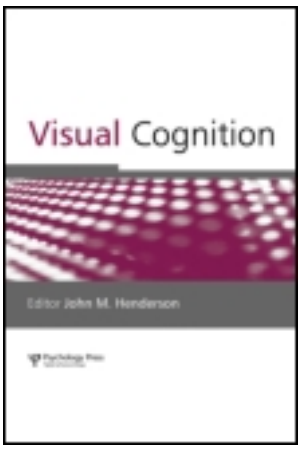

Visual Cognition
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pvis20

Incidental encoding of numerosity in visual long-term memory

Jiaying Zhao [a] & Nicholas B. Turk-Browne [a]
[a] Department of Psychology, Princeton University, Princeton, NJ, USA

Available online: 25 Aug 2011

To cite this article: Jiaying Zhao & Nicholas B. Turk-Browne (2011): Incidental encoding of numerosity in visual long-term memory, Visual Cognition, 19:7, 928-955

To link to this article: http://dx.doi.org/10.1080/13506285.2011.598482

VISUAL COGNITION, 2011, 19 (7), 928–955

Psychology Press
Taylor & Francis Group

Incidental encoding of numerosity in visual long-term memory

Jiaying Zhao and Nicholas B. Turk-Browne

Department of Psychology, Princeton University, Princeton, NJ, USA

The visual system can readily extract numerosity information from brief experiences. This numerical perception is characterized by diminishing accuracy as numerosity increases, and impaired discrimination for similar quantities and large magnitudes. Here we assess whether these properties apply more broadly to numerosity in visual long-term memory. In surprise memory tests, we observed: Remarkable accuracy in estimating the number of repetitions of an exemplar image (Experiment 1a), that this accuracy decreased but remained high when estimating over categories (Experiments 1b and 1c), that numerical discrimination from memory exhibited psychophysical distance and size effects (Experiment 2), that these effects may derive from stored representations rather than post hoc approximation (Experiment 3a), and that they can reflect total elapsed experience in addition to discrete counts (Experiment 3b). Similar to how numerosity is readily extracted during visual perception, our results suggest that numerosity is encoded incidentally in visual long-term memory.

Keywords: Distance effect; Incidental encoding; Numerosity; Size effect; Visual long-term memory.

The distinction between processes geared towards external sensory input (perception) and those that operate over internal mental representations (cognition) is intuitive and historical. However, while the targets of such processes may be different, their underlying computations may overlap in meaningful ways. This approach has been pursued in process models of cognition that emphasize the consequences of perception for memory (Johnson, 1983; Kolers & Roediger, 1984). As a recent example of this perspective, the posterior parietal cortex may implement general attentional

Please address all correspondence to Jiaying Zhao, Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08540, USA. E-mail: jiayingz@princeton.edu

For helpful conversations, we thank Dan Osherson and Mason Simon. Portions of this work were presented at the 2010 meeting of the Cognitive Science Society.

http://www.psypress.com/viscog DOI: 10.1080/13506285.2011.598482

processes involved in searching through both visual input and long-term memory (Cabeza, Ciaramelli, Olson, & Moscovitch, 2008). Beyond relying on overlapping neural representations, here we explore one way in which perceptual and mnemonic processes may share functional properties and constraints. In particular, we focus on numerical cognition—a topic studied extensively in the domain of perception, but less so in the domain of memory.

Numerical perception

The visual system is remarkably good at "seeing" numerical information in the world. In a typical study of numerical perception, arrays of items (e.g., dots) are briefly presented and participants report the number of items in each display. Several features of this *immediate* numerical perception have been discovered based on developmental, behavioural, and neuroscientific findings (see reviews by Ansari, 2008; Dehaene, Dehaene-Lambertz, & Cohen, 1998; Feigenson, Dehaene, & Spelke, 2004). Adults are very accurate and fast at judging small quantities (six or fewer), a process termed "subitizing" (Kaufman, Lord, Reese, & Volkman, 1949). However, for larger quantities, judgements are less accurate but still quite good, a process termed "approximation" (Mandler & Shebo, 1982; Trick & Pylyshyn, 1994). Numerical judgements are subject to Weber's law: As numerosity increases, judgements become less and less accurate following a logarithmic function with fixed Gaussian noise (Izard & Dehaene, 2008). It was initially believed that there is a single estimation process shared for small and large numerosities, and that subitizing reflects the high precision for small numerosity built into Weber's law (Dehaene & Changeux, 1993; Gallistel & Gelman, 1991). However, recent evidence has suggested that subitizing might rely on a separate process dedicated to small numerosity (Dehaene & Cohen, 1994; Feigenson et al., 2004; Piazza, Mechelli, Butterworth, & Price, 2002; Revkin, Piazza, Izard, Cohen, & Dehaene, 2008).

Numerical perception is often examined in tasks requiring the comparison of numerosities. In such tasks, two displays are briefly presented and participants judge which display contained more items. When discriminating between numerosities, performance improves as the numerical distance between numerosities increases (*distance effect*; Dehaene, Dupoux, & Mehler, 1990; Moyer & Landauer, 1967). When distance is held constant, performance declines as the absolute magnitudes of the two numerosities increase (*size effect*; Barth, Kanwisher, & Spelke, 2003; Izard & Dehaene, 2008; Whalen, Gallistel, & Gelman, 1999). Thus, numerical discrimination tracks the ratio between two numerosities (e.g., Banks, Fujii, & Kayra-Stuart, 1976; Buckley & Gilman, 1974; Hintzman, Yurko, & Hu,

1981; Holyoak, 1977; Kosslyn, Murphy, Bemesderfer, & Feinstein, 1977; Moyer, 1973; Parkman, 1971).

Numerical memory

Long-term memory for numerical information is an important part of everyday cognition: Have we visited your or my parents more often? How many books did I borrow from the library? Which bank has the most ATMs in town? In contrast to immediate perceptual judgements, this kind of numerical judgement is based on episodic memory aggregated over a longer timescale. Numerical judgements from long-term memory have previously been examined in the context of event frequency (Blair & Burton, 1987; Brown, 1995, 1997; Hasher & Zacks, 1979; Hintzman & Block, 1971; Howell, 1973; Means & Loftus, 1991, Menon, 1993). In a typical study of this type, lists of words are studied and some of the words are presented repeatedly. Later, participants must recall how many times they had seen a particular word. The corresponding estimates follow a logarithmic function of actual numerosity, and increase with the spacing between repetitions (Hintzman, 1969).

Several accounts have been offered to explain such performance: The strength hypothesis posits that numerical estimates about event frequency are based on the strength of its memory trace, which is determined by the number of event repetitions (Hintzman, 1969). The multiple-trace hypothesis posits that numerical estimates reflect the number of stored memory traces of an event rather than the strength of any single memory trace (Hintzman & Block, 1971). Finally, a more recent hypothesis posits that numerical estimates are influenced by the number of contexts in which events occur (Brown, 1995, 1997, 2002): A category label (*CITY*) paired with multiple contexts (*Boston*, *London*, and *Cleveland*) is judged to have lower numerosity than one paired with a single context an equal number of times (*London*, *London*, *London*).

The current study

Since the event frequency literature largely predates the numerical perception literature, our understanding of how numerosity information is stored and retrieved from long-term memory may benefit from the approaches that have been developed for the study of numerical perception. The overall goal of our study is thus to bridge these two domains, and test the similarities and differences between numerical judgements from visual long-term memory (VLTM) and those from immediate visual perception.

Here we consider VLTM to be a type of episodic memory for visual information that is formed over repeated experiences and persists for an

extended timescale. Visual short-term memory (VSTM) is certainly important for numerical perception tasks, where two arrays must be held in mind and compared, or where a single array is presented briefly and subsequently tested after a short delay. Memory is also involved in numerical perception tasks in terms of maintaining task rules, such as stimulus–response mappings, but such memory is unrelated to the numerosity of particular items. Our study extends beyond this prior involvement of memory in numerosity judgements by focusing on VLTM. Indeed, VLTM and VSTM are different in several ways, including: (1) VLTM has a remarkably large capacity (e.g., Brady, Konkle, Alvarez, & Oliva, 2008; Standing, 1973), whereas VSTM is severely capacity limited to about four items (e.g., Luck & Vogel, 1997); and (2) VLTM lasts for a long time (at least a week; Shepard, 1967), whereas VSTM decays rapidly (in about 10 s; Zhang & Luck, 2009).

To our knowledge, this study is the first attempt to apply tools from numerical perception to the study of VLTM. We examine whether psychophysical properties of numerical perception—precise estimates of small quantities, the distance effect, and the size effect—generalize to VLTM. Importantly, unlike previous studies of numerical perception, numerosity in our study derives from aggregating over extended time periods and interruptions by other stimuli. Moreover, in typical numerical perception tasks, observers know in advance that they will be asked about numerosity (since they complete multiple trials); in our study, numerosity is never mentioned until after all stimuli have been encountered, ensuring that judgements reflect incidental encoding. In addition, we examine whether numerosity can be tracked for individuals (e.g., an exemplar repeated multiple times), and for features shared among individuals (e.g., a category from which multiple exemplars are repeated). Finally, we examine how numerical judgements are formed in VLTM, such as whether numerosity is stored directly in memory or calculated post hoc from a set of retrieved memories.

As an outline, we first test the accuracy of numerical estimates from VLTM (Experiments 1a, 1b, and 1c); we then test psychophysical properties of numerical comparison from memory (Experiment 2); and finally we explore potential mechanisms for how numerical estimates can be generated from memory (Experiments 3a and 3b). In all cases, numerical estimates are obtained from a surprise test, after a large number of objects have been incidentally encoded into VLTM.

EXPERIMENT 1A

The purpose of this experiment is to test the accuracy of unexpected numerosity judgements from VLTM.

Method

Participants. Participants in all experiments had normal or corrected-to-normal vision, and provided informed consent. All experiments were approved by the Institutional Review Board for human subjects at Princeton University. In the first three experiments (1a, 1b, and 1c), a total of 60 Princeton University undergraduates (20 per experiment) participated in exchange for course credit (39 female, mean age = 19.3 years).

Materials. Stimuli were chosen from an image set containing 60 object categories, and 10 exemplar images per category. Categories included types of animals, plants, and everyday artifacts. To manipulate numerosity, 50 categories were pseudorandomly assigned to a numerosity between 1 and 10 such that each numerosity level was represented by five categories. One exemplar image was randomly chosen from each of these categories, and was presented separately the corresponding number of times during encoding. For example, if the categories *dog*, *bear*, *car*, *flower*, and *horse* were chosen at numerosity level "3", then one exemplar image from each category would be presented a total of three times, each time intermixed with images from this and other numerosity levels. There were 10 numerosity levels (1–10), five categories per level, and each category was presented the corresponding number of times, resulting in a total of: $5 \times 1 + 5 \times 2 + 5 \times 3 + 5 \times 4 + 5 \times 5 + 5 \times 6 + 5 \times 7 + 5 \times 8 + 5 \times 9 + 5 \times 10 = 275$ images. The order of images was randomized for each participant, but images could not repeat back-to-back. In addition to the 275 images of interest, 10 images appeared at the beginning and another 10 images appeared at the end to buffer against primacy and recency effects. These 20 filler images consisted of two exemplar images from each of the 10 remaining categories (from 60 initially), one exemplar presented at the beginning and the other at the end. Since each filler image was presented once, numerosity level 1 was more frequent than other levels and may have stood out. However, as shown later, participants reliably overestimated small numerosities, suggesting that they were not overly biased to respond "1".

Apparatus. Participants were seated in a darkened room 70 cm in front of a Viewsonic CRT monitor running at 100 Hz. Experimental stimuli were presented using Matlab (Mathworks, Natick, MA) and the Psychophysics Toolbox version 3 (Brainard, 1997; Pelli, 1997). Each image subtended 12.2 degrees of visual angle along its longest axis on the screen.

Procedure. Participants were informed that the experiment consisted of two parts, but were not told that their memory for numerosity would be tested. In the first phase, they were instructed to view each object and to

determine whether it corresponded to a *natural* or *artificial* thing by pressing one of two keys. This cover task prevented participants from adopting an explicit strategy such as counting, and was orthogonal to the primary manipulation because responses were balanced within and across numerosity levels. The same task was used in all experiments, and we do not report performance in the Results sections because accuracy was consistently very high (mean accuracy $> 94\%$ in every experiment).

On each trial, a colour photograph appeared for 2000 ms. Participants entered their response during this time by pressing one of two buttons with their index fingers of both hands (assignment of responses to buttons was counterbalanced). The image disappeared after 2000 ms regardless of whether the participant had responded, followed by a blank screen for 1000 ms. This part of the experiment contained 295 trials and lasted about 15 min. Participants then took a break and completed an unrelated distractor task for 15 min. The purpose of the break was to guarantee that subsequent judgements would be based on long-term memory.

Participants were then given instructions for the second phase of the experiment. They were again presented with photographs of single objects, but now estimated how many times they had seen the object in the first phase of the experiment. They responded between 1 and 10 by pressing a number key from "1" to "0" on the keyboard number line (with "0" used for "10"). The 50 categories were presented in a random order during this part. The estimated numerosity was compared to the actual numerosity from the first phase. Filler pictures were not tested in the second phase. We note that participants often expressed surprise when receiving these instructions. In postexperiment debriefing, no participant reported being aware during the first part that their memory for numerosity would be tested in the second part. These responses suggest that effects we observed reflect incidental encoding of numerosity in VLTM.

Results

We compared estimated numerosity from the second phase against the objective numerosity from the first phase. At every numerosity level we averaged across the five categories for that level, separately for each participant, and then compared these estimates to objective numerosity. The grand mean across participants is shown in Figure 1A.

To quantify performance, estimated numerosities were modelled as a function of objective numerosities using linear regression. Since estimated and objective numerosities ranged from 1 to 10, perfect performance would correspond to a slope of 1 and an intercept of 0 (i.e., estimated = objective). In contrast, chance performance (i.e., guessing) would lead participants to randomly (and thus uniformly) distribute their responses, resulting in a slope of

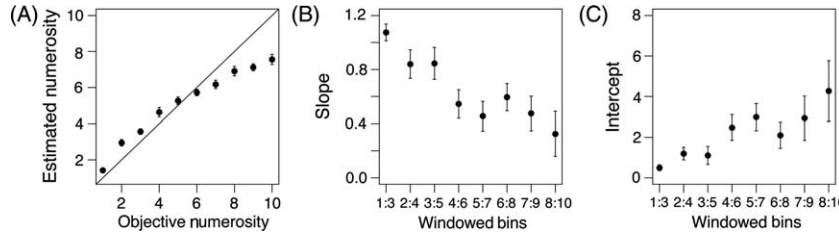

Figure 1. Results from Experiment 1a. (A) Mean estimated numerosity plotted against the number of times each image was presented during the first phase (objective numerosity). (B) Mean slope of a linear model applied to the data in Figure 1A over windows of three numerosity levels (e.g., "1:3" reflects window from 1 to 3 on the x-axis of Figure 1A). (C) Mean intercept of a linear model applied over the same windows. Error bars (often quite small) reflect 1 standard error of the mean.

0. If participants randomly distributed their estimates across all response options, the expected intercept would be 5.5 ([10 + 1]/2).[1] We can therefore judge the accuracy of estimated numerosity on a continuum from perfect performance to chance performance (i.e., slope: 1 → 0, intercept: 0 → 5.5). The linear regression analysis was performed in each participant to obtain a sample of slope and intercept values across participants. The mean of the slope values was 0.64 ($SD = 0.12$, median = 0.64), reliably above chance performance, $t(19) = 24.7$, $p < .01$. The mean of the intercept values was 1.59 ($SD = 0.71$, median = 1.53), again reliably better than chance, $t(19) = 24.5$, $p < .01$.

Prior research has indicated that performance might decline as numerosity increases. In particular, estimates may be precise for small quantities, but systematically underestimate objective numerosity for larger quantities (e.g., Izard & Dehaene, 2008). Accordingly, the extent to which estimated numerosity mirrors perfect versus chance performance may change along the objective numerosity axis. For example, over a low window (e.g., 1:3) slope and intercept values may be close to 1 and 0 respectively, whereas over a high window (e.g., 6:8) slope and intercept values may be close to 0 and 5.5 (or higher), respectively.

We thus ran a linear regression across all possible windows of three contiguous numerosity levels for each participant. The same linear regression analysis as before was run separately on windows [1:3], [2:4], . . . [8:10]. For each window, the mean slope and the intercept values are shown in Figures 1B and 1C. To quantify our results, one-way ANOVAs were performed over slopes and intercepts. There were reliable main effects of

[1] Participants performing at chance may prefer certain responses, and so the chance intercept could differ from 5.5. In such cases, however, we would still expect a slope of 0. Thus, we can assess chance performance irrespective of such response biases. We will nevertheless use 5.5 as a benchmark, since we have no *a priori* reason to believe that such biases would systematically favour low or high responses across participants.

objective numerosity on both measures: Slope, $F(7, 133) = 4.38$, $p < .01$; intercept, $F(7, 133) = 2.34$, $p < .05$. Post hoc Tukey HSD tests revealed that the mean slope for window [1:3] ($M = 1.08$, $SD = 0.28$) was reliably higher than for the rest of the windows, $t(19) = 2.2$, $p < .05$, and the mean intercept for the same window ($M = 0.50$, $SD = 0.65$) was reliably lower than the rest of the intercept values, $t(19) = 2.8$, $p < .05$.

These results suggest that performance starts off near perfect, and declines as a function of objective numerosity. This function may be continuous, with estimated numerosity having a logarithmic or power law relationship to objective numerosity. Such a relationship would be consistent with classic psychophysical laws relating physical and perceived stimulation (Fechner, 1860/1999; Stevens, 1961), and a similar pattern has been observed in a variety of studies of numerical cognition (e.g., Allik & Tuulmets, 1991; Attneave, 1953; Durgin, 1995; Izard & Dehaene, 2008; Nieder & Merten, 2007). This pattern is often interpreted as evidence of analogue magnitude coding in the approximate number system (e.g., Brannon, 2006). In our case, estimated numerosities were well described by both a power law model and a logarithmic model, with a slight advantage to the power law model (Table 1).

To examine whether estimated numerosity resulted from counting, we tested whether the coefficient of variation (CV) at each numerosity level (standard deviation/mean estimated numerosity across trials) could be predicted from mean estimated numerosity. A common property of numerical estimation is that the standard deviation of estimates increases with numerosity (i.e., scalar variability; Cordes, Gelman, Gallistel, & Whalen, 2001). Therefore, the slope relating CVs to estimated numerosity in the absence of counting should be close to zero (Izard & Dehaene, 2008). Overall, there was a slightly negative slope (mean = −0.03), which was reliably below zero in the group, $t(19) = 4.75$, $p < .01$. Further investigation revealed that only five of 20 participants had slopes that were reliably negative ($p < .05$). Combined with the results of Experiment 3a, the weak

TABLE 1
Estimated numerosity modelled by power law and logarithmic models for Experiments 1a, 1b, and 1c

	Power law: y = k* x^a			*Log model:* y = a* log(x) + e		
	k	a	R^2	a	e	R^2
Exp. 1a	1.86	.62	.99	2.69	1.07	.98
Exp. 1b	2.65	.34	.98	1.37	2.48	.95
Exp. 1c	2.31	.46	.99	1.96	1.89	.98

departure from scalar variability observed here in a subset of participants does not provide convincing evidence for explicit counting.

Discussion

Experiment 1a demonstrates that participants can make remarkably accurate numerosity judgements from VLTM, consistent with what has been observed in studies of word frequency judgements (e.g., Hintzman, 1969). Overall, the mean slope reported here is comparable to that obtained for judgements about word frequency from long-term memory (0.64 vs. 0.62/0.66; Brown, 2008). Importantly, our experiment differed from such past studies because participants were not instructed to memorize items for a later test, and instead encoded them incidentally during a cover task. Moreover, in neither this nor the next experiments were participants presented with an explicit category (or context) label during encoding that grouped items and served as a cue for later numerosity judgements. Indeed, the fact that participants were unaware that their memory would be tested and yet were able to achieve high accuracy, suggests that numerosity can be encoded automatically. Our results are consistent with past findings that the encoding of numerosity or frequency requires little effort (Hasher & Zacks, 1979, 1984; cf. Naparstek & Henik, 2010). Despite high overall accuracy, performance started off near perfect, and declined as a function of objective numerosity.

EXPERIMENT 1B

The decline in accuracy for large numerosities in Experiment 1a may be due to the fact that we repeated identical images many times. Such repetition could lead to habituation or reduced attention that would impair further encoding. To test this explanation, here we replicate Experiment 1a, but present multiple exemplars of the same category once, rather than the same exemplar multiple times. If performance in Experiment 1a was affected by habituation, this increased novelty may improve encoding and produce more accurate numerical estimates.

In addition, we interpreted the results of Experiment 1a as reflecting numerical memory for the exemplar images that were repeated multiple times, but participants may have also represented the numerosity associated with the basic-level category from which exemplars were drawn. Indeed, observers may naturally group exemplars into categories in VLTM, as evidenced by interference in recognizing one exemplar from VLTM when many other exemplars from the same category are also stored (Konkle, Brady, Alvarez, & Oliva, 2010). By presenting exemplars only once in this experiment, we can titrate numerical memory associated with categories.

Method

During the first phase, the procedure was identical to Experiment 1a with one important exception: Instead of presenting the same exemplar image from each category *n* times, *n* distinct exemplars were randomly drawn from each category and presented only once. For example, if the category *dog* was assigned to the numerosity level "3", then images of three different dog breeds would each be presented once. During the second phase, the procedure was identical to Experiment 1a except for one change: Category names (e.g., "dog") rather than exemplar images were used to elicit estimates of how many images of that category had been presented in the first phase. Thus, 50 category names were presented in a random order during the second phase.

Results

Data were analysed in the same manner as Experiment 1a (Figure 2). The mean slope relating estimated to objective numerosities across participants was 0.33 ($SD = 0.16$, median = 0.34), reliably *lower* than the mean slope ($M = 0.64$) in Experiment 1a, $t(38) = 6.9$, $p < .01$. The mean intercept across participants was 2.72 ($SD = 1.18$, median = 2.58), reliably larger than the mean intercept ($M = 1.59$) in Experiment 1a, $t(38) = 3.7$, $p < .01$.

We again explored the stationarity of estimates by computing the slopes and intercepts of linear functions over three-level windows of objective numerosity. Despite the relatively poorer performance in this experiment, visual inspection of Figures 2B and 2C revealed a qualitative difference between windows [3:5] and [4:6]. One-way ANOVAs did not reveal a main effect of numerosity on intercept values, $F(7, 133) = 0.99$, $p > .44$, or a main effect of numerosity on slope values, $F(7, 133) = 1.27$, $p > .27$.

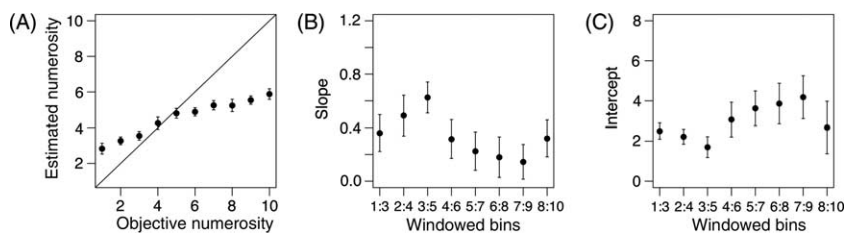

Figure 2. Results from Experiment 1b. (A) Mean estimated numerosity plotted against the number of exemplars of each category from the first phase (objective numerosity). (B) Mean slope of a linear model applied to the data in Figure 2A over windows of three numerosity levels (e.g., "1:3" reflects window from 1 to 3 on the x-axis of Figure 2A). (C) Mean intercept of a linear model applied over the same windows. Error bars reflect 1 standard error of the mean.

Discussion

Providing multiple exemplars for numerosity estimation did not improve accuracy. In fact, performance was worse than in Experiment 1a, where judgements were based on the number of repetitions of a single exemplar. This suggests that habituation or diminished attention cannot fully explain the results of Experiment 1a. The worse performance in this experiment could reflect poor encoding of images presented only once, or source confusion during retrieval in response to a category label (Johnson, Hashtroudi, & Lindsay, 1993). For example, "dog" may retrieve more than three exemplars, with reduced performance reflecting an inability to distinguish exemplars intruding from prior experience. The worse performance could also be due to the use of the availability heuristic when estimating category numerosities (Pandelaere & Hoorens, 2006).

Despite worse performance, overall accuracy was still above chance. This finding is remarkable because it shows that participants were able to maintain numerical memory for 50 categories in parallel, based only on incidental encoding, and aggregated across multiple distinct exemplars presented only once. In contrast, in the perceptual domain, participants can enumerate up to only three colour categories from a single glance (Halberda, Sires, & Feigenson, 2006). Moreover, when stimuli are presented serially over a few minutes, up to three object types can be enumerated successfully, but not four or five types (Feigenson, 2008). This "capacity" difference implies that numerical representations in working memory, even outside of the typical range of VSTM, are more capacity limited than VLTM (Brady et al., 2008).

These findings demonstrate that numerical information can not only be recovered from individual visual stimuli/exemplars, but also from conceptual abstractions over individuals. There are many possible connections between object features, and thus tracking the numerosity of every possible abstraction may be susceptible to combinatorial explosion. The categories used in the current experiment may be the first abstraction to be processed and updated since they were often basic categories (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Such biases in how objects are represented may constrain the number of abstractions that need to be updated.

EXPERIMENT 1C

Although we have attributed the worse performance in Experiment 1b versus Experiment 1a to weaker encoding, it remains possible that more accurate numerical representations existed in memory, but that their *expression* was hampered by a less informative category-label retrieval cue. To examine this possibility, here we pair the first phase of Experiment 1a (multiple repetitions

of a single exemplar) with the second phase of Experiment 1b (category-label probes). If the decline in performance between Experiments 1a and 1b is due solely to the less informative retrieval cue, then the results should mirror Experiment 1b. If instead the decline reflects a difference in the encoding of numerosity for exemplars vs. categories, then the results should mirror Experiment 1a.

Method

The first phase was identical to Experiment 1a: The exemplar image from each category was repeated based on the numerosity level. The second phase was identical to Experiment 1b: Participants were cued by a category name (e.g., "dog") and were asked to estimate how many times they had seen an image from that category.

Results

Data were analysed in the same manner as Experiment 1a (Figure 3). The mean slope relating estimated to objective numerosities was 0.46 ($SD = 0.15$, median = 0.48), reliably higher than the mean slope ($M = 0.33$) in Experiment 1b, $t(38) = 2.6$, $p < .01$, and reliably lower than the mean slope ($M = 0.64$) in Experiment 1a, $t(38) = 4.2$, $p < .01$. The mean intercept across participants was 2.31 ($SD = 0.82$, median = 2.06), which was not statistically lower than the mean intercept ($M = 2.72$) in Experiment 1b, $t(38) = 1.3$, $p > .05$, but was reliably higher than the mean intercept ($M = 1.59$) in Experiment 1a, $t(38) = 2.9$, $p < .01$.

To explore the stationarity of estimates, we computed slopes and intercepts for linear models over windows of three numerosity levels. Replicating the findings from Experiment 1a, one-way ANOVAs revealed main effects of numerosity on both measures: Slope, $F(7, 133) = 3.19$, $p < .01$; intercept, $F(7, 133) = 2.76$, $p < .05$. Moreover, post hoc Tukey HSD tests revealed that the slope values for window [1:3] ($M = 0.90$, $SD = 0.61$) were reliably higher than those of other windows, $t(19) = 2.83$, $p < .05$. The intercept values for the same window ($M = 1.26$, $SD = 1.20$) were reliably lower than the other intercept values, $t(19) = 2.78$, $p < .05$.

Discussion

This experiment confirmed that judgements of numerosity are more accurate for multiple repetitions of the same exemplar than for single presentations of multiple exemplars of the same category. The category label did somewhat impair performance with respect to Experiment 1a, but, critically, cannot entirely explain the poorer performance in Experiment 1b. Finally, this experiment provided a replication of Experiment 1a with a retrieval cue that

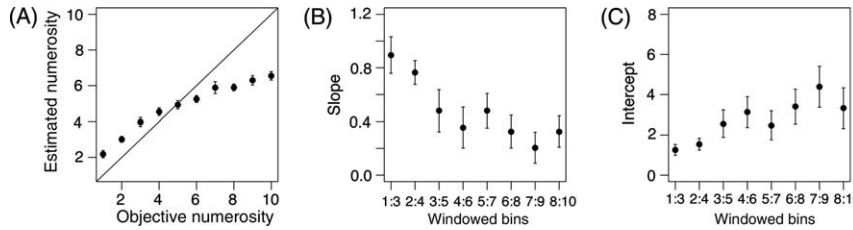

Figure 3. Results from Experiment 1c. (A) Mean estimated numerosity plotted against the number of exemplars of each category from the first phase (objective numerosity). (B) Mean slope of a linear model applied to the data in Figure 3A over windows of three numerosity levels (e.g., "1:3" reflects window from 1 to 3 on the x-axis of Figure 3A). (C) Mean intercept of a linear model applied over the same windows. Error bars reflect 1 standard error of the mean.

was not the actual encoded stimulus, but rather an abstracted property of the stimulus.

EXPERIMENT 2

While the previous experiments focused on judgements of absolute magnitude, numerical cognition is often concerned with judgements about relative magnitude: Which apartment has more space? Which route is faster? What store has the best selection? Here we examine how participants judge which of two objects had appeared more times during incidental encoding into VLTM. Based on the numerical perception literature (Barth et al., 2003; Izard & Dehaene, 2008), we predict that it will be easier to discriminate between objects with bigger differences in objective numerosity (distance effect), and, at a given distance, that it will be easier to discriminate between objects with lower numerosities (size effect).

Method

Participants. Twenty naïve Princeton University undergraduates participated in exchange for course credit (11 female, mean age = 19.6 years).

Procedure. The first phase was identical to Experiment 1a, with one exemplar image repeated multiple times. The second phase was changed to accommodate a discrimination task. On every test trial, two images were displayed side-by-side straddling fixation for 2000 ms. Participants judged which of the images they remembered seeing more times during the first phase by pressing one of two buttons for left and right. Based on objective numerosity levels from incidental encoding, we paired images together so as to fully cover the space of possible distances and proportional distances (for the size effect). At distance 1, we paired an image that was presented *n* times

with one that was presented $n+1$ times. Since numerosity levels range from 1 to 10, there were nine pairs at distance 1 (e.g., 1 vs. 2, 2 vs. 3, etc.). The same pairing method was applied to distances 2, 3, 4, and 5, which resulted in 8, 7, 6, and 5 pairs, respectively. Longer distances were not included due to the small number of possible trials. Thus, a total of 35 pairs were generated. The order of pairs was randomized for each participant, and the position on the screen of the image with the larger numerosity level was randomized.

Results

To test for a distance effect, we pooled all pairs of each distance (regardless of position on the objective numerosity axis) and computed mean accuracy and response time (RT; see Buckley & Gillman, 1974). The results for accuracy and RT are shown in Figure 4. We found that accuracy had a reliable positive correlation with distance across participants, mean Fisher's $z_r = .21$, $t(19) = 5.72$, $p < .01$, and that RT had a reliable negative correlation with distance across participants, mean Fisher's $z_r = -.08$, $t(19) = 2.15$, $p < .05$.

To test for a size effect, we pooled all of the pairs with a base (minimum) numerosity of 1–5 on the objective numerosity axis (regardless of distance) and computed mean accuracy and RT (Dehaene et al., 1998). We excluded base numerosities 6–9 because the distributions of possible distances for these numerosities were different from each other and from 1–5 (which each had one pair for every distance). Therefore this analysis tests for a pure effect of size equating for distance. We found that accuracy had a reliable negative correlation with size across participants, mean Fisher's $z_r = -.25$, $t(19) = 7.04$, $p < .01$, and that RT had a reliable positive correlation with size across participants, mean Fisher's $z_r = .13$, $t(19) = 2.48$, $p < .05$. To demonstrate the robustness of the size effect, we also ran a correlation analysis between accuracy and the base numerosity *within each distance* (as opposed to collapsing across distance, as earlier). Since there was only one trial for each size/distance pair (and thus the dependent variable, accuracy, was binary 1/0), we used a nonparametric Spearman's rank correlation. We found a reliable negative correlation between accuracy and base numerosity for all distances (all ps $< .05$). In other words, the size effect was observed at every distance.

The previous analysis required that we make an assumption about the relationship between objective numerosity and participants' subjective representations. However, we can also access these representations by using the estimated numerosity data from Experiment 1a as a proxy for their (psychologically transformed) magnitude representations. For every pair, we thus computed the difference in estimated numerosity (averaged across all participants from Experiment 1a) for each number in the pair. We found that accuracy had a reliable positive correlation with the difference in estimated

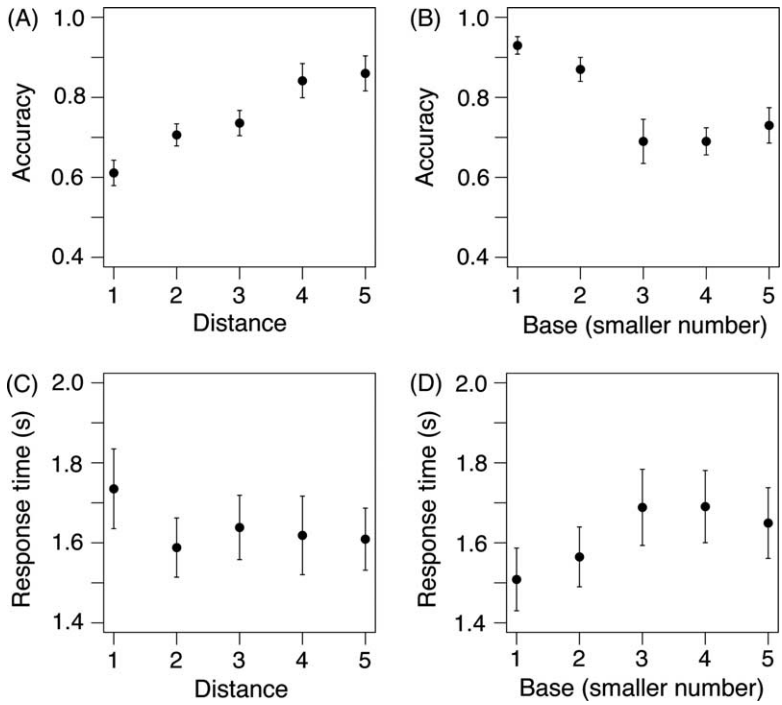

Figure 4. Results from Experiment 2. (A) Mean discrimination accuracy as a function of distance between two numerosities in a pair. (B) Mean discrimination accuracy as a function of base (smaller) numerosity in a pair. (C) Mean RT as a function of distance. (D) Mean RT as a function of base numerosity. Error bars reflect 1 standard error of the mean.

numerosity across participants, mean Fisher's $z_r = .29$, $t(19) = 8.15$, $p < .01$, and that RT had a reliable negative correlation with estimated difference across participants, mean Fisher's $z_r = -.12$, $t(19) = 3.10$, $p < .01$. Moreover, for every participant we compared the strength of correlation between the distance in objective numerosity and accuracy (mean $z_r = .21$) with the strength of correlation between the distance in estimated numerosity (from Experiment 1a) and accuracy (mean $z_r = .29$). Estimated numerosity was a better predictor of the distance effect than objective numerosity, $t(19) = 5.55$, $p < .01$.

Discussion

Despite an entirely different paradigm and timescale, our results were in line with previous findings on distance and size effects in numerical perception. When judging which exemplar appeared more times, participants were more accurate and faster when the difference between the numbers of repetitions was larger. Holding the distance constant, participants were also more

accurate and faster when the base numerosity of presentations going into the distance computation was smaller.

These findings also help interpret the results of Experiment 1a. Specifically, the nonlinear estimated numerosity in Experiment 1a could in principle reflect the fact that responses were bounded between 1 and 10. According to this account, participants may have had linear representations in mind, but they were artificially compressed by the response options. This would predict, however, that the estimated numerosities from Experiment 1a would not be good indices of the numerical representations underlying discrimination in this experiment (and that objective numerosities may provide better, linear indices). Instead, we found that estimated numerosities from Experiment 1a were better predictors of discrimination performance, supporting the interpretation that numerosity in VLTM scales subadditively.

EXPERIMENT 3A

We have shown that the psychophysical properties of numerical estimation (Experiments 1a, 1b, and 1c) and numerical discrimination (Experiment 2) from VLTM are similar to those typically observed from immediate visual perception. In the following two experiments we investigate *how* numerical estimates are generated from VLTM. Numerical estimates may be calculated retrospectively during test by retrieving multiple stored memories when probed with an object, and approximating numerosity from that set (the *calculation* hypothesis)—similar to approximating over a set of discrete perceptual objects (e.g., Franconeri, Bemis, & Alvarez, 2009). Alternatively, numerical estimates may be automatically stored and updated in memory as a byproduct of the repeated encoding of an object, and directly accessed during test with no need for further approximation (the *readout* hypothesis).

An analogous distinction has been made in the word frequency literature. The multiple-trace theory (Hintzman & Block, 1971) predicts that frequency judgements are made by estimating over the set of traces for a particular word (similar to the calculation hypothesis). Conversely, the propositional-encoding and numerical-inference hypotheses (Hintzman, 1976; Howell, 1973) state that frequency judgements are based on the retrieval of a counter that is updated continuously during encoding (similar to the readout hypothesis). The relationship between RT and frequency has been used to distinguish between these accounts (Hockley, 1984): According to the multiple-trace theory, RT should increase with frequency because more traces must be retrieved; whereas if frequency is already stored propositionally or numerically in memory, it is unclear why RT should vary with frequency. In the case of word frequency estimation, RT does in fact scale with frequency, consistent with the multiple-trace theory (e.g., Brown, 1995; Hockley, 1984).

Using the same logic, Experiment 3a examines the relationship between RT and objective numerosity in a speeded version of our task. Insofar as estimated numerosity is calculated during retrieval from VLTM, RT should increase when more memories are retrieved (i.e., for larger objective numerosity). This would be consistent with the perceptual literature in which RT scales with the number of objects being enumerated (e.g., Atkinson, Campbell, & Francis, 1976; Kaufman et al., 1949; Lemer, Dehaene, Spelke, & Cohen, 2003; Mandler & Shebo, 1982; Simon, Peterson, Patel, & Sathian, 1998; Trick & Pylyshyn, 1994), and with multiple-trace theory and the word frequency literature cited earlier. Instead, if estimated numerosity is directly read out from memory, RT should not increase with objective numerosity. This would be consistent with the propositional-encoding/numerical-inference hypotheses, and highlight a difference between numerosity judgements for verbal and visual materials.

Method

Participants. Forty-four naïve Princeton University undergraduates participated in Experiments 3a (30 participants) and 3b (14 participants) for course credit (27 female, mean age = 20.18 years).

Procedure. The procedure was identical to Experiment 1a, except that in the second phase participants were instructed to respond as fast as possible (while remaining accurate).

Results

Our earlier experiments did not emphasize response speed, and thus do not provide a clean test of the calculation hypothesis. Collapsing across objective numerosity, RTs in this experiment were much faster than in Experiment 1a (Ms = 1179.43 vs. 1959.50 ms), $t(48) = 5.55$, $p < .01$. Mean estimated numerosity and RT are plotted as a function of objective numerosity in Figure 5. The mean slope relating estimated to objective numerosity was 0.60 ($SD = 0.14$, median = 0.61). The mean intercept was 1.97 ($SD = 0.88$, median = 1.79). Neither of these values differed from Experiment 1a: Slope, $t(48) = 1.23$, $p > .22$; intercept, $t(48) = 1.62$, $p > .11$, suggesting that there was no speed–accuracy tradeoff.

The primary question explored in this experiment was whether it would take longer to produce larger numerical estimates. We thus examined within-subject correlations between objective numerosity and RT (Figures 5B and 5D), and found no reliable relationship, mean Fisher's $z_r = .04$, $t(29) = 1.49$, $p > .14$. We reasoned that estimated numerosity may be a more sensitive test

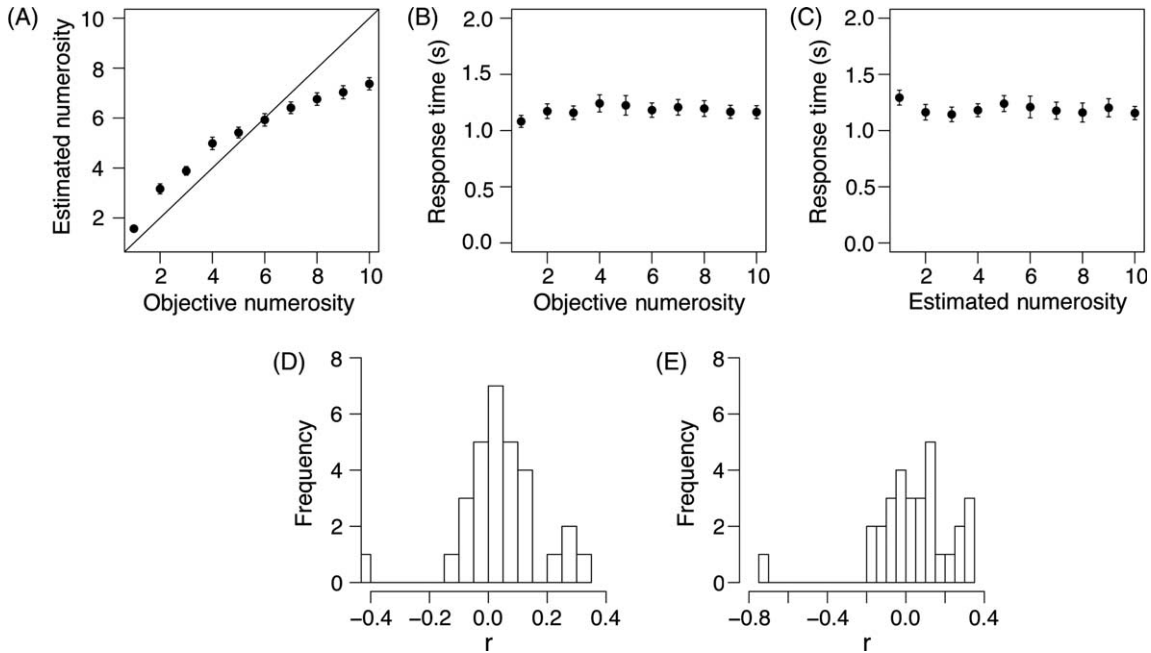

Figure 5. Results from Experiment 3a. (A) Mean estimated numerosity plotted against objective numerosity. (B) Mean RT plotted as a function of objective numerosity. (C) Mean RT plotted as a function of estimated numerosity. Error bars reflect 1 standard error of the mean. (D) Distribution of within-subject correlation coefficients between RT and objective numerosity. (E) Distribution of within-subject correlation coefficients between RT and estimated numerosity.

of this hypothesis (Figures 5C and 5E), but again there was no reliable relationship, mean Fisher's $z_r = .04$, $t(29) = 1.04$, $p > .30$.

Discussion

The amount of time that participants spent estimating numerosity was not related to the actual or reported number of times that the image had been repeated. Inspection of Figure 5 reveals that, in contrast to previous numerical perception studies, responses were neither slower for numerosity levels 5–10 versus 1–4 (e.g., Kaufman et al., 1949), nor progressively slower above the subitizing limit (e.g., Trick & Pylyshyn, 2004). The results are consistent with the readout hypothesis: That numerosity judgements are based on numerical representations stored in VLTM that are accessed during retrieval without further approximation. They also contrast with the word frequency literature, where an effect on RT has been observed (e.g., Hockley, 1984). This discrepancy could be due to the use of visual versus verbal stimuli and/or incidental versus intentional encoding tasks. As always, caution is needed in interpreting null effects. For example, although this is the largest sample size in any of our experiments and numerical estimates were quite accurate, our 10-alternative response task may not have been sensitive enough to detect subtle RT effects.

EXPERIMENT 3B

The results of Experiment 3a fail to support the hypothesis that numerosity from VLTM is calculated at the time of retrieval. The alternative readout hypothesis raises many questions about the nature of numerical representations in VLTM: For example, how are numerical representations updated by visual experience? Numerical representations could reflect a count of discrete stimulus occurrences, and/or a continuous sum of total elapsed stimulus time. To test these possibilities, here we manipulate the duration of stimulus presentations during encoding. If numerical representations are solely determined by a discrete count, then increasing stimulus duration should not influence numerical estimates. If numerical representations are also determined by the total experience with a stimulus, then extended viewing may result in greater numerical estimates during retrieval.

Method

As in Experiment 1a, each objective numerosity level was represented by five object categories. One exemplar image was randomly chosen from each category and was presented the corresponding number of times. The presentation duration of every repetition of that image was fixed at 1, 2, 3,

4, or 5 s (with each duration represented by one image from each category). For example, if the categories *dog*, *bear*, *car*, *flower*, and *horse* were chosen at a particular numerosity level, the exemplar image from each of the categories might always be presented for 3, 1, 5, 2, and 4 s, respectively. Duration assignment was randomized across participants and within each numerosity level. The procedure was identical to Experiment 1a, except that the experiment lasted longer due to the duration manipulation (20 min).

Results

Mean estimated numerosity is plotted as a function of objective numerosity and stimulus duration in Figure 6. Again replicating previous experiments, the linear regression between estimated and objective numerosity revealed a mean slope of 0.62 ($SD = 0.11$, median = 0.60), which did not differ from Experiment 1a, $t < 1$. The mean intercept was 2.24 ($SD = 0.87$, median = 2.19), which was reliably higher than Experiment 1a, $t(32) = 2.40$, $p < .05$. This difference may reflect the greater mean stimulus duration in this experiment.

The primary question that we explored in this experiment was whether image duration would increase estimated numerosity. To assess the relationship between image duration and estimated numerosity, we collapsed across objective numerosity and computed the mean estimated numerosity for each duration. We found that stimulus duration had a reliable positive correlation with estimated numerosity across participants, mean Fisher's $z_r = .08$, $t(13) = 3.88$, $p < .01$. We then ran a linear regression analysis for each participant to quantitatively assess the impact of duration on estimated numerosity. The mean beta coefficient across participants was 0.14

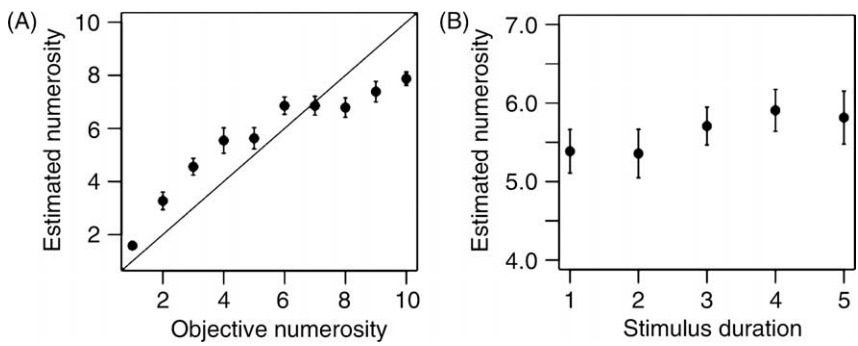

Figure 6. Results from Experiment 3b. (A) Mean estimated numerosity plotted against objective numerosity, collapsing across stimulus duration. (B) Mean estimated numerosity plotted against the duration of each presentation of a stimulus, collapsing across objective numerosity. Error bars reflect 1 standard error of the mean.

($SD = 0.15$, median $= 0.1$) and was reliably greater than zero, $t(13) = 3.59$, $p < .01$. Thus, each additional second of stimulus duration counted as 14% of a repetition.

Discussion

When estimating numerosities, participants were influenced by how long the image had been presented during incidental encoding. This finding supports the idea that numerical representations can be influenced by the total amount of experience with a stimulus, in addition to a discrete count of the number of stimulus occurrences.

Since previous studies have found Stroop-like interference of numerical cues with duration processing (e.g., Dormal, Seron, & Pesenti, 2006), the effect observed here might be due to Stroop-like interference of duration on numerosity. However, in our experiment the duration was experienced well before numerosity judgements (15–30 min), and there were many intervening items and durations presented before numerosity judgement. For these reasons, we do not think the effect of duration on numerosity can be explained by Stroop-like interference.

It should be noted that stimulus duration does not have the same effect on estimated numerosity as the discrete count: The observed slope for duration corresponded to a smaller change in estimated numerosity than the change associated with an additional repetition. Specifically, in Experiment 1a each objective numerosity level (2 s stimulus repetition) increased estimated numerosity by .64, whereas in Experiment 3b an equivalent duration step (2 s presentation time) caused a more modest change of .28. Thus, although total elapsed experience plays a role, numerical estimates from VLTM may also be influenced by the number of repetitions per se.[2]

GENERAL DISCUSSION

Summary

Across six experiments, we demonstrated that numerosity judgements from VLTM can be accurate, and that accuracy is especially high for a smaller number of repetitions. Moreover, judgements for multiple repetitions of the same exemplar image were more accurate than those for single presentations of multiple exemplars from a category. This decrement could not be

[2] We infer a role for discrete repetitions from the fact that duration caused an increase in estimated numerosity of half the size of an extra repetition (with a matching increase in duration). The independent contribution of repetitions could be quantified by holding the total duration of an item constant, and manipulating the number of repetitions. We thank a reviewer for this suggestion.

explained solely by the informativeness of the retrieval cue used to probe numerical memory. When discriminating between two numerosities, performance increased with their distance from each other, but decreased as the magnitude or absolute size of the numerosities increased. These findings are largely in agreement with studies of numerical perception. We then explored how numerical estimates can be generated from long-term memory. The amount of time it took to make numerical judgements did not correlate with magnitude, suggesting that these judgements were unlikely to be based on post hoc calculation over a set of retrieved episodes. Second, numerical judgements correlated with the presentation duration of images, suggesting that these judgements were at least partially based on the total amount of experience with a stimulus. In all experiments, participants expressed surprise at being tested for numerosity, demonstrating that precise estimates of numerosity can be encoded and updated automatically in VLTM.

Exemplar vs. category repetitions

Performance in estimating the number of exemplars of a category (Experiment 1b) was not superior to estimating the number of repetitions of the same exemplar image (Experiment 1a). This result held even when memory was probed with the same instructions and category-label retrieval cue (Experiment 1c). This rules out the possibility that the observed coding of magnitude in memory reflected a failure to attend to or encode multiple repetitions of the same stimulus due to habituation (see Grill-Spector, Henson, & Martin, 2006; Turk-Browne, Scholl, & Chun, 2008). Our result is consistent with the finding that when a word category label (e.g., CITY) was paired with the same context exemplar (e.g., London, London, London), its frequency was estimated to be higher than when it was paired with different context exemplars (e.g., Boston, London, Cleveland) on each repetition (Brown, 1995). Note, however, that participants in our study first encountered the category label in the second phase, purportedly after category-specific numerosity had already been represented.

The fact that estimating the number of exemplars of a category was *worse* than estimating the number of exemplar repetitions was surprising to us. One possible explanation is that stimulus repetition interacts with numerical memory at the categorical level. In particular, presenting exemplars once each in Experiment 1b may have resulted in weak memories such that subsequent retrieval based on a category label was more susceptible to failures of source monitoring (Dougherty & Franco-Watkins, 2003; Johnson et al., 1993). In other words, due to the relative weakness of experimental memories associated with the cue, participants may have been less able to screen out matching extraexperimental memories (e.g., knowing that I saw three dogs in the experiment, separate from the five seen earlier in the day).

The greater strength of the experimental memories in Experiment 1c due to multiple stimulus repetitions may have mitigated such intrusions. This account could be tested directly by including both repetitions of individual exemplar images and multiple exemplars from the same category.

Another possibility is that numerical representations in long-term memory are more precise when attached to specific versus abstract features of a stimulus. In other words, numerical memory may be stored and updated primarily at the level of individual stimuli rather than at higher levels of the category hierarchy. If true, this numerical bias for exemplars could be analogous to the processing priority received by basic object categories (e.g., Rosch et al., 1976). Nevertheless, the fact that numerical estimates for categories were even moderately accurate suggests that summary features (in this case numerosity) can be attached to multiple levels of the conceptual hierarchy in VLTM.

It remains possible that our superordinate natural/artificial cover task focused attention at a level of categorization higher than what was probed in Experiment 1b. At the same time, exemplar numerosity was encoded in spite of the superordinate task. Future research could: (a) manipulate cover tasks to explore whether the automatic encoding of numerosity in VLTM is affected by task demands, or (b) change the retrieval cue to probe different levels of categorization. A final and related possible explanation for the exemplar superiority in numerical visual memory is that there may have been variability in representativeness of exemplars we chose for a particular category. Thus, exemplars may have differentially contributed to the numerosity associated with a category, and this may have been lessened when a single exemplar was presented repeatedly (and could thus serve as the sole anchor for that category).

Exact and approximate representations of numerosity

Estimated numerosity was highly accurate for a small number of repetitions (generally up to five repetitions), after which estimates seemed to plateau. Moreover, performance declined—changing from near-perfect to more chance-like—over windows moving continuously from low to high quantities. Our results suggest that incidental encoding in long-term memory may rely on an exact system for small quantities coupled with an approximate system for larger quantities (e.g., Feigenson et al., 2004). At the same time, it can be difficult to conclusively interpret this pattern of results as emerging from an abrupt break between two systems because the pattern can also be represented by a continuous logarithmic or power function. Such functions have been used to argue for the existence of a single process that operates over both small and large quantities (e.g., Dehaene & Changeux, 1993; Durgin, 1995).

To adjudicate between these accounts, one can rely on the fact that the logarithmic function implies equal discriminability within sets of numbers that have the same ratios—whether 1–8 or 10–80 in steps of 10 (decades)—but that for the exact system, numbers 1–4 are special. A recent study used this logic and found support for an exact number system: Participants were much faster and more accurate at naming numbers 1–4 than 5–8, but there was no difference between 10–40 and 50–80 (Revkin et al., 2008). In this context, the lack of an RT effect in Experiment 3a is quite surprising. Despite being more accurate for numerosities 1–4, the fact that our participants were not faster in this range suggests that numerical memory for small quantities may not be equivalent to subitizing.

The distance and size effects we observed are in line with studies of immediate visual perception (Barth et al., 2003; Whalen et al., 1999). The prevailing explanation of these effects is that representations of large numbers are approximate and imprecise, and, as a result, that discrimination between two close large numbers is difficult or impossible (leading to chance discrimination performance). Indeed, accuracy in Experiment 2 dropped to 50% for pairs with both magnitudes above 8. These results provide evidence for an analogue representation of numerosity in VLTM, where the number of items in the set is represented by a magnitude that is a linear function of the cardinal value of the set, and that discrimination is a function of the log ratio of the quantities (see reviews by Brannon, 2006; Gallistel, 1990; Wynn, 1998). Such representations appear universal across development and species, and here we demonstrate that they can be aggregated incidentally over much longer time scales during visual experience.

The source of numerosity in long-term memory

Our initial experiments demonstrated that numerical memory is robust, and that it exhibits similar psychophysical properties to numerical perception. The remaining experiments explored the nature of these representations and, in particular, where they come from. Experiment 3a tested whether numerical representations are calculated during retrieval. However, we did not observe an effect of numerosity on RT despite clear accuracy effects. This result is incompatible with a multiple-trace model in which discrete episodes can be individuated, but where the number of such episodes is not an intrinsic part of the representation and must be calculated (Hintzman, 1976).

The lack of an RT effect is more consistent with the hypothesis that numerosity estimates are based on a readout of existing representations stored in VLTM. Several models may help explain the nature of such representations. One possibility is that numerical estimates are based on the strength of memory for the repeated stimulus, which could be influenced both by the number of presentations and the duration of exposure to the

stimulus (Hintzman, 1969, 1976). Our findings are also compatible with an accumulator model (Meck & Church, 1983), where the accumulator is based on a count of discrete experiences and/or a sum of their durations. The original findings that supported this theory focused on the count and total duration of a continuous train of auditory noises. Applying this theory to our data would require positing multiple accumulators (i.e., one for each exemplar and category), which could operate in parallel across delays and interruptions (cf. Halberda et al., 2006).

Many questions remain: For example, although estimated numerosity exhibited nonlinearity, it is unclear whether nonlinear estimates arise because of marginal gains in strength as a function of repetition, or because of an internal psychophysical function mapping magnitude representations stored in memory to retrieved and reported estimates.

Conclusions

We found that numerical memory has detailed similarities to numerical perception. In addition, some of the novel findings from our study—such as differences between exemplar and category numerosity, and effects of stimulus duration—suggest interesting avenues for future research on numerical cognition. For example, it may be fruitful to examine whether task-irrelevant stimulus dimensions that influence numerical memory also influence numerical perception (e.g., duration), and vice versa (e.g., area; Hurewitz, Gelman, & Schnitzer, 2006). Overall, our findings suggest that analogous mechanisms may operate in numerical perception and numerical memory.

REFERENCES

Allik, J., & Tuulmets, T. (1991). Occupancy model of perceived numerosity. *Perception and Psychophysics*, *49*, 303–314.

Ansari, D. (2008). Effects of development and enculturation on number representation in the brain. *Nature Reviews Neuroscience*, *9*, 278–291.

Atkinson, J., Campbell, F. W., & Francis, M. R. (1976). The magic number 4 ± 0: A new look at visual numerosity judgements. *Perception*, *5*, 327–334.

Attneave, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology*, *46*, 81–86.

Banks, W. P., Fujii, M., & Kayra-Stuart, F. (1976). Semantic congruity effects in comparative judgments of magnitudes of digits. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 435–447.

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition*, *86*, 201–221.

Blair, E., & Burton, S. (1987). Cognitive processes used by survey respondents to answer behavioral frequency questions. *Journal of Consumer Research*, *14*, 280–288.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*, 14325–14329.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Brannon, E. M. (2006). The representation of numerical magnitude. *Current Opinion in Neurobiology*, *16*, 222–229.

Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1539–1553.

Brown, N. R. (1997). Context memory and the selection of frequency estimation strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 898–914.

Brown, N. R. (2002). Encoding, representing, and estimating event frequencies: A multiple strategy perspective. In P. Sedlmeier & T. Betsch (Eds.), *Frequency processing and cognition* (pp. 37–53). Oxford, UK: Oxford University Press.

Brown, N. R. (2008). How metastrategic considerations influence the selection of frequency estimation strategies. *Journal of Memory and Language*, *58*, 3–18.

Buckley, P. B., & Gilman, C. B. (1974). Comparison of digits and dot patterns. *Journal of Experimental Psychology*, *103*, 1131–1136.

Cabeza, R., Ciaramelli, E., Olson, I. R., & Moscovitch, M. (2008). The parietal cortex and episodic memory: An attentional account. *Nature Reviews Neuroscience*, *9*, 613–625.

Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin and Review*, *8*, 698–707.

Dehaene, S., & Changeux, J. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience*, *5*, 390–407.

Dehaene, S., & Cohen, L. (1994). Dissociable mechanisms of subitizing and counting: Neuropsychological evidence from simultanagnosic patients. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 958–975.

Dehaene, S., Dehaene-Lambertz, G., & Cohen, L. (1998). Abstract representations of numbers in the animal and human brain. *Trends in Neurosciences*, *21*, 355–361.

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 625–641.

Dormal, V., Seron, X., & Pesenti, M. (2006). Numerosity-duration interference: A Stroop experiment. *Acta Psychologica*, *121*, 109–124.

Dougherty, M. R. P., & Franco-Watkins, A. M. (2003). Reducing bias in frequency judgment by improving source monitoring. *Acta Psychologica*, *113*, 23–44.

Durgin, F. H. (1995). Texture density adaptation and the perceived numerosity and density of texture. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 149–169.

Fechner, G. T. (1999). *Elemente der Psychophysik*. Bristol, UK: Thoemmes Press. Repr. (Original work published 1860).

Feigenson, L. (2008). Parallel non-verbal enumeration is constrained by a set-based limit. *Cognition*, *107*, 1–18.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*, 307–314.

Franconeri, S. L., Bemis, D. K., & Alvarez, G. A. (2009). Number estimation relies on a set of segmented objects. *Cognition*, *113*, 1–13.

Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: Bradford Books/MIT Press.

Gallistel, C. R., & Gelman, R. (1991). Subitizing: The preverbal counting process. In F. Craik, W. Kessen, & A. Ortony (Eds.), *Essays in honor of George Mandler* (pp. 65–81). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*, 14–23.

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science*, *17*, 572–576.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, *108*, 356–388.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *The American Psychologist*, *39*, 1372–1388.

Hintzman, D. L. (1969). Apparent frequency as a function of frequency and the spacing of information. *Journal of Experimental Psychology*, *80*, 139–145.

Hintzman, D. L. (1976). Repetition and memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 47–91). New York, NY: Academic Press.

Hintzman, D. L., & Block, R. A. (1971). Repetition and memory: Evidence for a multiple-trace hypothesis. *Journal of Experimental Psychology*, *88*, 297–306.

Hintzman, D. L., Yurko, D. S., & Hu, J. M. (1981). Two-digit number comparison: Use of place information. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 890–901.

Hockley, W. E. (1984). Retrieval of item frequency information in a continuous memory task. *Memory and Cognition*, *12*, 229–242.

Holyoak, K. J. (1977). The form of analog size information in memory. *Cognitive Psychology*, *9*, 31–51.

Howell, W. C. (1973). Representation of frequency in memory. *Psychological Bulletin*, *80*, 317–331.

Hurewitz, F., Gelman, R., & Schnitzer, B. (2006). Sometimes area counts more than number. *Proceedings of the National Academy of Sciences*, *103*, 3486–3489.

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

Johnson, M. K. (1983). A multiple-entry, modular memory system. In G. W. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 81–123). New York, NY: Academic Press.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*, 3–28.

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

Kolers, P. A., & Roediger, H. L. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior*, *23*, 425–449.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*, 558–578.

Kosslyn, S. M., Murphy, G. L., Bemesderfer, M. E., & Feinstein, K. J. (1977). Category and continuum in mental comparisons. *Journal of Experimental Psychology: General*, *106*, 341–375.

Lemer, C., Dehaene, S., Spelke, E., & Cohen, L. (2003). Approximate quantities and exact number words: Dissociable systems. *Neuropsychologia*, *41*, 1942–1958.

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

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

Means, B., & Loftus, E. F. (1991). When personal history repeats itself: Decomposing memories for recurring events. *Applied Cognitive Psychology*, *5*, 297–318.

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*, 320–334.

Menon, G. (1993). The effects of accessibility of information in memory on judgments of behavioral frequencies. *Journal of Consumer Research*, *20*, 431–440.

Moyer, R. S. (1973). Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception and Psychophysics*, *13*, 180–184.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature*, *215*, 1519–1520.

Naparstek, S., & Henik, A. (2010). Count me in! On the automaticity of numerosity processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1053–1059.

Nieder, A., & Merten, K. (2007). A labeled-line code for small and large numerosities in the monkey prefrontal cortex. *Journal of Neuroscience*, *27*, 5986–5993.

Pandelaere, M., & Hoorens, V. (2006). The effect of category focus at encoding on category frequency estimation strategies. *Memory and Cognition*, *34*, 28–40.

Parkman, J. M. (1971). Temporal aspects of digit and letter inequality judgments. *Journal of Experimental Psychology*, *91*, 191–205.

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

Piazza, M., Mechelli, A., Butterworth, B., & Price, C. J. (2002). Are subitizing and counting implemented as separate or functionally overlapping processes? *NeuroImage*, *15*, 435–446.

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological Science*, *19*, 607–614.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*, 156–163.

Simon, T. J., Peterson, S., Patel, G., & Sathian, K. (1998). Do the magnocellular and parvocellular visual pathways contribute differentially to subitizing and counting? *Perception and Psychophysics*, *60*, 451–464.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, *25*, 207–222.

Stevens, S. S. (1961). To honor Fechner and repeal his law. *Science*, *133*, 80–86.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, *101*, 80–102.

Turk-Browne, N. B., Scholl, B. J., & Chun, M. M. (2008). Habituation in infant cognition and functional neuroimaging. *Frontiers in Human Neuroscience*, *2*, 1–11.

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*, 130–137.

Wynn, K. (1998). Psychological foundations of number: Numerical competence in human infants. *Trends in Cognitive Sciences*, *2*, 296–303.

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

Manuscript received October 2010
Manuscript accepted June 2011