Atten Percept Psychophys (2017) 79:1336–1351
DOI 10.3758/s13414-017-1296-7

Numerosity estimates for attended and unattended items in visual search

Troy D. Kelley[1] · Daniel N. Cassenti[1] · Laura R. Marusich[1] · Thomas G. Ghirardelli[2]

Published online: 20 March 2017

Abstract The goal of this research was to examine memories created for the number of items during a visual search task. Participants performed a visual search task for a target defined by a single feature (Experiment 1A), by a conjunction of features (Experiment 1B), or by a specific spatial configuration of features (Experiment 1C). On some trials following the search task, subjects were asked to recall the total number of items in the previous display. In all search types, participants underestimated the total number of items, but the severity of the underestimation varied depending on the efficiency of the search. In three follow-up studies (Experiments 2A, 2B, and 2C) using the same visual stimuli, the participants' only task was to estimate the number of items on each screen. Participants still underestimated the numerosity of the items, although the degree of underestimation was smaller than in the search tasks and did not depend on the type of visual stimuli. In Experiment 3, participants were asked to recall the number of items in a display only once. Subjects still displayed a tendency to underestimate, indicating that the underestimation effects seen in Experiments 1A-1C were not attributable to knowledge of the estimation task. The degree of underestimation depends on the efficiency of the search task, with more severe underestimation in efficient search tasks. This suggests that the lower attentional demands of very efficient searches leads to less encoding of numerosity of the distractor set.

Keywords Visual search · Attention: Divided Attention and Inattention · Memory: Visual working and short-term memory

✉ Troy D. Kelley
troy.d.kelley6@mail.mil

[1] US Army Research Laboratory, Human Research and Engineering Directorate, Adelphi, MD 20783, USA

[2] Department of Psychology, Goucher College, Baltimore, MD, USA

Incidental memory for numerosity of items in visual search tasks

For over 30 years, the visual search task has been a staple laboratory technique for studying attention, because it provides a measure of the cognitive mechanisms underlying visual information processing. In the typical visual search task, the observer must determine if a pre-specified target is present among a set of distractors. A (typically single) target is presented on one-half of the trials, and the required response is a button press indicating whether the target is present or absent. The total number of items in the display (i.e., the *display size*[1]) is varied, and response time (RT) and accuracy are measured.

The linear function relating RT to display size is generally considered diagnostic of visual information processing. The intercept is a measure of the non-search components (e.g., low-level perception), and the slope is a measure of the cost of processing each additional item in the display. Therefore, the slope is an index of processing efficiency and provides evidence for the difficulty of selecting the target from among the non-target (or distractor) items in the display.

If a participant can distinguish the target item from the distractors on the basis of a single feature (e.g., a salient difference in color, orientation, or shape), the search for the target will be most efficient (Wolfe, 1998a). In this type of highly efficient *feature search*, the function relating RT to display size will be relatively flat; that is, it will have a slope that is not significantly different from zero. If the target has a feature that is different enough from the distractors, the target appears to *pop out* from the rest of the display (Treisman & Gelade, 1980).

[1] Throughout this paper we will use both display size and numerosity interchangeably to refer to the number of elements in the search display in order to be consistent with both the visual search and numerical estimation literature. The two terms mean precisely the same thing here.

In a *conjunction search*, target identification requires the conjoining of features, such as a particular color and a particular orientation (e.g., a red vertical line displayed among distractors consisting of both red horizontal lines and green vertical lines). Results from a conjunction search show a linear increase in the slope of the line relating RT to display size. Depending on the nature of the underlying feature differences, the RT by display size slopes can show substantial variation, and therefore these searches can be more or less efficient (Wolfe, 1998b).

In a *spatial configuration search*, the targets and distractors possess the same basic features (e.g., items that are defined by the conjoining of two perpendicular lines such as searching for a rotated letter T among rotated letter Ls). The defining difference between them is how the features are configured. These searches are generally quite difficult, and in some cases attention must be allocated to each item individually in order to identify a target (Gilden, Thornton, & Marusich, 2010). The slope relating performance to display size is linear and typically steep, on the order of 20–30 ms per item on target present trials and 40–60 ms per item on target absent trials (Wolfe, 1998a). Spatial configuration searches are the least efficient of the three types of searches described here.

The flat RT by display size slopes typical of a feature search have historically been taken as evidence that distractors were not processed in any meaningful way during these highly efficient searches, and that targets were identified "preattentively" (e.g., Treisman & Gelade, 1980). If the number of distractors does not affect RT, the implication is that early, bottom-up processes segment the target from the distractor field and that top-down allocation of attention to distractors is unneeded. In these cases, observers would presumably have little awareness of, or memory for, the distractors in a feature search.

Indeed, past work has postulated amnesic, or memory-less search (Horowitz & Wolfe, 1998). In a series of studies, Horowitz and Wolfe (1998, 2001, 2003) explicitly tested the premise of memory for the *location* of distractors using the randomized search paradigm. In this methodology, search displays are presented in either a dynamic or static condition. In both conditions, the content of the display remains the same, but in the dynamic condition the locations of the stimuli change from frame to frame or roughly every 100 ms. The assumption is that if observers use memory to avoid returning to previously attended and rejected distractors, performance in the static condition will be much faster than in the dynamic condition, when it is impossible to keep track of the locations of previously attended targets. However, even in some of the least efficient spatial configuration searches, Horowitz and Wolfe found no difference in the RT by display size slopes for the dynamic and static conditions. The researchers concluded that observers could not, or did not, keep track of the locations of rejected distractors even in static displays where it was possible to do so.

In contrast, other researchers have found evidence supporting memory for the *location* of distractors during a visual search. (Kristjansson, 2000; Peterson, Kramer, Wang, Irwin, & McCarley, 2001; Shore & Klein, 2000). Kristjansson used a modification of the randomized search task and found that dynamic search slopes were significantly steeper than static search slopes, at least for larger display sizes, supporting the role of memory in search (but see Horowitz & Wolfe, 2003). Additionally, Peterson et al. showed that the pattern of re-fixation of distractors during a visual search task (i.e., the number of distractors *revisited*) did not fit the predictions of the memory-less search model. These findings led Le-Hoa Võ and Wolfe (2015) to a revised Guided Search model that incorporated inhibition of return to reduce the likelihood that previous distractors were revisited. Interestingly, several revisitation studies found that inhibition of return to distractor locations was greater in inefficient searches than in efficient searches in inefficient searches (e.g. Klein, 1988; Klein & MacInnes, 1999). These findings suggest that memory for distractors in visual search may be moderated by, among other things, the efficiency of the search.

In the current work we explicitly test the hypothesis that memory for distractors may be poorest in highly efficient searches, but that in less efficient conjunction or spatial configuration search where distractors are processed to a greater extent, they may be better represented in memory. The focus of the previously cited work was re-visitation as an indication of the memory for the location of distractors in the display, while memory for general properties of the distractor set was not assessed directly. Here we examine the type of memories created for the numerosity of items during visual search. We anticipate that findings of memory for item numerosity, as a general property, may differ substantially from memory for item location during visual search.

Our research was conducted as part of the development of models of visual search using the Adaptive Control of Thought–Rational (ACT-R; Anderson et al., 2004) and later for the development of the Symbolic and Sub-symbolic Robotics Control System (SS-RICS; Kelley, 2006). Our initial attempts to model visual search tasks in ACT-R yielded counterintuitive results. For example, at the time (2006), the ACT-R architecture only created declarative memories for items that were directly attended. A pop-out search task could be re-created using ACT-R, but the model would yield only one declarative memory element, the target. If a modeler was to query ACT-R as to the number of items on the screen, the ACT-R model would produce "one," for the target. It is clear that human observers performing a pop-out search would recall that there were more items on the screen than just a single target, even if they were unable to recall an accurate estimate. This counterintuitive behavior of the ACT-R model led us to an investigation of human memory for distractor numerosity in visual search.

To assess participants' memory for the numerosity of items in the display, we asked them to estimate the number of items presented as one of several possible probe questions appearing after the search response on a subset of trials. Such numerical estimation tasks have been used previously to investigate numerosity perception and the representation of numerosity (e.g., Hollingsworth, et al., 1991). A complete review of numerosity for attended items is beyond the scope of this manuscript, but one finding from the numerical estimation literature is particularly germane; namely that in displays exceeding the subitizing limit of 1 to 5 items (e.g., Mandler & Shebo, 1982), humans can quickly report the approximate number of items in a display without counting, but these reports tend to be underestimations (Hollingsworth et al., 1991). It is important to note that in those and many other studies of numerical estimation (e.g. Izard & Dehaene, 2008; Gebuis & Reynvoet, 2012; Utochkin, 2015), participants were explicitly instructed to attend to the items, while we are investigating memory for the numerosity of items which are not attended to directly.

Specifically, our hypothesis is that observers have at least a general memory (i.e., better than chance) for the number of items in a visual display, and that the accuracy of the representation of the whole display will be related to the efficiency of the search, where efficiency can be seen as a measure of attention. As search efficiency increases, less attention is required to segment a target from distractors, leading to less accurate memories for the search display. Here we hypothesize that participants engaged in a feature search will have some memory of the numerosity, but will consistently underestimate the distractors (consistent with the numerical estimation literature). We further predict that in a less efficient conjunction search, the degree of underestimation will be reduced and reduced further in the least efficient spatial configuration search.

In this paper, we describe three phases of experiments. The first set (Experiments 1A, 1B, and 1C) explored the accuracy of numerosity judgments in feature search, conjunction search, and a spatial configuration search where participants' primary task was visual search and not attending to the items in the display. The second set of experiments (2A, 2B, and 2C) provided a comparison and control to the first set for *attended* items in the display. Using the same visual stimuli as the first set of experiments, participants only performed the numerical estimation task within an allotted amount of time (they did not search for targets). The final phase (Experiment 3) was designed as an additional control to the first set of experiments. In the first set of experiments, subjects could have gained awareness in the course of the experiment that the numerical estimation task was a secondary task. Therefore, in Experiment 3, we studied numerical estimation performance within a search task where participants made only a *single* numerosity judgment with no foreknowledge of this request. This eliminated the possibility that subjects inadvertently discovered the secondary task of attending to the numerosity of the distractors, as in the first set of experiments.

Experiment set 1

1A. Feature search

In Experiment 1A, participants performed a standard feature search and answered probe questions about the display. These probe questions included estimates of the number of items that had appeared on the display, as well as several other filler questions. The infrequency of the probe questions and the use of filler questions were designed to discourage participants from explicitly attending to the number of items in the display.

Method

Participants For all of the experiments contained in this manuscript participants were asked to sign fully informed consent volunteer agreement affidavit. The statement explained to participants the purpose, procedures, benefits, risks, and confidentiality of the data collected during the experiment. Twenty civilian employees of the US Army Research Laboratory volunteered for this experiment without compensation. In this and all laboratory experiments reported here, all participants were screened for at least 20/40 visual acuity (corrected or uncorrected) in both eyes and normal color vision.

Stimuli The stimuli consisted of US Army symbols. The target was a blue US Army symbol for an Infantry Platoon and distractors were red US Army symbols for Mechanized Infantry Platoons, presented on a white background. Sample stimuli are presented in Fig. 1. Targets were defined for the participants as a symbol that appeared in a different color from the rest of the items in the display (e.g., the blue symbol among red symbols). All non-target items were homogenous distractors.

Each symbol measured 1.7° by 1.0°. Stimuli were drawn to a grid and did not overlap. The boxes in the grid were 1.7° by 1.7°, which provided space around the upper and lower border of each symbol stimuli in the grid. The symbols were placed randomly within each grid location so that the stimuli did not appear to be in columns or rows. Each display consisted of three to 50 items. There was never more than one target presented.

Apparatus [2] Stimuli were presented via a Windows-based computer with a 19-in. CRT monitor set to a resolution of 640 × 480 pixels. This resolution yields a dot pitch of 0.31 mm/pixel. Stimulus presentation was controlled and

[2] Experiment 1A was conducted at the US Army Research Laboratory. Experiments 1B, 1C, and 2A-C, were conducted at Goucher College. Data collection for Experiment 3 was administered from the US Army Research Laboratory and conducted online via Amazon Mechanical Turk. The apparatus and set up in the two labs were nearly identical and the software was identical. The only difference in the set up was that the viewing distance at ARL was fixed at 45 cm while it was 57 cm in the Goucher lab and not controlled in Experiment 3. For this reason, stimulus and display dimensions are given in degrees visual angle throughout.

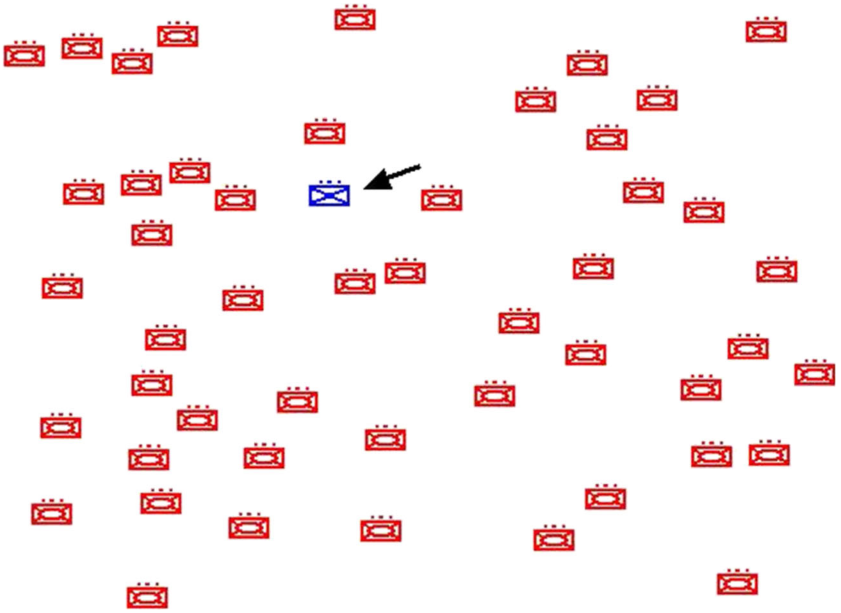


Fig. 1 Screenshot of a visual search display with 49 distractors and one target in Experiment 1. The arrow points to the target

participant responses (RT and accuracy of target detection for each trial) were collected using a program written with the E-Prime experimental presentation software version 1.1 (Psychology Software Tools, Inc.) and a response box (SR Box-Deluxe, Psychology Software Tools, Inc.). The response box had five buttons each labeled with a number from 1 to 5, ordered left to right, and was connected to the computer via the serial port.

Each participant was seated in a comfortable height-adjustable chair with their chin in a chin rest (Model#14302, Lafayette Instrument) affixed to the table in front of the participant. The response box was positioned behind the chin rest. The experiment took place in a dimly-lit experimental room.

Procedure Participants completed an informed consent form and then received written and oral instructions. The instructions emphasized maintaining fixation on the fixation cross prior to each trial and to start responding as quickly and as accurately as possible when they detected the presence of the target or decided there was no target present. Participants were instructed to press the rightmost button if the target was present and to press the leftmost button if there was no target present using the index fingers of each hand. They were also instructed that on some trials, a question about the display would appear after the search display.

Each trial began with the presentation of a small black fixation cross (a plus sign in 18-point Courier New font measuring 1.4° horizontally and vertically) for 2,000 ms, followed by the search display. The search display remained visible until the participant responded. Accuracy and RT feedback immediately followed the response and was displayed for 1,500 ms. If responses took more than 2,000 ms, participants were given feedback by the computer to work more quickly on future trials. Data from these trials were excluded from the analysis.

On trials with a question, the question was displayed in the center of the screen with all corresponding options and response keys listed directly below the question. All text was in 18-point Courier New font. Participants were instructed to answer the questions as accurately as possible by pressing the corresponding button on the response box. Following the question response, the next trial began automatically with the appearance of the fixation cross.

There were a total of 1,200 trials, given in four sessions. This included 900 non-probe trials, so that probe questions were asked on 25% of trials. A target was present on one-half of the trials. All participants received 30 practice trials in their first session and one practice trial for each of the remaining three sessions.

There were five possible probe questions with options. Among these, we were interested in only the first question, the *numerical estimation* question, which asked participants how many items were present on the previous screen. Participants could choose between five categories: 1–10, 11–20, 21–30, 31–40, and 41–50. The remaining probe questions were filler questions used to disguise the true purpose of the experiment (e.g., in what quadrant of the screen was the target located, what color was the target, what color were the distractors, etc.). When a probe question referred to a target, it was always a target-present trial. When the probe question was the numerical estimation question, there were always 6, 16, 26, 36, or 46 items. Each of these display sizes was presented 15 times per participant for this question. The display size in all other trials was chosen randomly with limits between three and 50 items.

Results

In this and all experiments reported here, all practice trials were eliminated from all analyses. In addition, all 240 trials in which the search task took more than 2,000 ms were eliminated from all analyses (1.01% of all trials). This was done to eliminate trials in which participants may have been explicitly counting items.

Search task Error rates were 2.4% on target present trials and 1.3% on target absent trials. An analysis of variance (ANOVA) with display size as the sole factor run on error rates revealed no significant effects for either target absent $F(4, 76) = 1.42, p > .05$, or target present trials $F(4, 76) = 1.56, p > .05$.

We binned display size into five categories (3–10 items, 11–20 items, 21–30 items, 31–40 items, and 41–50 items). We analyzed mean RT on correct trials as a function of display size, separately for target present and target absent trials. An ANOVA revealed a significant effect of display size on RT for target absent trials, $F(4, 76) = 12.53$, $MSE = 8209$, $p < .001$, but not for target present trials, $F(4, 76) = 2.25$, $MSE = 1178$, $p > .05$. The left panel of Fig. 2 (1A) shows mean correct RT as a function of target condition (present or absent) and display size for the search task.

We calculated the mean slope of the RT × Display size function and found the slope to be −0.3 ms per item for target present trials and −1.2 ms per item for target absent trials. Slopes were significantly lower than zero for present trials, $t(19) = -2.19$, $SE_M = 0.14$, $p < .05$, and for absent trials, $t(19) = -4.85$, $SE_M = 0.25$, $p < .001$.

Numerical estimation We analyzed numerical estimation only for trials in which the search response was correct (98% of trials) and the target was present (50% for all experiments). We first assessed if participants performed better than chance, or if they were merely guessing the display size. On each estimation, participants chose among five options (1–10, 11–20, 21–30, 31–40, or 41–50). The actual number of items displayed on the previous screen was always 6, 16, 26, 36, or 46. As a result, each response option represents a different *absolute error*, the absolute value of the difference between the correct number of items and the midpoint of the response range chosen. For example, if a participant chose "11–20" after a display size of six items, the absolute error would be $| 6 - 16 | = 10$. A probability value was derived for each absolute error value given the number of items that actually appeared. The details of this derivation appear in Appendix 1. The probability of each absolute error was multiplied by its corresponding difference score and resulted in an average guess difference of 16. This number represents chance performance, the expected absolute error resulting from random guessing. We calculated the mean absolute error ($M = 10.44$) and compared it to the value hypothesized by random guessing (16) using a single-sample t-test. We found a significant difference, $t(19) = 10.05$, $p < .001$, indicating that participants did not merely guess the display size and performed better than chance.

In order to examine the type of errors made (over or underestimation), we calculated the *signed error* of the response from correct. The absolute value of the signed error indicates how accurate participants were in estimating the numerosity, with larger signed errors indicating worse performance. The sign tells us the direction of the error. A signed error was positive when the response overestimated the correct answer and negative when the response underestimated the correct answer. The nature of the response (i.e., choosing among five number ranges) meant that there were unequal chances to commit an underestimate or an overestimate depending on the previous display size. For example, on a display size of 6, there were no response options lower than the correct one, so it was not possible to underestimate. Similarly, it was not possible to overestimate on display sizes of 46. Because the opportunity to make positive vs. negative

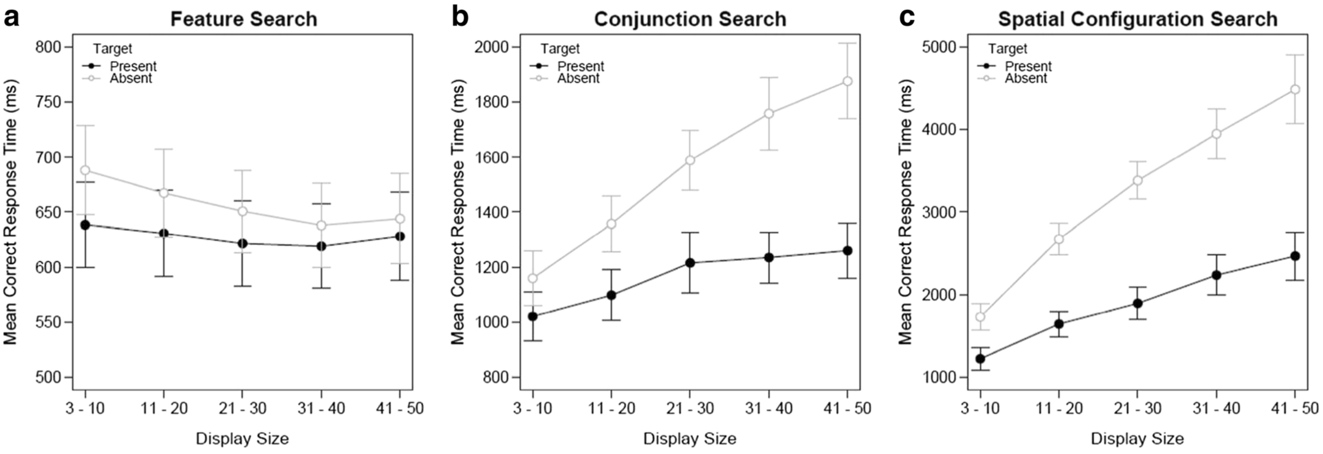


Fig. 2 Mean correct search response time (RT) as a function of target condition and display size in Experiment Set 1. Error bars represent standard error of the mean

signed errors was only balanced in the aggregate and not within individual display size conditions, we averaged the mean signed error across all display sizes. These overall signed error means were used in a single-sample t-test and compared with a test statistic of zero, which represented a balance of underestimates and overestimates with correct responses averaging only zeroes. For further information on the distribution of participant responses and signed error at each set size for this and subsequent experiments, see Appendix 2. The mean signed error ($M = -8.44$) in this experiment was significantly less than zero, $t(19) = -9.28$, $SE_M = 0.91$, $p < 0.001$, meaning that on average, participants underestimated the number of items in the display. In addition, we calculated the signed error at just display size 26, the only display size where participants had equal opportunity to over or underestimate. Here again, we found significant underestimation, $M = -7.57$, $t(19) = -5.69$, $p < 0.001$.

Discussion

Time to find a color feature target decreased as a function of display size. This pattern has been observed before and is known as the *absence advantage* (e.g., Humphreys, Quinlan, & Riddoch, 1989; see also Bacon & Egeth, 1991). Search accuracy was unaffected by display size. This pattern of results is typical of feature searches, suggesting that participants were performing the visual search task as instructed.

When asked to estimate the number of items in the display, participants performed significantly better than chance, indicating that they did not merely guess and that they retained some information about distractors, even in a feature search. However, they consistently underestimated the display size, in terms of both absolute and percent signed error. The results indicate that participants retain some information about distractor items in a visual search display because they performed at better than chance levels when asked to report the numerosity of items in the display.

Duncan and Humphreys (1989) have proposed that the mechanism by which efficient search takes place is that homogenous distractor items are grouped together and rejected as a whole rather than by any serial means of rejection. This account may help explain why participants had some notion of the number of items in the display and is consistent with previous studies showing that factors that facilitate grouping of distractors can result in increased search efficiency. For example, Bacon and Egeth (1991) manipulated the density of distractors around a target in a feature search task. They found decreasing search RT as a function of increasing display size, and therefore increasing distractor density. This suggests that increased grouping of homogenous distractors results in more efficient search, possibly because of the ease with which they are rejected from search. This argument is similar to that proposed by Duncan and Humphreys. As mentioned in the Introduction, our hypothesis is that search efficiency and numerical estimation accuracy are inversely related. The remaining experiments were designed to further test this hypothesis.

1B. Conjunction search

In Experiment 1A, we found that participants consistently underestimated the numerosity of items in the display. Experiment 1B investigated the underestimation effect when performing a less efficient search. If the underestimation found in Experiment 1A was the result of a very efficient search for the target, then a less efficient search should result in better estimations. In Experiment 1B, we used a typical conjunction search as the primary task and again measured participants' estimation of numerosity.

Method

Participants Participants were 21 undergraduate students at Goucher College who participated in exchange for extra credit in one of several psychology classes. One participant was removed for failing to meet the requirements for visual acuity and one for having a color deficiency, leaving a final sample of 19 participants. Nine participants searched for a red vertical bar and ten participants searched for a green horizontal bar among green vertical and red horizontal bars.

Stimuli The stimuli consisted of red vertical bars, green vertical bars, red horizontal bars, and green horizontal bars. Each item was approximately 1.3° by 0.1°. The stimuli were constrained as in Experiment 1A. As in Experiment 1A, each display consisted of three to 50 items, and there was never more than one target presented.

Apparatus The apparatus was the same as in Experiment 1A.

Procedure The procedure was the same as in Experiment 1A except that the instructions now described the targets and distractors in the conjunction search. In addition, we reduced the total number of trials to 400 trials per participant. Probe questions appeared on 75% of trials, but numerical estimation questions still were presented on only 15% of trials. A target was present on one-half of the trials. All participants received 30 practice trials.

Results

Search task Error rates were 4.1% for target-absent trials and 11.8% for target-present trials. ANOVA with display size as the sole factor on error rate as a function of display size revealed no significant effect for target absent trials $F(4, 72) = 1.35$, $p > .05$, but did find a significant effect for target present trials $F(4, 72) = 6.53$, MSE = 0.018, $p < .001$. On target present trials, participants were more accurate on trials with smaller display sizes.

We binned display size into the same five categories as in Experiment 1A and ran the same ANOVA. The ANOVA revealed a significant effect of display size on RT for target present trials, $F(4, 72) = 11.73$, $MSE = 205656$, $p < .001$ and for target absent trials, $F(4, 72) = 30.56$, $MSE = 1624297$, $p < .001$. The middle panel of Fig. 2 shows mean correct RT as a function of target condition and display size for the search task.

The mean slope of the RT × Display size function was 6.8 ms per item for target present trials and 18.9 ms per item for target absent trials. Slopes were significantly different from zero, for present trials, $t(18) = 4.99$, $SE_M = 1.37$, $p < .001$ and for absent trials, $t(18) = 7.56$, $SE_M = 2.51$, $p < .001$.

Numerical estimation We used the same procedure as in Experiment 1A to determine if participants were merely guessing the numerosity of the display. As in Experiment 1A, a one-sample t-test of mean absolute error ($M = 9.18$) found a significant difference from 16, $t(18) = -12.98$, $p < .001$, indicating that participants did not merely guess. As in Experiment 1A, we calculated a mean signed error and analyzed the results using a single sample t-test. The results again showed an overall tendency to underestimate ($M = -5.76$), $t(18) = -7.44$, $SE_M = 0.77$, $p < 0.001$. With just the display size of 26 items ($M = -4.53$), there was also significant underestimation, $t(18) = -4.78$, $p < 0.001$.

Discussion

Experiment 1B showed the typical conjunction search result of increased search RT with increased display size. Participants were more accurate on the search task with smaller display sizes, but because they were also faster on these trials, this pattern showed no speed-for-accuracy trade-off.

Participants again performed at better than chance when estimating the display size in Experiment 1B. They underestimated the numerosity of the display, however, the magnitude of underestimation in Experiment 1B (−5.76) was smaller than in Experiment 1A (−8.44; see below for cross-experiment inferential statistics). The decreased underestimation is consistent with the hypothesis that participants in this experiment created more detailed memories than in the feature search study and therefore should have been better at individuating items.

1C. Spatial configuration search

Experiment 1B showed that participants were more accurate at estimating the numerosity of the display when the search task was less efficient. Experiment 1C was designed as an extension of the reasoning in Experiment 1B, but used a spatial configuration search task. That is, if reduced search efficiency leads to the underestimation effect, then by further reducing the efficiency of the search, we should observe a better estimation of items in the display, and an even smaller underestimation effect than that observed in Experiment 1B.

In a spatial configuration experiment, targets and distractors differ from one another only in subtle configuration changes, resulting in search that is attention demanding. Experiment 1C was the same as Experiment 1B except for the following changes. Instead of the green or red and horizontal or vertical lines, block numbers 2 and 5 were used as distractors and targets, the traditional stimuli for spatial configuration searches.

Method

Participants Participants were 23 undergraduate students at Goucher College who volunteered in exchange for extra credit in one of several psychology classes. Two participants were removed for failing to meet the requirements for visual acuity. One other participant was removed for failing to complete the experiment, leaving a final sample of 20 participants.

Stimuli The stimuli consisted of black block number 2 and block number 5. Each item was approximately 1.3° by 1.0°. Each display consisted of three to 50 items. Eleven participants searched for the block number 2 among the block number 5 distractors and nine searched for the block number 5 among block number 2 distractors. Stimuli were constrained as before, and again each display consisted of three to 50 items. There was never more than one target presented.

Apparatus The apparatus was the same as in Experiments 1A and 1B.

Procedure The procedure was the same as in Experiment 1B with the exception that the instructions now described the new stimuli in the spatial configuration search. In addition, the two questions asking about the color of the target and of the distractor were replaced with questions about what numeral the target and distractor resembled, respectively.

Results

Search task The same ANOVA with the display-size factor as in Experiments 1A and 1B was run on binned RT for all correct trials and revealed a significant effect of display size on RT for target absent trials, $F(4, 76) = 67.04$, $MSE = 23500309$, $p < .001$ and for target present trials, $F(4, 76) = 45.27$, $MSE = 5140424$, $p < .001$. The right panel of Fig. 2 shows mean correct search RT as a function of target condition and display size.

The mean slope of the RT × Display size function was 32.2 ms per item for target present trials and 69.7 ms per item for target absent trials. Slopes were significantly different from zero for present trials, $t(19) = 7.32$, $SE_M = 4.39$, $p < .001$ and for absent trials, $t(19) = 8.87$, $SE_M = 7.87$, $p < .001$.

Error rates were 4.3% overall for target-absent and 6.5% for target-present. ANOVA on error rate as a function of display size revealed no significant effects for absent trials $F(4, 76) < 1$, but did reveal a difference for target present trials $F(4, 76) = 6.61$, $MSE = 0.015$, $p < .001$. For target present trials, participants were again more accurate on trials with a smaller display size.

Numerical estimation As in the previous experiments we performed the analysis to determine if participants were merely guessing the display size. The t-test of the mean absolute error ($M = 7.27$) found a significant difference from 16, $t(19) = -16.30$, $p < .001$, indicating that participants performed at better than chance when estimating display size. Again the mean signed error showed an overall tendency to underestimate ($M = -4.25$), $t(19) = -4.68$, $SE_M = 0.91$, $p < 0.001$. With just the display size of 26 ($M = -3.28$), there was also significant underestimation, $t(19) = -3.51$, $p < .01$.

Discussion

Participants showed the typical RT pattern for an inefficient (spatial configuration) search. As in Experiment 1B, participants were more accurate with smaller display sizes showing no speed-accuracy trade-off. They were again better than chance at estimating the numerosity of the display. Participants in this experiment demonstrated a tendency toward small underestimations, with an average absolute error of about −4 items.

The critical question is whether search efficiency influences memory for distractor numerosity; that is, do participants perform better at estimating the number of distractors when the search is more difficult? In order to assess differences in estimation errors across the three search experiments, we used a one-way ANOVA across search types. We found that search type did have a significant effect on estimation errors, $F(2,56) = 5.90$, $MSE = 16.06$, $p < .01$, with the largest estimation errors in the most efficient feature search, decreased error in conjunction search, and the smallest estimation errors in the least efficient spatial configuration search (see left side of Fig. 3). Linear regression across feature, conjunctive, and spatial configuration searches with estimation error showed a significant slope parameter (2.15), $t(57) = 3.41$, $SE = 0.63$, $p < .01$, bearing out the conclusion. These results support the hypothesis that less efficient search leads to better set-size estimation.

A possible alternative explanation for the difference in estimation performance across the three experiments concerns display time. That is, participants made better estimates when performing a more difficult search; however, because the display remained visible until a response was made, the display was presented for a longer period of time. It may be the case

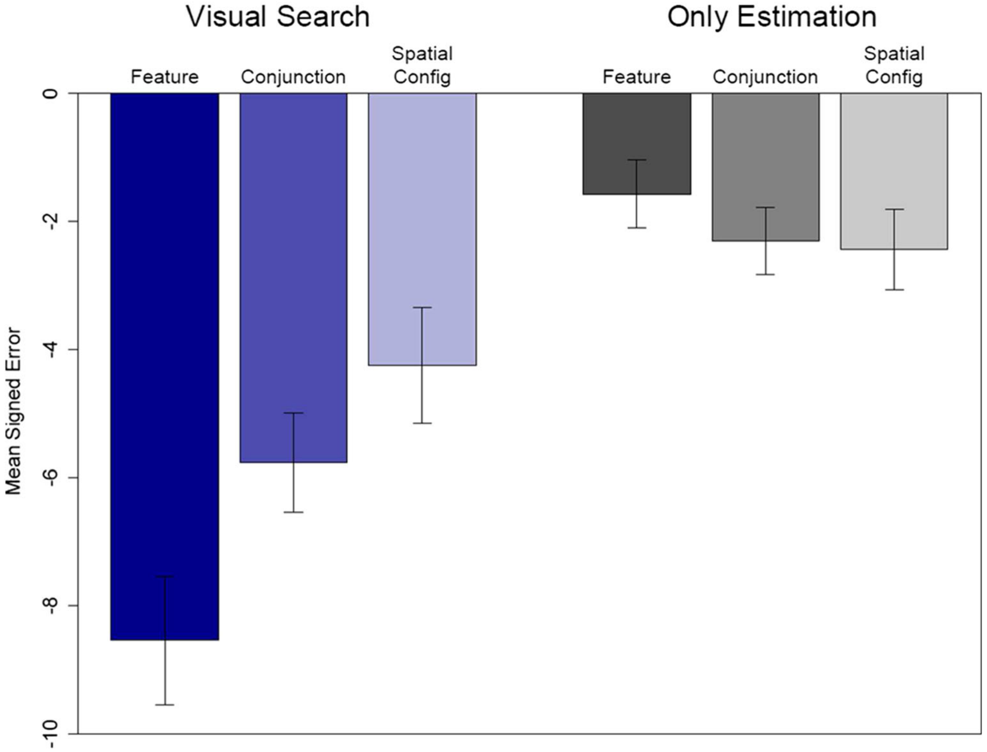


Fig. 3 Mean signed error across Experiment Sets 1 and 2. Error bars represent standard error of the mean

that estimation performance naturally improves when participants spend longer looking at the display. We assessed correlations between individual participants' RT and estimation error for each study in Experiment Set 1 and found no evidence that participants with longer RTs made smaller estimation errors (Experiment 1A: $r = -0.34$, $p > .05$; Experiment 1B: $r = -0.21$, $p > .05$, Experiment 1C: $r = 0.38$, $p > .05$). This suggests that our findings are not explained by longer looking times alone. However, we conducted a more explicit test to rule out this potential confound in Experiment Set 2, described below.

Experiment set 2

In the first three experiments, the primary task was a visual search task, using a standard feature search (1A), a standard conjunction search (1B), and a spatial configuration search (1C). Experiments 2A, 2B, and 2C were designed to assess participants' estimation performance using the same stimuli as in Experiments Set 1, however, the primary task was to estimate the number of items in the display. The search task was not performed. The purpose of this was to test if the variation in underestimation severity across Experiment Set 1 was due only to the nature of the stimuli used, and not the different attentional demands required by performing the search task with those stimuli.

In addition, Experiment Set 2 afforded the opportunity to rule out the potential confound of variable display times in different search types, described above. In Experiment Set 2, visual displays were presented for fixed durations in order to investigate whether the underestimation effect observed was related to search efficiency or to the amount of time the display was presented. The durations were chosen to approximate the RTs (and therefore display durations) from Experiments 1A, 1B, and 1C respectively. These experiments were designed to measure estimation performance when observers were deliberately estimating the number of items in the display and to test the effect of the amount of time available to view a display on this performance.

Experiment 2A

Method

Participants Fourteen undergraduate students from Goucher College volunteered in exchange for extra credit in one of several psychology classes. Eight military personnel were recruited from US Army Research Laboratory and volunteered without compensation. One undergraduate student and one military personnel were removed from the analysis due to failure to meet the visual acuity requirements leaving a final sample of 20 participants.

Stimuli and apparatus The stimuli and apparatus were the same as in Experiment 1A except that the SR Box was abandoned in favor of a standard keyboard because RT was not emphasized.

Procedure The procedure was the same as in Experiment 1A except for the following ways. First, participants were instructed to estimate the total number of items in the display and were not instructed to perform a visual search for a target. The search response was therefore eliminated. A target was presented in all displays. Second, the duration of the displays were either 250, 650, or 1,050 ms, manipulated within-participants as opposed to being displayed until the participant responded. These durations were chosen because they approximated the mean RT and 1.5 standard deviation units above and below the mean RT from Experiment 1A. One-third of the trials were presented at each duration. Each presentation of the search display was followed by the same numerical estimation question display from Experiments 1A–C, and participants were asked to press a key to respond. After each response, the next trial began with the 2,000 ms presentation of the fixation cross. Overall there were 75 trials, with five trials of each combination of numbers of distractors (6, 16, 26, 36, and 46 items) and display duration (250 ms, 650 ms, or 1,050 ms). Accuracy feedback was provided after all trials were completed.

Results

Numerical estimation A one-way ANOVA with the three display duration conditions was run on signed error. The ANOVA for signed error revealed no significant effect of display duration, $F(2, 38) = 2.96$, $MSE = 1.45$, $p > .05$. The same test of mean signed error from the earlier experiments showed significant underestimation at all display durations. Comparing mean signed error in this control experiment with the search experiment (Experiment 1A) that used the same stimuli, we found that estimation performance was significantly better in Experiment 2A ($M = -1.57$) than in Experiment 1A ($M = -8.44$), $t(38) = -6.51$, $SE_M = 1.05$, $p < .001$.

Discussion

The purpose of Experiment 2A was to measure numerical estimation performance when participants were deliberately estimating the number of items and to test whether the amount of time the display was viewable affected estimation performance. Participants still underestimated, however, the reduced underestimation in Experiment 2A compared to Experiment 1A indicates that when participants did not perform the search task, they showed better performance in estimating the numerosity of the display. Thus, we conclude that the estimation performance in Experiment 1A was dependent on attentional factors associated with the search task.

Experiment 2B

Method

Participants Twenty-four undergraduate students from Goucher College volunteered in exchange for extra credit in one of several psychology classes.

Stimuli and apparatus The stimuli and apparatus were the same as in Experiment 1B.

Procedure The procedure was the same as in Experiment 2A except that a 2,150 ms display duration was added to account for the longer search times needed for conjunction searches. This time was added to the 250, 650, and 1,050 ms durations from Experiment 2A and there were now 100 trials (five trials at each display size and display duration combination).

Results

Numerical estimation The same one-way ANOVA as in Experiment 2A (only with four display durations) was run on signed error. The ANOVA revealed no significant effect of display duration on signed error, $F(3, 69) < 1$, $MSE = 1.45$. The same signed error test showed significant underestimation across all display durations. Again, we compared mean signed absolute error in this control experiment with the search experiment that used the same stimuli and found that estimation performance in Experiment 2B ($M = -2.32$) was significantly better than in Experiment 1B, ($M = -5.76$), $t(41) = -3.87$, $SE_M = 0.89$, $p < .001$.

Discussion

As in Experiment 2A, the reduced underestimation in Experiment 2B compared to 1B suggests that participants are better at estimating the numerosity of the display when the requirements of the search task are removed. With display duration showing no effect on numerical estimation, the evidence suggests estimation performance is driven by attentional factors and not the time available to view the display.

Experiment 2C

Method

Participants Fourteen undergraduate students from Goucher College volunteered in exchange for extra credit in one of several psychology classes. One participant was removed from the analysis for failure to meet the visual acuity requirements leaving a final sample of 13 participants.

Stimuli and apparatus The stimuli and apparatus were the same as in Experiment 1C.

Procedure The procedure was the same as in Experiment 2A except for the duration of the displays were either 250, 650, 1,050, 1,850, or 3,750 ms, and the number of trials was increased to 125 in order to have five trials at each display size and display duration combination.

Results

Numerical estimation The same one-way ANOVA as in Experiment 2A (only with five display durations) was run on signed error. The ANOVA revealed no significant effect of display duration, $F(4, 48) < 1$, $MSE = 0.976$. The same signed error test from Experiments 2A and 2B was run on the duration conditions and showed a tendency for underestimation at all display durations. There was no difference between mean signed error in Experiment 2C ($M = -2.79$) and Experiment 1C, ($M = -4.25$), $t(31) = -1.17$, $SE_M = 1.25$, $p > .05$.

Discussion

Unlike Experiments 2A and 2B, we found no significant difference in numerical estimation performance between participants who also performed the search task and those who did not. As with Experiments 2A and 2B, we found no effect of display duration, suggesting that the underestimation effect is not dependent on to the time available to view the display.

Underestimation across types of searches

We conducted a 2 (task type) by 3 (visual stimuli) ANOVA across the six experiments reported above. The findings showed a significant main effect of task type ($F(1,111) = 42.85$, $p < 0.001$), indicating that estimation errors were much smaller in the attended estimation tasks than in the search tasks, and a marginal main effect of visual stimuli, $F(2,111) = 2.46$, $p = 0.09$. Critically, there was a significant interaction between these two variables, $F(2,111) = 5.86$, $p < .01$. As shown in Fig. 3, this interaction indicates that the type of stimuli presented had an effect on estimation error, but only when engaged in the search task, not when the goal was to estimate quantity. The decreasing underestimation effect with increasing search efficiency is shown in the linear regression results above. The same linear regression with task type in the numerical estimation task did not demonstrate a significant slope ($p > 0.25$). This suggests that search efficiency, and not simply the type of visual stimulus observed, determines the extent of underestimation.

This analysis, together with the comparisons of mean signed error between each search experiment and its matched control experiment using the same stimuli, lends support to the claim that search efficiency affects numerical estimation performance. When attention is directed to the whole display as in the control experiments, or because a demanding search task requires it as in Experiment 1C, participants create a more accurate representation of the display in memory, which increases the ability to estimate the number of items present. When performing an efficient search as in Experiment 1A, participants do not have an accurate representation of the display and perform worse in the numerical estimation task. A search with an intermediate efficiency such as the conjunction search in Experiment 1B, leads to intermediate numerical estimation performance.

Experiment 3

In Experiments 1A, 1B, and 1C, the search performance results were consistent with those typically found in feature, conjunction, and spatial configuration searches, respectively. This provided evidence that observers' search behavior was not significantly modified by the addition of occasional probe questions, and that they were not devoting attentional resources to estimating numerosity in addition to searching for targets. However, it was true that participants were aware throughout the experiment that this question might be asked of them. The goal of Experiment 3 was to provide a check that the results found in Experiment Set 1 could not be attributed to an awareness of a numerical estimation task. In this final study, we recruited a large number of online participants to complete a visual search study, with only one numerical estimation question occurring at the end of the search trials (similar to the strategy used by Krueger (1982)).

Method

Participants 347 participants completed the online study on Amazon's Mechanical Turk in exchange for $0.25.

Stimuli The stimuli were similar to those in Experiments 1A and 2A. The target was a red US Army symbol for an Infantry Platoon and distractors were blue Infantry Platoon symbols. Each display contained one or zero targets, and 5, 10, 15, 20, 25, 30, 35, 40, or 45 distractors.

Procedure Participants completed a brief consent form and then were shown an instruction page. These instructions showed an example target-present display indicating which item was the target. Participants were instructed to press the "m" key if a target was present and the "z" key if no target was present, and to respond as quickly and accurately as possible. Participants were not instructed that they would be asked to estimate the number of items on the screen at this point.

Each trial display remained visible until the participant responded with one of the two designated keypresses, after which there was a 500-ms interval before the next display appeared. Participants completed a total of 72 search trials, eight at each set size, with half of the trials containing a target. One additional target present search display was presented, with 5, 15, 25, 35, or 45 distractors. Following this display, participants were asked to estimate the total number of items they had just seen on the screen, with the information that the correct number was at least six and no more than 46. Participants typed their estimate into a text box and then received a completion code to receive credit for their participation. Note that this input response was different than the categorical responses in Experiment Set 1 and 2 to understand whether the results were skewed by the categorical restrictions on response.

Results

The data from seven participants were removed from all analyses due to not reaching 75% search criteria.

Search task As in Experiment Set 1, we binned display size into five categories (3–10 items, 11–20 items, 21–30 items, 31–40 items, and 41–50 items) and analyzed mean RT on correct trials as a function of display size, separately for target present and target absent trials. An ANOVA revealed no effect of display size on RT for target present trials, $F(4, 1356) = 2.29$, $p > 0.05$, and no significant effect for target absent trials, $F(4, 1356) = 1.25$, $p > .05$.

The mean slope of the RT × Display size function was 1.2 ms per item for target present trials and −1.5 ms per item for target absent trials. Slopes were significantly different from zero, for present trials, $t(339) = 2.03$, $SE_M = 0.61$, $p < .05$, but not for absent trials, $t(339) = 1.48$, $SE_M = 1.01$, $p > .05$.

Numerical estimation We examined participants' accuracy in estimating the number of items in the final search display. Across all conditions, the average absolute estimation error was −3.63, a significant underestimation, $t(339) = -4.34$, $p < 0.001$. With just the display size of 26 ($M = -5.48$), there was also significant underestimation, $t(78) = -5.11$, $p < 0.001$. Given the difference between the effect size in Experiment 1A and Experiment 3, an independent-samples t-test was run on signed error between the 26-item display size condition in Experiment 1A (−7.57) and the response from the participants who saw a 26-item display size in Experiment 3, but no difference was detected ($p > 0.35$).

Discussion

Experiment 3 was designed as a final check that the results found in Experiment Set 1 were not attributable to participants' knowledge that they might be asked to recall the number of items in a previous display throughout the experiment. We found evidence of an underestimation effect, as we did in Experiment 1A, which used a similar feature search. Although the magnitude of underestimation was somewhat smaller in Experiment 3, the results provide support that the presence of an underestimation effect in Experiments 1A–C is not explained solely by foreknowledge of the estimation task.

General discussion

Experiments 1A, 1B and 1C represented the classic categories of visual search: feature, conjunction, and spatial configuration, respectively with the addition of numerical estimates of distractors. Our results indicated that participants demonstrated better estimation of item numerosity in less efficient search tasks. In Experiments 2A, 2B, and 2C, we used the same stimuli as in Experiment Set 1, but only instructed participants to estimate the number of items on the screen. Numerosity estimates in these studies were more accurate overall than those generated when participants also performed the search task, but the critical finding was that the type of stimuli did not affect estimates when the sole task was estimation. Additionally, estimates were not affected by the amount of time participants had available for the task. Taken as a whole, these experiments indicate that the efficiency of a visual search task influences memory for numerosity, with more accurate estimations as efficiency decreases. We added an additional control study (Experiment 3) to ensure that the effect was maintained when participants were completely unaware they would be asked to estimate the number of items in a display and were allowed to type in an estimation number rather than selecting a category.

The goal of these experiments was to understand the influences on estimation of item numerosity within visual search, where estimation is not the primary task..These studies were conducted in addition to four support studies, (Cassenti, Kelley, & Ghirardelli, 2010) as part of a general research program to study human attention and to develop computational models of memory mechanisms for robotics applications (Kelley, 2006, 2014). We assessed numerosity of unattended items as a dependent variable to gain insight into the nature of memory encoding during visual search tasks, where the goal of the task was unrelated to the number of items. We found that participants consistently performed better than chance on the estimation tasks, suggesting that there is some memory encoding of item numerosity during the visual search process. However, there was an overall tendency toward underestimation in all our experiments, with the magnitude of the underestimation dependent upon search efficiency. The results suggest that with less efficient searches, more attention is required to find a target among a set of distractors, leading to more accurate memories of properties of those distractors.

In our previous visual search studies (Cassenti, et al., 2010), we manipulated distractor size, distance between target, and distractors, and clustering variables and found that each of these factors affected the magnitude of underestimation for unattended items. The overall finding was participants make greater underestimation errors when properties of the distractors facilitate Gestalt grouping. The current work provides a theoretical basis for all of our findings: Gestalt principles play a major role in determining the efficiency of a visual search task (Duncan & Humphreys, 1989; Humphreys et al., 1989; Gilden et al., 2010), which in turn influences the accuracy of numerosity estimations. When distractors are not easily grouped together, more attention is required to segment them from the target, leading to better memory for those distractors.

The theoretical question of what cognitive mechanism(s) underlies difficult visual searches has been one of major interest since the field's inception. Early work posited distinct parallel and serial processes, with parallel processes characterized by flat RT-set size slopes and serial processes by increasing slopes (e.g., Sternberg & Scarborough, 1969; Treisman & Gelade, 1980). As the field matured, (e.g., Townsend, 1971; Townsend & Wenger, 2004) demonstrated that data generated from typical visual search studies could not distinguish between a serial search mechanism and a limited-capacity parallel mechanism. Wolfe (1998a, b) and others concluded that without clear evidence for two mechanisms, all types of search could be explained by a parallel mechanism that varies in an efficiency continuum. The simplest feature searches were on the highly efficient side of this continuum, while more difficult conjunctive and spatial configuration searches were inefficient. More recently, however, Thornton and Gilden (2007) used a multiple target methodology (originally proposed by Townsend, 1990) in combination with sequential sampling models, which was capable of making the distinction between parallel and serial. It demonstrated, that while most classic search types were explained by a parallel process, a small set of search types were distinguishable as a serial process (see also Gilden, Thornton, & Marusich, 2010).

The three classes of search used in our experiments are all best described by Thornton and Gilden's parallel model, with varying degrees of capacity limitation, or search efficiency. From our experiments, it is clear that the differences in reaction times across the three experiments represent different degrees of efficiency. Our results indicate that this difference in efficiency affects the observer's ability to accurately estimate the number of search items displayed. Our hypothesis that the difficulty of the search influences memory for distractors does not make specific predictions for serial searches versus highly capacity-limited parallel searches. It would be informative to conduct additional studies using the same paradigm comparing very inefficient parallel searches with serial searches. It may be the case that the one-at-a-time processing of distractors in serial search leads to an improvement in numerical estimation over and above what would be predicted just by inefficiency of the search.

Previous studies of memory for distractors in visual search tasks can be divided into three categories: (a) memory for distractor location, (b) individual properties or identities of the distractors, and (c) the properties of the entire set of distractors. Studies of distractor location have been the most prevalent (e.g. Dodd, Castel, & Pratt, 2003; Horowitz & Wolfe, 1998, 2001, 2003; Klein, 1988; Klein & MacInnes, 1999; Kristjansson, 2000; Peterson et al., 2001; Shore & Klein, 2000; Takeda & Yagi, 2000), primarily using revisitation and inhibition of return paradigms to assess whether memory for distractor location exists. A smaller body of work has explored memory for individual distractor properties. Recent findings suggest that there is memory for individual target identities (Williams, Henderson, & Zacks, 2005), but that distractor identity is not remembered as well as distractor location (Beck, et al., 2006).

Our focus was on numerosity, a characteristic of the distractor set as a whole, rather than on aspects of individual distractor elements. Numerosity is an interesting property of visual displays because it is general to a set of items, yet the judgments themselves are built upon consideration of individual items. Future work may explore whether our finding that search efficiency affects memory for distractor numerosity extends to other distractor properties. There is some evidence that memory for distractor location is similarly affected by search efficiency (Klein, 1988; Klein & MacInnes, 1999), and recent work by Chong & Treisman (2005) indicates that search efficiency may influence memory for both properties of individual distractors (e.g., color, shape) or for properties of the distractor set besides numerosity (e.g., average size).

In the experiments reported here, we focused on assessing numerosity estimates within the context of visual search tasks, but our findings are also relevant within the context of the broader numerosity literature. Numerosity estimates for *attended* items, as opposed to *unattended* items, have had a long history of psychological research. The underestimation effect has been well documented for attended items. Jevons (1871) documented accurate numerosity judgments for very small sets of items, with a tendency toward underestimation developing as the number of items approached 10 and above. As the research in numerosity progressed, it was later found that smaller numbers of items could be automatically and accurately estimated (Taves, 1941), a process that was eventually called "subitizing" (Kaufman et al., 1949). Above the subitizing range of 1–4 items, estimation becomes more difficult and less accurate. One might assume that when estimating larger numbers of items, the errors would be evenly distributed around the actual mean of the items in a display (i.e., Weber's Law, enunciated by Ernst Heinrich Weber, 1795–1878) however, this was found to not be the case. Instead, there is a tendency toward underestimation (Burgess & Barlow, 1983; Ginsburg 1978; Ginsburg, 1991; Ginsburg & Nicholls, 1988; Indow & Ida, 1977; Krueger, 1982). Individuals generally underestimate the number of elements in a display even when only a single judgment was produced (e.g., Krueger, 1982). Additionally, recent studies have found that numerosity estimations, when attending directly to the items, are influenced by statistical regularities and grouping principles (Utochkin, 2015) and that size estimates are affected by the efficiency of a visual search task (Chong & Treisman, 2005). Our results demonstrate that underestimation occurs even when estimation is not the primary task, although the effect is attenuated when more attention is deployed to the items (e.g., in more difficult search tasks).

It is our hope that these studies will contribute to our understanding of the role of memory in visual search, particularly how the efficiency of visual search could be represented computationally. A better understanding of the computational mechanisms of visual search would help to inform the cognitive modeling community and further help our understanding of visual processing mechanisms.

Author Note Portions of this work have been presented at the 47th Annual Meeting of the Psychonomic Society, the 28th Annual Conference of the Cognitive Science Society, and the 77th Annual Meeting of the Eastern Psychological Association. We would like to thank Mike Byrne for his early input into these studies, Howard Egeth for comments on an earlier version of the manuscript, and Jean Breitenbach, Molly Abend, Mouna Attarha, Megan Davis, Paige Kretschmar, Ben Piper, Ann Mae Stellwag, Melissa Talleda, and Monica Zilioli for help with data collection.

Appendix 1

Guess analysis

We analyzed whether or not participants guessed by calculating the *absolute error.* An absolute error is the distance between the actual number of distractors on the screen and the number given as the answer by each participant. For example, in each case where there is a question concerning the total number of items (target and distractors) in the preceding display, the number of items is always known to be in the middle of the range of values listed. In other words, if the answer to a question is 1−10, there were six items on the screen, five of which were distractors. In order to obtain the discrepancy score, the number of known distractors is subtracted from the middle value in the range and results in a delta. For example, if the answer is 1−10 and the participant actually answered 11−20 for the question, the delta would be 10 (or 16 minus 6). If we let AE = the absolute error then, the absolute

Table 1 Absolute error associated with each display size and response choice combination

Display size	Response choices				
	1−10	11−20	21−30	31−40	41−50
6	0	10	20	30	40
16	10	0	10	20	30
26	20	10	0	10	20
36	30	20	10	0	10
46	40	30	20	10	0

errors for each actual number of distractors and possible answer are given in Table 1.

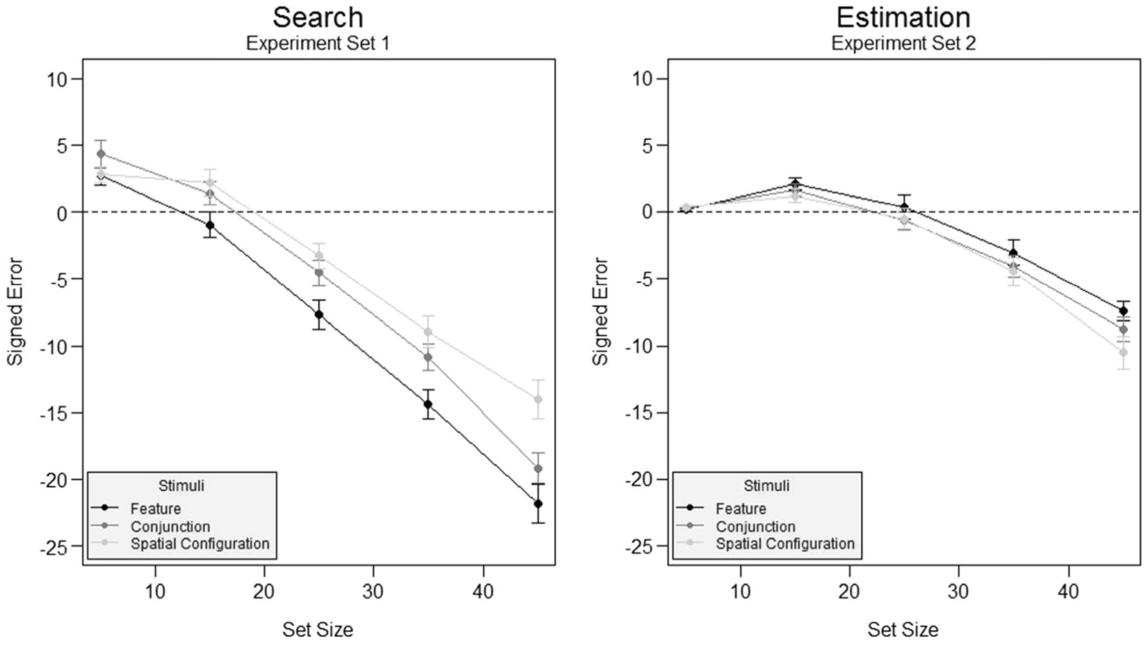
For the trials with the numerical estimation question in the search experiments, there were 15 trials at each of the five possible display sizes, so the probability for each of the possible absolute errors under the assumption that participants are simply guessing the actual number of items are given in the table below. Consequently, the mean value of AE under the assumption that participants guessed is equal to (0)(.2) + (10)(.32) + (20)(.24) + (30)(.16) + (40)(.08) = 16.

Now we can test the null hypothesis that AE = 16 versus the alternative hypothesis that AE≠16 using a standard t -test. Rejecting the null hypothesis, results in a conclusion that the participants are retaining information about the preceding display. Failing to reject the null hypothesis, results in the conclusion that participants are responding to the numerical estimation questions randomly.

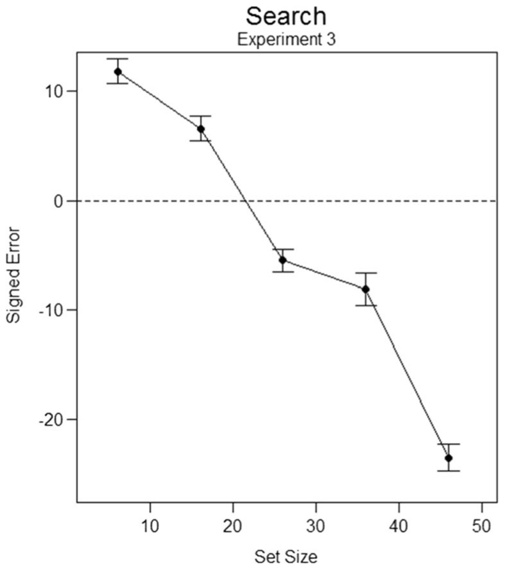
Appendix 2

Signed error and display size

Say here that in our experiments, participants were limited in the responses they could make (no lower than six and no higher in 46). As a result, the highest possible positive signed error and the lowest possible negative signed error were different for each set size. It was only when looking over all set sizes that the opportunity to over- versus under-estimate was balanced. For this reason, we analyzed the average signed error across all set sizes. In the figure below we show the average signed error by set size for each study in Experiment Sets 1 and 2. As is expected from the experimental design, the errors tend to be positive at the smaller set sizes and negative at the larger set sizes. However, it is clear that the data lean more heavily into the negative region of the graph, especially in the search experiments. In addition, estimates are consistently negative at set size 26, where the opportunity to make positive vs. negative signed errors is equal.



The critical finding for these studies is that underestimates become less severe with less efficient search. This pattern is also visible when looking across set size in the above figure: there is a clear effect of the stimuli used on signed error in the search experiments (but not in the control estimation experiments). Below we show a similar plot for signed error in Experiment 3, where participants made only one estimation at the end of a series of feature search trials. The pattern of data is consistent with that of Experiment Set 1.



Below, we show tables of the distribution of estimation responses by display size for each study in Experiment Sets 1 and 2, where participant responses were multiple choice.

Table 2 Response distribution with number of distractors and signed error in Experiment 1A

Number of distractors	Signed error								
	−40	−30	−20	−10	0	10	20	30	40
5					130	40	6	0	1
15				40	84	26	3	0	
25			25	80	33	13	1		
35		11	41	34	9	1			
45	17	46	54	24	11				

Table 3 Response distribution with number of distractors and signed error in Experiment 1B

Number of distractors	Signed error								
	−40	−30	−20	−10	0	10	20	30	40
5					193	68	18	5	1
15				34	186	57	7	1	
25			15	131	111	24	4		
35		15	73	134	47	16			
45	19	52	116	82	16				

Table 4 Response distribution with number of distractors and signed error in Experiment 1C

Number of distractors	Signed error								
	−40	−30	−20	−10	0	10	20	30	40
5					221	57	8	1	1
15				32	175	70	13	0	
25			8	105	139	25	3		
35		4	52	144	57	13			
45	2	25	103	95	48				

Table 5 Response distribution with number of distractors and signed error in Experiment 2A

Number of distractors	Signed error								
	−40	−30	−20	−10	0	10	20	30	40
5					294	6	0	0	0
15				8	226	62	4	0	
25			2	50	189	54	5		
35		2	4	119	133	42			
45	0	2	33	150	115				

Table 6 Response distribution with number of distractors and signed error in Experiment 2B

Number of distractors	Signed error								
	−40	−30	−20	−10	0	10	20	30	40
5					468	9	3	0	0
15				19	374	78	8	1	
25			3	108	287	82	0		
35		5	18	190	223	44			
45	1	7	86	224	162				

Table 7 Response distribution with number of distractors and signed error in Experiment 2C

Number of distractors	Signed error								
	−40	−30	−20	−10	0	10	20	30	40
5					343	4	2	1	0
15				10	279	57	4	0	
25			1	61	235	51	2		
35		0	8	158	154	30			
45	0	5	72	194	79				

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111,* 1036–1060.

Bacon, W. F., & Egeth, H. E. (1991). Local processes in pre-attentive feature detection. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 77–90.

Beck, M. R., Peterson, M. S., Boot, W. R., Vomela, M., & Kramer, A. F. (2006). Explicit memory for rejected distractors during visual search. *Visual Cognition, 14,* 150–174.

Burgess, A., & Barlow, H. B. (1983). The precision of numerosity discrimination in arrays of random dots. *Vision Research, 23,* 811–820.

Cassenti, D. N., Kelley, T. D., & Ghirardelli, T. G. (2010). Underestimating numerosity of items in visual search tasks. *Perceptual and Motor Skills, 111,* 379–398.

Chong, S. C., & Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics, 67*(1), 1–13.

Dodd, M. D., Castel, A. D., & Pratt, J. (2003). Inhibition of return with rapid serial shifts of attention: Implications for memory and visual search. *Perception & Psychophysics, 65,* 1126–1135.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96,* 433–458.

Gebuis, T., & Reynvoet, B. (2012). The role of visual information in numerosity estimation. *PLoS ONE, 7*, e37426.

Gilden, D. L., Thornton, T. L., & Marusich, L. R. (2010). The serial process in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 533–542.

Ginsburg, N. (1978). Perceived numerosity, item arrangement, and expectancy. *The American Journal of Psychology, 91,* 267–273.

Ginsburg, N. (1991). Numerosity estimation as a function of stimulus organization. *Perception, 20,* 681–686.

Ginsburg, N., & Nicholls, A. (1988). Perceived numerosity as a function of item size. *Perceptual and Motor Skills, 67,* 656–658.

Hollingsworth, W. H., Simmons, J. P., Coates, T. R., & Cross, H. A. (1991). Perceived numerosity as a function of array number, speed of array development, and density of array items. *Bulletin of the Psychonomic Society, 29,* 448–450.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature, 394,* 575–577.

Horowitz, T. S., & Wolfe, J. M. (2001). Search for multiple targets: Remember the targets forget the search. *Perception & Psychophysics, 63,* 272–285.

Horowitz, T. S., & Wolfe, J. M. (2003). Memory for rejected distractors in visual search? *Visual Cognition, 10,* 257–298.

Humphreys, G. W., Quinlan, P. T., & Riddoch, M. J. (1989). Grouping processes in visual search: Effects with single- and combined-feature targets. *Journal of Experimental Psychology: General, 118,* 258–279.

Indow, T., & Ida, M. (1977). Scaling of dot numerosity. *Perception & Psychophysics, 22,* 265–276.

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition, 106,* 1221–1247.

Jevons, W. S. (1871). The power of numerical discrimination. *Nature, 3,* 281–282.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology, 62,* 498–525.

Kelley, T. D. (2006). Developing a psychologically inspired cognitive architecture for robotic control: The Symbolic and Sub-symbolic Robotic Intelligence Control System (SS-RICS). *International Journal of Advanced Robotic Systems, 3,* 219–222.

Kelley, T. D. (2014). Robotic dreams: A computational justification for the post-hoc processing of episodic memories. *International Journal of Machine Consciousness, 6,* 109–123.

Klein, R. M. (1988). Inhibitory tagging system facilitates visual search. *Nature, 334,* 430–431.

Klein, R. M., & MacInnes, W. J. (1999). Inhibition of return is a foraging facilitator in visual search. *Psychological Science, 10,* 346–352.

Kristjansson, A. (2000). In search of remembrance: Evidence for memory in visual search. *Psychological Science, 11,* 328–332.

Krueger, L. E. (1982). Single judgments of numerosity. *Perception & Psychophysics, 31,* 175–182.

Le-Hoa Võ, M., & Wolfe, J. M. (2015). The role of memory for visual search in scenes. *Annals of the New York Academy of Sciences, 1339,* 72–81.

Mandler, G., & Shebo, B. J. (1982). Subitizing: an analysis of its component processes. *Journal of Experimental Psychology: General, 111,* 1–22.

Peterson, M. S., Kramer, A. F., Wang, R. F., Irwin, D. E., & McCarley, J. S. (2001). Visual search has memory. *Psychological Science, 12,* 287–292.

Shore, D. I., & Klein, R. M. (2000). On the manifestations of memory in visual search. *Spatial Vision, 14,* 59–75.

Sternberg, S., & Scarborough, D. L. (1969, July). Parallel testing of stimuli in visual search. In *International Symposium on Visual Information Processing and Control of Motor Activity, Sofia, Bulgaria*.

Takeda, Y., & Yagi, A. (2000). Inhibitory tagging in visual search can be found if search stimuli remain visible. *Perception & Psychophysics, 62,* 927–934.

Taves, E. H. (1941). Two mechanisms for the perception of visual numerousness. *Archives of Psychology (Columbia University)*, 1–47.

Thornton, T. L., & Gilden, D. L. (2007). Parallel and serial processes in visual search. *Psychological Review, 114,* 71–103.

Townsend, J. T. (1971). A note on the identifiability of parallel and serial processes. *Perception & Psychophysics, 10*(3), 161–163.

Townsend, J. T. (1990). Serial vs. parallel processing: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science, 1*(1), 46–54.

Townsend, J. T., & Wenger, M. J. (2004). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review, 111*(4), 1003.

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

Utochkin, I. S. (2015). Visual enumeration of spatially overlapping subsets. *The Russian Journal of Cognitive Science, 3,* 4–20.

Williams, C. C., Henderson, J. M., & Zacks, R. T. (2005). Incidental visual memory for targets and distractors in visual search. *Perception & Psychophysics, 67,* 816–827.

Wolfe, J. M. (1998a). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13–73). London: Psychology Press.

Wolfe, J. M. (1998b). What can 1,000,000 trials tell us about visual search? *Psychological Science, 9,* 33–39.