A Selective Sampling Account of Forming Numerosity Representations

Yonatan Vanunu*[1,2,3] and Roger Ratcliff[3]

1. Tel Aviv University
2. The University of Chicago
3. The Ohio State University

Author Notes

Acknowledgements: This work was supported by funding from the National Institute on Aging (Grant numbers R01-AG041176 and R01-AG057841). We thank Jared M. Hotaling for useful feedback on this project.

Author Contributions: Y.V. and R.R. conceptualized the project and wrote the paper; Y.V. designed the experiments, developed the computational models, analyzed the eye-tracking data, and performed the statistical analyses.

Competing interests: The authors declare no competing interests.

Additional information: The data and codes that support the findings of this study are available at https://osf.io/8mgdn/. Portions of the data and concepts discussed in this manuscript were presented at the 64th Annual Meeting of the Psychonomic Society in November 2023 and the Society for Mathematical Psychology meeting in July 2022. The current version of the manuscript was also posted as a pre-print at *PsyArxiv*. Correspondence and requests for materials should be addressed to Yonatan Vanunu. Email: yyv1984@gmail.com.

Yonatan Vanunu https:\\orcid.org\0000-0003-1960-4480

Roger Ratcliff https:\\orcid.org\0000-0001-9657-0814

Abstract

Two leading models of numerosity judgments describe numerical representations as Gaussian distributions on a mental number line. The linear model posits that both numerosity and variability increase linearly with number, while the logarithmic model assumes logarithmic scaling with constant variability. In this study, we use the selective sampling account, which proposes that information is gathered selectively based on goals and available resources, to explore the cognitive processes underlying variations in variability and scaling. In intermingled displays of blue and yellow dots (B/Y task), participants relied on incomplete representations of dots positioned near the center, where spatial resolution is highest, leading to increasing variability with set size. In contrast, spatially separated displays (L/R task) facilitated more comprehensive sampling, resulting in approximately constant variability across set sizes. Behavioral patterns and modeling analyses suggest that linear and logarithmic scaling capture sensitivity differences shaped by the display format and spatial resolution demands. Eye-tracking data further support our account, emphasizing the role of selective attention in forming numerical representations and providing a unified framework for understanding variability and scaling across tasks.

Key words: numerosity judgment; selective sampling; selective attention; diffusion model; approximate number system; spatial resolution.

Main

Forming numerosity representations is a fundamental cognitive skill shared by humans and animals (Brannon & Terrace, 1998; Dehaene, 2001; Nieder, 2020; Nieder & Dehaene, 2009). For example, numerosity representations are formed when estimating how many people are ahead of you in line or which bush of berries an animal should harvest first. Previous research has shown that the ability to form numerosity representations can predict the development of math skills among children and adults (Halberda, Mazzocco, & Feigenson, 2008; Hyde, Khanum, & Spelke, 2014; Park & Brannon, 2013; 2014), which are critical in day-to-day activities such as accounting, time management, and logistics.

The current study aims to provide a unified framework for understanding numerical judgments across different tasks. We use the selective sampling account (Vanunu, Hotaling & Newell, 2020; Vanunu, Hotaling, Le Pelley & Newell, 2021) to explain behavioral differences in common numerosity discrimination tasks, focusing on how selective attention and task demands shape these judgments. Specifically, we explore how the distribution of attention across the visual display impacts key aspects of numerical judgment—namely, *variability*, *scaling*, and *perceptual effects*. Our framework proposes that judgments are often based on incomplete numerosity representations, with certain display properties more likely to be represented because they attract attention. Consequently, variability, scaling, and perceptual effects in forming non-symbolic numerosity representations are closely tied to the scope and randomness of this process.

The Approximate Number System:

The most prominent model for numerosity judgments is the *Approximate Number System* (ANS) (Dehaene, 2003). According to the ANS, each number is represented by a Gaussian distribution on a mental line, where the mean represents the number and the standard deviation

(SD) represents the variability that arises in forming non-symbolic numerical representations. The ANS models are often consistent with Weber-Fechner's law, which states that as stimulus intensity increases, the size of the just-noticeable difference between stimuli increases so that the ratio of the difference in intensity to the intensity ($\Delta S/S$) remains constant. Two ANS models capture this rule: the *linear ANS* model, in which both the mean and SD increase on a linear scale; and the *log ANS* model, in which the mean of the Gaussian distribution increases on a logarithmic scale while the SD remains constant. Thus, the two models differ in two properties: variability (SD) and scale, as outlined in Figure 1A.

It was previously argued that the linear and log ANS models could not be discriminated using response choice behavioral data (Dehaene, 2003). However, recent work has shown that these two models can be distinguished. Different numerosity discrimination tasks produce radically different patterns of RTs, and if the two models are integrated with a diffusion model (Kang & Ratcliff, 2020; Ratcliff & McKoon, 2018; 2020), the result is two integrated models, each of which predicts one of the different RT patterns.

Ratcliff and McKoon (2018) used such models to examine choice and RT from two common numerosity discrimination tasks: the left/right (L/R) task, in which two arrays of dots were presented side by side, and participants determined whether there were more dots on the left or right side of the screen; and the blue/yellow (B/Y) task, in which a single array of intermingled blue and yellow dots was presented, and participants determined whether there were more blue or yellow dots on the screen (see Figure 1B for example displays from each task). Results showed that in the L/R task, for a small constant difference in numerosity, as overall numerosity in the two arrays increased, RT increased and accuracy decreased—a pattern best accounted for by the

logarithmic account. Conversely, RT increased as accuracy increased in the B/Y task—a counterintuitive pattern best accounted for by the linear account.

This article presents a novel account to explain the cognitive processes underlying variations in variability and scaling when forming non-symbolic numerosity representations across common numerosity discrimination tasks. We aim to understand why variability increases with set size in some tasks (B/Y task) but remains approximately constant in others (L/R task), and which features of the task determine whether a linear or logarithmic scale is used. Furthermore, our framework seeks to account for the influence of non-numeric perceptual properties, such as visual size and spatial position, on numerosity judgments—factors that traditional ANS models do not adequately address (DeWind, Adams, Platt & Brannon, 2015; Gebuis, Kadosh & Gevers, 2016; Gebuis & Reynvoet, 2012a; 2012b; Leibovich, Katzin, Harel & Henik, 2017).

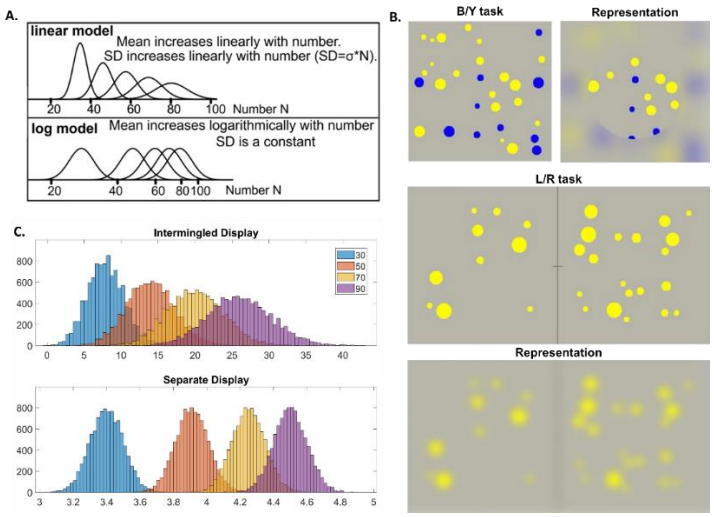


Figure 1. A) An illustration of the linear and log ANS models. B) Example stimuli from the B/Y task (top panels) and L/R task (bottom panels), where the dot areas are proportional to numerosities, along with the predicted numerosity representation based on the selective sampling account. C) The simulation of evidence distribution across different numerosities, suggesting that

incomplete yet highly sensitive representations mirror the linear ANS model, while comprehensive but less sensitive representations mirror the log ANS model.

The Selective Sampling Account:

Because human processing capacity is limited, individuals must selectively sample information based on their goals, available resources, and the format in which the information is displayed (Vanunu et al., 2020; Vanunu et al., 2021). This principle of selective sampling is rooted in the concept of bounded rationality (Simon, 1954) and has long been supported by rule-based heuristic models in judgment and decision-making research (Brandstätter, Gigerenzer, & Hertwig, 2006; Gigerenzer, 2004; Gigerenzer & Gaissmaier, 2011; Payne, Bettman, & Johnson, 1993; Tversky, 1972). However, unlike traditional heuristic models, the selective sampling account employs a more flexible mechanism based on probabilities rather than deterministic rules, enabling it to accommodate a range of sampling policies within a unified framework.

Vanunu et al. (2021) used the *Selective Sampling and Integration Model* (SSIM) to demonstrate the interaction between top-down and bottom-up attention in information sampling and subsequent choice. In their study, participants were briefly presented with an array of payoffs (numbers) displayed in varying font sizes, and they had to choose between a guaranteed payoff or a gamble on winning one of the displayed payoffs with equal probabilities. The SSIM effectively explained complex choice patterns by showing that payoffs with large values (reflecting top-down goals) or font sizes (indicating bottom-up saliency) were prioritized during sampling. However, the bottom-up prioritization only occurred if the salient item also aligned with the participant's goals, thereby positioning top-down processes as the final arbiter for choice. Eye-tracking results supported these findings, indicating that items prioritized during sampling also received prolonged

fixation, thereby providing strong evidence for the SSIM's utility in representing the sampling processes underlying attention and choice in this context.

Predictions from the Selective Sampling Account:

In numerosity discrimination tasks, the selective sampling account suggests that choices are often based on incomplete representations of the dot arrays and provides specific predictions regarding variations in variability and scaling across tasks. Variability is expected to increase with set size if judgment is based on a stochastic subsample of the information. However, it should remain relatively constant when the subsample is comprehensive—i.e., when the majority of the dots in the display are included in the sample. This occurs because incomplete samples can be drawn in many different ways from the same display, resulting in greater variability in forming representations as set size increases. In contrast, comprehensive samples reduce variability by encompassing most or all of the display. Further, we propose that linear or logarithmic scaling reflects sensitivity to the visual display, corresponding to high and low sensitivity, respectively. We assume that the display format dictates the sensitivity demand. High sensitivity is needed to distinguish between blue and yellow dots in a brief, intermingled display (B/Y task). In contrast, numerical judgments in the L/R task, which involve two distinct spatial locations, allow for spatial-based discrimination without requiring the same sensitivity level as color-based discrimination within overlapping arrays.

Importantly, the selective sampling account suggests that variations in variability and scaling are interconnected, shaped by both sample size and inherent differences in spatial resolution—i.e., the ability to distinguish between points in space—across the central and peripheral regions of the display (Loschky, Nuthmann, Fortenbaugh, & Levi, 2017; Rodieck, 1998; Shapiro et al., 2010). The human visual system processes different parts of the visual field

unevenly, with central dots receiving disproportionately more processing, thereby enhancing spatial resolution and sensitivity in that region. On the intermingled B/Y display, which presumably requires high sensitivity, numerical judgments likely rely on a subsample from the center, where spatial resolution is highest. These incomplete yet highly sensitive representations lead to increased variability with set size and a linear scale, aligning with the linear ANS model. In contrast, the L/R task, featuring two spatially separated arrays and presumably requiring less sensitivity, allows for more comprehensive sampling, including peripheral dots with lower spatial resolution. These comprehensive yet less sensitive representations result in more stable variability across set sizes and a logarithmic scale, aligning with the log ANS model. Figure 1B illustrates these task-specific predictions.

To demonstrate how variations in sample size align with ANS models assumptions about variability and scaling, we simulated the distribution of *evidence*—i.e., the number of dots in the representation—across different numerosities and sampling policies. Over a million iterations, each dot was independently sampled with uniform probability, and the number of sampled dots was scaled according to the anticipated sensitivity of the representation, with Gaussian noise added to reflect constant variability (i.e., internal noise). In the intermingled display, each dot had a 30% chance of being included in the subsample, generating incomplete representation, and a linear scale was used to represent high visual sensitivity. In the separate display, all dots were sampled with certainty to represent comprehensive sampling, and a logarithmic scale was used to reflect low sensitivity. The resulting distributions in Figure 1C mirror the ANS distributions in Figure 1A.

Finaly, the selective sampling account may also explain perceptual effects on numerosity judgment by considering how attention is distributed across the display. We propose that certain dots in the display are more likely to attract attention than others due to their visual size and spatial

position, making them more likely to be included in the numerosity representation, while dots that are less noticeable in a brief display are likely to be excluded from the representation. As a result, arrays with more attended dots (e.g., larger dots) would be perceived as having a larger numerosity compared to arrays with fewer attended dots.

Models Description

The ANS-diffusion models:

The diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008) assumes that evidence for a decision is noisy and accumulates over time until there is enough to make a decision. The process starts at a point z and terminates when the amount of accumulated evidence is 0 or a (a represents boundary separation). The rate at which evidence is accumulated is called the drift rate (v), and it is assumed to vary from trial to trial, distributed normally with a standard deviation η. Figure A1A in Appendix A illustrates three types of paths in the diffusion process: a fast correct decision, a slow correct decision, and an error, along with the non-decision components of RT.

In the *ANS-diffusion model*, the drift rate and between-trial SD in the drift rate (η) are specified by the numerical representations assumed in the ANS models. The drift rate is the difference between the means of two Gaussian distributions or the difference between the logarithms of the means, which represent the perceived numerosity of each set. The choice of scale is determined by a dummy parameter, K, which takes the value of 1 for the linear scale and 0 for the logarithmic scale, applied across all trials completed for each individual. In the current

modeling work, *K* serves as a *sensitivity parameter* that operates as a step function, distinguishing between highly sensitive representations when *K*=1 and less sensitive representations when *K*=0.[1]

The resulting difference is then scaled by a drift-rate coefficient (*v*), as illustrated in Eq. (1). The between-trial SD in the drift rate derived from the differences in scores from two Gaussian distributions (*η)*, is derived from the SDs in the two distributions by taking the square root of the sum of the squares of the variances. This SD in the difference scores is then adjusted by a scaling factor (σ_1), as shown in Eq. (2). Additionally, the model assumes a level of variability (η_0) that is independent of the specific characteristics of the display. This constant variability represents source of noise in drift rate across trials that affect the decision process, regardless of the numerosity or distribution of the dots on any given trial.

Eq. (1): $$V = v \times \left[K \times (N_L - N_S) + (1 - K) \times log\left(\frac{N_L}{N_S}\right)\right]$$

Eq. (2): $$\eta = \eta_0 + \sigma_1\sqrt{N_L^2 + N_S^2}$$

Here, N_L and N_S represent the large and small numerosities; v is a drift-rate coefficient for the numerical difference; *K* is a binary parameter for sensitivity to the visual display (i.e., scale); σ_1 is the SD scaling factor; and η_0 is constant variability.

Ratcliff & McKoon (2018) tested the linear and log ANS-diffusion models on data from the L/R task and the B/Y task. They concluded that behavior in the L/R task was consistent with forming separate representations of the two arrays on a logarithmic scale with constant variability (i.e., because the log ANS-diffusion model provided a better fit for most participants, and the SD scaling factor σ_1 was close to zero). In contrast, behavior in the B/Y task was consistent with

[1] This approach is similar to using a step function to control sensitivity in binary choice, as opposed to a temperature parameter in a logistic transformation. Notably, we also tested an alternative mechanism where *K* was estimated as a continuous weighting function between the two scales, but the results indicated an inferior fit in both tasks.

forming a representation based on the difference in the number of blue and yellow dots on linear scale with variability increasing as set size (the number of dots in the display) increased.

To account for perceptual effects on numerosity judgments, variants of the ANS-diffusion models were developed, in which the drift rate was derived from a regression model based on the perceptual conditions. It achieves this by integrating the effect of perceptual variables into the estimates of drift rate. For example, two drift-rate coefficients (v in Eq. 1) were calculated for settings where the surface area of dots was proportional to numerosity or remained constant regardless of numerosity, to account for the different patterns of responses observed between conditions. However, it is critical to acknowledge that the ANS-diffusion models operate under the assumption that there is not a specific processing explanation for the existence of two distinct drift rate coefficients. Our approach aims to elucidate the impact of perceptual conditions and their combinations on the cognitive processes underlying numerosity discrimination.

The Selective Sampling and Diffusion Model (SSDM)

The Selective Sampling and Diffusion Model (SSDM) is an extension of the SSIM that explains both choice and response time by incorporating the selective sampling account as a representation model within the diffusion model framework. It represents processing in two components: *selective sampling*, where information is gathered according to goals and task demands to form a representation of the stimulus (i.e., evidence for choice) with varying sensitivity, and *evidence accumulation*, where this representation drives the diffusion process to account for choice and RT.

Selective sampling is represented by a *Probabilistic Sampling Function* (PSF), which assigns a discrete probability of being sampled to each dot in the display based on its ranking on a specified scale. For instance, we can estimate the fit of a PSF that assigned probabilities according

to the dots' size, proximity to the center of the screen, or a combination of both (i.e., two-dimensional PSF). The best-fitting PSF provides an estimate of the most likely sampling policy an individual employs across trials for a specific task.

The PSF can adopt various shapes, governed by the combination of three free parameters: α defines the function's curvature, θ defines its symmetry, and β controls the area under the curve. Figure A1B in the Appendix illustrates example shapes the one- and two-dimensional PSF can take. The PSF is derived from a quadratic function in two stages. The function's shape, q(X), is defined in Eq. (3) using the product of the two terms. Then, q(X) is converted into a vector p(X), in which each element represents the probability of sampling a dot, denoted as Xi, based on its rank i within a display, as shown in Eq. (4):

$$\text{Eq. (3):} \quad q(X_i) = \alpha \times (w_i - \theta)^2$$

$$\text{Eq. (4):} \quad p(X_i) = q(X_i) - min_X(q[X_i]) + \beta$$

here, *w* is a rescaling of ranks into evenly spaced values between -1 and 1, relative to the set size *N* (e.g., in a display with 20 dots on the left and 25 dots on the right, rescaling would be conducted for *N* = 45 ranks). The shape of the function is determined by three parameters, each ranging from -1 to 1. α defines the function's curvature. A positive α value indicates a U-shaped curve that prioritizes the extreme ranks in sampling, whereas a negative α value indicates a bell-shaped curve that prioritizes mid-range ranks in sampling (shown by the blue and red curves in Figure A1B, respectively). θ defines the function's symmetry, with negative or positive θ values indicating a preference for sampling higher or lower ranks, respectively (depicted by yellow and purple curves in Figure A1B, respectively). β controls the area under the curve, specifically the function's lowest point after normalization. This normalization involves subtracting the minimum point in q(X) from

all values in q(X). Therefore, a β value of 1 implies that all dots are sampled. To ensure that the PSF values are probabilities and stay within the 0 to 1 range, any values outside these limits are adjusted to the nearest boundary, so that if the PSF value exceeds 1, it is set to 1, and if it is less than 0, it is set to 0.

The ranking scale that governs the distribution of probabilities among items must be adjusted to the specific research question at hand. For example, sampling may be influenced by the visual size of the dots, their spatial positioning, or a combination of both properties. By modifying the PSF to allocate probabilities based on the dots' radius ranks, proximity-to-center ranks, or a two-dimensional space incorporating both aspects, researchers can determine which attribute plays a more significant role in forming numerosity representations. See Figure A1B in the Appendix for an example two-dimensional PSF that assigns probabilities according to a combination of the dots' radius ranks and proximity-to-center ranks.

To accommodate sampling within a two-dimensional space, the formula for the product q(X) was extended to produce a quadratic surface in Eq. (5).

$$\text{Eq. (5):} \quad q(X_{ij}) = \alpha_C \times ([M_{Cij} - \theta_C]^2) + \alpha_R \times ([M_{Rij} - \theta_R]^2)$$

Here, *i* represents the rank based on proximity to the center and *j* denotes the rank based on dot radius. M_C and M_R are matrices that represent the variations in these two dimensions—proximity to center and dot radii, respectively—with the ranks rescaled to *N* evenly spaced values between -1 and 1. This rescaling process mirrors the approach used with the rescaled vector *w* in Eq. (1), where M_C is a replication of *w* across *N* rows and M_R is a replication of *w* across *N* columns. By summing their squared values, we generate a 3D quadratic surface, with α and θ determining its curvature and symmetry, respectively. Eq. (4) then transforms q(X) into a probability matrix p(X)

of sampling each item X_{ij}, with the parameter β setting the surface's altitude. Therefore, the y-axis of the quadratic surface corresponds to the dots' ranks of proximity to the center, the x-axis to the dots' radius ranks, and the z-axis to the probability of each dot being sampled (i.e., from dark blue to light yellow). For those interested in exploring or simulating different shapes for the PSF, MATLAB code is available at https://osf.io/8mgdn/.

For the display of dots, on each trial, the PSF algorithm simulates sampling based on probabilities derived from the equations and model parameters. These are assigned according to the dots' ranking in size and proximity to the center of the screen, considering all dots in the display. This process generates two subsamples of dots from each array to compare (i.e., left vs. right or blue vs. yellow). Then the linear or logarithmic difference between the numbers of dots in the subsamples is used to compute the drift rate, determined by the binary sensitivity parameter K, as shown in Eq. (1) for the ANS-diffusion model. Notably, K is fixed to 0 or 1 across trials.

Given that each dot has an independent probability of being sampled, the size of the subsample may differ across different samples from identical displays. This variation constitutes the between-trial SD in the drift rate (η). Specifically, for each trial, we simulated the sampling process ten thousand times to generate a distribution of plausible representations. During each simulation, a random number, ranging from 0 to 1, is produced for every dot in the display. If the random number is less than the dot's assigned PSF value, then that dot is included in the subsample. Consequently, lower assigned PSF values, which signify smaller probabilities of sampling, often lead to a wider distribution of plausible representations (see Figure 1C). The mean of this distribution is used to establish the drift rate, while the SD of the distribution determines the between-trial SD in the drift rate (η).

The selective sampling account thus embodies a stochastic process influenced by selective attention and task demands, occurring before any deliberation. This often leads to the formation of incomplete numerosity representations that vary across identical displays, which is consistent with the between-trial SD assumptions of the diffusion model (Ratcliff, 1978). This methodology highlights the role of selective attention and cognitive limitations in forming non-symbolic numerosity representations, emphasizing that individuals might form different representations (i.e., different subsamples) from the same stimulus due to the inherent variability in how information is attended and sampled.

We initially applied the SSDM to existing data from the B/Y and L/R tasks (Ratcliff & McKoon, 2018) to generate predictions about how information is sampled in these tasks. Subsequently, we tested the sampling schemes derived from these fits in Experiments 1, 2, and 3 by manipulating the properties identified by the SSDM as being prioritized during sampling in the reanalysis. Furthermore, we employed eye tracking in Experiments 2 and 3 to obtain more direct measures of attention (assuming that where they were looking corresponds to where they were attending), thereby offering further empirical support for the SSDM's predictions.

It is important to note that our goal in model comparisons is not to show that the SSDM provides a better fit to the data than the ANS-diffusion model, as both models share common features likely to yield similar fits. Instead, we aim to demonstrate that the SSDM achieves a comparable fit to the ANS-diffusion model while offering novel insights into the distribution of attention and its impact on key aspects of numerical judgment, thereby advancing our understanding of this process.

Reanalysis of Data from Ratcliff & McKoon (2018)

In both the L/R and B/Y tasks, the set size of each array and the difference between them were manipulated to create 20 displays (with 15/10, 20/15, 25/20, 30/25, 40/35, 20/10, 30/20, 40/30, 30/10, 40/20 dots and their opposites). Additionally, the surface area of dots was adjusted to be either proportional to numerosity or equal, regardless of numerosity. The display duration was brief, allowing for only a single fixation, which ensured that participants had to make their numerosity judgments based on a quick glance, rather than prolonged scrutiny or counting. For more details about the experimental design and findings see Ratcliff & McKoon (2018).

Several properties of the display could influence sampling. In this reanalysis, we focused on dot sizes and proximity to the center of the screen because we assume they represent the interplay between bottom-up (Itti & Koch, 2000; Treisman & Gelade, 1980) and top-down attention (Posner, 1980; Wolfe, Cave & Franzel, 1989) in sampling. Sampling larger dots aligns with bottom-up attention to salient stimuli, indicating a natural tendency to notice and prioritize visually salient features. Conversely, sampling dots closer to the center is consistent with a top-down goal to capture as much information as possible from a brief stimulus display. This is because the center of the display is richer in information: it provides the optimal vantage point for processing both central *and* peripheral information with a single fixation (Tatler 2007; Tatler, Baddeley & Gilchrist 2005).

To explore these dynamics, we tested three variants of the SSDM with PSFs that assign sampling probabilities based on either the dots' radius ranks, proximity-to-center ranks, or a two-dimensional space composed of both. We compared the goodness-of-fit among the SSDM variants and the ANS-diffusion model to determine if perceptual biases in numerosity judgments can be accounted for by selective sampling—a process model—rather than by estimating a distinct drift-rate coefficient for each experimental condition. Additionally, comparing SSDM versions that

assign PSF values according to the dots' ranking on different perceptual properties (i.e., size, position, or both) may inform us which property exerts a more significant influence on the sampling process and, by extension, on the participants' numerosity judgments.

The models were fitted to each participant's data separately. We estimated five quantiles of the RT distributions for each experimental condition (i.e., 0.1, 0.3, 0.5, 0.7, and 0.9, resulting in six bins per distribution), separately for correct and error responses. Model goodness-of-fit was evaluated using G^2 values, with model complexity controlled through cross-validation analysis. See Appendix B for a detailed report on the fitting methods and Appendix C for the results of an alternative model comparisons analysis using the average AIC scores.

Results:

Findings from the model comparisons are presented in Table 1. The group-level PSFs, averaged across participants for each task, are shown in Figure 2A. For an individual-level PSF display, calculated as the proportion of times each rank had the maximum probability across participants, refer to Figure A2 in Appendix D. The estimated between-trial SD (η) among set size conditions and tasks is depicted in Figure 2B. The average difference in η between consecutive set-size conditions (10/15 vs. 20/15, 20/15 vs. 25/20, etc.) was compared to zero using one-sample t-test, to test for monotonic changes in variability with set size. A visual comparison of how well the model's predictions align with the observed data points is illustrated by the *Quantile-Probability Functions* (QPFs) in Figure A3 in Appendix E.

In the L/R task, the two-dimensional SSDM provided the best fit for both the complete dataset and the testing dataset in the cross-validation analysis. The PSF derived from this model revealed a distinctive sampling bias characterizing participants' behavior in the L/R task: There was a pronounced preference for sampling the largest dots in the display. This preference for larger

dots explains the impact of dot area on numerical judgments, suggesting that an array with larger dots might be perceived as more numerous than an array with smaller dots due to this sampling bias. Furthermore, the group-level PSF indicates that dots ranked mid-range in proximity to the center were sampled more frequently than those closest and farthest from the center. This pattern highlights a strategic focus on dots located in the central region of each side of the L/R display.

In the B/Y task, the ANS-diffusion model was identified as the best fit for both the complete dataset and the testing dataset in the cross-validation analysis. This outcome is not surprising given that both the ANS and the selective sampling frameworks can account for an increase in variability with set size. Nevertheless, the SSDM showed a good fit to the data, a fit that is comparable to the ANS-diffusion model (Figure A3 in the Appendix), while providing an account of how the perceptual properties of the display influence variability. Similar to findings from the L/R task, participants' behavior in the B/Y task was characterized by a preference for sampling the largest dots, which accounts for the effect of the dots' surface area on numerosity judgments. Additionally, the PSF revealed a pronounced preference for sampling dots located closer to the center of the intermingled B/Y display, potentially uncovering a new perceptual bias in numerosity judgments for that task—*the proximity-to-center bias*.

Table 1. Model comparisons from the reanalysis, showing the fit of the models in mean G^2 values to the complete data (CD); the proportion of participants that were best described by a linear scale ($K = 1$); and models fits in mean G^2 values from a cross-validation analysis (CV).

Model	L/R task			B/Y task		
	G^2_{CD}	p(linear)	G^2_{CV}	G^2_{CD}	p(linear)	G^2_{CV}
1D-SSDM (radius)	279.1	0.31	380.4	262.8	0.94	340.5
1D-SSDM (center)	298.6	0.31	258.9	333.2	0.94	266.2
2D-SSDM (radius & center)	275.7	0.25	240.2	259.6	0.88	228.5

ANS-diffusion	287.0	0.50	241.2	256.1	0.94	224.2

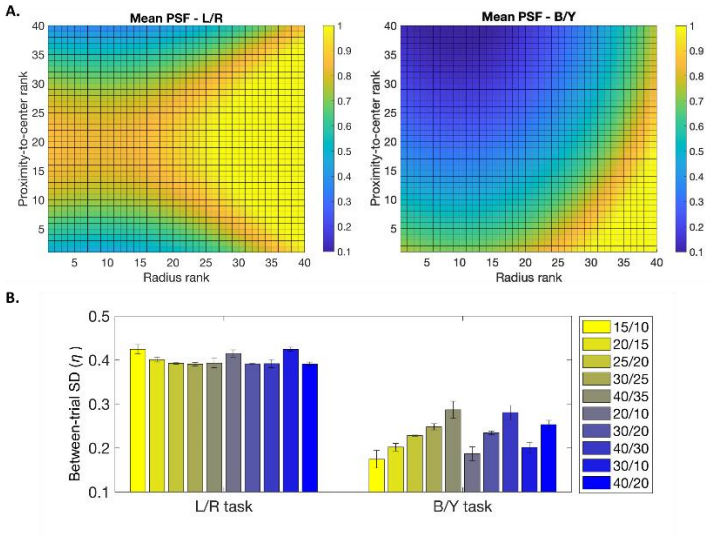


Figure 2. A) The PSFs from the two-dimensional SSDM in the L/R task (left panel) and the B/Y task (right panel) collapsed across participants. The colors represent the PSF values, ranging from dark blue (low probabilities) to light yellow (high probabilities). Dots are ranked from the smallest or closest to the largest or farthest, with the number of ranks determined by the set size condition. The shape of the PSF does not change with the number of ranks but just the number of bins. In both tasks, the group-level PSFs suggest a preference for sampling larger dots. However, while in the B/Y task there is a preference for sampling dots that are ranked closest to the center of the screen, in the L/R task, it shifts to dots that are close to the center of their respective side of the screen, indicated by a preference to sample the mid-range ranks on the proximity-to-center scale. B) The between-trial SD in drift rate (η) among set size conditions and tasks, estimated from the distribution of evidence formed by the selective sampling account. Findings show an increase in variability with set size in the B/Y task and approximately constant variability in the L/R task.

Lastly, we found that the behavior of the majority of participants was better accounted by a logarithmic scale in the L/R task and a linear scale in the B/Y task, suggesting higher sensitivity in the latter. Additionally, the PSF values derived from the two-dimensional PSFs were higher in the L/R task than in the B/Y task, suggesting that the sample size was larger in the former (M = .75, SD = .18) compared to the latter (M = .53, SD = .12). Consequently, the average difference in η between set-size conditions was significantly above zero in the B/Y task (M = .04, SD = .03; $t(15) = 5.49$, $p < .001$), indicating that variability increased significantly with set size. In the L/R task, however, the average difference in η between set-size conditions was not significantly different from zero (M = -0.01, SD = .05; $t(15) = -0.98$, $p = .342$), suggesting approximately constant variability or non-systematic changes in variability across set sizes (Figure 2B). This aligns with the primary assumptions of the linear and logarithmic ANS models, respectively.

The results from the reanalysis provide several key insights: i) The SSDM achieved a level of fit comparable to the ANS-diffusion model, demonstrating that common biases in numerical judgments can be explained by a stochastic sampling process. ii) The SSDM findings were consistent with previous classifications of the ANS models across tasks. The representation in the L/R task was more comprehensive but less sensitive than in the B/Y task, explaining the higher performance observed in the former. Specifically, in the L/R task, comprehensive sampling resulted in more stable variability with increasing set size on a logarithmic scale, whereas in the B/Y task, variability increased with set size on a linear scale. iii) In the B/Y task, attention likely focused more on the center of the display, whereas in the L/R task, attention was directed to the center of each side of the display. This indicates a task-dependent allocation of attention influenced by the spatial arrangement, which can be inferred using the selective sampling account. iv) The SSDM effectively captured the impact of dot surface area on numerosity judgments by prioritizing

larger dots in sampling. Interestingly, larger dots were more likely to be prioritized over smaller dots in the B/Y task than in the L/R task due to the smaller sample size, explaining the larger effect of surface area found in the B/Y task.

Experiment 1

Experiment 1 was conducted to empirically test one of the key insights from the reanalysis, which suggests a proximity-to-center bias in the B/Y task. That is, a distinctive preference to sample dots located closer to the center of an intermingled array rather than those at the periphery when forming numerosity representations, potentially exerting significant effects on behavior. To test this hypothesis, Experiment 1 incorporated systematic differences in the spatial position of dots with respect to the center of the B/Y display. The primary aim was to determine whether participants' numerical judgments supported or refuted the SSDM's prediction of a proximity-to-center bias, thus critically assessing a major prediction of the model.

If participants tend to form an incomplete numerical representation predominantly composed of dots located closer to the center of the display, we predict that they should identify the more numerous color more accurately when its dots are, on average, positioned closer to the center of the display than when they are positioned further away from the center. Following the same logic, if the dots of the less numerous color are situated closer to the center of the screen, we anticipate that participants will make more errors, often identifying the color with dots closer to the center as more numerous. These trends are expected to be amplified by set size, because selective sampling should be more pronounced in larger set sizes. In the SSDM, this behavior should be reflected in sampling policies that show a higher probability of sampling dots at the

center compared to those at the periphery, along with linear scaling to represent high sensitivity, as presumably required in intermingled displays.

We manipulated the dots' centrality by adjusting the average distance of dots from the center of the B/Y display for each color separately, setting up conditions in which either the dots representing larger numerosity or those representing smaller numerosity were, on average, positioned closer to the center (*large-N-closer* and *small-N-closer* conditions in Figure 3A). Additionally, the study encompassed six distinct set size conditions: 10/15, 20/25, 35/40, and their reverse counterparts, to test if the proximity-to-center bias is moderated by set size. The display's duration was brief (300 ms), allowing a single fixation per trial. A simulation of our predictions is reported in Appendix F, using the group-level PSF estimated from the B/Y task reanalysis to model evidence as a function of centrality and set size.

Methods:

Transparency and openness. This research was performed in accordance with the Declaration of Helsinki and was approved by The Ohio State University Institutional Review Board (protocol #2003B0201). Sample size was chosen to exceed the common sample size in similar studies (Kang & Ratcliff, 2020; Ratcliff & McKoon, 2018; 2020; Vanunu et al., 2020; Vanunu et al., 2021). Informed consent was provided by all participants. The data that support the findings is available at (https://osf.io/8mgdn/). This study was not preregistered.

Participants & Design. 36 students from the Ohio State University (16 females, age: 18–23 years, M=19.18) participated in exchange for course credit. We used a 2 (centrality: large-N-closer, small-N-closer) × 3 (set-size: 10\15, 20\25, 35\40) within-subjects design. Two participants

were excluded due to technical issues and one participant was excluded due to chance level performance[2].

Materials. The stimuli for Experiment 1 were derived from the B/Y task described in Ratcliff & McKoon (2018), with several adjustments to suit the experimental goals. In each trial, blue and yellow dots were pseudo-randomly scattered across a display "window" measuring 400×400 pixels, which translated to a visual angle of 15×15 degrees when viewed from a standard distance of 53 cm. This window was centrally positioned on a gray background of 640×640 pixels, all set against the center of a screen with a resolution of 1,280×960 pixels. A minimum spacing of 5 pixels was maintained between the edges of any two dots.

We manipulated the number of blue and yellow dots to produce six conditions: configurations of 15/10, 25/20, 40/35 (blue/yellow respectively), and their reverse combinations. To investigate whether participants exhibit a proximity-to-center bias in their sampling strategy—prioritizing dots located closer to the center of the display—we manipulated the spatial positioning of the dots based on their numerosity. Specifically, dots in one color were placed such that their average Euclidean distance from the center was 50 pixels less than dots in the other color, effectively positioning either the more numerous or the less numerous color closer to the center on average (large-N-closer and small-N-closer conditions, respectively). For example, in the '15/10 large-N-closer' configuration depicted in Figure 3A, 15 blue dots and 10 yellow dots were displayed, with the blue dots (representing the larger numerosity) having an average Euclidean distance from the center that was approximately 50 pixels shorter than that of the yellow dots.

[2] We defined chance level performance to occur if accuracy was at chance in at least 5 of the 6 conditions

In Experiment 1, we made a great effort to control for perceptual properties that have been previously found to influence numerosity judgment. To control for the surface area of dots in each color, the radii of the dots were drawn from a Gaussian distribution with means: 9 pixels for the larger-numerosity color and 10 pixels for the smaller-numerosity color. The SD was set to 1.5 pixels for both. This process was repeated until a set of radii was found that yielded approximately equal surface areas for the two arrays, ensuring that the surface area did not confound the numerosity judgment. Further, to control for the size of the convex hull (the visual envelope encompassing the dot arrays) while manipulating the distance of dots from the center, dots were initially placed randomly within the designated area until a configuration with roughly equal convex hull sizes for each color was achieved. After establishing these hulls, the positions of dots forming the convex hulls were fixed, and the positions of the remaining dots were adjusted to meet the condition of centrality.

Notably, these stringent constraints on display generation inevitably influenced the random distribution of dots across the screen, resulting in certain areas being more likely to contain dots than others. Specifically, due to the manipulation of distance from the center while preserving the convex hull sizes, dots were more frequently positioned at the center and the periphery of the window. This left four central areas in each quarter of the display window less densely populated with dots across trials. Nevertheless, across 20 blocks of 96 trials, this distortion was not noticeable.

Procedure. On each trial, an array of yellow and blue dots was displayed for 300 ms to allow for a single fixation. Participants were instructed to make numerosity judgments about whether there were more blue or yellow dots in the array. They could respond at any time during the display or after; however, if their response time was below 280 ms or above 1500 ms, a warning

message appeared on screen informing them that they responded “too fast” or “too slow,” respectively. Responses within 280-1500 ms received feedback about whether it was correct or an error. The number of trials in each condition was balanced within blocks. However, because response time varied between participants, the number of completed blocks differed between participants (testing stopped after 50 minutes). On average, 19.8 blocks (SD = 0.5) of 96 trials were completed, with a total of 1904.7 trials (SD = 50.5).

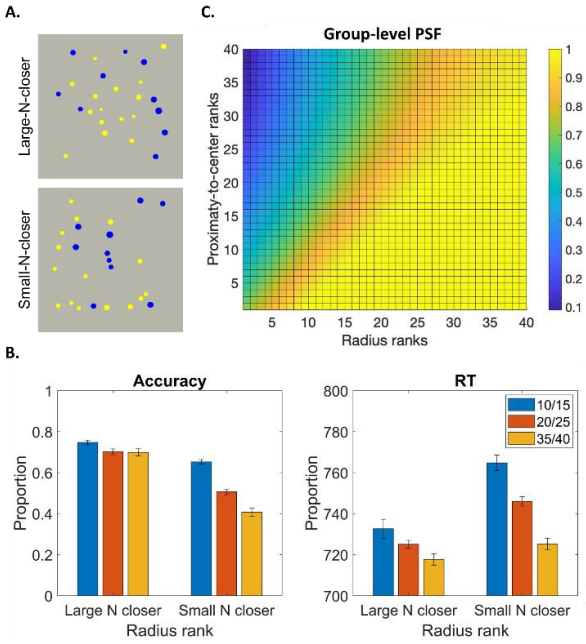


Figure 3. Behavioral and modeling results from Experiment 1: A) screen shots from the 15\10 set-size condition in a B/Y display when the larger- or smaller-numerosity color is closer to the center of the screen. B) Average accuracy ratings and RTs between the set-size and centrality conditions. Error bars correspond to the within-subjects standard error. C) The group-level PSF from the two-dimensional SSDM, with a color scale representing PSF values. Dots are ranked from the smallest or closest to the largest or farthest, with the number of ranks determined by the set size condition.

The results suggest a preference for sampling larger dots and dots that are closer to the center of the screen, while smaller dots in the periphery are often not sampled.

Results:

Behavioral results. The analysis of accuracy and RT was conducted using logit and linear mixed-effects models, respectively, with dots' centrality and set size serving as fixed effects, while variability between participants was accounted for with a random intercept effect. The data were consolidated across set sizes with opposite coloring configurations (e.g., 15/10 and 10/15) due to symmetric behavior observed between these conditions. The results are illustrated in Figure 3B, showing patterns of behavior that were previously captured by the linear account: as set size increased, accuracy and RT decreased.

For accuracy, we found significant main effects of set size ($\chi^2(2) = 978.25$, $p < .001$) and centrality ($\chi^2(1) = 2411.20$, $p < .001$), along with a two-way interaction between these two factors ($\chi^2(2) = 334.32$, $p < .001$). These findings suggest that participants often identified the color closer to the center as being more numerous, which resulted in below-chance performance in the small-N-closer condition for the largest set size ($t(31) = -4.55$, $p < .001$). Such findings robustly support our hypothesis that participants employ an incomplete sampling process that is biased towards dots located at the center of the display. Regarding RT, significant main effects were identified for set size ($\chi^2(2) = 154.57$, $p < .001$) and centrality ($\chi^2(1) = 129.33$, $p < .001$), along with a two-way interaction between these variables ($\chi^2(2) = 31.32$, $p < .001$).

Modeling results. We compared model fits between the two-dimensional SSDM and the ANS-diffusion model, where two drift rate coefficients were estimated for each centrality condition, as implemented in the reanalysis. Results from the model comparisons are shown in Table 2; the group-level PSF derived from the best-fitting SSDM is presented in Figure 3C (see

Figure A2 in Appendix D for the individual-level PSF); and the models' fit to data is illustrated in Figure A3C in Appendix E. See Appendix G for results from a more elaborated variant of the ANS-diffusion model, which aims to account for the interaction between set size and centrality by incorporating additional terms into the regression model used to compute the drift rate.

Findings show that the SSDM was the best-fitting model for both the complete dataset and for the testing dataset in the cross-validation analysis, outperforming the ANS-diffusion model variant that factored in multiple contributions to the drift rate based on the experimental conditions. Thus, the selective sampling account successfully explained the centrality effect and its interaction with set size through a process model, in which dots located nearer to the center of the display and larger dots are prioritized in sampling. This sampling policy resulted in an incomplete sample size (M = .67, SD = .23), and increased variability with set size across participants (η = [0.30, 0.35, 0.41], SD = [0.35, 0.32, 0.31] for the 10/15, 20/25 and 35/40 conditions, respectively). The average difference in η between set size conditions was significantly larger than zero (M = 0.05, SD = 0.06; $t(31) = 5.07$, $p < .001$), indicating a significant increase in variability with set size.

Table 2. Model comparisons from Experiment 1 and 2, showing for each model the fit in mean G^2 values to the complete data (CD), the proportion of participants that were best described by the linear scale, and mean G^2 values in a cross-validation analysis (CV). In Experiment 1, we compared the SSDM and the ANS-diffusion model. In Experiment 2, we also compared two variants of the SSDM: one that allocated probabilities according to the proximity of dots to the center of the screen (1-center SSDM) and another that allocated probabilities based on the proximity of dots to the center of their respective side of the screen (2-centers SSDM).

	Model	G^2_{CD}	p(linear)	G^2_{CV}
Exp. 1	SSDM	108.05	0.75	99.37

	ANS-diffusion	150.93	0.66	118.68
Exp. 2	2-centers SSDM	78.24	0.32	85.59
	1-center SSDM	79.44	0.24	94.92
	ANS-diffusion	88.34	0.4	88.68

Note: 1-center and 2-centers represent the proximity-to-center rank and proximity-to-two-centers rank, respectively.

Experiment 2

In Experiment 1, we found within the model framework that participants in the B/Y task were likely to form incomplete representations, which explained one ANS property—the source of increasing variability with set size—and confirmed key predictions from the reanalysis. However, it remains unclear why numerical judgments between tasks align with linear or logarithmic scaling—i.e., the second ANS property. We hypothesized that the type of scale reflects sensitivity to the visual display, influenced by display format and spatial resolution demands. An intermingled display presumably requires higher spatial resolution to distinguish between the blue and yellow dots, which is predominantly available at the center. In contrast, when the arrays are spatially separated, more comprehensive representations that include dots from the periphery, where spatial resolution is low, can be formed. Consequently, behavior in the B/Y task was best described by increasing variability with set size (incomplete samples) on a linear scale (high sensitivity), while behavior in the L/R task was best described by approximately stable variability with set size (comprehensive samples) on a logarithmic scale (low sensitivity).

In Experiment 2, we explored these hypotheses by introducing distractor dots of a different color into the L/R display (e.g., yellow target dots and blue distractors, as shown in Figure 4A). We anticipated that, as in the B/Y task, high spatial resolution would be necessary to effectively

distinguish between target and distractor dots. Consequently, we predicted that linear scaling would better account for participants' behavior than logarithmic scaling, which is associated with low sensitivity to the visual display when more comprehensive representations are formed. This setup aimed to assess whether the need for high sensitivity in distinguishing elements within an intermingled display would shift scaling from a logarithmic to a linear scale, mirroring the effects observed in the B/Y task.

Experiment 2 aimed to test another key insight derived from the reanalysis, which suggested a marked tendency for participants to prioritize larger dots over smaller dots in sampling, thereby accounting for the surface area effect via selective sampling. If this is true, the side of the screen containing larger dots would be perceived as more numerous, even when total surface area is controlled. To test this hypothesis, we varied the distribution of dot sizes on each side of the L/R display by manipulating the SD of the Gaussian distributions from which dot radii were derived, creating either a wide or a narrow distribution of dot sizes. Consequently, the side of the screen featuring a broader distribution of radii contained both the largest and smallest dots in the display, while maintaining an approximately equal surface area across dots on each side. See Figure 4A for an example display.

We predicted that participants would struggle to correctly identify the array with a greater number of dots if it presented a wide distribution of sizes (i.e., *large-N-wider* condition), due to a potential under-sampling of the smallest dots, which would diminish the perceived numerosity of the display. Following the same logic, we anticipated that accuracy would increase if the array with fewer dots featured a wide distribution of sizes (i.e., *small-N-wider* condition).[3] Additionally,

[3] Under-sampling the smallest dots or over-sampling the largest dots does not imply that sampling is symmetric between sizes. In cases where sampling is symmetric or balanced across sizes, we would not expect to find differences between the radius distribution conditions.

we predicted a larger difference in accuracy between the radius distribution conditions in larger set sizes, because selective sampling should be more pronounced under limited capacity. A simulation of our predictions is reported in Appendix F, using the group-level PSF estimated from the reanalysis of the L/R task to model evidence as a function of radius distribution and set size.

The SSDM also allows us to draw inferences about the distribution of attention across the display by altering the ranking scale in the PSF. For instance, a key insight from the reanalysis suggests that, in the L/R task, participants focus their attention towards the center of each side of the screen rather than the central point of the screen. This hypothesis can be examined by contrasting two variants of the SSDM: one that assigns probabilities based on proximity-to-center ranks and another based on *proximity-to-two-centers ranks*—namely, the proximity of each dot to the center of its respective screen side. For example, if dots A and B are positioned 20 and 40 pixels from the center of the left side of the screen, and dots C and D are positioned 30 and 50 pixels from the center of the right side of the screen, then their proximity-to-two-centers ranks will be 1, 3, 2, and 4 for dots A, B, C, and D, respectively. To corroborate the expected patterns of sampling with more direct measures of attention, we recorded participants' eye movements.

If participants direct their attention toward the center of each side of the L/R display, rather than keeping their gaze at the central point of the screen, the 2-centers SSDM is expected to provide a better fit to the data than the 1-center SSDM. Accordingly, we expect to observe more frequent and longer fixations at the two center positions on each side of the screen compared to other regions, excluding the central position where participants are required to focus to initiate the trial. In both SSDM variants, we anticipate finding a clear preference for sampling larger dots over smaller ones, which would potentially correspond with longer fixations on larger dots.

Methods:

Transparency and openness. This research was performed in accordance with the Declaration of Helsinki and was approved by The Ohio State University Institutional Review Board (protocol #2003B0201). Sample size was chosen to exceed the common sample size in similar studies. Informed consent was provided by all participants. The data that support the findings is available at (https://osf.io/8mgdn/). This study was not preregistered.

Participants & Design. 32 students from the Ohio State University (13 females, age: 18–25 years, M=21.13) participated in exchange for course credit. Seven participants were excluded due to technical issues with the eye tracker. We used a 2 (radius distribution: large-N-wider, small-N-wider) × 3 (set-size: 10\15, 20\25, 35\40) within-subjects design.

Materials. In Experiment 2, the stimuli were adapted from the L/R task in Ratcliff & McKoon (2018), with the addition of distractor dots to the display. For each block, the color of the target dots was randomly assigned to either blue or yellow, with the distractor dots assuming the opposite color. Each trial featured target and distractor dots for both the left and right sides of the display, distributed pseudo-randomly across a 600×600 pixels display window on each side. This area corresponded to 14.8×14.8 degrees of visual angle when viewed from a distance of 65 cm (the position of the chin rest), positioned at the center of each side of the screen with a resolution of 1680×1050 pixels. A minimum spacing of 30 pixels was maintained between the edges of adjacent dots. The display background was grey, covering the entire screen.

The manipulation of set size for the target dots in Experiment 2 was similar to that of Experiment 1. The radii of the target dots on each side of the screen were determined using two Gaussian distributions, with a mean of 11.5 pixels for the side displaying larger numerosity and 12.5 pixels for the side with smaller numerosity to control for surface area. The standard deviations

of these distributions were varied: in the large-N-wider condition, the side with a greater number of dots featured a wider distribution of dot sizes (SD = 4 pixels), resulting in sizes that ranged from 6 to 18 pixels. Conversely, the side with fewer dots had a narrow distribution of sizes (SD = 1 pixel), with sizes between 9 and 15 pixels. This configuration was inverted in the small-N-wider condition, as shown in Figure 4A.

Experiment 2 introduced a more randomized dot placement policy while ensuring that the differences in dot area and convex hull size between the left and right sides of the display were balanced for the conditions of radius distribution and set size. Furthermore, the number of distractor dots was adjusted according to the set size conditions: 10 for 15/10, 20 for 25/20, and 30 for 40/35. Finally, the distractor dots' radii were extracted from a Gaussian distribution with a mean of 11.5 pixels and an SD of 1 pixel, with radii ranging from 9 to 14 pixels.

Procedure. Participants were informed that they would be making numerosity judgments to determine whether there were more target dots on the left or the right side of the display, while ignoring the distractor dots. At the beginning of each block, they were told the color of the target dots. Each trial commenced with the presentation of a fixation cross; participants were instructed to focus on the cross to start the trial, which was then followed by a 500 ms display of the dots. The display duration was extended compared to previous studies for two reasons: i) the inclusion of distractors resulted in a more complex visual setup, and ii) we aimed to allow participants to make more than one fixation during the task, thereby enhancing the diagnostic value of the eye-tracking data collected.

Participants could respond at any time after stimulus onset; however, if their response time fell outside the 300 to 2500 ms window, an on-screen warning indicated they had responded "too

fast" or "too slow." Responses made within this timeframe were followed by feedback indicating whether the answer was correct or incorrect. The trials within each condition were evenly distributed across blocks. Due to variability in response time and calibration time across participants, the total number of completed trials varied. On average, participants finished 16.96 blocks (SD = 3.54) of 48 trials (except the final block), totaling approximately 741.70 trials (SD = 170.97). The task was performed under uniform luminance conditions for all participants.

Eye tracking. Participants' eye movements were recorded using an EyeLink 1000 system, operating at a sampling rate of 1000 Hz. This system was paired with a 22-inch monitor, set at a resolution of 1680×1050 and a refresh rate of 60 Hz. To ensure consistent head positioning and gaze direction, participants' head movements were stabilized with a chinrest positioned 65 cm from the screen. At the start of each session, the eye tracker was calibrated for each participant to ensure accuracy, and drift corrections were performed after each block.

Each trial was initiated once participants maintained their gaze on a fixation cross for a continuous period of 0.5 seconds, ensuring that their gaze was centrally aligned when the stimuli were presented. The area of interest (AOI) for the fixation cross was defined as a 100×100 pixels square centered on the screen. To establish a consistent AOI size for dots of varying sizes and positions, we added the minimum diameter of the dots to the minimum spacing between dot edges (12 pixels + 30 pixels = 42 pixels), resulting in a standardized AOI of 42×42 pixels for each dot. This configuration allowed for the smallest and largest dots to occupy 6% and 58% of their respective AOIs, ensuring that the AOIs of adjacent dots did not overlap, as depicted in Figure 5A.

Two metrics were used to understand participants' visual attention and physiological responses during the task: the number and positions of fixations and the proportional gaze duration

on each dot. A fixation was identified when the gaze position remained constant for over 20 ms. A change in gaze position was recorded if the absolute difference in gaze coordinates between two consecutive samples (time t and t+2) exceeded 10 pixels on either the x-axis or y-axis. For each participant, we determined the proportion of fixations by mapping the fixation points that fell within a 50-pixel radius of each fixation position. The total number of fixations across the visual field was divided by the number of trials to calculate the average fixation proportion.

Each dot within the display was categorized according to two criteria: radius rank and proximity-to-two-centers rank, the latter being an evaluation based on the dots' closeness to the center of their respective side of the display. The average time spent gazing at dots of each rank was calculated separately for each participant and set-size conditions (due to different number of ranks), followed by normalizing the gaze duration against the total gaze time spent on all dots within the respective condition. Following normalization, these proportions were averaged across all participants. Gaze outside of the specified AOIs was excluded from the analysis.

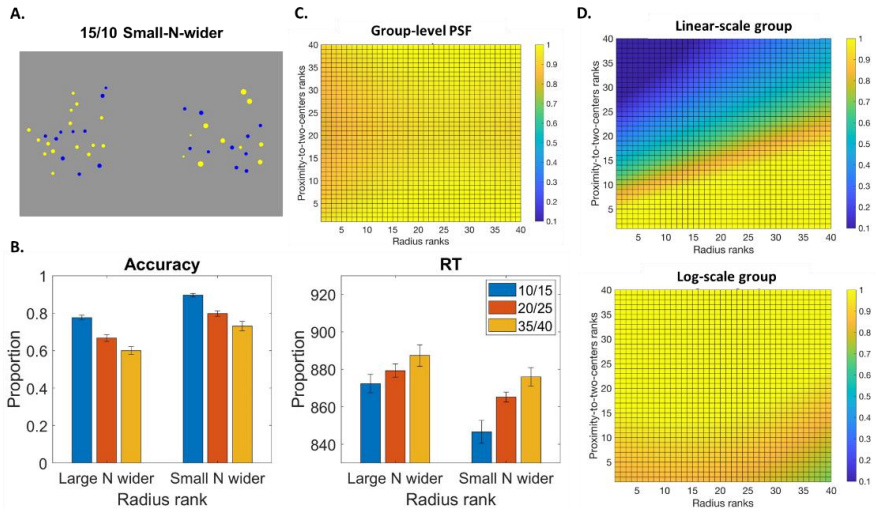


Figure 4. Behavioral and modeling results from Experiment 2: A) screen shots from the 15\10 set size condition in a L/R display (we also tested 25\20 and 40\35), in which the side with the smaller

numerosity had a wider radius distribution (blue dots are distractors). B) Mean accuracy and RTs as a function of set size and radius distribution. Error bars correspond to the within-subjects standard error. C) The group-level PSF across participants from the two-dimensional SSDM with two centers. Dots are ranked from the smallest to the largest, and from closest to farthest from the center of their respective screen side, with the number of ranks determined by the set size condition. The results suggest a balanced and comprehensive sampling policy across radius and proximity-to-two-centers ranks. D) The group-level PSF split into two groups according to the best-fitting scale (linear, log), showing incomplete sampling policy from the center in linear-scale group and a more balanced and comprehensive sampling policy in the log-scale group.

Results:

Behavioral results. We analyzed the data using similar mixed-effects models as in Experiment 1, but with radius distribution as a fixed effect instead of centrality. The results are depicted in Figure 4B, showing patterns of behavior that were previously captured by the log account: as set size increased, accuracy decreased and RT increase, contrary to our predictions.

For accuracy, we found main effects of set size ($\chi^2(2) = 432.91$, $p < .001$) and radius distribution ($\chi^2(1) = 431.06$, $p < .001$) and a two-way interaction of both ($\chi^2(2) = 8.49$, $p = .014$). These results align with our hypotheses: participants exhibited greater accuracy when the array with smaller numerosity featured a wider distribution of dot sizes (small-N-wider condition) compared to when the array with larger numerosity had the wider distribution (large-N-wider condition). Additionally, the difference between the radius distribution conditions appears to be more pronounced in the larger set sizes ($\Delta M_{10/15} = 0.12$, $\Delta M_{20/25} = 0.13$, $\Delta M_{35/40} = 0.13$), affirming our prediction of greater biases in larger set sizes due to limited capacity. These findings indicate that participants' sampling strategies may have neglected smaller dots in forming numerosity

representations, especially when there was a large number of dots in the display competing for attention. Consequently, arrays containing the smallest dots were perceived as less numerous, despite controlling for surface area.

Regarding RT, we observed main effects of set size ($\chi^2(2) = 38.13$, $p < .001$) and radius distribution ($\chi^2(1) = 30.26$, $p < .001$), showing that RT generally decreased as accuracy increased, therefore, adhering to patterns that were previously captured by the logarithmic account. This observation contradicts our initial prediction that the inclusion of distractors in the L/R display would lead participants to base numerical judgments on representations with high spatial resolution, which is presumably captured by linear scaling. One plausible explanation for this discrepancy is that participants were able to selectively ignore the distractors during sampling. This ability is consistent with sampling policies governed by top-down attention (Vanunu et al., 2021), and aligns with previous findings from a similar task in which participants were capable of suppressing or ignoring unattended dots in a different color (Cai, Hofstetter, Harvey, & Dumoulin, 2022; Ratcliff & McKoon, 2018, Experiment 3).

Modeling results. We evaluated two variations of the two-dimensional SSDM that allocate sampling probabilities based on the radius ranks, in conjunction with either the proximity-to-center ranks (1-center SSDM) or the proximity-to-two-centers ranks (2-centers SSDM) and the ANS-diffusion model. We compared the goodness-of-fit for these two SSDM variants to investigate participants' attentional strategies in the L/R task—specifically, whether participants were more likely to shift their focus between each side of the screen or to maintain their gaze centrally between the two arrays. Results from the model comparisons are shown in Table 2; the group-level PSF derived from the best-fitting SSDM version is presented in Figure 4C (see Figure A2 in Appendix D for the individual-level PSF); and the models' fit against data is shown in Figure A3D.

Results from the ANS-diffusion model variant that aims to account for the interaction effect between radius distribution and set size are detailed in Appendix G.

Findings show that the 2-centers SSDM provided the best fit for both the complete dataset and the testing dataset in the cross-validation analysis. This supports our predictions derived from the reanalysis of data from Ratcliff & McKoon (2018), which suggests that participants in the L/R task pay more attention to dots positioned near the center of each side of the screen, rather than those near the central point of the screen. This is despite the requirement for participants to fixate their gaze at the center at the start of each trial.

Consistent with the RT data but contrary to our expectations, the inclusion of distractors did not seem to prompt participants to rely on information processed with high spatial resolution, presumably captured by the more sensitive linear scale. Instead, the behavior of the majority of participants was more accurately described by a logarithmic scale, suggesting low sensitivity to the visual display. Consequently, sample size across participants was high (M = .71, SD = .23) while variability remained approximately constant with set size (η = [0.32, 0.31, 0.33], SD = [0.233, 0.25, 0.30] for the 10/15, 20/25 and 35/40 conditions, respectively). The average difference in η between set size conditions was not significantly different than zero (M = 0.01, SD = 0.08; $t(24) = 0.54$, $p = .590$), suggesting that variability remained relatively stable across set sizes.

Interestingly, the group-level PSF derived from the 2-centers SSDM indicated a balanced sampling policy across ranks, suggesting that participants prioritized larger dots near their respective side's center, as well as dots located at the periphery of the arrays (Figure 4C). However, a group-level display that collapses PSF values across participants can show misleading results because it may obscure meaningful individual differences. Therefore, when the group-level PSF indicates a balanced policy, it is also important to examine the individual-level display, as depicted

in Figure A2 in the Appendix, where it appears that some participants prioritized dots closer to the center while others prioritized dots at the periphery.

Further analysis revealed that differences in sampling behavior between dots at the center and the periphery may be related to spatial resolution demands. When participant data were grouped according to the scale that best described their performance—with 7 participants best accounted for by a linear scale and 18 by a logarithmic scale—opposite patterns emerged. Participants in the linear-scale group formed incomplete samples that prioritized dots at the center, whereas those in the log-scale group exhibited a more balanced and comprehensive sampling process that included dots at the periphery, as illustrated in Figure 4D. These conflicting sampling policies between participants who were best described by a linear or a logarithmic scale align with the assumption that scale is related by sample size, as well as with the characterizations of the linear and log accounts. Namely, the linear account is associated with higher sensitivity to incomplete samples of dots from the center, whereas the logarithmic account is associated with lower sensitivity to more comprehensive samples that include dots from the periphery.

Eye-tracking results. Two metrics were examined: the number and positions of fixations and the proportion of gaze duration for each dot in the display according to its radius rank or proximity-to-two-centers rank. The results are shown in Figure 5. On average, participants performed 1.84 fixations (SD = 0.54) during the stimulus display. Fixations at the center of the screen were most common, as required by the task. However, on many trials (M = 61%, SD = 0.32), participants moved their gaze to the center of one side, especially the left side (Figure 5B). Moreover, the proportion of gaze duration was greater for dots located closer to the center of their respective side (Figure 5C, top panel), supporting our predictions and the SSDM findings, which indicated that central dots were prioritized during sampling.

Curiously, gaze duration for dots at the periphery (i.e., ranked furthest from the respective center) was the lowest, and gaze duration among radius ranks seems to be random (Figure 5C, bottom panel), which contrast part of the trends identified in the resulting PSF, whereby larger dots and dots at the periphery were often included in the subsample. However, it is important to note that within a brief display period (500 ms), participants typically have enough time to perform two fixations: one at the center of the screen, at the fixation cross, and another at the center of one of the sides. Consequently, gaze duration among radius ranks was largely influenced by the random placement of dots around these two focal points, and gaze durations for dots at the periphery often approached zero.

This raises the question: How was sampling achieved outside of the focal area? One possible explanation pertains to the role that spatial resolution presumably plays in forming numerical representations. Specifically, to sample dots from the periphery during a brief display, participants must gather as much information as possible from the periphery under conditions of low spatial resolution. This approach is consistent with logarithmic scaling that represents diminished sensitivity.

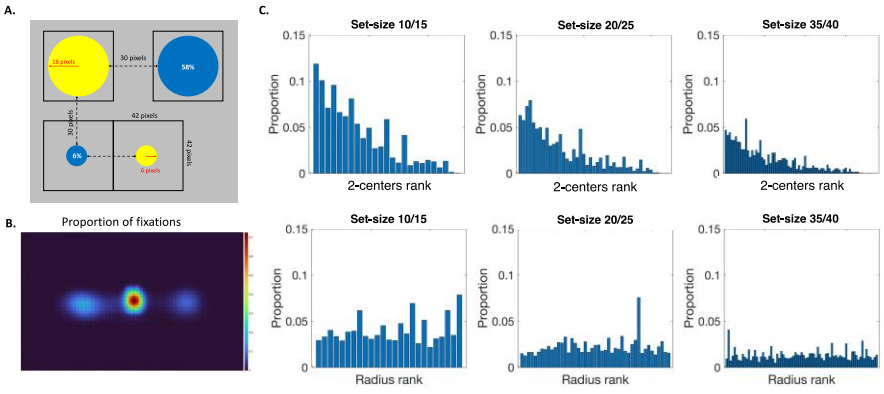


Figure 5. A) An example placement of dots with the largest radius (18 pixels) and the smallest radius (6 pixels) in Experiment 2. The minimum distance between dots is 30 pixels (dashed

arrows). The solid squares represent the Areas of Interest (AOIs), 42x42 pixels in size around the center of each dot. This size ensured that the AOIs of adjacent dots would not overlap. The largest and smallest dots occupy 58% and 6% of the AOI, respectively. B) The proportion of fixations according to screen position ranging from zero (dark blue) to one (dark red). C) the proportion of gaze duration according to proximity-to-two-centers ranks and radius ranks.

Experiment 3

Converging evidence from behavioral and computational analyses suggests that an intermingled array of blue and yellow dots produces numerosity representations on a linear scale, primarily relying on a subsample of dots from the center with downstream effects on behavior. Specifically, we found that participants often identified the color with more centrally positioned dots as more numerous, even when this was a mistake—a behavior captured by the SSDM via higher sampling probabilities for centrally located dots. We hypothesize that this proximity-to-center bias arises because the intermingled display requires high spatial resolution to differentiate between the colors. However, when the arrays are spatially separated, as in the L/R display, the need for central high spatial resolution is reduced and can be traded off with lower spatial resolution information that includes dots from the periphery. Thus, we hypothesize that the increased variability with set size, linear scaling, and the proximity-to-center bias observed in the B/Y task stem from the intermingled display format.

Experiment 3 was designed to isolate the effect of an intermingled (versus separate) display format on the formation of non-symbolic numerosity representations and its downstream consequences on accuracy and RT. To achieve this goal, one group of participants performed a B/Y task with an intermingled display, while a second group performed a similar task with one

notable difference: the blue and yellow dots were spatially separated within an identically sized display, as depicted in Figure 6A. In both groups, set size and dots' centrality were manipulated in the same manner as in Experiment 1, making display format the sole distinguishing factor between groups. Lastly, we recorded eye movements to corroborate the expected patterns of sampling between groups with a direct measure of attention.

We hypothesized that the proximity-to-center bias is limited to intermingled displays, which presumably require high spatial resolution to accurately determine the more numerous color. Therefore, we predicted that spatially separating the blue and yellow dots within an identically sized array would reduce the reliance on central information, as it enables spatial-based rather than color-based discrimination. Consequently, spatial separation should lead to less pronounced differences in behavior between the centrality conditions.

The expected differences in behavior between the display format groups should be captured in the SSDM analysis across three dimensions: First, a comparison of the group-level PSF between display format groups should reveal a stronger preference for sampling dots at the center in the intermingled format relative to the separate format. However, unlike in Experiment 2, the display time was very brief (300 ms). Therefore, it is unlikely that participants had enough time to make additional fixations (beyond the required central fixation) before the stimulus disappeared. This led us to predict a better fit for the 1-center SSDM over the 2-centers SSDM in both display format groups. Accordingly, we expect to find more frequent and longer fixations on dots closer to the center of the screen in both groups, rather than on dots near the center of each side.

Second, sample size is predicted to be larger (i.e., more comprehensive) in the separate versus intermingled formats due to reduced sensitivity demands in the former. Third, the representation scale is predicted to reflect sensitivity to the visual display. Therefore, a greater

proportion of participants in the intermingled format group is anticipated to be accounted for by a linear scale compared to those in the separate format group due to the visual complexity of the former. Following the results from Experiment 2, we do not expect to find differences in gaze duration between the radius ranks.

Methods:

Transparency and Openness. This research was conducted in accordance with the Declaration of Helsinki and received approval from the University of Chicago Institutional Review Board (IRB23-0402). The sample size was selected to exceed that commonly seen in previous studies. All participants provided informed consent. The data supporting the findings are available at [https://osf.io/8mgdn/]. This study was not preregistered.

Participants & Design. Seventy nine participants, recruited from the Centre of Decision Research's pool at the University of Chicago Booth School of Business (36 females; age range: 18–45 years, M = 23.92, SD = 4.79) for a monetary compensation of 12 USD, were randomly assigned to one of the display format groups (38 in the intermingled format group and 41 in the separate format group). Four participants were excluded due to chance level performance and two participants were excluded from the eye-data analysis due to technical issues with the eye tracker. We tested a 2 (centrality: large-N-closer vs. small-N-closer) × 3 (set size: 10/15, 20/25, 35/40) × 2 (display format: intermingled vs. separate) design. Centrality and set size were manipulated within subjects, and display format was manipulated between subjects.

Materials. In Experiment 3, stimuli were adapted from the B/Y task in Experiment 1, with an additional manipulation of the display format between subjects. In both display format groups, blue and yellow dots were distributed pseudo-randomly across a 750×750 pixels display window,

positioned at the center the screen with a resolution of 1920x1080 pixels. This area corresponded to 12.4×12.4 degrees of visual angle when viewed from a distance of 3 feet (position of the chin rest). A minimum spacing of 5 pixels was maintained between adjacent dots. The display background was grey, covering the entire screen. The display's duration was brief (300 ms).

In the intermingled format group, blue and yellow dots were intermingled within a single array, as in Experiment 1. In the separate format group, blue and yellow dots were spatially separated between two sides of an identically sized display window (i.e., identical visual angle) if split vertically into two halves. Set size and centrality were manipulated as in Experiment 1, and differences in the dots' surface area and convex hull between the yellow and blue arrays were controlled for. The means of the radii distributions were set to 9.5 pixels for the more numerous color and 10.5 pixels for the less numerous color to control for the surface area of the dots, with a standard deviation of 1.5 pixels in both distributions.

Procedure. Participants were informed that they would be making numerosity judgments to determine whether there were more blue or yellow dots in the display. The procedure of Experiment 3 was identical to that of Experiment 1, with the exception of calibrating the eye tracker as detailed in Experiment 2. Additionally, response times faster than 300 ms or longer than 2500 ms prompted a warning message: "too fast" or "too slow," respectively. Due to variability in response time and calibration time across participants, the total number of completed trials varied. On average, participants finished 6.49 blocks (SD = 2.02) of 96 trials (except the final block), totaling approximately 692.08 trials (SD = 199.44). The task was performed under uniform luminance conditions for all participants.

Eye-tracking. Participants' eye movements data were recorded using an EyeLink Portable Duo system, which operates at a sampling rate of up to 2000 Hz. This system was paired with a

23-inch monitor, set at a resolution of 1920x1080 and a refresh rate of 60 Hz. Participants' head movements were stabilized with a chinrest positioned 3 feet from the screen.

The AOI for the fixation cross and the time required for fixation to initiate the trial were maintained identical to those in Experiment 2. For dots of various sizes and positions, the AOI was set at 27×27 pixels around each dot, ensuring it encompassed the largest dots within the display, whereby the smallest and largest dots occupied 11% and 97% of their respective AOIs. Due to the minimal spacing between dots (5 pixels), achieving a unified AOI size for all dots without overlaps was impossible. Consequently, a single sample of participants' gaze could overlap with multiple AOIs of adjacent smaller dots, and it would be recorded accordingly. Lastly, the same metrics used for analyzing eye-tracking data in Experiment 2 were employed in Experiment 3, with the exception that gaze duration was measured in terms of proximity-to-center ranks rather than proximity-to-two-centers ranks.

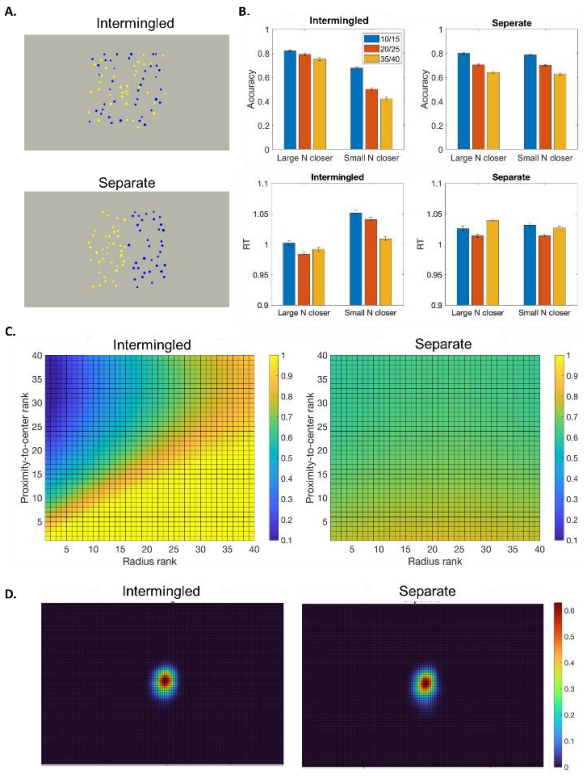


Figure 6. The behavioral and modeling results from Experiment 3: A) Example displays from the 25/20 large-N-closer condition in the intermingled and separate display formats, where both displays depict more yellow dots than blue dots, with the yellow dots also being closer, on average, to the center of the screen. B) Mean accuracy and RT as a function of set size, centrality, and display format group. Error bars correspond to the within-subjects standard error. C) The group-level PSFs in each display format group from the two-dimensional 1-center SSDM. The color scale represents the PSF values. Dots are ranked from the smallest or closest to the largest or farthest, with the number of ranks determined by the set size condition. The results suggest a preference for sampling larger dots and dots closer to the center of the screen under the intermingled format, while a more balanced sampling policy is observed under the separate format. D) The proportion of fixations according to screen position ranging from zero (dark blue) to one (dark red), showing similar results across the display format groups.

Results:

Behavioral results. We analyzed the data using similar mixed-effects models as in Experiment 1, but with display format group as an additional fixed effect. The data were consolidated across set sizes with opposite coloring configurations (e.g., 15/10 and 10/15) due to symmetric behavior observed between these conditions. The results are illustrated in Figure 6B.

Replicating findings from Experiment 1, we found significant main effects of set size ($\chi^2(2) = 1083.08$, $p < .001$) and centrality ($\chi^2(1) = 943.42$, $p < .001$) on participants' accuracy, along with a two-way interaction between these two factors ($\chi^2(2) = 50.47$, $p < .001$). Additionally, we found a significant main effect of display format group ($\chi^2(1) = 8.86$, $p = .003$), a two-way interaction between display format group and centrality ($\chi^2(1) = 912.86$, $p < .001$), and a three-way interaction among all variables ($\chi^2(2) = 53.24$, $p < .001$). These results suggest that the proximity-to-center

bias was moderated by the display format. The centrality effect ($\chi^2(1) = 1833.97$, $p < .001$) and its interaction with set size ($\chi^2(2) = 103.05$, $p < .001$) were replicated under the intermingled format, leading to below-chance performance in the small-N-closer condition for the largest set size ($t(36) = -3.41$, $p = .002$). In contrast, a null main effect ($\chi^2(1) = 2.88$, $p = .090$) and interaction ($\chi^2(2) = 0.70$, $p = .706$) were found under the separate format.

Consistent with our predictions, it appears that the proximity-to-center bias is limited to intermingled display formats that presumably require high spatial resolution to distinguish between the number of blue and yellow dots. This demand is less pronounced when the dots are spatially separated, presumably enabling participants to form numerosity representations based on the dots' spatial position rather than color.

Regarding RT, significant main effects were identified for set size ($\chi^2(2) = 26.82$, $p < .001$) and centrality ($\chi^2(1) = 97.23$, $p < .001$), along with two-way interactions between set size and centrality ($\chi^2(2) = 24.64$, $p < .001$), set size and display format group ($\chi^2(2) = 54.28$, $p < .001$), and centrality and display format group ($\chi^2(1) = 109.33$, $p < .001$), and a three-way interaction among all variables ($\chi^2(2) = 10.83$, $p = .005$). Notably, the monotonic patterns of RT that were previously found in the B/Y and L/R tasks—i.e., decreasing or increasing RT with accuracy, respectively—were only replicated in the small-N-closer condition in the intermingled-format group, while the RT patterns in the remaining conditions did not show monotonic trends with accuracy (Figure 6B).

Modeling results. We compared model fits between the ANS-diffusion model, the two-dimensional 1-center SSDM, and a 2-centers SSDM variant, which, in this context, allocated probabilities based on the dots' proximity to the center of their respective side of the display when split vertically into two halves. Results from the model comparisons are shown in Table 3; the group-level PSF derived from the best-fitting SSDM is presented in Figure 6C (see Figure A2 in

the Appendix for the individual-level PSF); and the models' fit to data is illustrated in Figure A3E. Results from the ANS-diffusion model variant that aims to account for the interaction effect between centrality and set size are reported in Appendix G.

Findings show that in both groups, the 1-center SSDM outperformed both the ANS-diffusion model and the 2-centers SSDM, which assumed participants allocated attention to two centers of the display window if split vertically in half, when fitted to the complete dataset and in a cross-validation analysis. This confirmed our hypothesis that selective sampling explains the formation of numerical representations through a process model with limited attentional capacity, enabling inferences about the distribution of attention—that is, confirming that participants are likely to remain fixated at the center of the display due to the brief display duration.

Furthermore, although differences were non-significant, the proportion of participants best accounted for by the linear scale (i.e., $K = 1$) was directionally larger in the intermingled format group (M = .60, SD = .50) compared to the separate format group (M = .40, SD = .50; t(73) = 1.74, p = .086). This trend aligns with our main hypothesis, which posits that the scale used in forming non-symbolic numerosity representations may reflect sensitivity to the visual display. A linear scale corresponds to higher sensitivity when distinguishing between objects in an intermingled format, while a logarithmic scale corresponds to lower sensitivity when forming comprehensive representations from a spatially separated format. See the General Discussion for possible reasons why a significant scale classification was not found between the display format groups.

Furthermore, although differences were non-significant, the proportion of participants best accounted for by the linear scale (i.e., $K = 1$) was directionally larger in the intermingled format group (M = .60, SD = .50) compared to the separate format group (M = .40, SD = .50; t(73) = 1.74, p = .086). This trend aligns with our main hypothesis, which posits that the type of scale used in

forming non-symbolic numerosity representations may reflect sensitivity to the visual display. A linear scale corresponds to higher sensitivity when distinguishing between objects in an intermingled format, while a logarithmic scale corresponds to lower sensitivity when forming comprehensive representations from a spatially separated format. See the General Discussion for possible reasons why a significant scale classification was not found between the display format groups.

Table 3. Model comparisons from Experiment 3, showing for each model the fit in mean G^2 values to the complete data (CD), the proportion of participants that were best described by the linear scale, and mean G^2 values in a cross-validation analysis (CV) between the display format groups.

Display format	Model	G^2_{CD}	p(linear)	G^2_{CV}
Intermingled	2-centers SSDM	86.04	0.62	92.48
	1-center SSDM	79.12	0.60	83.98
	ANS-diffusion	94.66	0.78	95.95
Separate	2-centers SSDM	75.07	0.63	78.21
	1-center SSDM	74.30	0.40	77.77
	ANS-diffusion	78.67	0.40	79.00

Note: 1-center and 2-centers represent the proximity-to-center rank and proximity-to-two-centers rank, respectively.

We also observed substantial variations in sampling policies between the condition groups (Figure 6C). In the intermingled format, participants often prioritized central dots over peripheral ones and larger dots over smaller ones, replicating the results from the B/Y task in the reanalysis and Experiment 1. In the separate format, the group-level PSF indicated a more balanced sampling policy across ranks. This suggests that when blue and yellow dots were intermingled in a single display, participants prioritized dots near the center, consistent with our hypothesis that spatial

resolution demands drive the proximity-to-center bias. However, when the dots were spatially separated and the demand for high spatial resolution was presumably reduced, all dots in the display were sampled with roughly equal likelihood.

Nevertheless, with a balanced group-level PSF, it is also important to examine individual-level data to identify meaningful differences in participants' sampling policies. As shown in Figure A2 in the Appendix, the balanced group-level PSF under the separate format was due to conflicting individual sampling policies within the condition group. While PSF values among the proximity-to-center ranks were relatively balanced across participants, accounting for null differences between the centrality conditions in the separate format group, opposite patterns emerged among the radius ranks. Specifically, most participants discounted the smallest dots in the display, while a smaller proportion prioritized the smallest dots over larger ones (see Appendix D for more details). Consequently, sample size was not significantly different in the separate format (M = 0.57, SD = 0.24) compared to the intermingled format (M = 0.64, SD = 0.28; $t(73) = 1.01$, $p = .315$), contrary to our prediction. In both groups, variability increased with set size (separate: η = [0.25, 0.27, 0.28], SD = [0.10, 0.09, 0.11]; intermingled: η = [0.28, 0.28, 0.32], SD = [0.16, 0.11, 0.11]).

Eye-tracking results. We analyzed the number and positions of fixations and the proportion of gaze duration for each dot in the display according to its radius rank or proximity-to-center rank. The proportion of fixations according to screen position is depicted in Figure 6D and the proportion of gaze duration among ranks are depicted in Figure A5 in Appendix H. In both groups, participants, on average, performed 1.06 fixations (SD = 0.15) at the center of the screen during the stimulus display. Additionally, the proportion of gaze duration replicated previous

results, where dots closer to the central point of the screen were fixated for longer in both groups, while differences among radius ranks appear to be random (Figure A5)

These findings provide further support for the conclusions drawn from the SSDM analysis using direct attention measures. In both display format groups, participants were more likely to attend to (and therefore sample) dots located at the center of the display, rather than near the two centers on each side if split in half, resulting in a better fit of the 1-center SSDM compared to the 2-centers SSDM across groups. However, in the intermingled format, which presumably demands higher spatial resolution, our account suggests that numerosity representations were often limited to overtly fixated central dots, leading to a proximity-to-center bias in numerosity judgments. In contrast, in the separate format with lower spatial resolution demands, our account implies that participants were likely able to include peripheral dots through covert fixation, though this came at the cost of sensitivity.

General Discussion

The current study addresses three key issues in forming numerosity representations that traditional Approximate Number System (ANS) models fail to explain: (i) why variability in representations increases with set size in some numerical judgments but remains relatively constant in others, (ii) what drives people to form numerosity representations on a linear or a logarithmic scale with constant or varying variability, and (iii) the underlying cognitive mechanism that gives rise to perceptual effects on numerosity judgments.

To answer these questions, we used the Selective Sampling and Diffusion Model (SSDM), based on the selective sampling account (Vanunu et al., 2020; Vanunu et al., 2021). This account posits that people often do not use all available information when making decisions but instead

selectively sample it according to their goals, available resources and the display format. In numerical judgments, selective sampling suggests that judgments are based on a subsample of information where perceptual properties of objects, such as proximity to the central point of the display and larger size, are prioritized. Supporting this idea, previous studies have shown that time constraints on stimulus presentation influence numerosity estimates, depending on the number of fixations (i.e., how much information is sampled) and cognitive capacity (Cheyette & Piantadosi, 2019; Cheyette & Piantadosi, 2020).

Following the results from the SSDM analysis and the empirical tests, we can infer that (i) variability arises from a stochastic sampling process driven by selective attention and task demands, leading to variability in numerosity representations for identical displays if the sampling process is incomplete; (ii) scaling may reflect sensitivity to the visual display, with high sensitivity corresponding to linear scaling and low sensitivity corresponding to logarithmic scaling, depending on the spatial resolution demands imposed by the display format (intermingled vs. separate); and (iii) the distribution of selective attention to perceptual properties of the display, such as visual size and spatial position, appears to be responsible for producing perceptual biases in numerosity judgments.

We initially tested the SSDM on existing datasets from common numerosity-discrimination tasks, where participants decided if there were more blue/yellow dots on the screen or more dots on the left/right side of the screen (Ratcliff & McKoon, 2018). We compared the SSDM to the ANS-diffusion model, which implemented a similar mechanism with the ANS as a representation model. Subsequently, we tested the SSDM's predictions by manipulating the perceptual properties prioritized during sampling in the reanalysis, specifically dots' centrality in the B/Y task (Experiments 1 and 3) and the distribution of dots' sizes in the L/R task (Experiment

2). We also introduced distractors in Experiment 2 to test if a linear scale would better account for behavior when an intermingled display of target and distractor dots presumably requires high spatial resolution to distinguish between them. This notion was further examined in Experiment 3, where we analyzed participants' behavior in identical numerosity-discrimination tasks differing only in the display format of dots: intermingled or spatially separated. Lastly, eye-movement recordings in Experiments 2 and 3 provided support with a direct measure of attention.

Findings across the reanalysis and all three experiments showed that the SSDM accounted for behavior better than or similarly to the ANS-diffusion model, while providing additional insights into how the distribution of overt and covert attention to perceptual properties of the display influenced the formation of numerosity representation. According to the SSDM findings, larger dots and dots positioned closer to the center of the array were frequently prioritized during sampling across datasets. However, the sampling process was often more comprehensive when dots were spatially separated compared to when they were intermingled. This difference in sample size explains variations in the linear or log ANS model classifications between the L/R and B/Y tasks. Specifically, why variability—one property of the ANS models—was approximately constant in the former but increased with set size in the latter.

The manipulation of perceptual properties in Experiments 1-3 confirmed the SSDM's predictions for the distribution of attention and the subsequent behavior. Specifically, we found that accuracy decreased with set size below chance level if the dots in the color with smaller numerosity were closer to the center of an intermingled B/Y display (Experiments 1 and 3), and increased if the screen-side with a smaller numerosity in the L/R display also had the dots with the smallest radii (Experiment 2). The selective sampling account explained these perceptual effects on numerosity judgments by prioritizing corresponding properties in sampling, such as screen

position and visual size. In contrast, the ANS-diffusion model could not account for these effects without additional assumptions.

Towards explaining variations in the second ANS property—i.e., the representation scale—the results of Experiment 3 suggests that the display format of the arrays and the subsequent demand for high spatial resolution influence whether numerosity judgments would align with linear or logarithmic scaling. According to the SSDM, when the blue and yellow dots were intermingled within a single array, the high spatial resolution required to distinguish between them led the majority of participants to form incomplete but sensitive representations from the center, resulting in the documented proximity-to-center bias and a better fit for the linear scale. Conversely, this demand was presumably reduced when the dots were spatially separated, allowing numerosity representations to form based on spatial position rather than color. This likely enabled participants to create less sensitive representations that included dots at the periphery, eliminating the proximity-to-center bias and resulting in a better fit for the logarithmic scale.

Curiously, the difference in the proportion of participants best accounted for by the linear or logarithmic scales was not as pronounced in Experiment 3 as in previous comparisons between the B/Y and L/R tasks. This discrepancy suggests that another perceptual factor may influence scaling beyond the intermingled versus separate formats. Specifically, the visual angle of the display may also affect the representation scale. A separate display with a larger visual angle is more likely to result in logarithmic scaling than an intermingled display with a smaller visual angle, as dots in the former are more likely to be positioned at the periphery of the visual field. Consequently, more participants were best accounted for by the logarithmic scale in the L/R tasks in Experiment 2 and the reanalysis, compared to the separate format group in Experiment 3, where the visual angle was up to three times smaller.

A smaller visual angle in the separate format may have introduced new strategies that were less applicable in the original L/R task with a wider display. For example, modeling results from Experiment 3 suggest that some participants in the separate format group prioritized smaller dots over larger ones, in contrast to previous findings where larger dots were typically prioritized across datasets. The smaller visual angle in the separate format group may have enhanced visual sensitivity, allowing some participants to detect subtle differences in average dot size between the sides. With surface area controlled, the array with the larger numerosity consistently had slightly smaller dots, on average, than the array with the smaller numerosity. Identifying this subtle difference by focusing on the smallest dots could serve as an optimal decision strategy, likely resulting in a smaller sample size with downstream effects on variability and scaling.

Theoretical implications. The current work significantly advances our understanding of how non-symbolic numerosity representations are formed by providing a cognitive mechanism to explain variations in prominent properties of numerical judgment—namely, variability and scaling—while integrating insights from the literature on numerical cognition, judgment and decision-making, attention, and vision. Furthermore, selective sampling accounts not only for known perceptual effects on numerosity judgment, such as surface area, but also reveals new effects, like the proximity-to-center bias and size distributions. This unified framework reflects the distribution of both overt and covert attention to the perceptual properties of the display.

Notably, the SSDM has the potential to account for perfect numerical estimations for very small set sizes (3-4 items) as a result of a full sample and low constant variability, whereas previous work suggested a separate system from the ANS to explain this behavior (Feigenson, Dehaene, & Spelke, 2004). The current work also paves the way for new research avenues to study how

attentional biases impact decisions involving numerical judgments, such as consumer decision-making, where product quantity is a major factor in choice (Vanunu & Donnelly, 2024).

Lastly, selective sampling is a theory for how people collect information from the world and is not limited to explaining numerosity judgments alone. Since the SSDM is derived from the well-established diffusion model, it should be capable of accounting for data as effectively as the diffusion model does when stimuli are constructed from multiple attributes requiring a division of attentional resources. Additionally, it explains how between-trial variability in evidence arises with respect to those attributes, making the SSDM a powerful tool for future analyses with the potential to uncover new effects within familiar experimental manipulations and tasks.

Conclusions. The present work elucidates variations in behavior between common numerosity-discrimination tasks using a single framework—the selective sampling account—suggesting that variability in numerical representations arises from task demands, particularly attentional limitations and spatial resolution requirements due to the arrangement of objects in the display. An intermingled display of blue and yellow dots presumably requires high spatial resolution to distinguish between them, potentially leading to the formation of incomplete representations from the center of the array, using a linear scale that reflects high sensitivity. This process increases variability with set size, consistent with the assumptions of the linear ANS model. Conversely, the demand for high spatial resolution is presumably reduced when the two numerosities are spatially separated in the display. Consequently, reducing the need for high spatial resolution potentially enables the formation of more comprehensive representations using a logarithmic scale that reflects low sensitivity, often resulting in constant variability with set size, consistent with the assumptions of the log ANS model.

References

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review, 113*(2), 409–432.

Brannon, E. M., & Terrace, H. S. (1998). Ordering of the numerosities 1 to 9 by monkeys. *Science, 282*(5389), 746–749.

Cai, Y., Hofstetter, S., Harvey, B. M., & Dumoulin, S. O. (2022). Attention drives human numerosity-selective responses. *Cell Reports, 39*(13), 111005.

Cheyette, S. J., & Piantadosi, S. T. (2019). A primarily serial, foveal accumulator underlies approximate numerical estimation. *Proceedings of the National Academy of Sciences, 116*(36), 17729–17734.

Cheyette, S. J., & Piantadosi, S. T. (2020). A unified account of numerosity perception. *Nature Human Behaviour, 4*(12), 1265–1272.

Dehaene, S. (2001). Précis of the number sense. *Mind & Language, 16*(1), 16–36.

Dehaene, S. (2003). The neural basis of the Weber–Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences, 7*(4), 145–147.

DeWind, N. K., Adams, G. K., Platt, M. L., & Brannon, E. M. (2015). Modeling the approximate number system to quantify the contribution of visual stimulus features. *Cognition, 142*, 247–265.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8*(7), 307–314.

Gebuis, T., Kadosh, R. C., & Gevers, W. (2016). Sensory-integration system rather than approximate number system underlies numerosity processing: A critical review. *Acta Psychologica, 171*, 17–35.

Gebuis, T., & Reynvoet, B. (2012a). The interplay between nonsymbolic number and its continuous visual properties. *Journal of Experimental Psychology: General, 141*(4), 642–653.

Gebuis, T., & Reynvoet, B. (2012b). The role of visual information in numerosity estimation. *PLOS ONE, 7*(5), e37426.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587–606.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*(1), 451–482.

Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature, 455*(7213), 665–668.

Hyde, D. C., Khanum, S., & Spelke, E. S. (2014). Brief non-symbolic, approximate number practice enhances subsequent exact symbolic arithmetic in children. *Cognition, 131*(1), 92–107.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10–12), 1489–1506.

Kang, I., & Ratcliff, R. (2020). Modeling the interaction of numerosity and perceptual variables with the diffusion model. *Cognitive Psychology, 120*, 101288.

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From “sense of number” to “sense of magnitude”: The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences, 40*, E164.

Loschky, L. C., Nuthmann, A., Fortenbaugh, F. C., & Levi, D. M. (2017). Scene perception from central to peripheral vision. *Journal of Vision, 17*(1), 6.

Nieder, A. (2020). The adaptive value of numerical competence. *Trends in Ecology & Evolution, 35*(7), 605–617.

Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *Annual Review of Neuroscience, 32*(1), 185–208.

Park, J., & Brannon, E. M. (2013). Training the approximate number system improves math proficiency. *Psychological Science, 24*(10), 2013–2019.

Park, J., & Brannon, E. M. (2014). Improving arithmetic performance with number sense training: An investigation of underlying mechanism. *Cognition, 133*(1), 188–200.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. New York, NY: Cambridge University Press.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*(1), 3–25.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922.

Ratcliff, R., & McKoon, G. (2018). Modeling numerosity representation with an integrated diffusion model. *Psychological Review, 125*(2), 183–217.

Ratcliff, R., & McKoon, G. (2020). Decision making in numeracy tasks with spatially continuous scales. *Cognitive Psychology, 116*, 101259.

Rodieck, R. W. (1998). *The first steps in seeing.* Sinauer Associates.

Shapiro, A., Lu, Z. L., Huang, C. B., Knight, E., & Ennis, R. (2010). Transitions between central and peripheral vision create spatial/temporal distortions: A hypothesis concerning the perceived break of the curveball. *PLOS ONE, 5*(10), e13296.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14), 4–4.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*(5), 643–659.

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

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review, 79*(4), 281–299.

Vanunu, Y., & Donnelly, K. (2024). Spatial position affects quantity judgments and product preference.

Vanunu, Y., Hotaling, J. M., & Newell, B. R. (2020). Elucidating the differential impact of extreme outcomes in perceptual and preferential choice. *Cognitive Psychology, 119*, 101274.

Vanunu, Y., Hotaling, J. M., Le Pelley, M. E., & Newell, B. R. (2021). How top-down and bottom-up attention modulate risky choice. *Proceedings of the National Academy of Sciences, 118*(39), e2025646118.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 419–433.

Appendix A – Examples of the Diffusion Process and Probability Sampling Functions

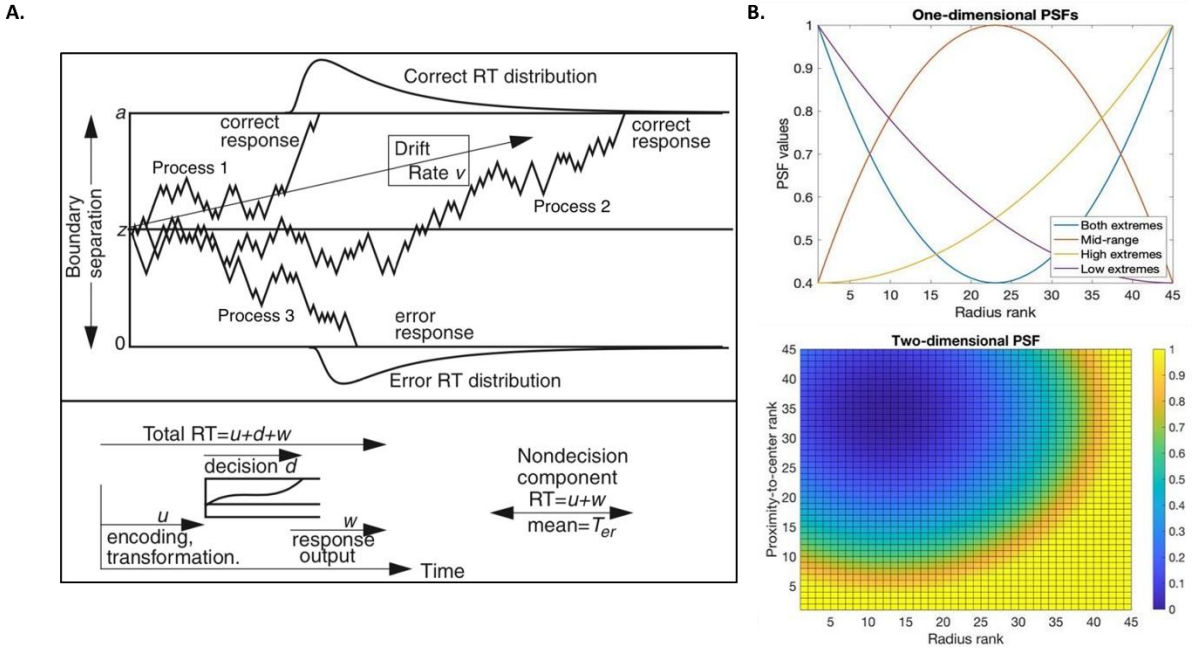


Figure A1. A) Example paths in the diffusion process illustrating a fast correct decision, a slow correct decision, and an error, along with the additional non-decision components of the reaction time (RT). Figures are taken from Ratcliff & McKoon (2018). B) Example Probability Sampling Function (PSF) shapes for the one-dimension PSF (top panel) and for the two-dimensional PSF (bottom panel). The colors in the latter represent the PSF values, where light yellow and dark blue describe high and low probabilities, respectively. The number of ranks depicted in the example PSFs is arbitrary. In practice, the number of ranks was determined by the set size *N*, ranging from the smallest or closest dots at rank '1' to the largest or farthest dots at rank '*N'*. The parameter values that produced the one-dimensional PSF (α, θ and β) are [.6, 0, .4] for the blue line, [-.6, 0, .4] for the red line, [.15, -1, .4] for the yellow line and [.15, 1, .4] for the purple line. The parameter values that produced the two-dimensional PSF (α_C, α_R, θ_C, θ_R and β) are [1, 1, 0, -.5, .5].

Appendix B - Fitting Methods

The diffusion model consists of nine parameters: drift rate (V); boundary separation (a); starting point (z); nondecision time (T_{er}); between-trial SD in drift rate (η), starting point (s_z), and nondecision time (s_t); within-trial variability (s); and a contaminant parameter (p_0). Note that the within-trial variability component (s) was fixed at 0.1 as is standard. The starting point was set to $a/2$ because there was no bias between left and right or blue and yellow responses. The drift rate and its between-trial SD were computed from the representation models—i.e., the ANS or the selective sampling account—and the rest of the parameters were estimated from the data.

The ANS representation model had five free parameters: two drift-rate coefficients for each experimental condition (e.g., v_1 and v_2 for the proportional and equal area conditions, respectively);

a SD scaling factor (σ_1); systematic variability (η_0); and a binary sensitivity parameter (K) that determined the representation scale. The selective sampling account as a representation model had six free parameters in the one-dimension version and eight free parameters in the two-dimension version: α_C, α_R, β, θ_C and θ_R defined the shape of the two-dimensional PSF; a drift-rate coefficient (v); systematic variability (η_0); and a binary sensitivity parameter (K) for the representation scale. Overall, the ANS-diffusion model, the one-dimensional SSDM, and the two-dimensional SSDM, had 10, 11, and 13 free parameters, respectively.

All models were fitted separately to each individual. We used a standard explicit solution of the Ratcliff two-choice diffusion model to fit the data (Ratcliff & Tuerlinckx, 2002). For each experimental condition, we used five quantiles of the RT distributions for both correct and error responses (i.e., 0.1, 0.3, 0.5, 0.7, and 0.9; 6 bins for each distribution). We split the correct and error RT distributions between the experimental conditions (e.g., twenty conditions in the reanalysis based on set size, set-size differences, and dot area). We used the 'cdfdif' function in the DMAT toolbox (Vandekerckhove & Tuerlinckx, 2008) to compute the predicted cumulative probability of a response occurring for each RT quantile in each experimental condition. To compute the distribution of plausible subsamples within each condition, for each trial within that condition, the SSDM simulated sampling ten thousand times. This was based on the PSF values, which were assigned according to the size and position ranking of each dot in the display for the respective trial.

The free parameters of the models were optimized using a simplex minimization routine, which searched for a set of parameters that minimized the differences between the observed and the expected frequencies represented by a as G^2 values in Eq. (A1):

Eq. (A1): $$G^2 = 2 \times \sum O \times log\ (O \backslash E)$$

where the observed frequencies (O) were calculated by multiplying the total number of observations by the proportions of responses between the data quantiles for each condition. The expected frequencies (E) were calculated by multiplying the total number of observations with the proportion of responses in the predicted RT distribution that lay between the data quantiles. Contributions were computed separately for correct and error responses in each condition, and these were summed to an overall G^2 value, which the simplex routine aimed to minimize.

Lastly, to assess the generalizability of the models with varying levels of complexity (i.e., the number of free parameters) and to account for the possibility that more complex model may overfit the data, we performed a cross-validation analysis. This process entailed initially training the models on 50% of the data, which was randomly selected from each block. Subsequently, we applied the best-fitting parameters derived from this training phase to evaluate the models' accuracy in predicting the outcomes of the remaining trials (i.e., the testing dataset). A complex model that overfit the training data is expected to produce a lower G^2 value on the testing data compared to a model that does not overfit the training data (Browne, 2000; Busemeyer & Wang, 2000). Notably, the conclusion drawn from the cross-validation analysis does not change when model comparisons are performed using AIC score (see Appendix C for a full report).

Appendix C – Model Comparisons using AIC Scores.

Table A1. Results from the model comparisons using AIC scores replicate the findings from the cross-validation analysis across data sets.

Model	Reanalysis		Exp. 1	Exp. 2	Experiment 3	
	B/Y	L/R			Inter.	Sepa.

1-center 2D-SSIM	545.2	577.4	242.1	184.9	183.2	174.6
2-centers 2D-SSIM	N/A	N/A	N/A	182.5	198.1	176.1
Simple ANS-diffusion	532.2	594.0	321.9	196.7	209.3	177.3
Complex ANS-diffusion	N/A	N/A	250.9	258.4	183.5	180.8

Note: 1-center and 2-centers represent proximity-to-center rank and proximity-to-two-centers rank, respectively. Simple and complex ANS-diffusion refer to the model variants estimating two drift-rate coefficients for each perceptual condition and the variant including an interaction term, respectively. Inter. and Sepa. represent the intermingled and separate format groups.

Appendix D - individual level PSF display

There are two ways to display the resulting PSFs across participants. For a group-level display, we calculate the average PSF values across participants in each condition. For an individual-level display, we calculate the proportion of times each rank had the maximum probability across PSFs in each condition. However, a group-level PSF display can sometimes be misleading as it may obscure meaningful individual differences. For example, the surface in the right panel of Figure A1B illustrates a sampling function that prioritizes the largest dots and dots at the center in a two-dimensional space. If this function is combined with an opposite PSF that prioritizes the smallest dots and dots at the periphery, it would produce a group-level PSF that appears to balance sampling across ranks, while in reality, both high- and low-ranks are prioritized over middle ranks. Therefore, it is crucial to examine the PSF at the individual level in such cases.

Figure A2 shows the individual-level PSFs across tasks and experiments. The panels on the left represent sampling policies from numerosity-discrimination tasks with an intermingled display of blue and yellow dots, while the panels on the right represent sampling policies from tasks with two arrays of dots that are spatially separated on the screen. Findings from the

intermingled displays were consistent across datasets, showing a preference for sampling larger dots over smaller dots and dots at the center over those at the periphery, as indicated by the large mass of yellow bars in the top corner of each panel on the left. In the separate format, the largest dots were prioritized across datasets. However, a smaller portion of participants prioritized the smallest dots in the display, as indicated by the smaller mass of blue bars at the bottom of each panel on the right.

A sampling policy that prioritizes small dots over large dots in sampling may reflect a top-down strategy to compensate for an 'automatic' bottom-up response to the largest (i.e., most salient) dots in the display (Vanunu et al., 2021). Such a strategy should be optimal when the difference in surface area between the arrays is controlled, as the dots representing larger numerosity are, on average, slightly smaller. This strategy was more common under the separate format, possibly due to the reduced complexity of the separate display format, which allowed participants to notice the difference in the average size of dots on each side.

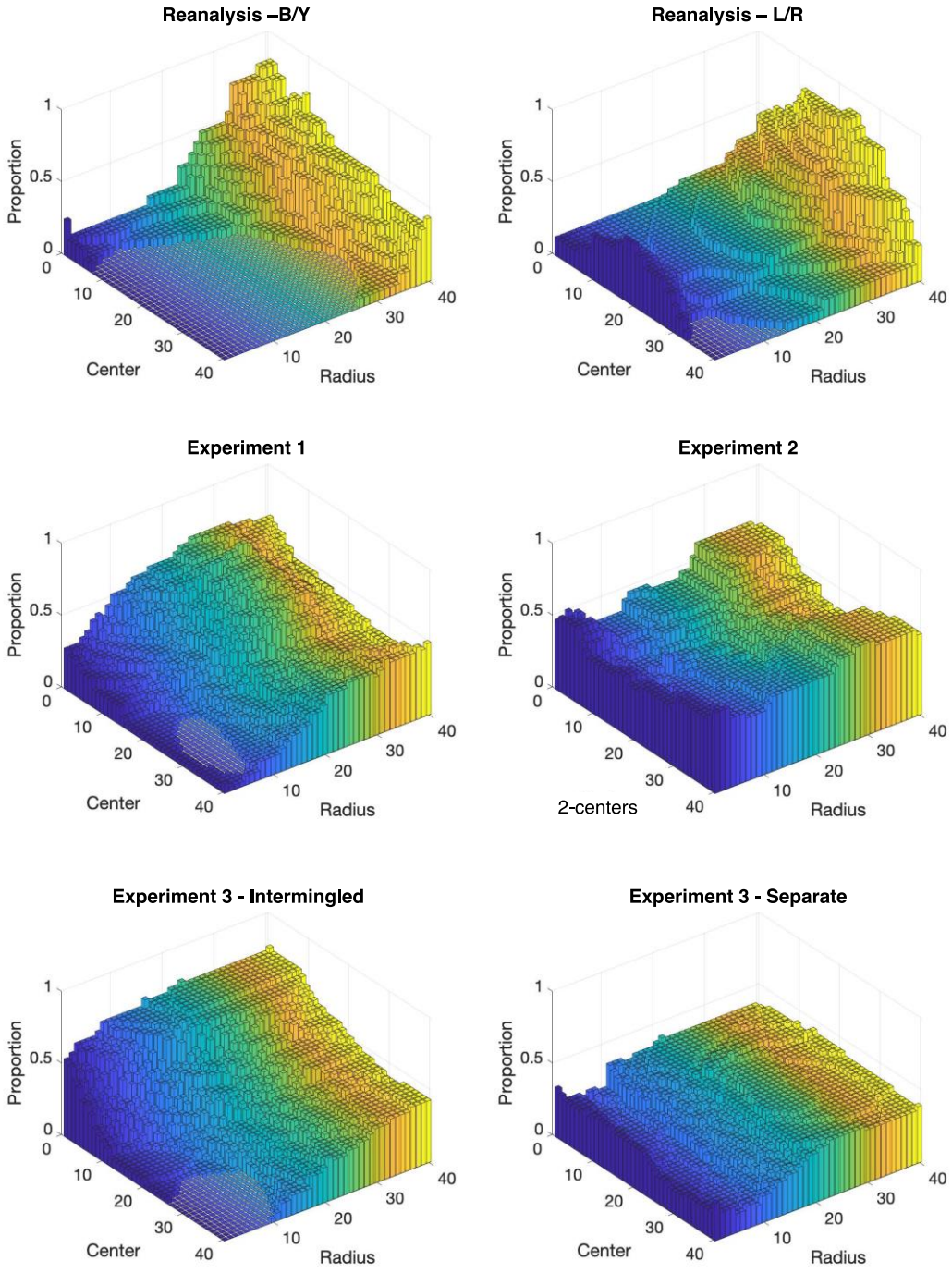
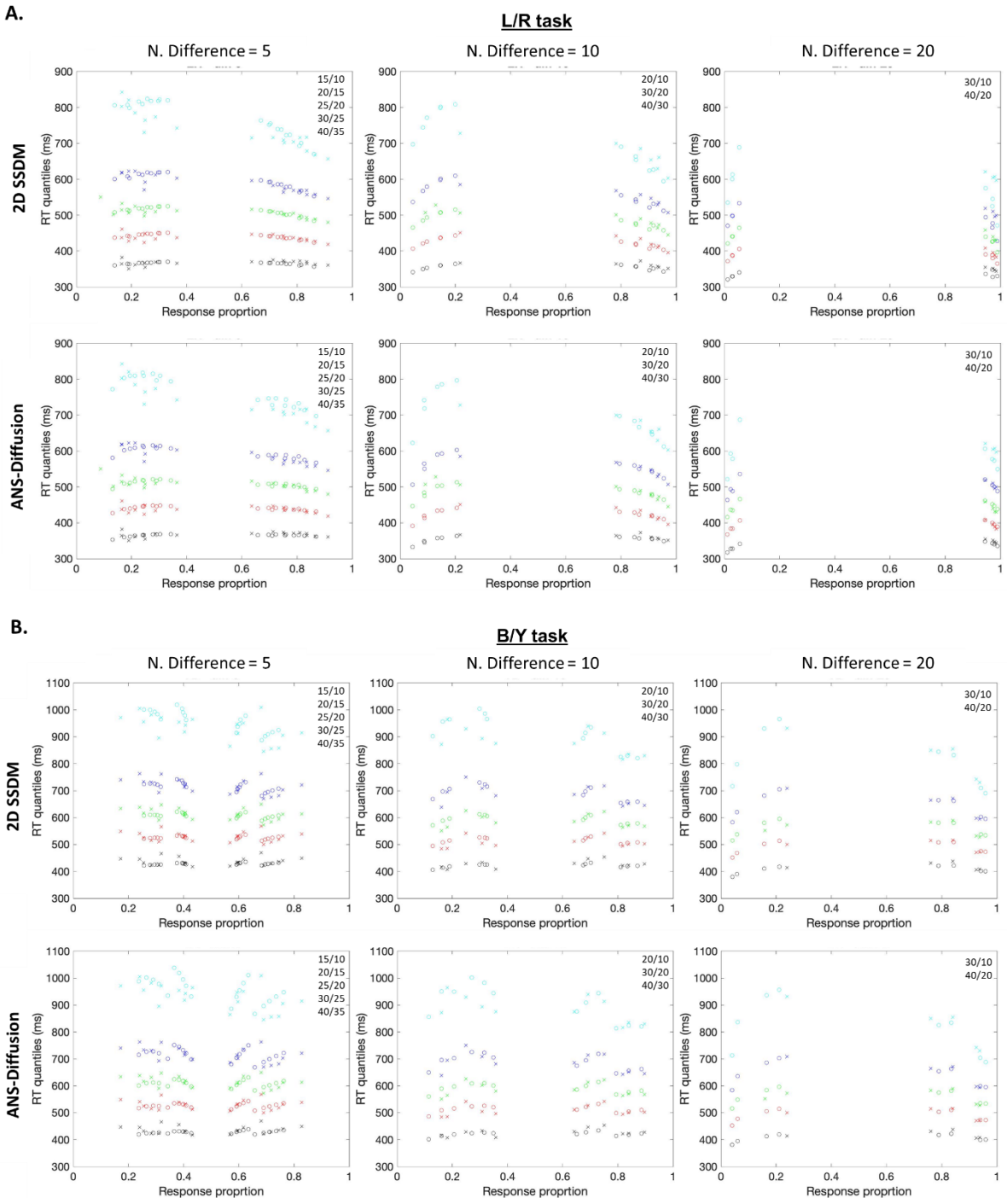


Figure A2. The individual-level PSFs between experiments and conditions groups.

Appendix E – Models Predictions versus Data

We use *Quantile-Probability Functions* (QPFs) to compare model predictions against data, representing the relationship between response time distribution and choice probabilities within a single figure. QPFs map five response time quantiles—.1, .3, .5, .7, and .9—for each experimental condition, as estimated in model fitting (see Appendix B), alongside the corresponding choice probabilities for correct and error responses. By mapping RT quantiles to the probability of each decision type, QPFs illustrate how decision speed varies with accuracy, highlighting speed-accuracy tradeoffs (Ratcliff & Rouder, 1998; Starns et al., 2012; Vanunu & Ratcliff, 2023).

Each QPF figure presents the proportion of correct and error responses on the x-axis among the set size conditions. The proportions of correct response are ordered from the smallest set size on the right (e.g., 10/15) to the largest set size in the middle (e.g., 35/40), while error proportions are ordered from the smallest set size on the left to the largest set size in the middle, creating a mirrored x-axis representation. The y-axis displays the five RT quantiles, ordered from the lowest at the bottom (black markers) to the highest at the top (teal markers). The 'x' markers indicate behavioral data, and the 'o' markers represent model fits. When fewer than five observations exist for some participants, the median (green x) is plotted, and if participants have no errors, no value is plotted for that condition. The QPFs from the reanalysis and the empirical studies are depicted in Figures A3A to A3E.


A.
L/R task
N. Difference = 5
N. Difference = 10
N. Difference = 20
2D SSDM
ANS-Diffusion
RT quantiles (ms)
Response proprtion
15/10
20/15
25/20
30/25
40/35
20/10
30/20
40/30
30/10
40/20
B.
B/Y task

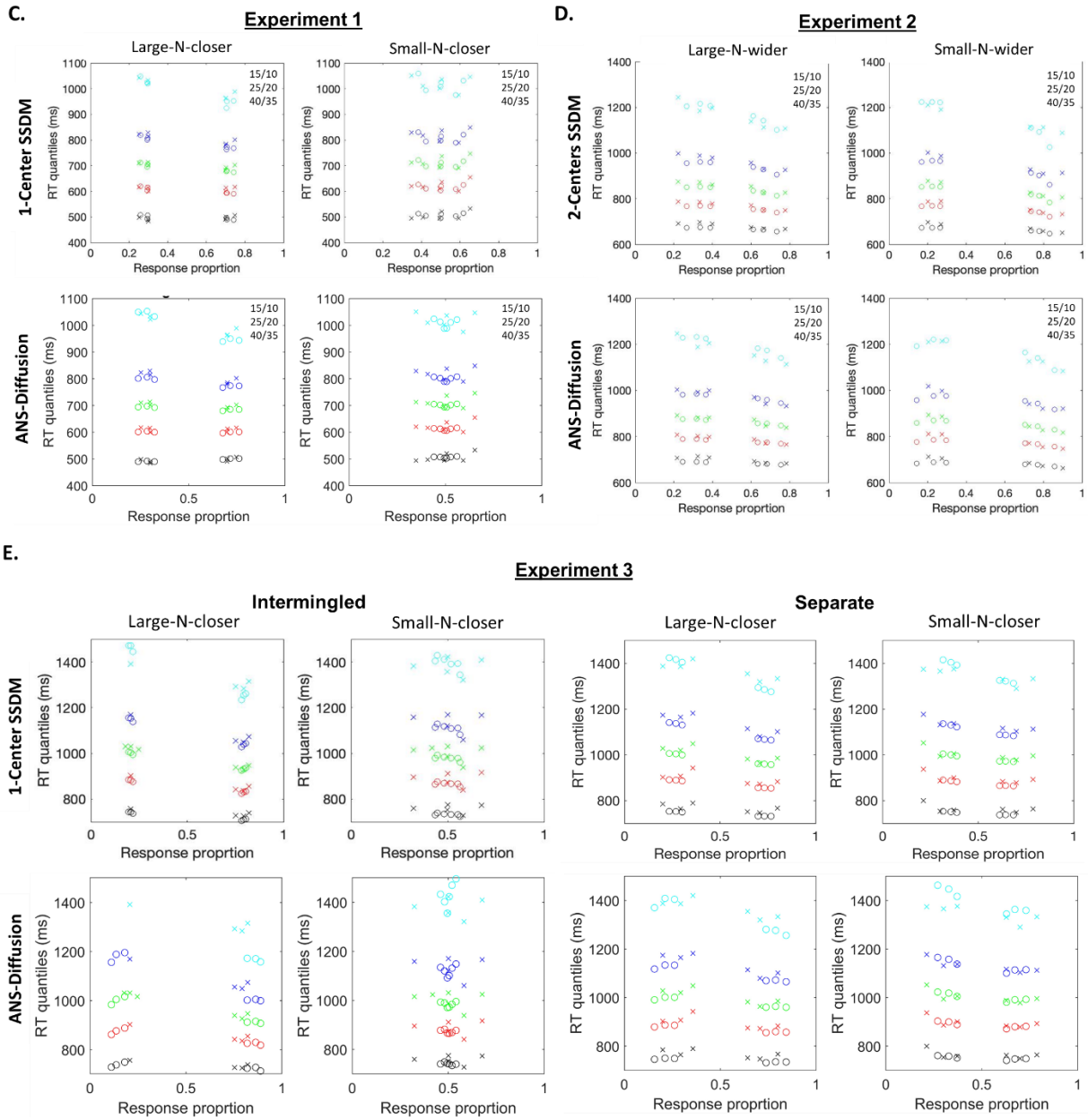


Figure A3. The Quantile-Probability Functions (QPFs) from the reanalysis and follow-up studies for the two-dimensional SSDM and ANS-diffusion model. Panels A and B depict the QPFs from the reanalysis, divided by tasks (L/R task in panel A and B/Y task in panel B) and by numerosity difference conditions. Panels C through E show the QPFs from Experiments 1, 2, and 3, respectively, divided by the perceptual conditions. Across all figures, x-axis values represent response proportions ordered from right to left by set-size conditions, as shown in the top-right

corner of each figure. The quantile response times, displayed from bottom to top, are the .1, .3, .5, .7, and .9 quantiles. "x" symbols represent behavioral data, and "o" symbols represent the models' theoretical fits. The QPFs for the reanalysis and Experiments 1 and 3 reflect model predictions from the 1-center SSDM, while those for Experiment 2 show predictions from the 2-center SSDM.

Appendix F – Simulated Predictions for Experiment 1 and 2

To illustrate our predictions for behavior in Experiments 1 and 2, we conducted a simulation using the modeling results from the reanalysis to inform behavior in these experiments. Specifically, we simulated the numerical differences between the larger numerosity (N_L) and the smaller numerosity (N_S) in each condition, based on the group-level PSFs estimated for the B/Y task (Experiment 1) and the L/R task (Experiment 2) in the reanalysis section. This process was repeated four times, with incremental increases of .25 in the PSF values for the B/Y task and .15 for the L/R task to simulate changes in evidence as the subsample becomes more exhaustive. For Experiment 1, we simulated evidence using the linear scale and for Experiment 2, we used both scales to predict outcomes based on the success of the distractor manipulation. The PSFs and simulated evidence are presented in Figure A4.

Across data sets, differences between the conditions attenuated as the subsample became more exhaustive, illustrating the impact of sample size on numerosity judgment. For Experiment 1 (Figure A4A), we found that evidence toward the larger numerosity ($N_L - N_S > 0$) increased with set size when the dots from the larger numerosity were, on average, closer to the center of the display (large-N-closer condition), while the opposite trend occurred when the dots from the smaller numerosity were closer (small-N-closer condition)—predicting a systematic error under the latter condition if numerosity judgments are based on a subsample from the center.

For Experiment 2 (Figure A4B), using a logarithmic scale, we found that evidence for the larger numerosity decreased with set size across both radius distribution conditions. However, evidence was greater when the smaller numerosity, rather than the larger, had a wider radius distribution, due to the under-sampling of smaller dots. If participants used a linear scale because of the distractors, the simulation predicts that evidence for the larger numerosity would decrease with set size in the large-N-wider condition but increase in the small-N-wider condition.

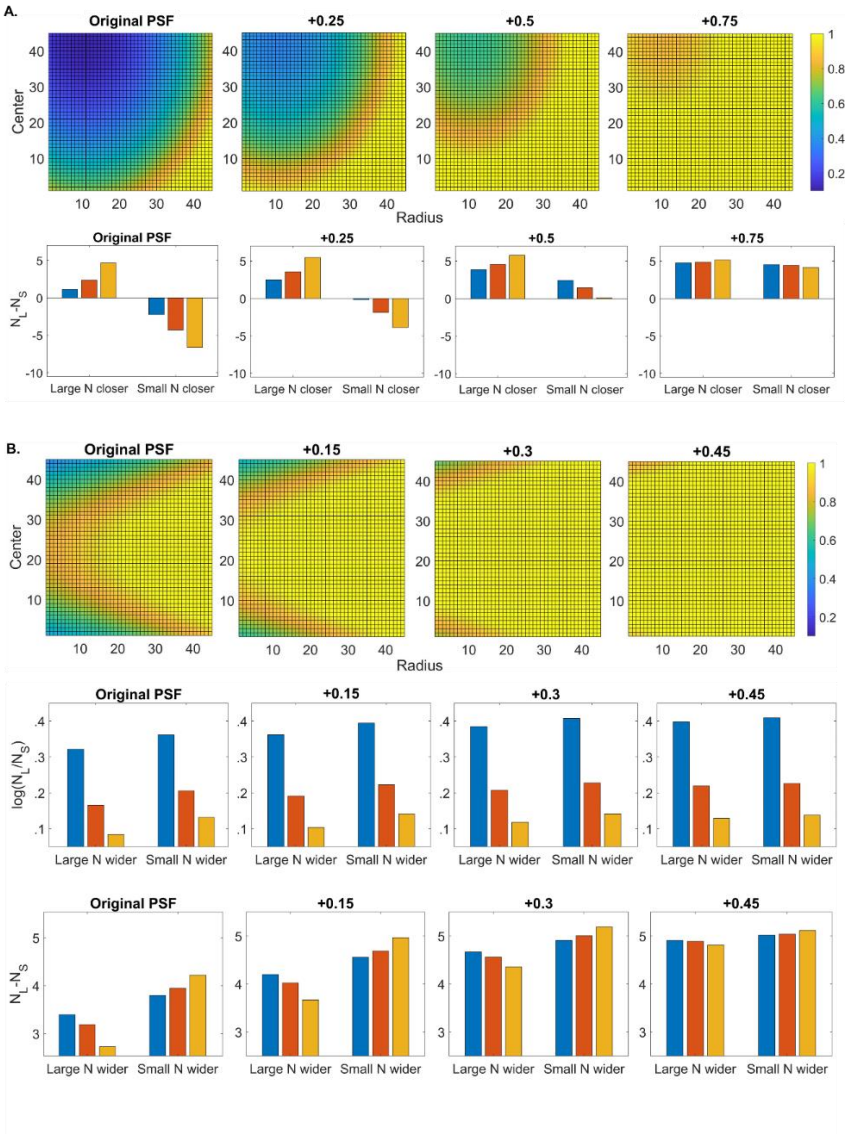


Figure A4. The group-level PSFs estimated in the reanalysis from the B/Y task (panel A) and the L/R task (panel B), with an incremental increase in PSF values, and the simulated evidence (e.g., N_L - N_S) using the group-level PSFs on data from Experiments 1 (panel A) and 2 (panel B).

Appendix G – Variants of the ANS-Diffusion Model to Account for Interaction Effects

To account for perceptual effects on numerosity judgment, Ratcliff and McKoon (2018) used a regression model to estimate drift rate, estimating separate coefficients for each surface area condition (proportional, equal; see Eq. 1). However, this model does not consider the possibility of an interaction effect between the perceptual properties of the display and set size. In more recent work, Kang and Ratcliff (2020) developed a variant of the ANS-diffusion model that includes multiple contributions to the drift rate based on set size conditions and their interaction with the perceptual properties. Following this approach, we developed a similar variant to account for the interaction effect of set size with centrality (Exp. 1 and 3) or set size with radius distribution (Exp. 2), as shown in Eq. A1:

Eq. (A1): $$V = v_1 \times \left[K \times (N_L - N_S) + (1 - K) \times log\left(\frac{N_B}{N_Y}\right)\right] + C \times v_2 \times (N_L + N_S)$$

Here, v_1 is a drift-rate coefficient for the set-size difference, and K is a binary sensitivity parameter that takes values of 1 or 0 for linear and logarithmic scaling, respectively, across trials. v_2 is a second drift-rate coefficient for the interaction effect between set size and the perceptual property (centrality or radius distribution). C is a dummy parameter that takes values of 1 or -1 for the large-N-closer and small-N-closer centrality conditions in Experiments 1 and 3, or for the large-N-wider and small-N-wider radius distribution conditions in Experiment 2. Multiplying a positive or negative value of C with v_2 and the sum of dots should reflect an increasing or decreasing drift rate with set size, respectively.

Results from the model comparisons across data set are presented in Table A2. Across data sets, we found inconsistencies when comparing the fit of the simple and complex ANS-diffusion model variants: the simple variant better accounted for data from a separated display format, while the complex model fit data from an intermingled display format more effectively. We chose to report the simpler variant in the main text to maintain consistency with the modeling work we follow (Ratcliff & McKoon, 2018).

Table A2. The simple and complex ANS-diffusion models' fit in mean G^2 values to the complete data (CD), the proportion of participants that were best described by the linear scale, and mean G^2 values in a cross-validation analysis (CV) between data sets. The 'simple' ANS-diffusion model represents the model used in Ratcliff & McKoon (2018) and the reanalysis, while the 'complex' ANS-diffusion model represents the model variant that includes an interaction term.

Data		Model	G^2_{CD}	p(linear)	G^2_{CV}
Experiment 1		Simple	150.93	.66	118.69
		Complex	115.47	.13	100.29
Experiment 2		Simple	88.34	.40	88.68
		Complex	119.18	.32	101.91
Exp. 3	Intermingled	Simple	94.67	.78	95.95
		Complex	81.77	.19	82.47
	Separate	Simple	78.67	.40	79.00
		Complex	80.40	.40	79.30

Note: 1-center and 2-centers represent the proximity-to-center rank and proximity-to-two-centers rank, respectively.

Appendix H – The Proportion of Gaze Duration in Experiment 3

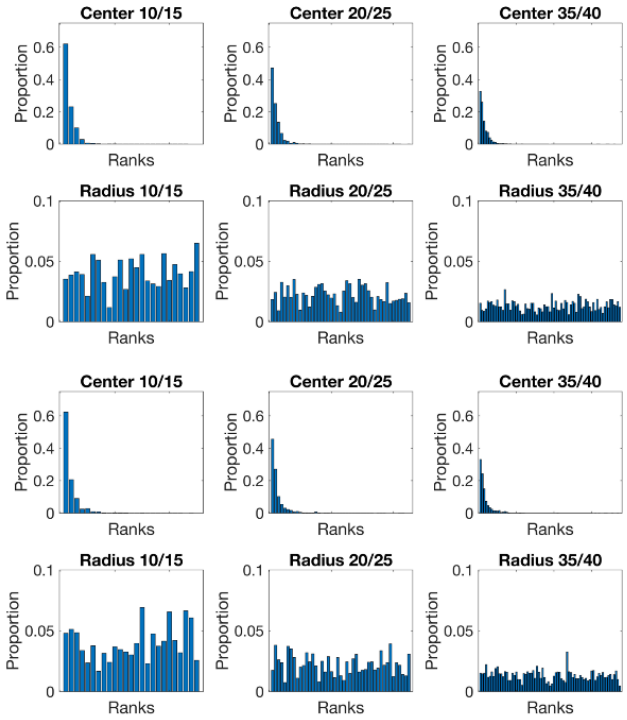


Figure A5. The proportion of gaze duration by proximity-to-center, radius ranks, and display format groups shows longer gaze durations for dots closer to the center of the screen, with no differences among radius ranks across groups.

References

Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, *44*(1), 108-132.

Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*(1), 171-189.

Kang, I., & Ratcliff, R. (2020). Modeling the interaction of numerosity and perceptual variables with the diffusion model. *Cognitive Psychology*, *120*, 101288.

Ratcliff, R., & McKoon, G. (2018). Modeling numerosity representation with an integrated diffusion model. *Psychological review*, *125*(2), 183–217.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological science*, *9*(5), 347-356.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic bulletin & review*, *9*(3), 438-481.

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive psychology*, *64*(1-2), 1-34.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior research methods*, *40*(1), 61-72.

Vanunu, Y., & Ratcliff, R. (2023). The effect of speed-stress on driving behavior: A diffusion model analysis. *Psychonomic bulletin & review*, *30*(3), 1148-1157