Incidental statistical summary representation over time

Chris Oriet
Department of Psychology, University of Regina, Regina, Saskatchewan, Canada

Kadie Hozempa
Department of Psychology, University of Regina, Regina, Saskatchewan, Canada

Information taken in by the human visual system allows individuals to form statistical representations of sets of items. One's knowledge of natural categories includes statistical information, such as average size of category members and the upper and lower boundaries of the set. Previous research suggests that when subjects attend to a particular dimension of a set of items presented over an extended duration, they quickly learn about the central tendency of the set. However, it is unclear whether such learning can occur incidentally, when subjects are not attending to the relevant dimension of the set. The present study explored whether subjects could reproduce global statistical properties of a set presented over an extended duration when oriented to task-irrelevant properties of the set. Subjects were tested for their memory of its mean, its smallest and largest exemplars, the direction of its skew, and the relative distribution of the items. Subjects were able to accurately recall the average size circle, as well as the upper and lower boundaries of a set of 4,200 circles displayed over an extended period. This suggests that even without intending to do so, they were encoding and updating a statistical summary representation of a task-irrelevant attribute of the circles over time. Such incidental encoding of statistical properties of sets is thus a plausible mechanism for establishing a representation of typicality in category membership.

Introduction

The ease with which people can detect an outlier suggests their representations of categories of objects incorporate some notion of what is average, and of the upper and lower limits in the normal variability within the dimension in question. When viewing a set of objects, a particularly large or small member of the set will stand out and is readily detected (e.g., Hodsoll & Humphreys, 2001), but as any owner of a large breed dog will attest, people do not need to see the object in the context of other set members to recognize it is unusually large. People know, for example that the Great Dane they encounter in the park is large for a dog, and, if they are familiar with the breed, they can also determine whether this particular dog is large for its breed. This example underscores the fact that in learning about the members that comprise a set of like items, individuals form a representation that includes characteristics such as the size of the average member and the upper and lower boundaries that typically define the limits of the set. Thus, an item that seems to lie outside of these bounds is readily recognized as atypical. In the present work, we explore the incidental learning of statistical characteristics of sets of items presented over an extended duration as a potential mechanism for learning about typicality in category membership.

Recently, Duffy, Huttenlocher, Hedges, and Crawford (2010) demonstrated that observers' performance in a length judgment task is influenced by the central tendency of the set, suggesting that they have learned about this property incidentally in the course of completing the task. In Duffy and colleagues' (2010) experiments, subjects were shown a single line for 1 s and asked to reproduce it from memory 1 s later by adjusting an anchor line to match the length of the target line. Over a number of trials, estimates regressed to the running mean (i.e., the mean length of the stimuli presented up to that point in the experiment; Weiss & Anderson, 1969): lines longer than the average line were remembered as shorter than they were, and lines shorter than the average line were remembered as longer. This was observed despite the fact subjects were never explicitly asked to report the mean size of the lines. A subsequent experiment ruled out use of the median or midpoint of the set. The authors also tested whether observers' judgments were influenced by the entire set of stimuli, or only by the most recent stimuli shown. Initially, items were drawn from a distribution with a small (or, for some subjects, large) mean.

Citation: Oriet, C., & Hozempa, K. (2016). Incidental statistical summary representation over time. *Journal of Vision, 16*(3):3, 1–14, doi:10.1167/16.3.3.

doi: 10.1167/16.3.3 Received October 2, 2015; published February 1, 2016 ISSN 1534-7362

Midway through the experiment, the experimenters covertly switched the distribution to one with a large (or small) mean. The results suggested that judgments were more strongly influenced by the running mean of the entire set than by the mean of the most recently experienced items.

Further evidence that subjects represent the central tendency of a set of items presented over an extended duration comes from the literature on statistical summary representations, a process by which observers rapidly summarize sets according to their statistical properties without representing individual members of the set (Ariely, 2001). Such statistical summaries are believed to be formed without intention (Dubé & Sekuler, 2015; Oriet & Brand, 2013), to not require access to limited capacity processing resources (Attarha, Moore, & Vecera, 2014), and to benefit from a global allocation of attention to the displays (Chong & Treisman, 2005). Whiting and Oriet (2011) presented subjects with sets of circles followed by two probe circles, and instructed subjects to choose the probe corresponding to the average size circle in the set. When visibility of the sets was limited by presenting them briefly and backwards masking them, or when the circles (unknown to subjects) were omitted entirely, subjects strongly favored the probe closer in diameter to the running mean. No such tendency was observed when the stimuli were clearly visible. Thus, when unsure of the average size, subjects invoked a strategy of referencing the central tendency of the distribution of items to improve their guesses, suggesting that, like Duffy and colleagues' (2010) subjects, they may have incidentally represented and updated a "running mean."

A mechanism for learning about the statistical properties of sets would be especially useful if it could allow for relatively fast learning, and if such learning could occur with minimal conscious effort. However, in the work reviewed above, subjects explicitly attended to the relevant dimension and performed a task that would benefit from encoding statistical properties of the sets, such as their smallest and largest members, and their average size. Rather than a random guess, a guess that falls within the bounds of items experienced or that is similar to the central tendency of the set improves the chance of making an accurate response. It might not be surprising, then, to find evidence that subjects computed a running average in this context because doing so would confer a strategic advantage in the task at hand. Even if computing the central tendency of the set is unintentional, doing so might nevertheless require attention to the measured dimension (i.e., size) of the individual items in the set in the first place in order to establish a reliable representation of central tendency.

In the present work, we tested whether attention to a particular dimension is necessary for extracting the statistical properties of this aspect of sets over an extended duration. We note that in other forms of incidental learning, learning requires attention to the relevant dimension. For example, in contextual cueing (Chun & Jiang, 1998), targets are found more quickly when embedded in a repeated (rather than novel) spatial context, despite subjects having no awareness of the repeated context. Of importance for the present investigation, this benefit of repetition is only observed for attended distractors. Jiang and Chun (2001) instructed subjects to search for a target among distractors in the target color while ignoring distractors in a nontarget color. The spatial context of either the attended or ignored distractors repeated on every trial. The benefit of repeating the spatial context of the distractors was observed only for attended distractors; targets were located no more quickly when the spatial configuration of the ignored distractors repeated than when it changed randomly on every trial. These results suggest that, with respect to some incidental learning at least, attention is required to establish a representation that can influence performance.

In contrast to incidental learning of spatial relations, there is evidence to suggest that subjects might be able to learn about statistical characteristics of distributions without attending to the tested dimension. Corbett and Melcher (2014) recently showed evidence of such learning for displays of items that retained their statistical properties for a run of trials. In that work, however, the statistical properties of the distribution of items could be learned within a single trial. Although performance improved with subsequent exposures, it was not necessary to integrate information across successive displays to learn about the statistical properties of the set. As such, it is unclear from their study whether subjects learned about the distribution in small "snapshots" over time, or formed a complete representation of the full set immediately, which was then strengthened with repeated encounters with the full set. In the present work, we explore the extent to which information about the statistical properties of a distribution can be accrued over multiple snapshots of the full distribution, when the full set of items is never seen all at once.

Estimates of certain perceptual attributes (e.g., heaviness [Helson, 1947]; size [Parducci, 1956]) are known to be influenced by the endpoints and central tendency of the distribution of stimuli tested. Using cards of varying sizes, with a varying number of dots printed on each card, Parducci (1956) found that subjects correctly recalled stimuli that had been presented 16 times more frequently than stimuli that had been shown only twice. Subjects also demonstrated sensitivity to the range of stimuli in the set, judging

stimuli that fell outside of the range as having been presented less frequently than those actually shown, even when attending to a different dimension (e.g., the size of the cards) than the one tested at recall (e.g., the number of dots).

Parducci (1959) also explored subjects' memories for distributions of sticks varying in length. Subjects were shown a series of 43 sticks, presented one at a time in either increasing or decreasing order of length, and were instructed to pay "close attention" as the sticks were presented. They were not given any further instruction until the full set had been presented. Subjects who were instructed to assign the sticks to one of five length categories (very short, short, medium, long, very long) did so with reasonable accuracy, correctly reproducing a positively skewed distribution.

With sticks varying along only one attribute (length), however, it is difficult to imagine what subjects would have attended to if not length. Moreover, the presentation of the sticks in a fixed (increasing or decreasing) order might have inadvertently oriented subjects to their lengths, and with length being the only thing differentiating one stick from another, it would be difficult not to notice—and then anticipate—this systematic change. Thus, it is likely in this work that subjects were attending to the ostensibly unattended attribute. Similarly, Weiss and Anderson (1969) noted that subjects asked to report the running average length of a sequence of six line segments were able to do so reasonably well, although their estimates overweighted the contribution of recent items and underestimated the true average overall. As is likely to be the case in the Parducci (1959) work, however, the measured attribute (length) was attended.

In the present study, we revisit the question of whether incidental representation of the central tendency and variability of a set requires attention to the measured attribute, which in this case, is size. Importantly, we were careful not to orient subjects to the size of the items, nor to any statistical properties of the set. Subjects were given no incentive to try to remember the items they were shown. To minimize inadvertent attention to size, we presented items in sets (rather than showing one item at a time), minimized the duration of presentation, and assigned subjects tasks that required attending to aspects of the sets other than size. We then tested their memory for statistical characteristics of the sizes of the entire distribution of items they were shown. Unlike in previous research (e.g., Parducci, 1956; 1959), we explicitly compared subjects' responses to the values expected if subjects had encoded a veridical representation of these statistical properties without intention. Our work thus allowed us to assess not only whether performance was influenced by statistical characteristics of the distributions, but also the precision with which such characteristics were represented in memory.

Subjects were presented with sets of items drawn from one of two distributions. Half of observers viewed sets of circles drawn from a leptokurtic, negatively-skewed distribution with low variability and half viewed a platykurtic, negatively-skewed distribution with higher variability. Upon completing this familiarization phase of the experiment, observers performed three size-matching tasks based on the displays seen in the first phase. First, they adjusted a randomly generated circle to match the average size circle in the set. Next, they repeated this task with a different circle, matching it to the smallest circle in the set. Following this task, they adjusted a third circle to match the largest circle in the set. Finally, they were asked to reproduce the distribution of the circles in the set by assigning 4,200 circles to 19 bins, each corresponding to a possible circle size in the set (similar approaches have been used to collect remembered frequency estimates in previous work; e.g., Parducci, 1956; 1959). Subjects were not told that they would be asked to make these judgments until the familiarization phase was complete.

The overarching question of interest in this study is whether observers can incidentally learn statistical properties of a set of items based on repeated exposure to subsets (samples) drawn from the full set of items (population) over an extended duration, similar to how learning about categories would occur outside of the lab. In particular, we tested whether observers could reproduce key statistical characteristics of the population: its mean, its smallest and largest exemplars, the direction of its skew, and the relative distribution of the items.

Method

Subjects

Undergraduate volunteers ($N = 152$; 120 women, 32 men) participated in this experiment. Subjects were randomly assigned to complete either an enumeration task (indicate whether each display contains fewer or more than 15 circles; $N = 50$), a color task (indicate whether any two circles shared the same color; $N = 50$), or to passively view the displays ($N = 52$). The latter task was included to ensure that the enumeration and color tasks did not inadvertently increase (or decrease) attention to size. Within each group, half of the subjects viewed sets with low variability and half viewed sets with high variability during the familiarization phase.

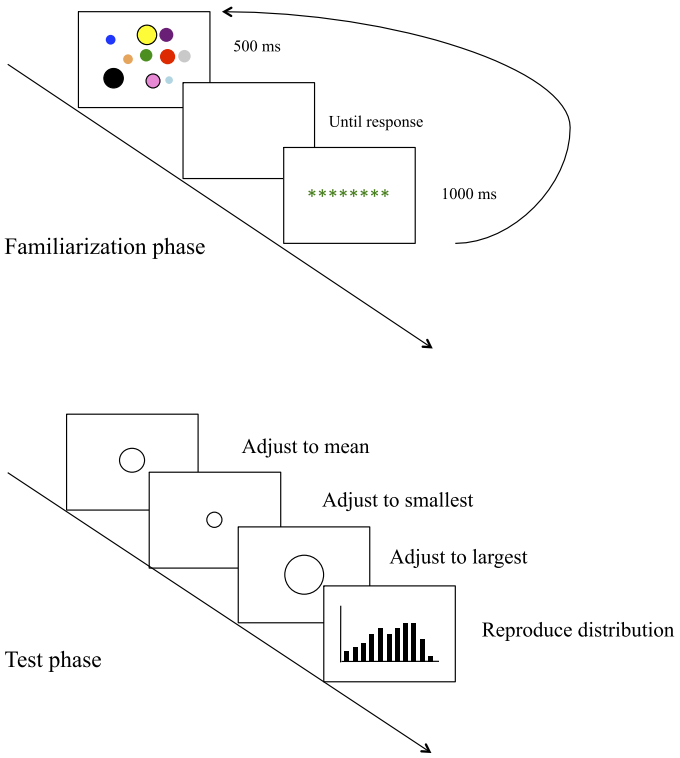

Figure 1. Sequence of events in familiarization phase and test phase. Displays not drawn to scale. See text for details.

Stimuli and apparatus

E-Prime software (Psychology Software Tools, Inc., Sharpsburg, PA) was used to control all stimulus presentation and timing functions. Stimuli were displayed on 27 in. iMac computers set to a screen resolution of 1920 × 1200 pixels and a vertical refresh rate of 60 Hz. In the familiarization phase, subjects viewed displays containing between 10 and 20, but never 15, colored circles presented within an invisible 5 × 5 grid (Figure 1), following Zhao, Goldfarb, and Turk-Browne (2013). For each set size, 24 different sets were created by randomly selecting circles varying in diameter (with a minimum of two just noticeable differences [JNDs] between any two circles based on a power function relating perceived and physical size with an exponent of 0.76; Teghtsoonian, 1965) from a set of 4,200 circles, which either ranged between 16 and 146 pixels in diameter (high variability condition; $M = 86.33$ pixels, $\sigma = 26.6$; skew = −0.10) or between 52 and 96 pixels in diameter (low variability condition; $M = 86.33$ pixels, $\sigma = 9.43$; skew = −1.09). The color of each circle was selected at random without replacement from 20 possible colors (color name and RGB values: red: 255,0,0; green: 0,255,0; blue: 0,0,255; yellow: 255,255,0; rose: 255,155,255; pink: 255,79,150; forest green: 0,128,0; navy: 0,0,102, peach: 255,155,117; dark purple: 108,0,108; cyan: 0,255,255; grey: 185,185,185; orange: 255,140,0; brown: 103,29,0; black: 0,0,0; violet: 137,59,195; dark grey: 77,77,77; sky blue: 0,153,153; burgundy: 165,0,33; teal: 0,200,90; and presented on a white screen: 255,255,255) with the constraint that 50% of the trial displays contained two contiguous circles of the same color.

The test phase contained two different display types. The first type of display was shown during the first three (adjustment) tasks of the test phase. These displays contained a single circle that was centered in the display. The diameter of the circle was randomly generated from between 21 and 137 pixels, excluding 52 or 96 pixels (i.e., the smallest and largest items in the low variability set).

The second display type was shown during the fourth test phase task. The display contained the x and y axes of an empty bar graph. Irrespective of variability condition, the y axis' scale was labeled 0 to 1,600 in increments of 100 circles and the x axis was labeled with the letters A through S. Each letter corresponded to one of the possible 19 circle sizes (ranging from 16 to 146 pixels) that subjects could have seen during the familiarization phase. Subjects adjusted the height of the bars corresponding to the number of circles of each size they believe appeared in the familiarization phase by pressing the letter corresponding to the bar they wished to adjust. When subjects selected a bar to adjust, a circle of the corresponding circle size appeared under the bar graph. Before adjustment, each bar location was empty, which was indicated by a counter set to zero displayed above the corresponding letter. As the height of the bar was adjusted, the counter below each bar was adjusted to reflect the change. Concurrently, a second counter was also adjusted to reflect the number of circles (out of 4,200) that had been accounted for in subjects' frequency estimates.

Procedure

Instructions for all tasks were displayed on the computer screen and reviewed by the experimenter. Responses were recorded by the computer program. During the familiarization phase, subjects viewed the displays of colored circles and completed an enumeration, color, or control task. Subjects completed 14 blocks of 20 trials (two repetitions of each set size per block) for a total of 280 trials. Set size and the location of the duplicate color circles (when present) were randomized within blocks, with the constraint that duplicate color circles occupied contiguous locations. Because the distributions were created precisely and the number of times each circle was seen was carefully controlled, no practice trials were completed. In the enumeration task, subjects judged whether a display of circles contained more than or

fewer than 15 circles. Subjects pressed the "1" key on the keyboard to indicate a judgment of fewer than 15 circles, or "2" to indicate more. The color task required subjects to judge whether any color repeated within a display. Subjects pressed the "1" key on the keyboard to indicate that there were not two circles of the same color and the "2" key otherwise. The passive viewing control task required subjects to view the displays without completing a task. Subjects in this condition pressed the "0" key on the keyboard when ready to view the next display. In all conditions, each display was shown for 500 ms, followed by a white screen displayed until subjects made a response. Subjects completing the enumeration or color task were given feedback after their judgment in the form of eight asterisks displayed in a line located in the center of the white screen. Correct judgments were followed by green asterisks, and incorrect judgments were followed by red asterisks. Subjects completing the control task saw eight black asterisks instead. Asterisks were displayed for 1000 ms, cleared, and then replaced immediately by the circles display for the next trial.

The test phase of the study consisted of four judgment tasks, carried out in the same order by all subjects. The first task required subjects to "draw" a circle with the same average diameter as all the circles they saw across all displays in the familiarization phase by adjusting a randomly-generated test circle until it matched this value. The second and third tasks required subjects to adjust a single circle until its size matched the smallest and largest circles, respectively, that appeared across all exposure displays. Pressing the "8" key on the keyboard made the test circle larger by 1 pixel and pressing the "2" key made it smaller by 1 pixel. When subjects were satisfied that the adjusted circle matched either the average size, the smallest size, or the largest size circle, they pressed the "0" key to lock in their answer and move on to the next task.

The fourth and final task required subjects to adjust a bar graph until it represented the distribution of the circles shown across all familiarization phase displays. Subjects adjusted the bars by either adding circles to or subtracting circles from each bar. The instructions emphasized the importance of adjusting the heights of the bars such that they represent the subject's recollection of the relative frequency of the various circle sizes shown and examples were provided (e.g., the tallest bar should represent the circle seen most frequently). To adjust a bar, subjects pressed the letter on the keyboard that corresponded to the bar they wanted to adjust. Subjects increased the number of circles in the selected bar by two circles when they pressed the "8" key or by 20 circles when they pressed the "7" key. Subjects reduced the number of circles in the selected bar by two circles when they pressed the "2" key or by 20 circles when they pressed the "3" key. Each subject adjusted the bar graph until it matched the distribution of circles shown across the familiarization phase. If subjects accurately encode and retain a representation of the distribution, they should produce negatively skewed distributions. Subjects in the low variability condition should leave 12 of the 19 bars empty.

Because subjects carried out the tasks in the same fixed order, it is possible that later judgments are contaminated by earlier ones. However, we opted not to counterbalance the tasks for a number of reasons. First, based on previous work (Duffy et al., 2010; Whiting & Oriet, 2011) we were fairly confident that subjects would at least represent the mean size of the sets with reasonable accuracy, so we opted to "protect" this judgment by requiring it to be completed first. We also felt that there was a greater chance that subjects would anchor their judgments of mean size to their recollections of the endpoints of the distribution (e.g., adjust the probe midway between the estimates of the smallest and largest items drawn a few moments earlier) than the other way around. This would be especially problematic in the present context because the distributions were negatively skewed, and the use of such a strategy would mask what might have been a veridical recollection of average size. If subjects' representations of average size are accurate, they should be larger than the midpoint value. Finally, because the distribution task took much longer to complete than the other three judgments, we opted to present it last to minimize the possibility that the task itself created retroactive interference for the drawing tasks.

Design

Familiarization phase

Subjects were randomly assigned to one of three conditions: color task, enumeration task, or passive viewing. Within each of these conditions, half of subjects viewed displays drawn from a low variability distribution and half viewed displays drawn from a high variability distribution, yielding a 3×2 between-subjects factorial design. Accuracy was measured for the color and enumeration groups.

Test phase

Subjects performed four tasks, yielding estimates of the recalled mean, smallest, and largest circles (collapsed into a single measure of perceived range), and of the mode, standard deviation, and skew of the distributions.

	Drawing task			Distribution task		
Variability	Mean	Range	Midpoint	Mode	*SD*	Skew
Color						
Low	84.9	60.1	80.8	86.6	22.3	−0.16
High	97.8	136.1	78.6	83.2	33.4	−0.01
Enumeration						
Low	85.4	69.1	86.3	78.0	19.0	−0.16
High	93.4	113.0	71.9	82.0	31.6	0.08
Control						
Low	85.1	55.2	79.5	73.6	17.9	−0.09
High	94.2	120.1	76.0	83.1	31.5	0.07
Actual						
Low	86.3	44.0	74.0	88.0	9.43	−1.09
High	86.3	130.0	81.0	92.0	26.6	−0.10

Table 1. Group means (in pixels), listed as a function of task type and variability for the three drawing task responses and three parameters of the distribution task. Actual values in the distributions shown are listed for comparison.

Results

Familiarization phase

All analyses were carried out with an α of 0.05, and all t tests were two-tailed. Mean accuracy was computed for each subject in the enumeration and color conditions only (as there was no task in the control group), as a function of task type and variability, and analyzed in a 3×2 ANOVA with these variables. Accuracy was much higher in the enumeration task ($M = 0.89$; $SD = 0.07$) than in the color task ($M = 0.61$; $SD = 0.11$), leading to a significant main effect of task type, $F(1, 96) = 225$, $MSE = 0.008$, $p < 0.001$, $\eta_P = 0.70$. The main effect of variability and the interaction were not significant in this analysis.

Test phase

For each subject, we computed the mean size of the average, smallest, and largest circles drawn in the test phase, as well as the mode, skew, and standard deviation of the distributions they produced. Results are listed in Table 1. The difference between the smallest and largest circles drawn by each subject was computed as a measure of the perceived range of circles shown. Scores on each measure were averaged across subjects within each task condition (enumeration, color, control) and variability condition (low vs. high) and analyzed in a series of 3×2 ANOVAs. We then followed up on these with unprotected single-sample t tests, comparing each value with its counterpart in the actual set of items shown to observers.

Drawing tasks

Perceived mean: Observers in the low variability condition adjusted the test circle to an average diameter of 85.1 pixels and in the high variability condition, to an average diameter of 95.1 pixels. The error of this estimate, computed by taking the diameter of the average sized circle drawn by observers and subtracting the actual average diameter of the circles shown, served as the dependent variable in a 3 (condition) $\times$ 2 (variability) ANOVA (group means are displayed in Figure 3). Negative values indicate the perceived average circle was smaller than it actually was. Error was greater in the high variability condition ($M = 8.79$; $SD = 31.1$) than in the low variability condition ($M = -1.12$; $SD = 26.3$), leading to a significant main effect of variability, $F(1, 146) = 4.44$, $MSE = 854$, $p < 0.04$, $\eta_p = 0.03$. Although accuracy in the familiarization phase suggests that the color task was considerably more difficult than the enumeration task, this difference in difficulty had no measurable effect on the drawing task, with neither the main effect of condition nor the interaction reaching significance, both Fs < 1. Indeed, error in estimates was consistently smaller in the low variability condition (ranging from −0.89 to −1.45 pixels across the three task types) than in the high variability condition (ranging from 7.11 to 11.4 pixels).

To test the precision of the estimates of mean size, we carried out one-sample t tests, comparing the observed error against 0. The error observed in the high variability condition estimates of mean size was reliably greater than zero, $t(75) = 2.46$, $p < 0.02$, 95% CI [1.65, 15.9], but the error observed in the low variability condition was not, $t(75) < 1$, $p > 0.63$, 95% CI [–7.22, 4.82].

Although the results in the low variability condition suggest that subjects retained a very precise representation of the mean size, it is possible that they based

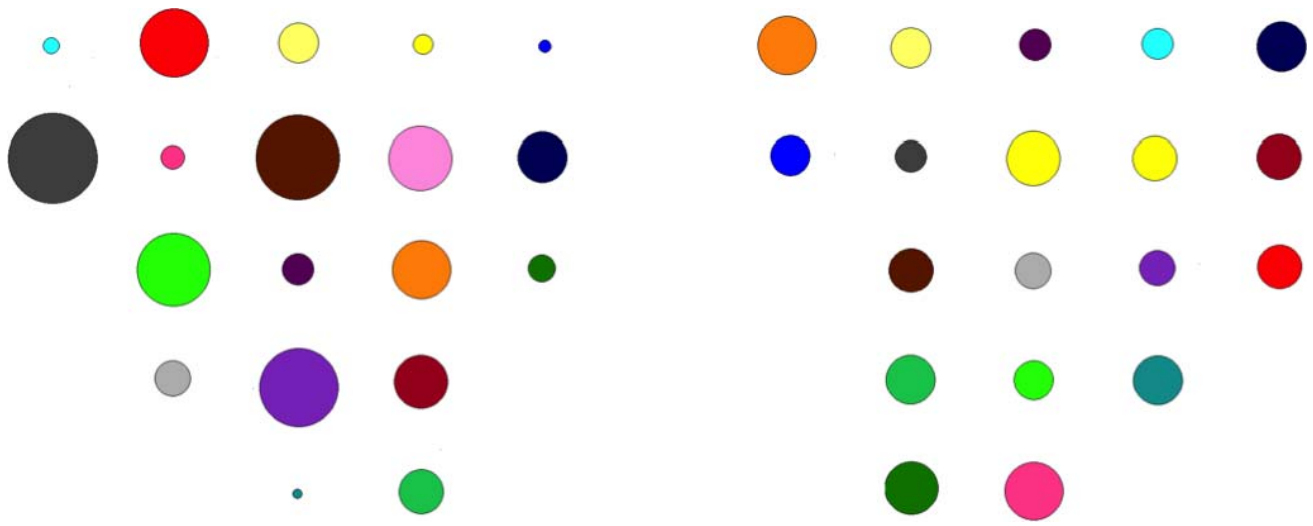

Figure 2. Left panel: Sample display in the high variability condition. Right panel: Sample display in the low variability condition.

their estimates on some other parameter of the set, such as the midpoint of the range of items shown, or the mode. To evaluate the viability of these alternatives, we computed measures of the agreement between the diameter of the average sized circle drawn by subjects and the *perceived midpoint* of the set, estimated from subjects' drawings of the smallest and largest circles in the set, and the agreement between the diameter of the average sized circle and the *perceived mode* of the set, estimated from the distributions produced by subjects in the test phase. Both sets of correlations were significant (mean/midpoint, $r = 0.44$, $p < 0.001$; mean/mode, $r = 0.20$, $p < 0.02$), so we then compared the error observed in judgments of mean size to the error that subjects would make if their estimate of the average circle size was based on the actual mean, the perceived midpoint, and the perceived mode. After excluding data from any subject who adjusted the smallest circle to a diameter larger than the average circle or larger than the largest circle in the set, data from 121 subjects were analyzed in a one-way repeated measures ANOVA. The effect of reference point was significant, $F(2, 240) = 6.25$, $MSE = 373$, $p < 0.003$, $\eta_p = 0.05$, and suggested that estimates would be more accurate if they were based on the actual mean (mean error = 0.79, $SD = 25.9$, $t(120) = 0.34$, $p > 0.73$, 95% CI [−3.87, 5.46]) than if they were based on subjects' own perceived midpoint ($M = -5.26$, $t(120) = -3.46$, $SD = 16.7$, $p < 0.01$, 95% CI [−8.26, −2.25]) or perceived mode ($M = -7.74$, $t(120) = -3.26$, $SD = 26.1$, $p < 0.01$, 95% CI [−12.4, −3.03]).

Perceived range: A measure of perceived range was computed by subtracting the diameter of the smallest circle from the diameter of the largest circle drawn by subjects in the test phase. Only data from the 121 subjects whose smallest circle was smaller than their average circle, and whose average circle was smaller than their largest circle were included in the analysis. This excluded data from 20 subjects in the low variability condition and 11 subjects in the high variability condition. Although the main effect of condition was significant in this analysis, it was not considered further because the actual ranges of the stimuli differed across the high and low variability conditions. The range was perceived as larger by subjects in the high variability condition ($M = 132$; $SD = 32.0$) than in the low variability condition ($M = 64$; $SD = 20.8$), leading to a significant main effect of variability, $F(1,115) = 197$, $MSE = 707$, $p < .001$, $\eta_p = .63$. Although subjects overestimated the range in both conditions, the observed average range of 132 pixels did not reliably differ from the actual range of 130 pixels in the high variability condition, $t(64) < 1$, $p > 0.58$, 95% CI [−5.76, 10.1], for the difference. Moreover, the diameter of the smallest and largest circles drawn by subjects was an average of 1.41 and 3.60 pixels larger than in the actual set of items shown in this condition, and neither value was statistically different from zero, both ts < 1, both ps > 0.32. The observed average range of 60.2 pixels in the low variability condition, however, was reliably greater than the actual range of 44 pixels, $t(55) = 7.33$, $SD = 20.8$, $p < 0.01$, 95% CI [14.7, 25.9] for the difference, with the smallest circle underestimated by an average of 5.93 pixels, $t(55) = -2.71$, $SD = 16.4$, $p < 0.01$, 95% CI [−10.3, −1.54], and the largest circle overestimated by an average of 14.4 pixels, $t(55) = 5.18$, $SD = 20.9$, $p < 0.001$, 95% CI [8.84, 20.0].

Distribution estimates: The distributions produced by subjects are shown in Figure 4 as a function of variability, superimposed on the actual distributions of circles presented to subjects. Kullback-Leibler's D, a measure of the extent to which observers' distributions diverge from the actual distributions they saw, are

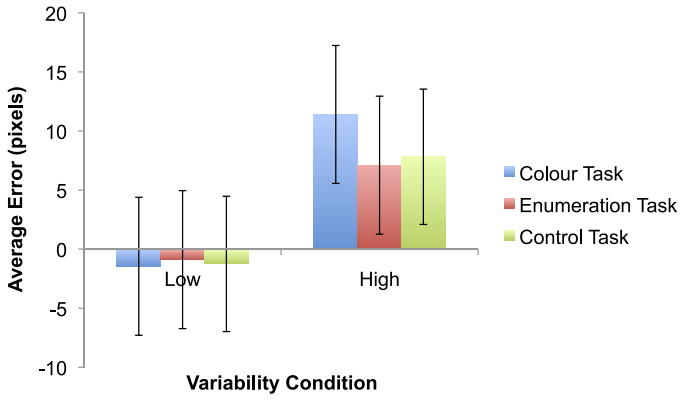

Figure 3. Average error in estimates of mean size as a function of task type and variability. Error bars correspond to the standard error of the mean.

shown in Figure 4 for each condition; D is expected to be close to zero if observers' reproductions of the distributions they saw are accurate (Kullback & Leibler, 1951). D values were strikingly consistent across the three tasks in the high variability condition ($D = 0.19$ to 0.21), highlighting the null effect of familiarization task difficulty. More revealing is the fact that D values for the low variability distributions ($D = 0.78$ to 1.47) were approximately 4 to 8 times greater than for the high variability distributions, and much less consistent across the three groups, underscoring that subjects in the low variability condition produced lower fidelity distributions than those in the high variability condition.

Perceived mode: Subjects were instructed to ensure that the highest bar in their distributions corresponded to the circle seen most frequently. Thus, the perceived mode was taken as the most frequently occurring value in the distributions produced by subjects. When two consecutive circle sizes in the set were judged to be equally frequent, the mode was computed as the average of the two values; when the circle sizes were not consecutive, or when more than two circle sizes occurred with equal frequency, the mode was undefined. A mode could be determined for all but two subjects ($N = 150$). The difference between the actual mode in each distribution (92 in the high variability condition and 88 in the low variability condition) and the mode as judged by subjects was computed and this error score served as the dependent variable in a 3 (condition) × 2 (variability) ANOVA. The mode was underestimated by an average of 8.93 pixels, and the

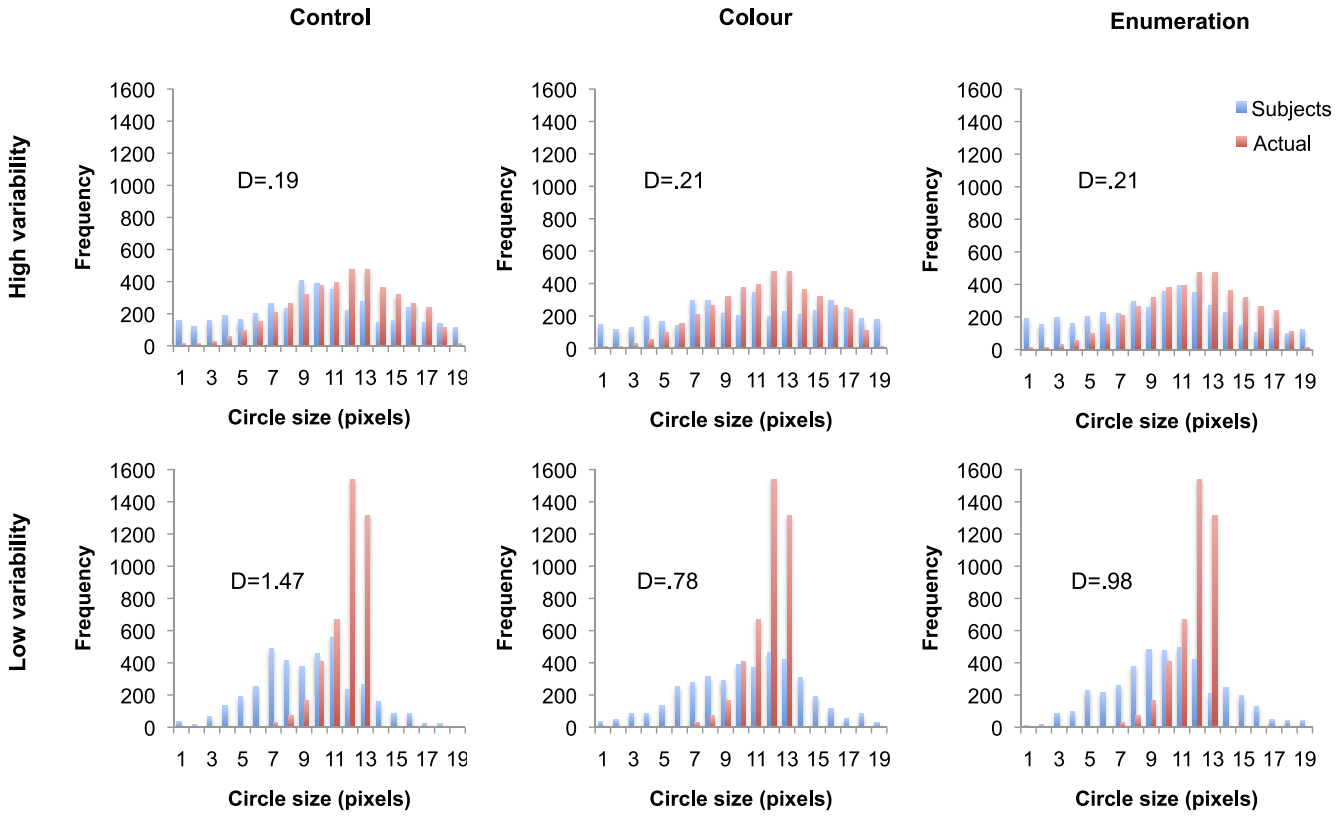

Figure 4. Average frequency of circle diameters as reported by subjects, as a function of task (control, color task, enumeration) and variability (high vs. low). D represents the Kullback-Leibler divergence between the actual distribution and subjects' reproductions of the distributions as they recall seeing them (Kullback & Leibler, 1951). Larger D values indicate greater disparity between the actual distributions and subjects' reproductions of them.

extent of this underestimation was similar across all groups with neither main effect nor the interaction approaching significance, all Fs < 1, all ps > 0.43. The error in judging the mode reliably differed from zero, $t(149) = -4.19$, $SD = 26.1$, $p < 0.001$, 95% CI [−13.1, −4.71]. Within the low variability condition, the median and modal value of the mode were the same (73), whereas in the high variability condition, the median value was 81 and the modal value was 73.

Perceived standard deviation: The standard deviation of the distributions created by each subject was analyzed in a 3 (condition) × 2 (variability) ANOVA. As expected, distributions created by those in the low variability condition had smaller standard deviations ($M = 19.7$; $SD = 6.85$) than in the high variability condition ($M = 32$; $SD = 7.09$), although both significantly overestimated the actual standard deviation in these conditions (low = 9.43; high = 26.6, both ps < 0.001).[1]

Perceived skew: The skew of the distributions created by each subject were calculated using the formula $skew = m_3/{m_2}^{3/2}$, where m_2 and m_3 represent the second and third moments of the distributions, respectively. The difference between this value and the actual skew of the distributions of circles shown to subjects was then analyzed in a 3 (condition) × 2 (variability) ANOVA. The average skew of the distributions created in the high variability condition was closer to the actual skew (mean error = 0.15; $SD = 0.51$) than in the low variability condition (mean error = 0.95; $SD = 0.51$), leading to a significant main effect of variability, $F(1, 146) = 92.7$, $MSE = 0.27$, $p < 0.001$, $\eta_P = 0.39$. Neither the main effect of condition nor the interaction were significant, both Fs < 1, both ps > 0.77. Although the estimated skew in the high variability condition was closer to the actual skew than in the low variability condition, the error in both estimates was reliably different from zero, both ts > 2.5, both ps < 0.02. Nevertheless, the estimated average skew in the low variability condition ($M = -0.14$; $SD = 0.51$) was significantly different from that expected if subjects had produced an unskewed distribution, $t(75) = -2.37$, p < 0.03, 95% CI [−0.25, −0.02] for the difference, but it was not different from an unskewed distribution in the high variability condition, $t(75) < 1$, $p > 0.40$, 95% CI [−0.07, 0.17] for the difference.

Discussion

The results can be summarized as follows. Increasing the variability of the size of a set of items decreased the precision of estimates of mean size, even when size was irrelevant to the task and learned only incidentally. Nevertheless, when the variability in the set was low, subjects adjusted the diameter of a test circle to within just a few pixels of the average diameter of the set. Increasing the variability of the set had the opposite effect on the precision with which subjects estimated the range of items in the set, with estimates closely matching the actual range in the high variability condition but overestimating the range in the low variability condition. Estimates of mean size were precise despite the fact that subjects recalled the negatively skewed distributions as symmetrical.[2]

Drawing task

Three aspects of the results of the drawing task are worth considering further. First, mean judgments were strikingly accurate given that the sets were negatively skewed, and guesses based on the midpoint of the set should have led to underestimation of the average circle. Contrary to this, estimates were not reliably different from the actual mean size in the low variability condition, and were *greater* by about 9 pixels than the actual mean size in the high variability condition. Similarly precise representations of mean size can be inferred from the first half of Duffy and colleagues' (2010) experiment 1 (i.e., before the distribution from which items were drawn changed covertly) and experiment 2, in which the estimated point of zero bias occurred within 2 or 3 pixels of the mean of the entire set of lines, although subjects attended to size in those experiments and were not tested for their recollection of the mean.

Second, the improved estimates of mean size in the low variability condition do not reflect smaller absolute differences in diameter among items near the mean in the low variability condition than in the high variability condition. This is because the high variability condition included the exact same stimuli as the low variability condition, plus 12 more (six smaller, six larger). Only the frequency of the items common to the two variability conditions differed.

Third, the improved estimates of the range (and of the smallest and largest items) in the high variability set cannot be attributed to subjects in this condition encountering the smallest and largest items more often in the high variability set than in the low variability set. Indeed, subjects in the high variability condition encountered just 14 instances of each of the smallest and largest items, yet were far more accurate in their estimates of range than subjects in the low variability condition who encountered 28 and 1,316 of the smallest and largest items in their set, respectively.

Although several other investigations have explored the role of variability in size averaging in static displays (e.g., Corbett, Wurnitsch, Schwartz, & Whitney, 2012; Im & Halberda, 2013; Marchant, Simon, & de Fockert,

2013; Tong, Ji, Chen, & Fu, 2015; Utochkin & Tiurina, 2014), few studies have examined the encoding of variability and central tendency in a set over an extended duration, and none have done so under conditions in which the tested attribute was unattended. The present findings suggest that, in addition to encoding the central tendency and variability of static displays, observers integrate information across multiple displays to represent more global properties of the sets from which the items presented in static displays are drawn. In contrast to previous work exploring subjects' memory for variability in static displays (e.g., Kareev, Arnon, & Horwitz-Zeliger, 2002) however, subjects in the present study tended to *overestimate* the variability of the distribution, particularly in the low variability condition. This might reflect the fact that the distributions are skewed, but in their reconstructions, subjects tended to make them symmetrical. Thus, in reproducing the smallest and largest circles in the set and the distribution of circles, subjects may have chosen endpoints to be equidistant from the (accurately recalled) mean, effectively "unskewing" the distributions and overestimating their variability as a result. Consistent with this interpretation, the error in estimating the size of the largest circle in the low frequency condition was greater (approx. 14 pixels) than the error in estimating the smallest circle (approx. 6 pixels).

Distribution task

Although subjects produced reasonable estimates of the average size and range of items when asked to draw circles with diameters corresponding to these values, their ability to reproduce key features of the distribution itself seems to be considerably more limited. Distributions produced by subjects in both variability conditions were too flat and too symmetrical. Moreover, distributions produced by the low variability group extended far beyond the range of sizes of stimuli actually seen, and, paradoxically, extended beyond their own recollections of the smallest and largest circles in the set as indicated by the drawing task. Some caution is warranted in interpreting these findings, however, as they may reflect demand characteristics owing to the fact that subjects in the low variability condition were asked to estimate the frequency of 12 stimuli they never actually saw, and, more generally, assigning items to 19 different size categories (cf. Parducci's studies [1956; 1959], which used just five categories) may have just been a very difficult task. It is worth noting, however, that although subjects in the low variability condition indicated they saw circles they never actually saw, the frequencies reported by those in the high frequency condition (who actually saw these stimuli) were typically much greater, consistent with Parducci's (1956) finding that subjects rated stimuli that fell within the range of stimuli they had seen as more frequent than those falling outside the set. Moreover, in their reproductions of the distribution of circles, about 93% of subjects in the low variability condition included circles that were smaller or larger than their own estimates of the smallest and largest circles in the set, as indicated in the drawing tasks they completed a few moments earlier. Only 57% of subjects in the high variability condition extended the range in this way, so it is not the case that every subject was compelled to place a circle in every bin available. If subjects were indeed "unskewing" the distributions as described above, the observed extension of the range of circle sizes beyond that reported in the drawing task would be expected.

Subjects appear to have little awareness of the mode of the distribution despite encoding a very precise representation of the mean. Although 68% of the circles shown to subjects in the low variability condition were between 88 and 96 pixels in diameter, the median (and modal) value of the mode in subjects' distributions in this condition was much lower (73 pixels). This is surprising given that circles with a diameter of 73 pixels were actually very rare, comprising only 9.7% of the circles actually shown. The modal mode was also 73 pixels for the high variability group (actual mode = 92 pixels), even though even fewer of the circles they saw (9.0%) corresponded to this diameter. Thus, to the extent subjects' perception of the mode can be validly inferred from their distribution diagrams, it does not appear that subjects formed a representation of the mode of the set. This could simply reflect a preference to place the mode near the midpoint of the items shown, but the failure to recall the most frequently occurring item is consistent with previous claims that subjects are not encoding individual exemplars in the sets (Allik, Toom, Raidvee, Averin, & Kreegipuu, 2013; Ariely, 2001; Corbett & Oriet, 2011; Šetić, Švegar, & Domijan, 2007).

Although subjects were unable to capture the key features of the distributions in their diagrams, there were nevertheless differences across the high and low variability groups that are worth noting. Subjects in the low variability condition did produce distributions with a lower standard deviation on average than in the high variability condition, consistent with their estimates of the range of the stimuli derived from the adjustment tasks. Further, those in the low variability condition produced distributions whose skew more closely approximated the actual distribution they saw than those in the high variability condition. The skew of the low variability group's distributions was negative (as

was the skew of the actual distributions) and reliably different from zero; the skew of the high variability group's distributions did not differ from zero. Finally, although they vastly underestimated the frequency with which they saw most circles, subjects in the low variability condition nevertheless correctly indicated that they were shown more items within the range of diameters actually shown to them (i.e., 52 to 96 pixels) than subjects in the high variability condition. Thus, although subjects' distributions bore little resemblance to the actual distributions of items they experienced, there were nevertheless differences as a function of variability in the direction expected if the range of items in each condition was encoded.

Could subjects perform as well as they did on these tasks if they had based their responses only on their recollection of the last display? To address this question, we carried out two sets of regression analyses. In the first, we regressed the mean size of the circles in the last display, and the average of the mean sizes of the preceding five displays (i.e., displays 15 to 19 in the last familiarization block), on subjects' perceived mean as indicated by the diameter of the circle they drew. In the second, we regressed the range of sizes in the last display, and across the preceding five displays, on subjects' perceived range as indicated by the difference between the largest and smallest circles they drew. If subjects are integrating information across displays in making their judgments, values obtained from displays 15 to 19 should emerge as stronger predictors of subjects' mean and range estimates than values obtained from the last display alone.[3] The model predicting perceived mean scores from the last display and the preceding five displays accounted for significant variability, adjusted $R^2 = 0.053$, $F(2, 151) = 5.19$, $p < 0.008$. In contrast to what is expected if only the last display was referenced, only the mean of displays 15–19 emerged as a significant predictor, $\beta = 0.27$, $t(151) = 3.22$, $p < 0.003$, with $\beta = 0.09$, $t(151) = 1.07$, $p > 0.28$ for the mean of the last display. Similarly, for the perceived range, the model again accounted for significant variance, adjusted $R^2 = .37$, $F(2, 151) = 44.5$, $p < 0.001$, and again only the range of items experienced over displays 15 to 19 emerged as a significant predictor, $\beta = 0.61$, $t(151) = 9.43$, $p < 0.001$, with $\beta = -0.02$, $t(151) = -0.26$, $p > 0.79$ for the range of the last display. Note that although increasing the number of displays naturally brings the mean and range closer to their actual values in the population, they will only be better predictors of observed performance if subjects' judgments also incorporate more information than is available in the last display. Thus, the outcome of these analyses suggest that it is unlikely that subjects are basing their responses on information contained in the last display only.

Statistical summary and categorization

The present findings suggest that even when directed to attend to nonstatistical properties of sets of items, statistical summaries are formed over time. This provides a natural explanation for how we come to learn about what typifies a category member on a particular dimension, and what should be considered the typical or expected range of variability on that dimension; Utochkin (2015) has also discussed the role of summary statistics in categorization, albeit from the intriguing perspective of understanding how summary statistics can be used to distinguish between categories rather than learning about the statistical properties of the categories themselves. Moreover, Dubé and Sekuler (2015) have argued that the abundance of demonstrations of perceptual averaging in visual short-term memory tasks is evidence for a role for perceptual averaging in perceptual categorization. The present finding that perceptual averages, and perhaps other summary statistics, are computed for exemplars encountered over longer durations supports this claim. This is an important theoretical consideration, as some particularly influential models of perceptual categorization, such as the generalized context model (e.g., Nosofsky, Little, Donkin, & Fific, 2011), propose that categorization relies on storing individual exemplars, rather than on comparison to a prototype abstracted over such exemplars. Although their model is designed to explain performance in visual short-term memory tasks, the present finding that subjects retain a very accurate summary representation of an irrelevant, unattended feature would seem to be at odds with the claim that no prototype is extracted as the set of exemplars is learned. The possibility that statistical summaries contribute to perceptual categorization is strengthened by the fact that the display duration (500 ms), inter-stimulus intervals (1000 ms), and feedback provided in the familiarization phase are similar to those used in perceptual categorization tasks, enabling comparisons between such tasks and ours. Future work should more directly address the role of summary statistics in longer-term acquisition of perceptual categories.

The category-adjustment model of Huttenlocher and colleagues (e.g., Huttenlocher, Hedges, & Vevea, 2000) provides a good account of some of our findings. According to this model, recalling an instance of a category is a reconstructive process that involves combining two sources of information: information from the stimulus itself and information from the category to which the stimulus belongs. As the strength of the representation of the former increases, reliance on the latter decreases. At the limit, a perfect representation of the stimulus necessitates no contribution of category information, and the complete absence of a representation of the stimulus requires full

reliance on category information. In such cases, guesses are likely to correspond to the central tendency of the category. However, according to the model, the extent to which instances of a category cluster will also influence performance. Specifically, greater variability in the set of instances within the category should increase variability in judgments because the central tendency of the category is less similar to the average category member than when variability is lower. This in turn should mean that observers rely on category information less in making their judgments.

The category-adjustment model was designed to explain performance in tasks in which observers make judgments about category membership rather than explicitly describe statistical properties of a distribution. However, the model can be extended to account for performance in the task used in the present study by assuming that subjects have conscious access to the representations that support their categorization responses (e.g., representations of central tendency and the boundaries of the set).

The model does not specifically address the attentional requirements of category learning. Nevertheless, the model proposes a trade-off between the strength of the stimulus representation and the extent to which observers rely on category information. Thus, one could infer that the absence of attention to the relevant attribute of the stimuli during the familiarization task should increase reliance on category-level information and in particular, should bias judgments toward the central tendency of the category. Of course, in the present task, this is beneficial: subjects are explicitly asked to estimate the mean size of the items comprising the distribution, and did so within a few pixels of the correct value. The present results are thus consistent with the model's prediction that observers will rely on the central tendency of the set when representations of individual items are weak.

The model also predicts that as variability increases and the central tendency of the set becomes less representative of individual exemplars, subjects should rely on it less in making judgments about category members. In the present study, subjects were asked to reproduce the smallest and largest members of the set. When the variability of the set was high, their judgments were accurate within a few pixels; as variability decreased, subjects underestimated the size of the smallest item and overestimated the size of the largest item. This finding appears to be at odds with the category-adjustment model, as instead of biasing their judgments *toward* the mean when recalling the smallest and largest items in the low variability condition, subjects appeared to bias them *away* from the mean. However, it is important to note that when reconstructing the distributions, subjects made them much flatter than they actually were. Hence, the overestimation of the range in the low variability condition may reflect subjects' attempts to reconcile their recollection of seeing many stimuli that were very similar to one another with an erroneous belief that they must have appeared with similar frequency.

Concluding remarks

An abundance of research suggests that subjects can represent statistical properties of small sets of items presented simultaneously, or over a period of a few seconds (Albrecht & Scholl, 2010; Chong & Treisman, 2005). A smaller set of studies shows that statistical properties can also be extracted for sets of items encountered over time, when subjects attend to the measured attribute. Our work extends this capacity to larger sets of items shown over an extended duration, and suggests that subjects incidentally encode and retain statistical summary representations of irrelevant attributes of the items comprising a set, such as mean size and the upper and lower bounds of the set. This type of statistical summary learning appears to have occurred without conscious intent. Future research should directly test whether such representations can be formed for natural categories, and in particular for those that carry survival value for an organism, such as the size of berries on a bush or variability in the speed of movement of a potential predator.

Keywords: statistical summary representations, ensemble coding, category learning

Acknowledgments

This research was supported by a Natural Sciences and Engineering Council of Canada Discovery Grant and Canada Foundation for Innovation Leaders Opportunity Fund grant to C. Oriet. We are grateful to Jingyu Li for assistance with data collection and analysis.

Commercial relationships: none.
Corresponding author: Chris Oriet.
Email: chris.oriet@uregina.ca.
Address: Department of Psychology, University of Regina, Regina, Saskatchewan, Canada.

Footnotes

[1] Distributions created by subjects in the color task were slightly more variable ($M = 27.9$; $SD = 8.44$) than

distributions created by subjects in the enumeration ($M = 25.3$; $SD = 9.29$) or control ($M = 24.7$; $SD = 10.0$) conditions, leading to a significant main effect of task type, $F(2, 146) = 3.10$, $MSE = 47.6$, $p < 0.05$, $\eta_P = 0.041$.

[2] Despite the greater difficulty of the color task, estimates of mean size were no less precise following this task than following enumeration or passive viewing. This may be an indication that the mean was encoded accurately irrespective of how attention was allocated to the displays (cf. Chong & Treisman, 2005). However, a limitation of our design is that, to avoid creating a dual-task situation (which would make the color and enumeration tasks different from the passive viewing task in two ways), we did not assess the distribution of attention in the familiarization phase and as such, no strong conclusions can be drawn.

[3] Note: Five displays were used because with this number, there was sufficient variability in the predictor values across subjects. When the preceding six displays were used, all but seven subjects in the high variability condition had experienced the full range of stimuli, leaving insufficient variability in this predictor for carrying out a multiple regression analysis.

References

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting statistical summary representations over time. *Psychological Science*, *21*, 560–567. doi:10.1177/0956797610363543.

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research*, *83*, 25–39. doi:10.1016/j.visres.2013.02.018.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*, 157–162.

Attarha, M., Moore, C. M., & Vecera, S. P. (2014). Summary statistics of size: Fixed processing capacity for multiple ensembles but unlimited processing capacity for single ensembles. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 1440–1449. doi:10.1037/a0036206.

Chong, S. C., & Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*, *67*(1), 1–13. doi:10.3758/BF03195009.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*, 28–71. doi:10.1006/cogp.1998.0681.

Corbett, J. E., & Melcher, D. (2014). Stable statistical representations facilitate visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 1915–1925. doi:10.1037/a0037375.

Corbett, J. E., & Oriet, C. (2011). The whole is indeed more than the sum of its parts: Perceptual averaging in the absence of individual item representation. *Acta Psychologica*, *138*, 289–301. doi:10.1016/j.actpsy.2011.08.002.

Corbett, J. E., Wurnitsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual Cognition*, *20*, 211–231. doi:10.1080/13506285.2012.657261.

Dubé, C., & Sekuler, R. (2015). Obligatory and adaptive averaging in visual short-term memory. *Journal of Vision*, *15*(4):13, 1–13, doi:10.1167/15.4.13. [PubMed] [Article]

Duffy, S., Huttenlocher, J., Hedges, L. V., & Crawford, L. E. (2010). Category effects on stimulus estimation: Shifting and skewed frequency distributions. *Psychonomic Bulletin & Review*, *17*, 224–230. doi:10.3758/PBR.17.2.224.

Helson, H. (1947). Adaptation-level as frame of reference for prediction of psychophysical data. *The American Journal of Psychology, 60*(1), 1–29.

Hodsoll, J., & Humphreys, G. W. (2001). Driving attention with the top down: The relative contribution of target templates to the linear separability effect in the size dimension. *Perception & Psychophysics*, *63*(5), 918–926. doi:10.3758/BF03194447.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, *129*, 220–241.

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception, & Psychophysics*, *75*, 278–286. doi:10.3758/s13414-012-0399-4.

Jiang, Y., & Chun, M. M. (2001). Selective attention modulates implicit learning. *The Quarterly Journal of Experimental Psychology: Section A*, *54*, 1105–1124.

Kareev, Y., Arnon, S., & Horwitz-Zeliger, R. (2002). On the misperception of variability. *Journal of Experimental Psychology: General*, *131*, 287–297.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Sciences*, *22*, 79–86.

Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: Effects of set size and item heterogeneity on average size perception.

Acta Psychologica, *142*, 245–250. doi:10.1016/j.actpsy.2012.11.002.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review*, *118*, 280–315. doi:10.1037/a0022494.

Oriet, C., & Brand, J. (2013). Size averaging of irrelevant stimuli cannot be prevented. *Vision Research*, *79*, 8–16. doi:10.1016/j.visres.2012.12.004.

Parducci, A. (1956). Incidental learning of stimulus frequencies in the establishment of evaluation scales. *Journal of Experimental Psychology*, *52*, 112–118.

Parducci, A. (1959). An adaptation-level analysis of ordinal effects in judgment. *Journal of Experimental Psychology*, *58*, 239–246.

Šetić, M., Švegar, D., & Domijan, D. (2007). Modelling the statistical processing of visual information. *Neurocomputing*, *70*, 1808–1812. doi:10.1016/j.neucom.2006.10.069.

Tong, K., Ji, L., Chen, W., & Fu, X. (2015). Unstable mean context causes sensitivity loss and biased estimation of variability. *Journal of Vision*, *15*(4): 15, 1–12, doi:10.1167/15.4.15. [PubMed] [Article]

Teghtsoonian, M. (1965). The judgment of size. *The American Journal of Psychology*, *78*, 392–402.

Utochkin, I. S. (2015). Ensemble summary statistics as a basis for rapid visual categorization. *Journal of Vision*, *15*(4):8, 1–14, doi:10.1167/15.4.8. [PubMed] [Article]

Utochkin, I. S., & Tiurina, N. A. (2014). Parallel averaging of size is possible but range-limited: A reply to Marchant, Simons, and De Fockert. *Acta Psychologica*, *146*, 7–18. doi:10.1016/j.actpsy.2013.11.012.

Weiss, D. J., & Anderson, N. H. (1969). Subjective averaging of length with serial presentation. *Journal of Experimental Psychology*, *82*, 52–63. doi:10.1037/h0028028.

Whiting, B. F., & Oriet, C. (2011). Rapid averaging? Not so fast!. *Psychonomic Bulletin & Review*, *18*, 484–489. doi:10.3758/s13423-011-0071-3.

Zhao, J., Goldfarb, L., & Turk-Browne, N. B. (2013). When numbers and statistics collide: Competition between numerosity perception and statistical learning. *Journal of Vision*, *13*(9): 1087, doi:10.1167/13.9.1087 [Abstract].