception is subject to attentional bottlenecks at multiple levels of processing, it is also the case that a great deal of visual computation goes on in parallel all over the visual field. Such parallel processing may naturally afford computing and storing multiple spatial relations between a pair of objects at the same time (see Hummel & Biederman, 1992). By contrast, verbal stimuli, whether spoken or read, are necessarily processed in a sequential fashion. It is interesting to wonder whether the "stacking" strategy the visual system seems to use with visual stimuli may also apply to more abstract (including verbal) materials for the purposes of "higher" cognition: If one is told that "John loves Mary" and

"John is taller than Mary," can one stack these two relations into the single proposition *loves-and-taller-than* (John, Mary)? Or must these two propositions, if presented verbally, always require four slots in WM?

Open questions and future directions

The data presented here do not challenge the well-supported conclusion that the capacity of WM is roughly three to five "items" across multiple domains in vision and cognition (e.g., Avons et al., 1994; Baddeley et al., 1975; Cowan et al., 1997; Halford et al., 2005; Hitch et al., 1996; Levy, 1971; Longoni et al., 1993; Luck & Vogel, 1997; Murray, 1968; Peterson & Johnson, 1971; Song & Jiang, 2006; Woodman & Vogel, 2008; for reviews, see Baddeley, 2003; Cowan, 2001), and they do not speak (at least not directly) to the issue of whether the capacity of WM is better conceptualized in terms of "slots" or some other, more general, notion of "resources" (e.g., Alvarez & Cavanaugh, 2004; Alvarez & Oliva, 2009; Bays & Husain, 2008). However, they underscore the importance of understanding the currency of WM—and understanding the operations that use the contents of WM for making judgments—for understanding the nature of WM and its role in perception and cognition.

The findings presented here are preliminary, in that they are based on a restricted set of stimuli and behavioral tasks, and many open questions remain. Our stimuli were intentionally designed to present novel shapes in an impoverished context, so it is unclear how our findings will generalize to more natural, or familiar, objects in more natural scenes. For example, how would subjects' performance differ if, say, the relative sizes of two objects in a display contradict or complement their relative sizes as real objects (e.g., as when the image of a mouse is larger than the image of a cat). We have also discussed our stimuli as collections of "objects," but it unclear whether, perceptually, they are more naturally thought of as separate objects or as parts of a single object. For example, among other things, in every one of our displays, every object appeared to be "touching" or overlapping at least one other object. Although we did not mention it in the results of Experiment 1, we analyzed subjects' performance as a function of whether the queried objects had been touching in the study display and found that this variable had no effect on performance. In contrast, Saiki and Hummel (1998) found evidence that whether two figures appear to touch does affect subjects' perception of the spatial relations between them (specifically, the perceptual binding of the relations to the parts so related). Our findings also do not speak to the question of whether our subjects' errors reflect failures of perception, encoding, or memory: For example, how would their performance change if we doubled or tripled the time they had to look at the displays? Finally, the results presented here do not speak to the question of how different kinds of WM may work together (or at odds) in the service of perceiving and remembering spatial relations: Are the same cognitive and neural resources responsible for representing both objects and the relations between them? To our knowledge, all of these questions remain largely, if not completely, open.

Appendix

The model is an eight-layer artificial neural network based loosely on Hummel and Biederman's (1992) JIM model of object recognition. Layers 3–7 of the hybrid model are numbered according to the corresponding layers of the JIM model. The model's third layer represents objects (random polygons from Experiments 1 and 3) in terms of their identity (one unit per polygon; "Shape Attrib. Units" in Figure A1, Layer 3), location in the horizontal and vertical dimensions of the visual field (10 units each in Layer 3), and size (10 units in Layer 3). Units in Layer 7 encode collections of polygons in specific arrangements corresponding to entire encoding displays from Experiments 1 and 3 (e.g., Fig. 2 in the main text; see also Figure A1). Units in Layer 4 compute (Layer 4) and represent (Layer 5) the relative sizes and locations of the polygons. Units in Layer 5 store polygons in specific relational role bindings into memory (e.g., "polygon1+*larger*+*left_of*+*below*"; Layer 6), pairs of polygons in specific stacks of relations (e.g., "*larger*+*left_of*+*below* (polygon1, polygon0)"; Layer 6.5), and whole configurations (i.e., encoding displays; Layer 7) into the model's memory.

Units in Layers 1 and 2 serve as "attentional control" units that represent and activate specific polygons at specific sizes and locations in the visual field (e.g., "polygon1+size=3+h_location=2+v_location=3"; Layer 2) and specific pairs of polygons (e.g., "polygon0 and polygon1"; Layer 1). The model attends to a pair of polygons by activating the corresponding Layer 1 unit, which activates the Layer 2 units to which it is connected (e.g., the Layer 1 unit for "polygon0 and polygon1" would activate the Layer 2 units for polygons 0 and 1). Layer 2 units mutually inhibit one another so that, in response to a fixed excitatory input (i.e., from a Layer 1 unit), they will oscillate out of synchrony with one other—for example, with the polygon0 unit firing first, followed by polygon1, followed by polygon0, and so on (see Hummel & Biederman, 1992; Hummel & Holyoak, 1997, 2003). In response to the activation of a

single Layer 1 unit, the result on Layer 3 is two mutually desynchronized patterns of activation: one representing polygon0 in terms of its identity/shape, size, and location, and the other representing polygon1 in terms of its identity/shape, size, and location. These synchrony/asynchrony relations, imposed by the units in Layer 2, are carried forward through Layers 3–6 and represent the bindings of polygons to their basic attributes (Layer 3) and their relations to one another (Layers 5 and 6). Units in Layer 6 learn to respond to conjunctions of units in Layers 3 and 5 (and thus represent polygons in specific relational roles), and units in Layer 6.5 learn to respond to specific conjunctions of units in Layer 6 (and thus represent pairs of polygons in specific relations). Layer 1 units (corresponding to pairs of polygons) are activated, one at a time (two Layer 1 units per display), with the result that the model processes and encodes polygons in pairs. Units in Layer 7 learn to respond to conjunctions of units in Layer 6.5 and thus come to represent pairs of pairs of polygons—that is, approximations of entire encoding displays (as elaborated in the main text).

All the units composing the model are basic leaky integrators (with the exception of those in Layer 1, whose activations are simply set by the user, and those in Layer 4, described below):

$$\Delta a_i = \gamma(a_i - n_i) - \delta a_i, \tag{1}$$

where a_i is the activation of unit i, n_i is the net (excitatory plus inhibitory) instantaneous input to i, and γ and δ are growth and decay rates, respectively. Units in Layer 4 act as AND-gates, which represent conjunctions of metric values and relational roles (e.g., *larger-and-size5*) and take as their activation the product of the activation of the corresponding metric unit in Layer 3 (here, *size5*) and the time-delayed activations of all other relevant Layer 3 units (here, *size1* . . . *size4*—i.e., all sizes smaller than *size5*). In this way, the units in Layer 4 form *comparitor circuits* that take metric values (e.g., specific sizes) as input and produce categorical relations (e.g., larger and smaller) as output on Layer 5 (see Doumas et al., 2008; Hummel & Biederman, 1992). All learning in the model is performed by the simple Hebbian rule:

$$\Delta w_{ij} = f(a_i a_j), \tag{3}$$

where w_{ij} is the (excitatory) connection weight from unit j to unit i. All inhibitory weights in the model have fixed values of −1.0.

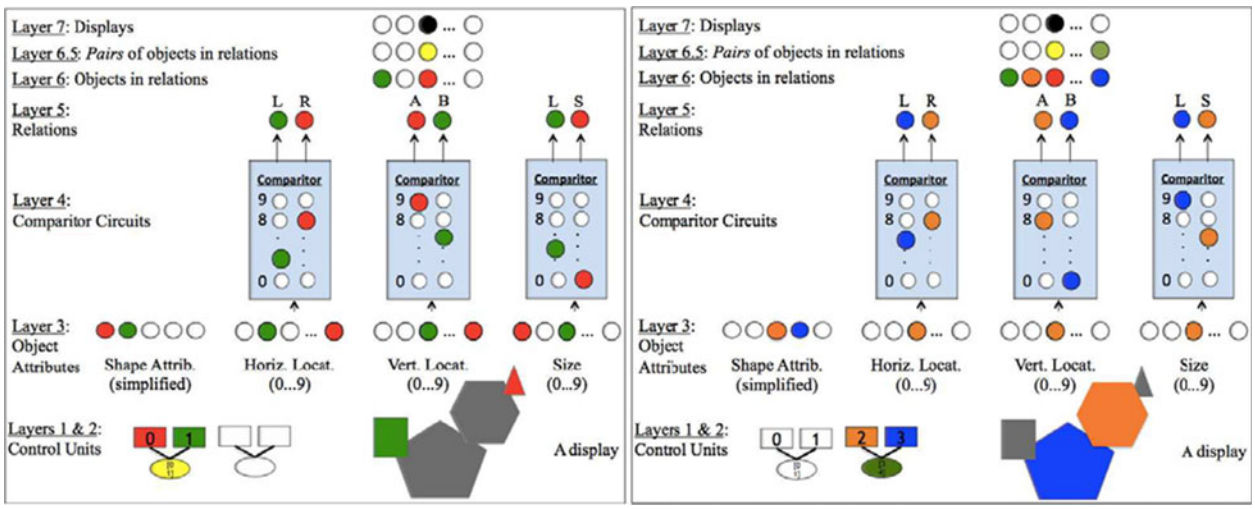

Fig. A1 The model represents objects in pairs. Units filled in the same color are firing (i.e., becoming active) in synchrony with one another; those in different colors are firing out of synchrony. Left panel: The encoding of the pair [0, 1] of the display illustrated in the lower right of the panel. Orange units represent polygon0, those in green polygon1. The yellow unit represents the [0, 1] pair, and the black unit the (emerging) display as a whole. Right panel: The encoding of pair [2, 3]. See the text for details

The model's basic large-scale operations consist of *encoding* (i.e., encoding a display into memory during the encoding phase of Experiments 1 and 3) and *retrieval* (i.e., probed with a pair of polygons, attempting to recover their relations, as in the query phase of Experiments 1 and 3).

During encoding, pairs of polygons are presented to the model (one pair at a time) by activating units in Layer 1. These units activate Layer 2 units, which become active ("fire") out of synchrony with one another, imposing patterns of activation on Layer 3, each of which represents a single polygon in terms of its

shape and metric properties (size and location). These patterns of activation propagate forward through the model's higher layers, with the result that the display is encoded in the mdoel's memory as two pair of polygons (Layer 7), with every relation encoded between the members of each pair (Layers 6 and 6.5).

Retrieval works just like encoding, with the following exceptions. (1) The units in Layer 2 activate representations of the polygons in terms of their shapes ("Shape Attrib." units in Layer 3), but not their locations or sizes. This convention corresponds to our practice (Experiments 1 and 3) of presenting the polygons at query centered on the screen and of equal sizes. (2) Rather than encoding the resulting patterns of activation in Layers 6 . . . 7, existing units in those layers (established during the corresponding encoding phase) are activated (by the Shape Attrib. units in Layer 3) and allowed to feed activation backward, from Layer 7 to Layer 6.5, from 6.5 to 6, and from 6 to 5, activating a representation of the *likely* relations between those polygons in the corresponding encoding phase. In other words, during retrieval, the model attemps to remember or infer what the relation between the polygons had been during the corresponding phase. The relations so activated during this phase are taken as the model's response on that retrieval trial. For example, if, queried with polygons 0 and 1 in Figure A1, the model activates *larger* in synchrony with 1 and *smaller* in synchrony with 0, then we take that pattern of activation as the model responding that 1 had been larger than 0 in the encoded stimulus.

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological science, 15*(2), 106–111.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences of the United States of America, 106*(18), 7345–7350.

Anderson, J., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher-level cognition and its relation to visual attention. *Human-Computer Interaction, 12*(4), 439–462.

Avons, S. E., Wright, K. L., & Pammer, K. (1994). The word-length effect in probed and serial recall. *Quarterly Journal of Experimental Psychology, 47*(A), 207–231.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36*(3), 189–208.

Baddeley, A. D., & Hitch, G. J. (1975). Working memory. *The psychology of learning and motivation, 8,* 47–89.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14,* 575–589.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321*(5890), 851–854.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*(2), 115–147.

Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163–183.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and brain sciences, 24*(1), 87–114. discussion 114–85.

Cowan, N., Wood, N. L., Nugent, L. D., & Treisman, M. (1997). There are two word length effects in verbal short-term memory: Opposed effects of duration and complexity. *Psychological Science, 8,* 290–295.

Doumas, A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological review, 115*(1), 1–43.

Falkenhainer, B., Forbus, K., & Gentner, D. (1989). The structure mapping engine: Algorithm and examples. *Artificial Intelligence, 41*, 1–63.

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition, 122*(2), 210–227.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature neuroscience, 6*(3), 316–322.

Halford, G. S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How many variables can humans process? *Psychological science, 16*(1), 70–76.

Heitz, R. P., Unsworth, N., & Engle, R. W. (2005). Working memory capacity, attention control, and fluid intelligence. *Handbook of understanding and measuring intelligence, 61–77.*

Hitch, G. J., Burgess, N., Towse, J. N., & Culpin, V. (1996). Temporal grouping effects in immediate recall: A working memory analysis. *Quarterly Journal of Experimental Psychology, 49A,* 116–139.

Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual Cognition, 8,* 489–517.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*(3), 480–517.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*(3), 427–466.

Hummel, J., & Holyoak, K. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*(2), 220–264.

Jung, W., & Hummel, J. E. (2009). Learning probabilistic relational categories. In B. Kokinov, K. Holyoak, & D. Gentner (Eds.), *New Frontiers in Analogy Research: Proceedings of the Second International Conference on Analogy.* Bulgaria: Sofia.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic bulletin & review, 9*(4), 637–671.

Levy, B. A. (1971). Role of articulation in auditory and visual short-term memory. *Journal of Verbal Learning and Verbal Behavior, 10,* 123–132.

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance, 20*(5), 1015–1036.

Logan, G. D., & Compton, B. J. (1996). Distance and distraction effects in the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 159–172.

Longoni, A. M., Richardson, J. T. E., & Aiello, A. (1993). Articulatory rehearsal and phonological storage in working memory. *Memory and Cognition, 21,* 11–22.

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

Morrison, R. G. (2005). Thinking in working memory. In K.J. Holyoak & R.G. Morrison (Eds.), *Cambridge Handbook of Thinking and Reasoning* (pp. 457-473). New York, NY: Cambridge University Press.

Morrison, R. G., Holyoak, K. J., & Truong, B. (2001). Working memory modularity in analogical reasoning. In *Proceedings of the twenty-third annual conference of the Cognitive Science Society* (pp. 663–668).

Murray, D. (1968). Articulation and acoustic confusability in short-term memory. *Journal of Experimental Psychology, 78*(4), 679–684.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155,* 23–36.

Peterson, L., & Johnson, S. (1971). Some effects of minimizing articulation on short-term Retention. *Journal Of Verbal Learning And Verbal Behavior, 10,* 346–354.

Roth, J. C., & Franconeri, S. L. (2012). Asymmetric coding of categorical spatial relations in both language and vision. *Frontiers in psychology, 3.*

Saiki, J., & Hummel, J. E. (1996). Attribute conjunctions and the part configuration advantage in object category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1002–1019.

Saiki, J., & Hummel, J. E. (1998). Connectedness and the integration of parts with relations in shape perception. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 227–251.

Simons, D. J., & Ambinder, M. S. (2005). Change blindness theory and consequences. *Current directions in psychological science, 14*(1), 44–48.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception-London, 28*(9), 1059–1074.

Song, J., & Jiang, Y. (2006). Visual working memory for simple and complex features: An fMRI study. *NeuroImage, 30*(3), 963–972.

Tomlinson, M. T., & Love, B. C. (2006). From pigeons to humans: Grounding relational learning in concrete examples. In Proceedings of the National Conference on Artificial Intelligence, 21 (1), 199. Menlo Park, CA; Cambridge, MA; London AAAI Press; MIT Press.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 353*(1373), 1295–1306.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology, 12*(1), 97–136.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. Journal of Experimental Psychology: *Human Perception and Performance, 27*(1), 92–114.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review, 1*(2), 202–238.

Woodman, G. F., & Vogel, E. K. (2008). Selective storage and maintenance of an object's features in visual working memory. *Psychonomic Bulletin Review, 15*(1), 223–229.