Grouping effects in numerosity perception under prolonged viewing conditions

Leo Poom[1*¶], Marcus Lindskog[1], Anders Winman[1], Ronald van den Berg[1¶]

[1] Department of Psychology, University of Uppsala, Uppsala, Sweden

* Corresponding author

¶ These authors contributed equally to this work.

E-mail: leo.poom@psyk.uu.se (LP)

ABSTRACT

Humans can estimate numerosities – such as the number sheep in a flock – without deliberate counting. A number of biases have been identified in these estimates, which seem primarily rooted in the spatial organization of objects (grouping, symmetry, etc). Most previous studies on the number sense used static stimuli with extremely brief exposure times. However, outside the laboratory, visual scenes are often dynamic and freely viewed for prolonged durations (e.g., a flock of moving sheep). The purpose of the present study is to examine grouping-induced numerosity biases in stimuli that more closely mimic these conditions. To this end, we designed two experiments with limited-dot-lifetime displays (LDDs), in which each dot is visible for a brief period of time and replaced by a new dot elsewhere after its disappearance. The dynamic nature of LDDs prevents subjects from counting even when they are free-viewing a stimulus under prolonged presentation. Subjects estimated the number of dots in arrays that were presented either as a single group or were segregated into two groups by dot proximity, size, color, or motion. Grouping by color and motion reduced perceived numerosity compared to viewing them as a single group. Moreover, the grouping effect sizes between these two features were correlated, which suggests that the effects may share a common, feature-invariant mechanism. Finally, we find that dot size and total stimulus area directly affect perceived numerosity, which makes it difficult to draw reliable conclusions about grouping effects induced by dot proximity and size. Our results provide new insights into biases in numerosity estimation and they demonstrate that the use of LDDs is an effective method to study the human number sense under prolonged viewing.

INTRODUCTION

Humans can estimate the quantity of a set of objects without explicitly counting them. Lately, this cognitive ability has received overwhelming attention in psychological research, much due to a suggested link between the acuity of human number sense and performance on arithmetic tasks [1], as well as the proposal of a dedicated approximate number system to explain this link [2]. However, research on the number sense dates back to as early as the 1870s, when W. Stanley Jevons published a paper in which he reported results of the following experiment, performed on himself:

> *"A round paper box 4½ inches in diameter, lined with white paper, and with the edges cut down so as to stand only ¼ inch high, was placed in the middle of a black tray. A quantity of uniform black beans was then obtained, and a number of them being taken up casually were thrown towards the box so that a wholly uncertain number fell into it. At the very moment when the beans came to rest, their number was estimated without the least hesitation, and then recorded together with the real number obtained by deliberate counting."* [3]

After repeating this laborious procedure for an impressive total of 1,027 trials, Jevons's main finding was that his estimation errors increased with the number of beans landing in the box. This finding has been replicated extensively and is now known to be consistent with Weber's law [4].

Biases in perceived numerosity

A wide range of biases have been identified in the number sense, many of which are related to the spatial arrangement of objects (see Fig. 1 for an example). Although Jevons [3] did not examine

effects of spatial arrangement in his study, he must have been aware of them when he cited the following quote by Hamilton [5]:

> *"If you throw a handful of marbles on the floor, you will find it difficult to view at once more than six, or seven at most, without confusion; but if you group them into twos, or threes, or fives, you can comprehend as many groups as you can units, because the mind considers these groups only as units; it views them as wholes, and throws their parts out of consideration."*

People seem to have known of the existence of interactions between number sense and spatial arrangement long before Hamilton. For example, as noted by Ginsburg [6], the Genesis contains a story in which it is told that Jacob divided the animals that he gifted to his brother Esau into smaller groups, with the intention to make the gift appear more numerous. Several studies have found evidence in favor of Jacob's hypothesis that perceived numerosity increases with the number of groups [7,8], but at least one other study has found an opposite effect [9]. Another class of studies has examined effects of spatial regularity and found that, in general, regularly spaced items are perceived as more numerous than randomly arranged, more clustered items [6,10–13]. Furthermore, both spatial and temporal order effects on perceived numerosity have been reported: stimuli presented on the left are on average perceived as more numerous than stimuli on the right [14] and when two stimuli are presented in sequence, the former tends to be perceived as less numerous than the latter [15]. In addition, it has been found that the total stimulus area and – conversely – object density affects numerosity estimates [16,17]: the more spread out a set of objects, the more numerous they generally appear to be. Finally, it has been found that reducing the symmetry in a stimulus increases both perceived numerosity [18] and

numerosity estimation accuracy [19], while physically connecting objects with line segments decreases perceived numerosity [20,21]. Besides these effects of spatial arrangement, several studies have found that perceived numerosity also depends on the size of the individual objects, although with mixed results: some of these studies report that larger elements are perceived as more numerous than smaller ones [22–25], while others report the opposite effect [6,26–30].

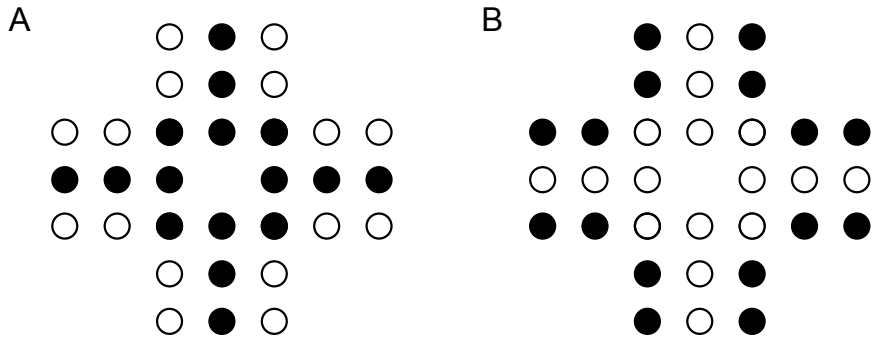

Figure 1 | The solitaire illusion. (A) While the figure contains equal amounts of black and white dots, the black ones appear to be more numerous to most observers. (B) Reversing the colors results in an opposite effect, which indicates that the relative overestimation of the central items is due to the spatial arrangement of the objects and not due to their color.

Numerosity judgments in the laboratory versus numerosity judgments in the wild

Modern-day psychologists have abandoned the practice of tossing beans into a box and instead tend to toss dots onto a computer screen. However, other methodological aspects have hardly changed since Jevons's landmark study [3]. In particular, most experiments still use static stimuli with extremely brief stimulus exposures. While this approach has revealed valuable insights into

basic aspects of the human number sense, it is unclear how representative they are for situations outside the laboratory. For example, when estimating the number of people in a crowd or the number of birds in a flock, the relevant visual information is usually dynamic and available for an extended duration. Indeed, outside the laboratory, scene dynamics – and not exposure time – often seems to be the key limiting factor when estimating numerosities. Since prolonged presentation might be necessary to accomplish reliable perceptual grouping [31,32] and possibly reduces bottom-up influences of basic stimulus features, it is unclear whether biases and other findings found using brief displays generalize directly to situations with prolonged stimulus viewing.

Study aims

The aim of the present study is to explore perceptual grouping effects on perceived numerosity in dynamic stimuli that are freely viewed for prolonged durations. To this end, we develop an experimental paradigm with limited-dot-liftetime displays (LDDs), in which a specified number of dots are visible at any moment, but each individual dot lives for only a brief duration. When a dot disappears, a new dot appears at another location. The dynamic nature of LDDs makes it impossible to count dots outside the subitizing range, even under prolonged viewing of the stimulus. Moreover, the dynamics make that the perceived stimulus area is close to the area specified by the experimenter, which removes unwanted trial-to-trial variability in perceived area, which is typically large when using static stimuli. We use this paradigm to measure perceived numerosity in arrays in which objects either form a single group or are segregated into two groups by the Gestalt laws of proximity, common size, common color, or common motion. Comparison of results across grouping features will enable us to determine whether any identified effects are feature-specific or driven by a feature-invariant grouping mechanism.

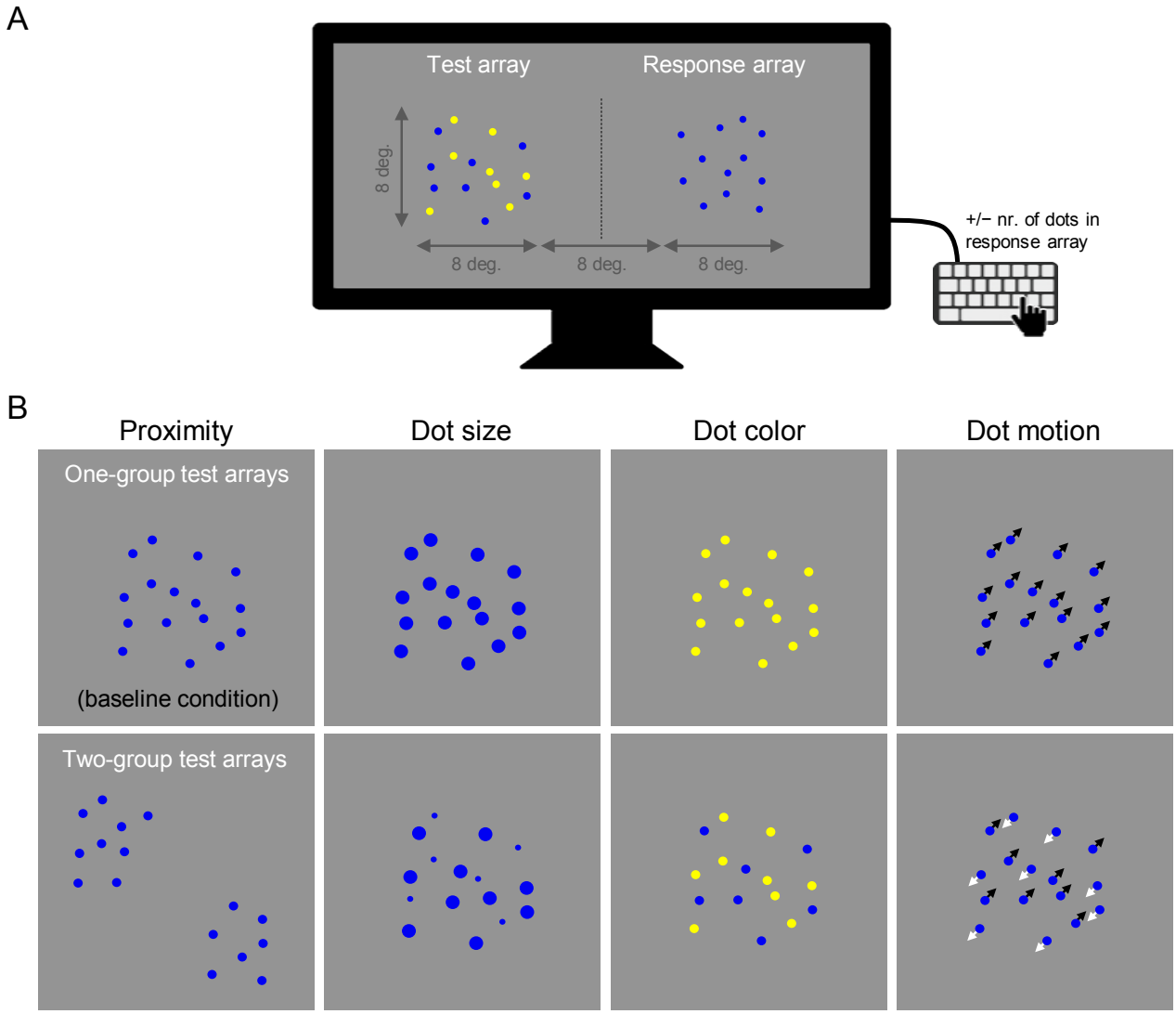

Figure 2 | Stimulus examples in Experiment 1. (A) A snapshot of an example stimulus in Experiment 1 (not to scale). In this example, the test array is composed of two groups, with dot color being the grouping feature. Subjects adjusted the number of dots in the response array through key presses. (B) Snapshots of the eight types of test array used in Experiment 1 (example videos of these stimuli are provided at https://osf.io/vbdf5/). In the one-group conditions (top row), test dots appeared as a single group. In the two-group conditions (bottom row), the dot array was segregated into two groups, based on proximity, size, color, or motion. The visual properties of the adjustable response array were the same throughout the experiment and matched the properties of dots in the one-group proximity condition (top left). Therefore, we refer to this condition as a baseline condition.

EXPERIMENT 1

Methods

Data sharing

The data from this experiment are available at https://osf.io/vbdf5/.

Participants

Fifty-six adults were recruited from the student population at the department of Psychology at Uppsala University (23±2.3 years of age, 31 females). Participants received either course credit or a cinema ticket for their participation. The data of one participant were excluded from the analyses, because of partial loss of data due to a technical error.

Stimuli and procedure

Each test array consisted of a set of dots and was presented on the left side of the screen. A response array with an adjustable number of dots was simultaneously presented on the right side of the screen (Fig. 2A). Participants were instructed to adjust the number of dots in the response array to match the numerosity of the test array. The arrays were horizontally separated by 8 degrees of visual angle and viewed at a distance of approximately 60 cm. A vertical dashed line was shown at the center of the screen to separate the screen into two stimulus areas. Dots in both arrays were presented within areas of 8 × 8 visual degrees, except when dots were grouped by proximity (see below). Dots were randomly scattered within the stimulus area, with the constraint that the minimum center-to-center distance between each pair of dots was 0.80 deg. All dots had a lifetime of 300 milliseconds, with randomized temporal phases to avoid simultaneous replacement. Immediately after a dot disappeared, another one would appear elsewhere.

The experiment contained four one-group conditions (Fig. 2B, top row). In the first of these, the visual properties of the test array were identical to those of the response array (Fig. 2B, top, left panel). We occasionally refer to this condition as the *baseline* condition. In the other three one-group conditions, dots in the test array differed from dots in the response array by their size (1.0 deg. *vs.* 0.50 deg. in the response array), color (yellow *vs.* blue in the response array), or motion (1.5 deg. per second *vs.* stationary in the response array). In addition, the experiment contained four two-group conditions, in which the dots in the test array were divided into groups by proximity, size, color, or motion (Fig. 2B, bottom). In the two-group *proximity condition,* dots were presented within two separate square areas of 32 deg^2 each that were separated by 8 deg. The entire test stimulus was rotated with a random angle around the midpoint between the two groups. In the two-group *size condition*, dots in one group were smaller (Ø=0.25 deg.) and dots in the other group larger (Ø=1.0 deg.) than dots in the response array (Ø=0.50 deg.). In the two-group *color condition,* dots in one group were yellow and dots in the other group blue. Finally, in the two-group *motion condition,* dots in one group travelled with a speed of 1.5 deg. per second in a random direction and dots in the other group travelled with the same speed in the opposite direction. On each two-group trial, one of the two groups was assigned half of the total number of dots plus or minus 0, 1, or 2, dots (randomly chosen) and the remaining dots were assigned to the other group. In each of the 8 described conditions, the test stimulus contained 16 dots on half of the trials and 20 on the other half.

The visual properties of the response array were fixed throughout the experiment: the dots were blue, 0.50 deg. in diameter, stationary during their lifetime, and presented as a single group. The initial number of dots in the response array was on each trial drawn from a uniform distribution on integers 8-24 when the test stimulus contained 16 dots and on integers 10-30 when the test stimulus contained 20 dots. Adjustments in the response stimulus were made by pressing

the "F" and "K" keys to respectively decrease or increase the number of dots. The number of dots in the response array was constrained to the range 0 to 50[a]. Responses were submitted by pressing space bar. No feedback was provided. Stimulus arrays were presented in random order with 8 repetitions, giving a total of 128 trials (8 conditions × 2 numerosities × 8 repetitions). Videos of stimulus examples are available at https://osf.io/vbdf5/.

Analyses

The dependent variable in all analyses is the final number of dots in the response array, which we refer to as the point of subjective equality (PSE). We analyze the data using ANOVAs, t-tests, and Pearson correlation tests. In addition to frequentist *p* values we also report various types of Bayes factors. The first type, denoted BF_{10}, specifies the ratio between the evidence for a hypothesis H_1 relative to another hypothesis H_0, where the latter typically is the "null" hypothesis of no effect. E.g., a finding of BF_{10}=3 in a t-test means that the data are 3 times more likely under the hypothesis that there is an effect (H_1) compared to the hypothesis that the effect size is 0 (H_0). When doing a directed test, we denote the Bayes factor as BF_{+0} (when "H_1: Group 1 > Group 2") or BF_{-0} (when "H_1: Group 1 < Group 2"). Finally, in the case of an ANOVA, $BF_{inclusion}$ denotes the evidence for an effect averaged across all hypotheses that include the effect relative to all hypotheses that do not include the effect[b]. For example, $BF_{inclusion}$=3 for a main effect of some factor F_1 in a multi-factor ANOVA indicates that the data are on average 3 times more likely under hypotheses that include a main effect of F_1 than under hypotheses without this main effect. We use the convention provided by Wagenmakers et al. [33] to label the strength of evidence

[a] Subjects rarely reported numbers close to the edges of the range; over 98% of the responses were in the range [10, 30].

[b] We do not include higher-order interactions in these analyses (by selecting the option "Across matched models" in JASP), because this provides a better comparison with p-values of main effects from a frequentist ANOVA.

provided by a Bayes factor as “extreme”, “very strong”, “strong”, “moderate”, “anecdotal”, or “none”. All analyses were performed using the JASP software package [34].

Results

Subjects matched the number of dots in an adjustable response array to the number of dots in a test array. Dots in the test array were either presented as a single group or segregated into two groups by proximity, size, color, or motion (Fig. 2B). The aim of the experiment is to examine feature-driven effects and grouping-driven effects on perceived numerosity. With “feature-driven effects”, we refer to differences in perceived numerosity that occur due to a feature difference between two conditions. These effects will be examined by comparing the one-group baseline condition – in which the test stimulus had the same visual properties as the response array (Fig. 2B, top left) – with the one-group conditions in which the test stimulus differed from the response array in dot size, dot color, or dot motion. With “grouping-driven effects”, we refer to differences in perceived numerosity that occur due to a difference in the number of perceptual groups between two conditions. These effects will be examined by comparing two-group conditions with corresponding one-group conditions. However, a complication in the latter analysis is that grouping-driven effects can only be reliably assessed when no feature-driven effect was found. The reason for this is that two-group arrays necessarily differ both in terms of its basic visual features and in the number of groups from one-group arrays (for example, when using dot size to segregate dots into two groups, then the two-group arrays contain dots with two different sizes while one-group arrays contain dots with a single size). Therefore, if a feature-driven effect is found, then any difference in perceived numerosity between two-group arrays and one-group arrays may be a purely feature-driven effect or a combination of feature-driven and grouping-

driven effects. We present the results per feature, starting with the features in which we found no feature-driven effects.

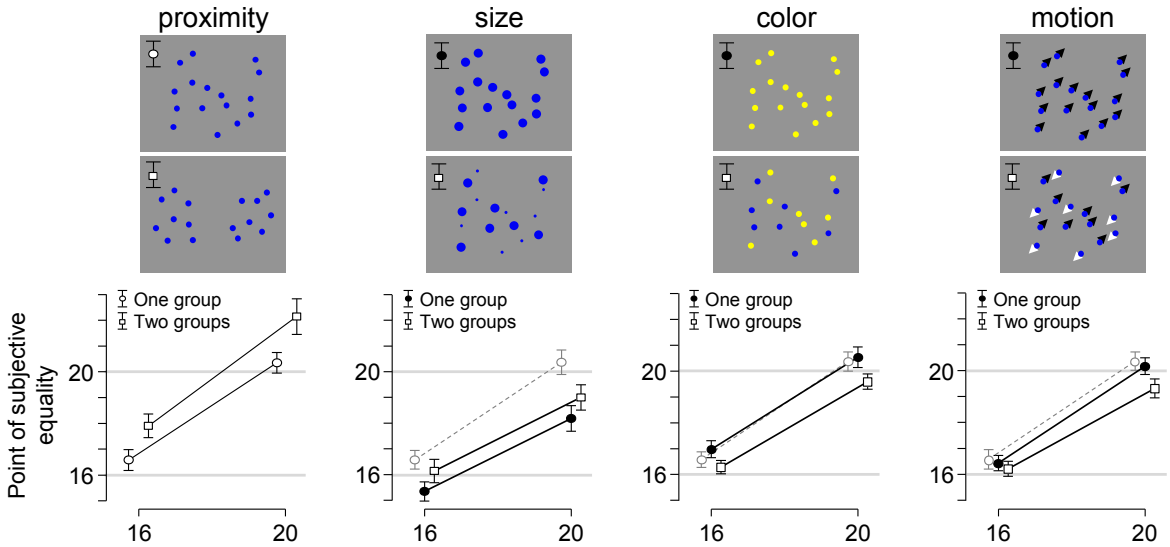

Figure 3 | Subject-averaged numerosity estimates across the 8 stimulus conditions in Experiment 1. Subjects matched the number of dots in an adjustable response array to the number of dots in 8 different types of test array. We refer to the estimates as the point of subjective equality. The visual properties of test array in the one-group proximity condition (top left) were identical to those of the response array. For comparison, we replotted the PSEs in this condition (black circles in left graph) in the results for the size, color, and motion conditions (gray circles). Segregating dots into two groups resulted in an increase in perceived numerosity in the proximity and size conditions and in a decrease in the color and motion conditions. Error bars indicate 95% confidence intervals.

Motion

The average reported number of dots is nearly identical between the one-group motion condition (Fig. 3, right graph, filled circles) and the stationary one-group baseline condition (Fig. 3, right graph, open circles). Indeed, t-tests provide moderately strong evidence in favor of the null hypothesis, both in trials with 16 dots (Δ=0.14, BF_{10}=0.18, p=0.26) and in trials with 20 dots (Δ=0.19, BF_{10}=0.19, p=0.24). This indicates that the presence of motion by itself did not affect perceived numerosity and, therefore, that any difference that we may find between one-group and

two-group conditions is likely to be due to grouping. We find that the average number of reported dots in the two-group motion condition is lower than in the one-group motion condition, both in trials with 16 dots ($\Delta=-0.22$) and in trials with 20 dots ($\Delta=-0.86$). A statistical analysis reveals extremely strong evidence for a decrease in trials with 20 dots ($BF_{-0}=2.06\cdot10^2$, $p<.001$), but results for trials with 16 dots are inconclusive ($BF_{-0}=0.43$, $p=.15$). In summary, dot motion by itself does not seem to affect perceived numerosity, but when using motion to segregate dots into groups, there is evidence that the average perceived number of dots decreases.

Color

The results for color are very similar to the results for motion. The average reported number of dots in one-group trials with yellow arrays was nearly identical to the average number of reported dots in the baseline condition (Fig. 3, third graph from the left; compare filled circles with open circles). Although the results for trials with 16 dots are inconclusive ($\Delta=0.41$, $BF_{10}=0.67$, $p=.96$), we find moderate evidence for the null hypothesis in trials with 20 dots ($\Delta=0.16$, $BF_{10}=0.17$, $p=.71$). Hence, we can be reasonably confident that any difference that we may find between two-group trials and one-group trials is likely to be grouping-driven rather than feature-driven. Just as was the case with motion-based grouping, we find that numerosity estimates were on average lower in trials where dots were grouped by color (Fig. 3, third column, squares) compared to one-group trials (Fig. 3, third column, filled circles). A t-test provides strong evidence for this effect in trials with 16 dots ($\Delta=-0.69$, $BF_{-0}=66.6$, $p<.001$) and extremely strong evidence in trials with 20 dots ($\Delta=-0.94$, $BF_{-0}=45.5\cdot10^2$, $p<.001$). In summary, the results for color are similar to the results for motion: this feature does not directly affect perceived numerosity, but when it is used to segregate a set of dots into two groups, the average reported number of dots decreases.

Dot size

To examine whether dot size directly affected perceived numerosity, we compare estimates in the one-group baseline condition (Fig. 2B, top left) with estimates in the one-group condition with larger dots (Fig. 2B, top, second column). This comparison reveals extremely strong evidence for that perceived numerosity was smaller in the condition with larger dots, both in trials with 16 dots ($\Delta=-1.22$, $BF_{-0}=7.89\cdot10^3$, $p<.001$) and in trials with 20 dots ($\Delta=-2.18$, $BF_{-0}=3.66\cdot10^5$, $p<.001$). As explained above, this feature-driven effect makes it difficult to assess grouping-driven effects, because any effect that we may find between two-group trials and one-group trials could be either due to dot size differences or due to a combination of dot size differences and a difference in the number of groups.

The number of reported dots in the two-group trials (Fig. 3, squares in second graph) is on average larger than in one-group trials with only large dots (Fig. 3, filled circles in second graph). This hypothesis is corroborated by moderately strong statistical evidence on both the trials with 16 dots ($\Delta=0.79$, $BF_{+0}=6.42$, $p=.006$) and trials with 20 dots ($\Delta=0.81$, $BF_{+0}=3.13$, $p=.013$). However, perceived numerosity in the two-group trials is on average *smaller* than in the one-group trials with intermediate dot size (Fig. 3, open circles in second graph). While the difference is not significant for trials with 16 dots ($\Delta=-0.43$, $BF_{-0}=0.84$, $p=.064$), there is extremely strong evidence for an effect in trials with 20 dots ($\Delta=-1.37$, $BF_{-0}=12.1\cdot10^2$, $p<.001$). These opposite effects can be coherently explained as an effect of dot size ("larger dots = smaller perceived numerosity"), but not as a pure grouping effect. If a grouping effect is present in these data – which we cannot rule out nor confirm – then it was obscured by a direct effect of dot size on perceived numerosity. In summary, while these results make it difficult to draw a conclusion about the effect of grouping by dot size on perceived numerosity, they do show clear evidence of

feature-driven effects of dot size in stimuli with prolonged viewing, similar to previous findings using brief, static displays.

Proximity

On average, subjects reported dots in two-group arrays as being more numerous than in one-group arrays (Fig. 3, left). Paired t-tests provide extremely strong evidence for this, both in trials with 16 dots (Δ=1.34, BF_{+0}=$33.2\cdot10^2$, p<.001) and in trials with 20 dots (Δ=1.81, BF_{+0}=$35.1\cdot10^2$, p<.001). We see two possible explanations for this effect. First, it could be an effect of grouping: when dots are divided into two spatially separated groups, they are perceived as more numerous than when presented as a single group. However, this would be opposite to the grouping effects that we found in the conditions where dots were grouped by motion or color. An alternative explanation is that the effect is caused by a difference in the total response array area between two-group and one-group trials (Fig. 3A, left column). The current data do now allow us to distinguish between these two hypotheses and will address this in Experiment 2.

Correlation analysis

We next examine whether there is evidence for correlations in effect sizes between the different features. To do so, we compute for each subject the trial-averaged percentage of overestimation in two-group arrays compared to one-group arrays, separately for each of the four grouping features and averaged across the two set sizes[c]. The results of a Bayesian analysis (Table 1)

[c] For example, if a subject reported for a particular feature on average 20 dots on 20-dot two-group trials and 19 dots on 20-dot one-group trials, then the subject overestimated two-group numerosities by 5% relative to one-group numerosities for this feature. The same procedure was used to compute the percentage of overestimation in the 16-dot trials for this feature and the average of the two percentages was then used in the correlation analysis. Negative numbers indicate underestimation.

reveal strong evidence for a correlation between color-based and motion-based grouping effects on perceived numerosity, but no other correlations[d]. This finding suggests that effects of color-based and motion-based grouping on numerosity estimates originate from a common, feature-invariant source, while the opposite effects found in conditions with proximity-based and dot-size-based grouping may have distinct origins.

Table 1

Correlations between influences of the four grouping features on perceived numerosity.

Bayesian Pearson Correlations

		Proximity	Size	Color	Motion
Proximity	Pearson's r	—			
	BF_{+0}	—			
Size	Pearson's r	0.146	—		
	BF_{+0}	0.498	—		
Color	Pearson's r	-0.172	0.111	—	
	BF_{+0}	0.079	0.364	—	
Motion	Pearson's r	0.095	0.009	0.378	—
	BF_{+0}	0.319	0.177	17.12 *	—

Note . For all tests, the alternative hypothesis specifies that the correlation is positive.
* $BF_{+0} > 10$

Interaction effects

While we are mainly interested in feature-driven and grouping-driven biases *within* features, for completeness we also perform a 2 × 2 × 4 (number of dots × number of groups × grouping feature) within-subjects ANOVA on the entire dataset (Table 2). The results indicate extremely strong evidence for an interaction between number of groups and grouping feature on perceived

[d] Results from a frequentist analysis are consistent with the Bayes Factors: the correlation between motion and color is significant (p=.004) and all others are non-significant (p>.21).

numerosity. Moreover, we find extreme evidence for main effects of both number of test dots and grouping feature, but no evidence for a main effect of number of groups. Finally, there is moderate evidence for an interaction between number of dots and grouping feature, with more dots resulting in larger differences.

Table 2

Results from frequentist and Bayesian repeated-measures ANOVAs performed on the data from Experiment 1 (higher-order interactions are excluded).

	Sum of Squares	**df**	**F**	**p**	**$BF_{inclusion}$**	**η^2**
#Dots	2582	1, 54	1077	< .001	$6.05 \cdot 10^{135}$	.95
#Groups	13.8	1, 54	3.24	.077	0.75	.057
Feature	389	3, 162	38.0	< .001[a]	$1.57 \cdot 10^{29}$	.41
#Dots × #Groups	0.421	1, 54	0.42	.521	0.13	.008
#Dots × Feature	38.5	3, 162	6.72	< .001	6.01	.11
#Groups × Feature	211	3, 162	23.0	< .001[a]	$2.20 \cdot 10^{13}$	.30
#Dots × #Groups × Feature	8.30[a]	3, 162	1.78	.154[a]	0.07	.032

[a] Sphericity assumption was violated, but Greenhouse-Geisser and Huynd-Feldt corrections did not substantially change the results.

Discussion

We find differences in perceived numerosity between two-group and one-group arrays in all four tested grouping features. However, there is variation across the tested features in both the effect direction and the conclusions that we are able to draw about the origin of the effects. Grouping by color and motion led in both cases to a decrease in the average perceived number of dots. For both features, we found evidence against feature-driven effects on perceived numerosity and no evidence in favor of such effects. Moreover, we found strong evidence for a correlation in the

effect sizes between these two features. Altogether, these results suggest that the grouping effects in the color and motion conditions are caused by a shared, feature-invariant mechanism. By contrast, grouping by dot size and proximity both *increased* perceived numerosity. However, for these two features we could not rule out that the effects may have been feature-driven rather than due to a difference in the number of groups. In the dot-size conditions, we actually found strong evidence in favor of such effects. This does not exclude the possibility that grouping-driven effects were also present in those data, but we are at present unable to separate the two types of effect. In the proximity conditions, the identified effect may have been caused by a difference in total array area between the two-group and one-group stimuli. Since Experiment 1 did not contain a manipulation of array area independent of the number of groups, we perform a second experiment to examine this possibility in more detail. A secondary aim of Experiment 2 is to test whether the evidence for the null effects for color and dot size are robust under small changes in experimental conditions.

EXPERIMENT 2

Methods

Data sharing

The data from this experiment are available at https://osf.io/vbdf5/.

Participants

Eighty-seven adults were recruited from the student population at the department of Psychology at Uppsala University (24.8±6.9 years of age, 59 females). Participants received either course credit or a cinema ticket for their participation

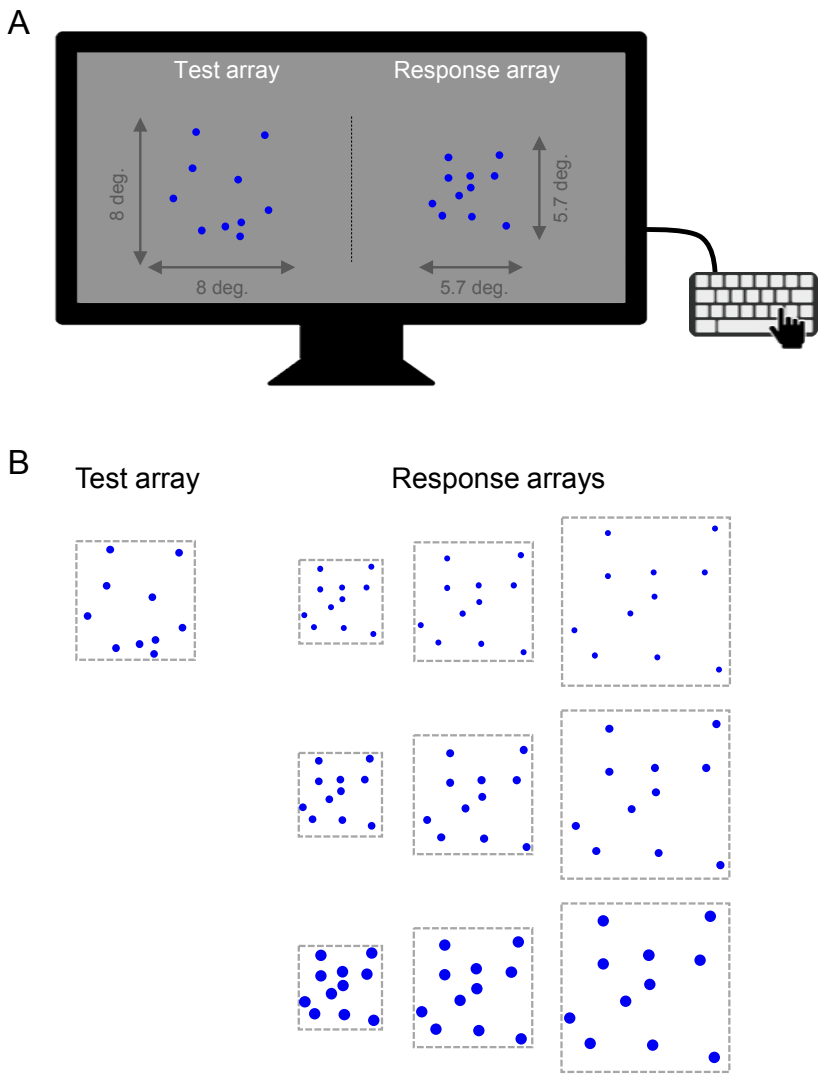

Figure 4 | Illustration of stimuli used in Experiment 2. (A) A still image of an example stimulus in which the area of the response array is half that of the test array (32 *vs.* 64 deg^2). (B) The dot size and area of the test array were fixed throughout the experiment, while the dot size and area of the response array were varied in a 3 × 3 factorial design.

Stimuli and procedure

As in Experiment 1, subjects adjusted the number of dots in a response array to match the number of dots in a test array (Fig. 4A). The stimuli and procedure were the same as in Experiment 1, except for the following differences. The test arrays contained 10, 15, or 20 dots and each dot had

a diameter of 0.20 deg. The visual properties of the test array were fixed throughout the experiment, while the total array area and dot size of the response array were varied in a 3 × 3 factorial design (Fig. 4B). Specifically, the total array area and dot size in the response arrays were half, the same, or double the area or size of the test array. Each subject was tested 3 times on each of the 27 combinations (3 numerosities × 3 array areas × 3 dot sizes). The order in which these 81 trials were presented was randomized per subject. For half of the subjects, the dots in the test array were yellow and for the other half they were blue (dots in the response array were always blue).

Results and discussion

We perform a 3 × 3 × 3 × 2 repeated-measures ANOVA with number of dots, array area, and dot size as within-subject factors, the color of the test dots as a between-subjects factor, and the reported number of dots as the dependent variable. The result reveals extremely strong evidence for a main effect of total array area on perceived numerosity ($BF_{inclusion}=\infty$, $p<.001$), with a positive relation: the larger the area of the response array, the smaller the setting of the number of response dots to obtain PSE with the test array (Fig. 5A). Hence, an increase in the response array area decreases the point of subjective equality with the test array, which means that the larger the area of an array, the larger the perceived number of dots. This is consistent with the results from Experiment 1, where spatial grouping of dots in the test array increased both the total array area and the average reported number of dots. Hence, it is plausible that the effect of spatial separation found in Experiment 1 was at least in part due to differences in array area, possibly on top of an effect due to a difference in the number of groups. Since we cannot disentangle these effects, it remains inconclusive what caused the increase in two-group proximity arrays compared to single-group arrays in Experiment 1.

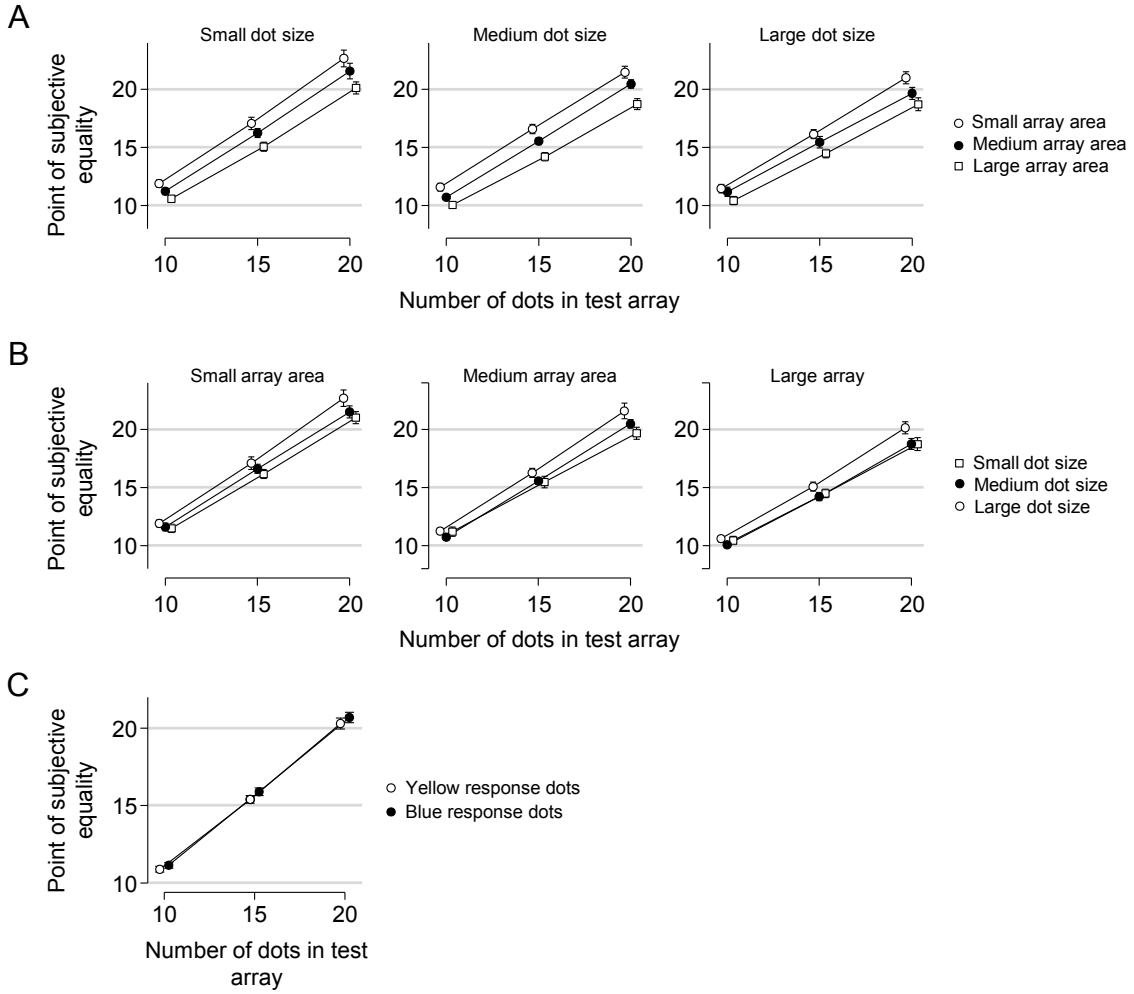

Figure 5 | Results of Experiment 2. (A) Reported number of dots as a function of the number of dots in the test array, split by response array area (different lines) and response array dot size (different panels). Results are pooled across the two groups of subjects with different response dot colors. (B) An alternative visualization of the data that more clearly shows the effect of dot size (different lines). (C) The same data plotted separately for subjects with yellow and blue response dots (collapsed over dot size and array area conditions).

We also find extremely strong evidence for a main effect of the size of the dots in the response array on reported numerosity ($BF_{inclusion}=\infty$, $p<.001$), with a positive relation between the two variables: the smaller the dots in the response array, the smaller the reported number of dots (as indicated by the PSE, Fig. 5B). This result suggests that compensation occurs since the

smaller dots in the response array are perceived as more numerous, which is consistent with the findings in Experiment 1 and provides further evidence that if there was a grouping-driven effect in the dot-size condition, then it was obscured by a feature-driven effect of dot size. Therefore, the results from the two-group dot-size condition in Experiment 1 are also inconclusive about the question how the number of groups affected perceived numerosity.

Furthermore, the ANOVA provides moderately strong evidence for the hypothesis that dot color does not affect the number of reported dots (Δ=0.38, $BF_{inclusion}$=0.143; Fig. 5C), which is consistent with our finding in Experiment 1. This is further evidence that the color grouping effect was a pure grouping-driven effect. However, we should note that the corresponding *p* value is 0.019, which is conventionally interpreted as evidence *against* the null hypothesis.

Finally, it is worth mentioning that the results provide extremely strong evidence *against* an interaction effect between array area and dot size on perceived numerosity ($BF_{inclusion}$=0.007, p=0.16), which suggests that these two biases act independently of each other. This finding is consistent with the lack of a correlation between effects of grouping by proximity and grouping by dot size in Experiment 1.

GENERAL DISCUSSION

There has lately been much research on the human number sense and on the various biases in this cognitive ability. Most previous experimental work has used static stimuli with extremely brief exposure times. While that approach has delivered important insights into the human number sense, little is known about how the human number sense works in situations where stimuli are dynamic and viewed for extended periods. Here, we developed an experimental paradigm that more closely mimics such conditions and we used it to assess grouping effects in numerosity judgment. We found that grouping by color and motion decreased the average number of reported

429 dots. Since we found evidence against direct influences of these features, we conclude that the
430 effects were solely due to grouping. Moreover, we found a correlation between the effect sizes
431 for these two features, which suggests that the effects may have a feature-invariant, higher-level
432 origin. We were unable to draw conclusions about effects of grouping by proximity or dot size,
433 because these features were found to have a direct effect on perceived numerosity. Overall, our
434 results suggest that perceived numerosity is negatively related to the number of groups, even
435 though these effects may be overridden by stronger feature-driven effects in the opposite
436 direction when grouping is established using spatial separate or object size as grouping feature.
437 Therefore, the story told in Genesis, where Jacob spread a herd of sheep in smaller groups to
438 appear more numerous [6], describes a hypothetical influence on perceived numerosity that is
439 maybe not caused by grouping, but more likely due to the resulting larger total stimulus area
440 formed by the spatially separate groups of sheep. Our finding that smaller dots tend to be
441 perceived as more numerous than larger dots is consistent with previous work that primarily used
442 static stimuli with brief exposure [6,26–30]. It is currently unknown what the origin of this
443 "smaller-dot-size-is-more" effect is. One possibility is that a given array area has more room for
444 small items than for large items and that humans use this prior information during inference.
445 However, other studies have found an opposite, "smaller-is-less" effect [22–25]. Findings from a
446 recent eye-tracking experiment suggests that this effect may be related to attentional capture
447 (Lindskog, Poom, & Winman, in preparation).

448 Our results also revealed an effect of total array area on perceived numerosity: the larger
449 the area, the larger the average perceived numerosity. This result is also in line with previous
450 studies [9,16,17]. A speculative explanation of the area-based bias is that when humans make
451 numerosity judgments, they take into account their previous experiences. If a person has
452 experience that large areas or volumes tend to contain more objects than small ones and that

453 smaller objects are typically more abundant within a given area or volume, then this could create
454 a bias similar to the one we found. Consistent with this speculation, is has been found that
455 increasing the size of a container increases the perceived number of beans in the container [35].
456 Moreover, this type of explanation is consistent with findings that biases may be reduced through
457 practice [29] and instructions given to the observers [36].

458 Besides contributing further empirical insights into the number sense, the present study also
459 introduced a novel experimental paradigm, in which stimuli were dynamic and viewed for
460 prolonged durations. Our main motivation for using this kind of stimulus was that we believe that
461 such stimuli are representative for a large class of numerosity judgments outside the laboratory.
462 However, another advantage of LDDs is that they provide a more accurate percept of the total
463 array area, thus providing the experimenter with stronger control over this experimental variable.
464 The experiments presented here do not address the issue whether numerosity judgment tasks with
465 LDDs target the same brain mechanisms as the traditional paradigm with brief and static stimuli.
466 Therefore, it could be informative if future studies would directly compare empirical properties
467 of number estimation in brief, static stimuli on the one hand and dynamic stimuli with prolonged
468 viewing on the other hand. For example, some studies have argued that numerosity estimates are
469 mainly derived from low-level visual cues [37,38]. It is unclear whether the use of longer
470 presentations would increases or decreases the tendency to use visual cues. On the one hand,
471 longer presentation times are expected to reduce the influence of rapid, bottom-up effects caused
472 by basic features. However, on the other hand, they may also lead to more accurate estimates of
473 the visual features, which could be a reason to rely on them more strongly. Since our paradigm
474 allows for the use of arbitrary stimulus times, without the risk that subjects will explicitly count
475 the stimuli, future studies could use tasks with LDDs obtain more insight into the relation
476 between stimulus duration and the use of visual cues in number judgment. Moreover, it could be

interesting to test whether the putative link between numerosity judgment accuracy and mathematical ability [1] generalizes to tasks with LDDs. Finally, other psychophysical methods than the adjustment procedure can be used with LDDs, such as the 2AFC method using fixed presentation times. Such procedures using LDD could also be useful when prolonged viewing is required, for example in functional imaging studies.

ACKNOWLEDGEMENTS

LP acknowledges support from the Swedish Research Council (Vetenskapsrådet; reg. nr. 2013-01005). RB acknowledges support from the Swedish Research Council (Vetenskapsrådet; reg.nr. 2015-00371) and Marie Sklodowska Curie Actions, Cofund (project INCA 600398).

REFERENCES

1. Chen Q, Li J. Association between individual differences in non-symbolic number acuity and math performance: A meta-analysis. Acta Psychol (Amst). 2014; doi:10.1016/j.actpsy.2014.01.016

2. Dehaene S. Varieties of numerical abilities. Cognition. 1992; doi:10.1016/0010-0277(92)90049-N

3. Jevons WS. The power of numerical discrimination. Nature. 1871;

4. Shepard RN, Kilpatric DW, Cunningham JP. The internal representation of numbers. Cogn Psychol. 1975; doi:10.1016/0010-0285(75)90006-7

5. Hamilton W. Lectures on Metaphysics and Logic. Edinburgh: Blackwood; 1859.

6. Ginsburg N. Effect of item arrangement on perceived numerosity: Randomness vs

499 regularity. Percept Mot Skills. 1976; doi:10.2466/pms.1976.43.2.663

500 7. Im HY, Zhong S hua, Halberda J. Grouping by proximity and the visual impression of
501 approximate number in random dot arrays. Vision Res. 2016;
502 doi:10.1016/j.visres.2015.08.013

503 8. Ginsburg N, Goldstein SR. Measurement of visual cluster. Am J Psychol. 1987;
504 doi:10.2307/1422403

505 9. Frith CD, Frit U. The solitaire illusion: An illusion of numerosity. Percept Psychophys.
506 1972;11: 409–410. doi:10.3758/BF03206279

507 10. Allik J, Tuulmets T. Occupancy model of perceived numerosity. Percept Psychophys.
508 Springer-Verlag; 1991;49: 303–314. doi:10.3758/BF03205986

509 11. Ginsburg N. Perceived numerosity, item arrangement, and expectancy. Am J Psychol.
510 1978; doi:10.2307/1421536

511 12. Ginsburg N. Numerosity estimation as a function of stimulus organization. Perception.
512 1991; doi:10.1068/p200681

513 13. Zhao J, Yu RQ. Statistical regularities reduce perceived numerosity. Cognition. 2016;
514 doi:10.1016/j.cognition.2015.09.018

515 14. Nicholls MER, Bradshaw JL, Mattingley JB. Free-viewing perceptual asymmetries for the
516 judgement of brightness, numerosity and size. Neuropsychologia. 1999;
517 doi:10.1016/S0028-3932(98)00074-8

518 15. van den Berg R, Lindskog M, Poom L, Winman A. Recent is more: A negative time-order

519 effect in nonsymbolic numerical judgment. J Exp Psychol Hum Percept Perform. 2017;43.
520 doi:10.1037/xhp0000387

521 16. Krueger LE. Perceived numerosity. Percept Psychophys. 1972; doi:10.3758/BF03212674

522 17. Piaget J. Quantification, conservation, and nativism. Science (80-). 1968;
523 doi:10.1126/science.162.3857.976

524 18. Apthorp D, Bell J. Symmetry is less than meets the eye. Curr Biol. 2015;
525 doi:10.1016/j.cub.2015.02.017

526 19. Howe E, Jung K. Judgment of numerosity: Effects of symmetry and goodness in dot
527 pattern arrays. Acta Psychol (Amst). 1987; doi:10.1016/0001-6918(87)90058-8

528 20. Franconeri SL, Bemis DK, Alvarez GA. Number estimation relies on a set of segmented
529 objects. Cognition. 2009; doi:10.1016/j.cognition.2009.07.002

530 21. He L, Zhang J, Zhou T, Chen L. Connectedness affects dot numerosity judgment:
531 Implications for configural processing. Psychon Bull Rev. 2009;
532 doi:10.3758/PBR.16.3.509

533 22. Gebuis T, Kenemans JL, de Haan EHF, van der Smagt MJ. Conflict processing of
534 symbolic and non-symbolic numerosity. Neuropsychologia. 2010;
535 doi:10.1016/j.neuropsychologia.2009.09.027

536 23. Gilmore C, Attridge N, Clayton S, Cragg L, Johnson S, Marlow N, et al. Individual
537 Differences in Inhibitory Control, Not Non-Verbal Number Acuity, Correlate with
538 Mathematics Achievement. PLoS One. 2013; doi:10.1371/journal.pone.0067374

24. Hurewitz F, Gelman R, Schnitzer B. Sometimes area counts more than number. Proc Natl Acad Sci. 2006; doi:10.1073/pnas.0609485103

25. Soltész F, Szucs D, Szucs L. Relationships between magnitude representation, counting and memory in 4- to 7-year-old children: A developmental study. Behav Brain Funct. 2010; doi:10.1186/1744-9081-6-13

26. Gebuis T, van der Smagt MJ. False approximations of the approximate number system? PLoS One. 2011;6: e25405. doi:10.1371/journal.pone.0025405

27. Ginsburg N, Nicholls A. Perceived numerosity as a function of item size. Percept Mot Skills. 1988; doi:10.2466/pms.1988.67.2.656

28. Miller AL, Baker RA. The effects of shape, size, heterogeneity, and instructional set on the judgment of visual number. Am J Psychol. 1968; doi:10.2307/1420810

29. Tokita M, Ishiguchi A. How might the discrepancy in the effects of perceptual variables on numerosity judgment be reconciled? Attention, Perception, Psychophys. 2010; doi:10.3758/APP.72.7.1839

30. Shuman M, Spelke E. Area and element size bias numerosity perception. Vision Sciences Society Annual Meeting Abstract. 2006. p. 777.

31. Poom L. Motion and color generate coactivation at postgrouping identification stages. Attention, Perception, Psychophys. 2011; doi:10.3758/s13414-011-0132-8

32. Razpurker-Apfeld I, Kimchi R. The time course of perceptual grouping: The role of segregation and shape formation. Percept Psychophys. 2007; doi:10.3758/BF03193775

33. Wagenmakers E, Love J, Marsman M. Bayesian inference for psychology. Part II: Example applications with JASP. Psychon Bull Rev. 2018;25: 58–76.

34. JASP Team. JASP (Version 0.8.4.0) [Computer program]. 2018.

35. Bevan W, Maier RA, Helson H. The influence of context upon the estimation of number. Am J Psychol. 1963; doi:10.2307/1419789

36. Teghtsoonian M. The judgment of size. Am J Psychol. 1965; doi:10.2307/1420573

37. Tokita M, Ishiguchi A. Effects of perceptual variables on numerosity comparison in 5-6-year-olds and adults. Front Psychol. 2013;4. doi:10.3389/fpsyg.2013.00431

38. Gebuis T, Reynvoet B. The interplay between nonsymbolic number and its continuous visual properties. J Exp Psychol Gen. 2012;141: 642–648. doi:10.1037/a0026218