Investigating Perceptual Biases in Icon Arrays

Cindy Xiong
University of Massachusetts Amherst
Amherst, Massachusetts, USA
cindy.xiong@cs.umass.edu

Ali Sarvghad
University of Massachusetts Amherst
Amherst, Massachusetts, USA
asarv@cs.umass.edu

Çağatay Demiralp
Sigma Computing
San Francisco, California, USA
cagatay@sigmacomputing.com

Jake M. Hofman
Microsoft Research
New York, New York, USA
jmh@microsoft.com

Daniel G. Goldstein
Microsoft Research
New York, New York, USA
dgg@microsoft.com

ABSTRACT

Icon arrays are graphical displays in which a subset of identical shapes are filled to convey probabilities. They are widely used for communicating probabilities to the general public. A primary design decision concerning icon arrays is how to fill and arrange these shapes. For example, a designer could fill the shapes from top to bottom or in a random fashion. We investigated the effect of different arrangements in icon arrays on probability perception. We showed participants icon arrays depicting probabilities between 0% and 100% in six different arrangements. Participants were more accurate in estimating probabilities when viewing the top, row, and diagonal arrangements, but they overestimated the proportions with the central arrangement and underestimated the proportions with the edge arrangement. They were biased to either overestimate or underestimate when viewing the random arrangement depending on the objective proportions, following a cyclical pattern consistent with existing findings in the psychophysics literature.

CCS CONCEPTS

• **Human-centered computing → Visualization techniques**.

KEYWORDS

Icon Arrays, Proportions, Perception, Bias, Communication, Crowdsourcing, Risk Visualization, Visualization of Probabilities

ACM Reference Format:
Cindy Xiong, Ali Sarvghad, Çağatay Demiralp, Jake M. Hofman, and Daniel G. Goldstein. 2022. Investigating Perceptual Biases in Icon Arrays. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3491102.3501874

1 INTRODUCTION

Imagine you had just extracted a key piece of probabilistic information from your data, perhaps that the rate of patients getting an allergic reaction to a treatment is 25%. Now it's time to share it with a group of stakeholders. How would you effectively communicate this probability to a general audience? Past work has demonstrated that icon arrays - also known as pictographs or information grids - are a simple and effective medium for communicating risks and probabilities to laypeople [18, 24, 33, 56, 67]. Icon arrays are composed of juxtaposed icons representing members in a group (e.g., men above 60 years old), and designers typically communicate information by changing the shape, color, or other visual properties of a subset of icons. For example, to express a 25% risk of getting an allergic reaction after a treatment, you might design a 10 by 10 grid of 100 circles with 25 of them emphasized in black.

Icon arrays have been shown to help people understand probabilistic information by leveraging the human perceptual system [47], especially those with lower numeracy skills [23, 31, 33]. Additionally, icon arrays present probabilistic information using natural frequencies (e.g., 10 out of 100) instead of using proportions or percentages (e.g., 10%). This design choice of replacing proportions or percentages with counts of X out of Y has been shown to more often elicit optimal statistical inferences and increase numerical comprehension [20, 34, 38]. Furthermore, the one-to-one mapping of the entity-to-discrete-icon induces a sense of recognition and self-identification [41], which tends to increase engagement with data [39]. For instance, when a patient reads an icon array conveying information regarding health risks, they can better relate to the information by projecting themselves as one of the icons [41].

Despite the strengths of icon array visualizations, designers still have many decisions to make before they can produce an optimal icon array design. For example, the iconicity level, which ranges from being abstract (e.g., a rectangular block) to concrete (e.g., a stick-figure or a picture of a human), can affect risk perception [63]. Lower iconicity is associated with lower risk perception, and higher iconicity with higher risk perception. However, while prior research provides useful insights, it remains unclear how different design choices can impact probability perception in icon arrays. In particular, we do not know if the way filled icons are spatially arranged in an icon array can cause systematic perceptual biases in the estimation of frequencies and proportions.

Stylistic choices for how icon arrays are filled might impact the viewer's perception of probability by changing the perceptual proxies the viewers use to extract numerical values from the icon arrays [36, 65]. Most existing work that investigates the effectiveness of icon arrays focuses on comparing icon arrays in the top-to-bottom arrangement or a random arrangement (see Top and Random in

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501874

Figure 2) to other representations, such as numbers in [23] and [3], without examining whether and how particular arrangement of icons in an array affects viewer perception [19, 57].

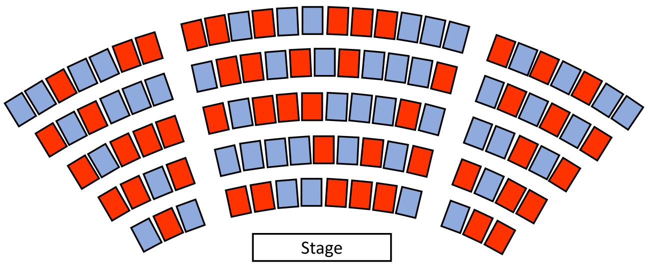

Figure 1: Risk theatre showing polling probabilities as seats in a theatre. Fifty seats are randomly colored blue and fifty seats are randomly colored red. Sitting in a blue seat represents the outcome that the Democratic candidate will win, while sitting in a red seat represents the outcome that the Republican candidate will win.

In the real-world, many designers tend to fill the grids in an ordered fashion, such as filling from the top or bottom (e.g., [40, 42]). More recently, some data journalists have been using icon arrays in the random arrangement, usually referred to as a 'risk theatre', to communicate projected U.S. election outcomes in order to help people make a more informed decision about whether they should go vote or not [29]. Some existing discussions have suggested that it helps people better grasp the concept of uncertainty and form a more accurate mental representation of the depicted probability [46, 50]. The risk theatre shows polling probabilities as seats in a theatre, as shown in Figure 1. These seats are randomly colored in blue or red, in which a blue seat represents that the Democratic candidate would win the election and a red seat represents that the Republican candidate would win. The proportion of blue and red seats corresponds to the likelihood of the Democratic and Republican candidate winning the election. The viewer is typically asked to imagine buying a ticket to the theatre. As they find their seat, sitting in a blue seat represents the outcome that the Democratic candidate has won, while sitting in a red seat represents the outcome that the Republican candidate has won. This technique helps the viewer "experience" the election outcome with a given probability, and may change perceptions of seemingly small probabilities (e.g., a 15% chance of a candidate winning may seem less negligible when presented as a risk theatre than numerically).

Our goal in this work is to systematically investigate if certain spatial arrangements of filled icons can cause systematic over or underestimation of probabilistic information. It would be undesirable, for instance, if a key probability a designer tried to convey gets overestimated or underestimated because the arrangement choice had a biasing effect on the viewer's perception. Understanding (and countering) such potential biases will allow us to make recommendations to designers on how to arrange filled grids in their icon arrays so their audiences can most accurately and confidently extract the proportion value. This is especially crucial in circumstances where objective information can not be easily obtained or where decisions have to be made rather quickly. For instance, a patient may need to promptly consent to a medical procedure by evaluating the associated risks presented as an icon array. In such a case, it is imperative to design and present information in a manner that minimizes the chance of misestimating risks and benefits.

Contributions: We contribute three empirical studies that investigate the effect of arrangement in icon arrays on a viewer's probability perception, creating a model that predicts perceptual bias in icon arrays across various levels of probability values. We found people to be very accurate in estimating proportions when viewing the top, row, and diagonal arrangements; in contrast, they overestimated the depicted probabilities with the central arrangement and underestimated them with the edge arrangement. People are also biased to either overestimate or underestimate when viewing the random arrangement depending on the objective proportions, following a cyclical pattern consistent with existing findings in the psychology literature [35, 58]. At the end of the paper, we propose several design guidelines to help icon array designers arrange the filled grids to most effectively communicate a probability value to their readers.

2 RELATED WORK

Icon arrays are often used in the medical domain for communicating the risk associated with medical procedures and treatments to patients. Several existing studies investigated the utility and effectiveness of icon arrays for communicating risk in comparison to other visual and non-visual formats (e.g., [8, 31, 49, 59, 61]). Icon arrays significantly increase patients' gist (general impression) and verbatim (specific numerical) understanding of the information compared to textual narratives, such as the risk of nausea is "low" [31, 59]. Empirical evidence has also shown that icon arrays outperform numerical representations of risk, such as the risk of nausea is "10%" or "10 out of 100" [8, 49, 61], in particular for people with low numeracy skills.

Prior studies have compared and contrasted the utility of icons arrays for communication of risk with basic statistical graphs, including line, bar, and pie charts [19]. One of the primary outcomes of these studies is the different affordances of graphical formats for risk communication. Line charts are the best for illustrating risk trends, such as survival or mortality over time [52]. Similarly, bar charts are suitable for comparing risks between different options [64], and icon arrays are useful for helping people understand risk in a general and precise way [28, 62, 64, 66]. Additionally, icon arrays have also been shown to improve users' trust in information compared to other visual and non-visual forms of risk communication [31].

2.1 Number and Area Perception

When a viewer reads an icon array, how do they extract a percentage value from it? The viewer could visually 'count' the number of filled and unfilled icons and then mentally compute a ratio , or use surrogate features such as the overall area or spatial frequency of the filled icons among unfilled ones to make a rough estimate. The answer to this question has been debated in the world of human perception and cognition for many years [10, 43]. It is a challenging question because perceiving numerosity (e.g., the number of items in a set) naturally correlates with perception of features like area and spatial frequency [11, 25], so it is nearly impossible to tease these two processes apart [43]. In icon arrays, for example, as the number of filled icons increase, the amount of dark area increases, and the spatial frequency of filled icons increases.

To resolve this debate, recent efforts have tried modeling numerosity perception tasks with deep neural networks [60]. They demonstrate that both numerosity perception and perception of continuous magnitudes, such as density or area, play an important part in these tasks, and in our process of making sense of our visual environment in general. However, other studies have shown that relying on features like area and spatial frequency rather than numerosity is more automatic and less effortful when making number estimations [43]. This is partly because our visual system is good at quickly generating summary information about a collection of objects [13, 48]. We can quickly retrieve the mean size or average density of a collection of objects and use this information to approximate the number of objects in an area [5, 13, 45]. For example, when asked to determine the number of circles in an area, our visual system can quickly sample the average size of all the circles. We can then calculate the approximate number of those circles that would fit in the area to figure out the number of total circles. We can also use the spatial frequency of the circles to make the same approximation. The more circles are densely packed in an area, the higher the total number of circles [10].

Despite the debate over the exact perceptual mechanisms at play during number estimation, there is ample evidence suggesting that at least one mechanism in number estimation involves the viewer visually segmenting a display into sub-groups [16, 21]. Objects sharing similar perceptual features (e.g. color or shape) tend to be chunked together, and objects that are distinct in features tend to be separated from others [30]. As a result, people's number estimation can be strongly influenced by whether similar objects in the collection can be easily perceived as a group or not, and how many perceivable groups there are in a collection of objects [2, 21]. For example, a number estimation becomes lower when nearby objects are viewed holistically as a unit [22, 27]. In the "solitaire illusion" from [22], although there are the same number of black and white circles, because the nearby black circles are viewed holistically as a unit, the area occupied by black circles appears bigger. As people rely on area as a surrogate feature to approximate the number of black circles, they mistakenly think there are more black circles than white circles.

In icon arrays, the way grids are filled and arranged can trigger different visual groupings. For example, as shown in Figure 2, in the first row, filling the array from top to bottom in the top arrangement segments the icon array into two groups by color (black on top, and white on the bottom). On the other hand, filling the array with the row arrangement can create up to ten groups (5 groups of black and 5 groups or white). This suggests that different arrangements can change the number of visual groups a viewer sees, and thus the overall proportion estimations. If a viewer relies on surrogate features like visual groups, area, or density to estimate the number of filled grids in an icon array, they can also become susceptible to visual inaccuracies or perceptual bias. For example, people are not particularly accurate at comparing the ratio of two areas [14]. They also tend to think objects that are regularly spaced are more numerous than objects that are randomly spaced [26]. In addition, as the number of quantities to perceive increases, the more imprecise our numerosity perception becomes [15]. Because the number of perceivable groups in an icon array depends on the objective probability it depicts (e.g., in the row arrangement shown in Figure 2, showing 30% has 3 clusters of black but showing 10% will only have 1), we hypothesize that arrangement will interact with objective probability in icon arrays to influence probability perception.

3 OVERVIEW OF EXPERIMENTS

In Experiment 1, we identified six icon array arrangements and examined how these arrangements affect the way people perceive proportion values. We sampled a subset of probability values to test (e.g., 5%, 10%, etc.) and observed that, in general, participants were more accurate in estimating proportions when viewing the top, row, and diagonal arrangements, but tend to misestimate the proportions when viewing the central, edge, and random arrangements. In Experiment 2, we expanded the probability values we tested to cover all the whole number values between 0% and 100% to more systematically investigate the perceptual bias in the *central and edge arrangements*. We reproduced the overestimation and underestimation biases we observed in Experiment 1 and extended our results to cover several variations of the central and edge arrangement designs. In Experiment 3, we expanded the probability values tested for the *random arrangement* and reproduced the misestimation observed in Experiment 1. We also tested the robustness of the effect with four aesthetic styles of icon arrays. We summarize our research questions for each experiment below:

Experiment 1: How does icon array arrangement (i.e., the six listed in Figure 2), affect people's perception of proportion values? Which arrangements warrant the most and least accurate proportion estimates?

Experiment 2: What is the effect of centrality on people's perception of proportion values? Specifically, how are people biased when viewing icon arrays in the central and edge arrangement?

Experiment 3: How do people perceive randomly arranged icon arrays? Do they use particular values as reference points in their estimation? Does this bias generalize to other visual designs of icon arrays?

4 EXPERIMENT 1 ARRANGEMENT EFFECTS

We begin with an investigation on how to visually arrange the icons in an array to elicit a more accurate perception of proportions, uncovering the potential perceptual biases associated with reading icon arrays.

4.1 Design Motivation

Most icon arrays that researchers have studied tend to fill the grids from the top or bottom (e.g., [3, 23]), as shown in the top row of Figure 2. A few explored a random arrangement (e.g., [19, 57]), as shown in the bottom row of Figure 2. However, there are many other ways to fill grids in an icon array. [4] has additionally tested icons with a central and edge arrangement, as shown in the second and third row of Table 2, but only with one condition where the icons depicted a 45% probability. To the best of our knowledge, no existing work examines how people perceive probabilities depicted

Arrangement	Description
Top	Icons are filled starting from the top-left corner with no gaps between filled icons until the target proportion of black/white icons is reached.
Edge	Icons in the outer-most layer of the arrays are filled first, then the second layer, and so on until the target proportion of black/white icons is reached.
Central	This arrangement is the opposite of the edge arrangement, where priority is given to filling inner-most layer of icons and the filling spirals outwards.
Diagonal	Priority is given to fill icons diagonally, starting from the top-left corner until the target proportion of black/white icons is reached.
Row	Priority is given to filling alternate rows, starting from the top row until the target proportion is reached. Once every other row is filled, the previously skipped rows get filled.
Random	Icons are filled randomly until the target proportion of black/white icons is reached.

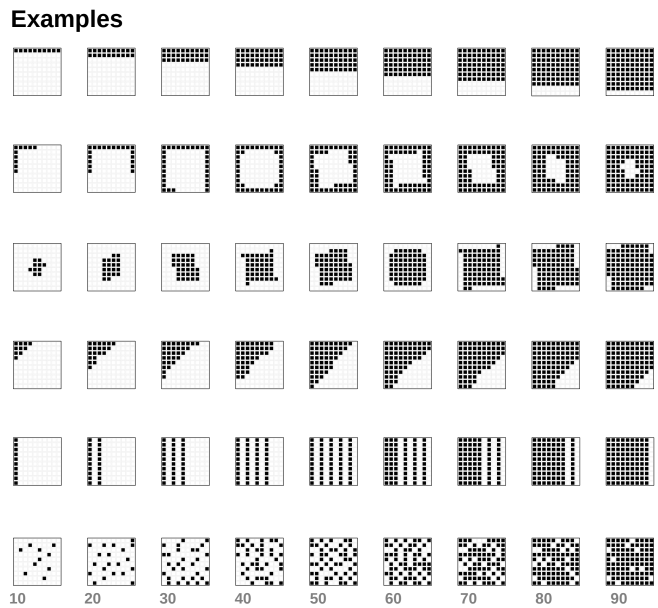

Figure 2: Six icon array arrangements used in Experiment 1, representing proportions from 10% to 90%.

by these arrangements across a wide range of objective probabilities. Additionally, there exist many other ways to arrange filled grids in icon arrays. To explore additional possibilities in this space, we conducted a pilot survey with students enrolled in an introductory data visualization class at the University of Massachusetts Amherst. We provided them with 6x8 arrays of white grids and asked them to fill in the grids to represent a proportion. The authoring team extracted key features from the student drawings and reflectively synthesized them, along with arrangements explored in existing studies, into six categories of common arrangements, as shown in Figure 2. These configurations will serve as a starting point for our investigation on the effect of icon array arrangement on perceived proportions. See our supplementary materials for the students' response.

4.2 Participants

Forty-nine undergraduate students from Northwestern University participated in the experiment in exchange for course credit in an introductory psychology class. Due to health and safety concerns, participants were sent a link to the experiment via Qualtrics [51] to complete the survey at home. Although this means that the size of the icon arrays participants viewed for the study might have varied due to the different machines on which they completed the study, earlier work (e.g., [32]) has demonstrated the viability of crowdsourcing graphical perception studies online, where participants engaged with the studies using screens with varying sizes, by reproducing results of prior laboratory experiments.

4.3 Experimental Stimuli and Design

We generated icon arrays following the six arrangements from Figure 2. The icon arrays of 10×10 grids are presented on a 720 by 720 pixel white background. Every grid has a grey outline and can be either filled (black) or unfilled (white). The number of filled grids represents proportion values, which range between 0% and 100%.

To keep the experiment length manageable, we pre-generated the icon arrays showing varying proportions for all six arrangements. Specifically, we covered the multiples of ten in the range from 0% to 100%, such as 10%, 20%, 30%, and 40%. We also covered the quartiles (25% and 75%), as well as boundary cases (5% and 95%). Because past work has demonstrated that people's perception tends to be extra susceptible to bias near the 50% mark, we added extra conditions to cover that range (45% and 55%). Additionally, recent work has found that people may be susceptible to rounding biases, preferring to make numerical estimates in perception tasks that end with 5's and 10's [12]. Therefore we also included conditions that ended on other numbers, specifically, 12%, 18%, 32%, 38%, 72%, 78%, and 92%. In total, we generated 132 icon arrays with 6 arrangements × 22 proportions.

4.4 Procedure

Participants started the experiments with a short introduction to the task and several practice trials. Every participant completed 132 trials for the experiment. For each trial, participants were presented with one icon array on the screen for one second, and then they were prompted to estimate the percentage of filled grids in the icon array via an uninitialized slider ranging from 0 to 100. They could change their response by dragging the slider handle and the value selected on the slider was displayed above the slider. The participants were also prompted to report how confident they were in their estimation on a similar slider ranging from 0, which means not at all confident, to 100, which means extremely confident. The time to respond was unlimited, and the participant proceeded to the next trial by clicking a 'next' button after they reported their confidence.

4.5 Results

Following procedures in similar perception-based studies with visualizations, we filtered for quality of responses by excluding the trials where the participant made an estimation with an error that is more than two standard deviations away from the mean estimation error. This left us with 94.88% of the responses. Overall, the average estimation error considering all trials was 0.67 (SD = 6.26, out of 100).

4.5.1 Proportion Estimation. Participants were generally accurate in their estimates of the objective probabilities presented via the icon arrays. Overall, as shown in Figure 3, considering the entire spectrum of objective probabilities tested, the participants were the most accurate in their estimation when the icon arrays were presented in the top and row arrangement. They were the least accurate when the icon arrays were presented in the edge arrangement, followed by the random and the central arrangements. We constructed a within-subject ANOVA to compare the overall estimation error for the six arrangement types ($F(5, 240) = 44.05$, p < 0.001). Post-hoc pair-wise comparisons with Bonferroni's adjustment suggests that all pair-wise comparisons of estimation error within the six arrangement types are significantly different, except the top and the row arrangement pair ($p = 1.00$), the top and the diagonal pair ($p = 0.27$), and the diagonal and the row pair ($p = 0.20$). More details can be found in the supplementary materials.

Participants also had a tendency to overestimate or underestimate depending on the icon array arrangement type and the objective probability. As shown in the line plots in Figure 3, participants overestimated most of the probability values when they were presented with the central arrangement (M_{error} = 1.89, SE_{error} = 0.23). The trend for the overestimation peaked when the objective probability is near 50 percent.

They mostly underestimated the probabilities in the edge arrangement (M_{error} = -2.96, SE_{error} = 0.25), and the random arrangement (M_{error} = -1.94, SE_{error} = 0.25), as shown in Figure 3. For the edge arrangement, the pattern of estimation error appeared to be a mirror image of that in the central arrangement, with the underestimation peaking near 50 percent objective probability. For the random arrangement, the pattern of estimation error appeared more cyclical. The underestimation peaked near 40 percent objective probability, and the estimation error seemed to have shifted to an overestimation near 80 percent.

Participants also slightly underestimated the probabilities in the diagonal arrangement (M_{error} = -0.86, SE_{error} = 0.18), the top arrangement (M_{error} = -0.21, SE_{error} = 0.10), and the row arrangement (M_{error} = -0.19, SE_{error} = 0.13). But overall, for these arrangements, participants were accurate in their probability estimations.

4.5.2 Estimation Confidence. A within-subject ANOVA revealed that participants reported their probability estimations with varying levels of confidence depending on both the icon array arrangement and the objective probability ($F(5, 240) = 111.2$, $p < 0.001$). As shown in Figure 3, participants were more confident when the objective probability was near 0 and 100 percent, and less confident when the objective probability was near 50 percent. Overall, participants were the most confident in their probability estimates when they viewed icon arrays in the top and row arrangements, and the least confidence when they viewed the random arrangement. Post-hoc pair-wise comparisons on the overall confidence levels with Bonferroni's adjustment suggest participants' confidence levels to be significantly different for every icon array pairs ($p < 0.001$), except the top and row arrangement pair ($p = 1.00$), the edge and diagonal pair ($p = 0.25$), the central and diagonal pair ($p = 0.42$), and the central and edge pair ($p = 0.001$). More details can be found in the supplementary materials.

How does reporting confidence relate to estimation error? We constructed a linear regression model predicting the absolute value estimation error with participants' reported confidence, and found that confidence and estimation error were negatively correlated (Est = -0.080, t = -41.29, $p < 0.001$, R^2_{adj} = 0.21). This suggests that participants seemed to have a good intuition about their estimations. They were more confident when their estimations were more accurate.

4.5.3 Potential Confounding Variables. As a sanity check, we examined whether the order in which participants viewed the icon arrays affected the amount of estimation error they made. A linear regression model predicting estimation error with the order in which the icon arrays were displayed suggests that there is no evidence of order effects on estimation errors (Est = 0.002, t = 1.14, p = 0.25, $R^2_{adj} < 0.001$). However, order seemed to have a significant effect on participants' reported confidence (Est = 0.07, t = 7.80, $p < 0.001$, R^2_{adj} = 0.0092), such that participants became more and more confident in their estimations as they progressed through the experiment.

We also examined whether our participants' ability to estimate probabilities varied with their age, research experience, and visual literacy (measured via the ten subjective questions from [24]). On average, our participants had 1.53 years of research experience (SD = 2.24) and they reported an average score of 42.7 out of 60 on the subjective visual literacy survey (SD = 6.83).

A linear regression model predicting estimation error with age, research experience, and visual literacy score suggested age (Est = 0.16, t = 2.72, p = 0.006) and visual literacy scores (Est = 0.10, t = 3.58, $p < 0.001$) to be significant factors, but we did not observe an effect of research experience (Est = 0.01, t = 0.39, p = 0.69). Older participants and participants with higher visual literacy scores tended to make more estimation errors. Although age, research experience, and visual literacy scores are statistically significant predictors to estimation error, their effect sizes are extremely small (η^2_{age} = 0.001, $\eta^2_{research} < 0.001$, $\eta^2_{literacy} < 0.001$, overall R^2_{adj} = 0.0015). We conducted a similar linear regression examining the effect of these factors on participants' reported confidence. All three factors were significant predictors, but again with small effect sizes (η^2_{age} = 0.03, $\eta^2_{research}$ = 0.01 , $\eta^2_{literacy}$ = 0.07 , overall R^2_{adj} = 0.11). Older participants (Est = 4.06, t = 14.97, $p < 0.001$), participants with more research experiences (Est = 0.80, t = 4.93, $p < 0.001$), and participants with higher reported visual literacy (Est = 1.14, t = 21.62, $p < 0.001$) tended to be more confident in their estimations.

However, considering that all of our participants were undergraduate students, the narrow range in their age, research experience,

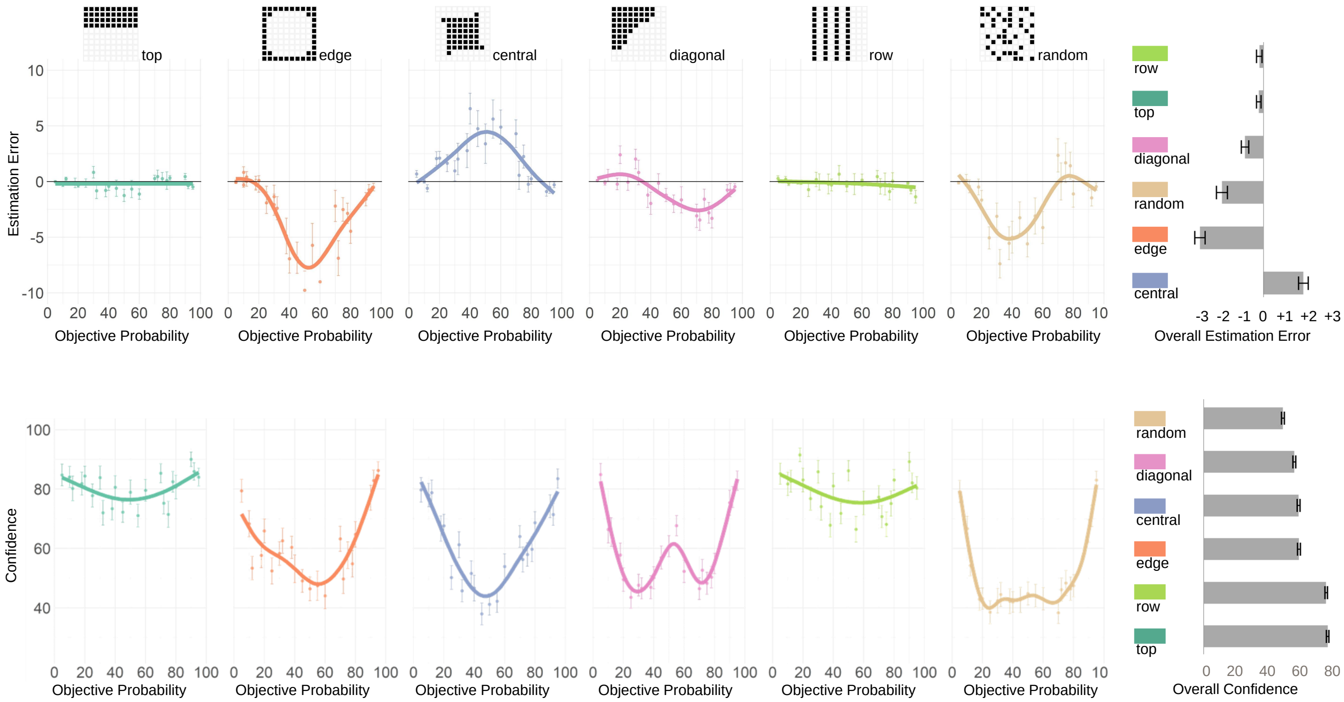

Figure 3: Top: Average estimation error as a function of the number of filled grids in an icon array for the six arrangements, and on the right, the overall estimation error for the six arrangements aggregating all levels of objective probabilities. Bottom: Average reported estimation confidence as a function of the number of filled grids in an icon array for the six arrangements, and on the right, the overall confidence for the six arrangements aggregating all trials. Error bars show standard error.

and visual literacy scores may limit the generalizability of these findings.

4.6 Discussion and Motivation for Exp 2 and 3

How filled grids are arranged within an icon array significantly influences people's perception of depicted probabilities. Participants were generally accurate in their estimations when icon arrays were in the row, top, and diagonal arrangements, but were biased when the icon arrays were presented in the central, edge, and random arrangements.

For the central and edge arrangements, participants consistently overestimated the proportions in the central arrangement and underestimated the proportions in the edge arrangement. This observation is consistent with the centrality effect observed in [4], where participants viewed a 5×8 matrix of 18 blue and 22 red squares. They underestimated the percentage of blue squares when the blue squares were placed in the middle of the matrix (similar to our central arrangement), and overestimated the percentage of blue squares when the blue squares were placed on the edge (similar to our edge arrangement). Our results demonstrate that this overestimation and underestimation bias seems to generalize beyond just the one scenario tested in [4]. Additionally, when we generated the stimuli in Experiment 1, we recognized that there are many ways to arrange icon arrays even within the broader definition of 'central' and 'edge' arrangements, as shown in Figure 4. The generalizability of this centrality effect, considering the potential randomness in arranging filled icons in the central and edge arrangements, has not been explored. **In Experiment 2**, we systematically study this effect of centrality with icon arrays across a range of values, adding randomness to patterns in each icon array.

For icon arrays in the random arrangement, we observed a clear cyclic bias pattern for estimation errors that is not present in the other arrangements. Participants in our experiment seemed to underestimate proportions closer to 40% and overestimate proportions close to 80%. This observation is consistent with existing findings in human perception research that contributed to the development of Steven's Power Law [58] and the cyclical power model [35], which suggest a non-linear relationship between the perceived stimulus value and the actual stimulus value. Usually, smaller values tend to be overestimated, while bigger values tend to be underestimated, creating a one-cycle bias pattern in value estimations [53, 55]. Recent work has demonstrated that when participants use a reference point in their value estimations, a repulsion effect can occur such that the estimations are perceived to be farther from that reference point [44]. This effect turns the one-cycle bias pattern into a repeating, multiple-cycle bias pattern where values slightly smaller than the reference point are treated similarly to 'large' values and get underestimated, and values slightly larger than the reference point are treated similarly to 'small' values and get overestimated. In Experiment 1, while the proportion estimation for the random arrangement does follow a cyclical pattern, because we only tested a limited number of objective proportions, it is unclear whether participants were using any reference points and how their bias pattern compare exactly to existing work in similar domains. This motivated us to conduct **Experiment 3**, where we took a deeper dive into investigating how people perceive random arrangements of icon arrays and what factors might play a role in biasing their proportion judgement.

5 EXPERIMENT 2 CENTRALITY EFFECTS

While the results from [4] suggest that centrality can have a large effect on perceived proportion, their experiments only tested two patterns — those in which focal squares are placed in the extreme edge or extreme center, and only for one proportion (45%). In this experiment, we systematically explored 101 proportions at four levels of centrality, as shown in Figure 4. We did this by considering random square grids generated from a Gaussian distribution over locations centered in the middle of a 10×10 grid. Specifically, the probability that any particular grid location is populated is given by

$$p_{ij} \propto exp\left(-d_{ij}^2/2\sigma^2\right) \tag{1}$$

where d_{ij} is the Euclidean distance from the center of a square to the middle of the grid and σ^2 is a free parameter, which we refer to as the **scale**. Given a scale parameter and the proportion of squares to populate, we sample square locations without replacement from this distribution. Small values of the scale parameter result in high centrality grids (see the top row of Figure 4), with squares clustered towards the middle, whereas large values of the scale parameter produce grids that are populated uniformly at random (see the bottom row of the top panel in Figure 4). We use the same distribution to produce grids with extremely low centrality (squares on the edge) by repeating this procedure to sample unpopulated (rather than populated) square locations.

Figure 4 shows a subset of the results of this procedure for different proportions and scales. The top half of the figure shows the central arrangement icon arrays, and the bottom half shows the edge arrangement icon arrays. We selected four scale values to test (σ^2 = 2.0, 3.5, 5.0, 6.5), covering a wide range of possible centrality, as represented in each row. Within each arrangement, the top row has the highest centrality (σ^2 = 2.0) and the bottom row has the lowest centrality (σ^2 = 6.5). The far left column contains empty icon arrays representing 0% objective probability, the next column contains icon arrays with 5 filled squares representing 5% objective probability, and so on. This provides a convenient way to interpolate between grids whose squares are concentrated in the center of the grid or towards the edges. Notice that icons with a very low or very high number of grids filled tend to look more similar as centrality is small at those values. In total, we generated 101 central arrangement and 101 edge arrangement icon arrays for each of the scale values, one for each of the whole number percentage values from 0 to 100, totaling 808 stimuli used. These icon arrays were presented on a 1024 by 600 pixel light gray background, just like the icon arrays in Experiment 1.

5.1 Participants and Procedure

We recruited 800 participants using Amazon Mechanical Turk (MTurk), a popular micro-task market that is regularly used for online experiments with human participants. We implemented our experiment as web applications hosted on a server external to MTurk. Participants accessed the experiments within an embedded frame presented on the MTurk site. We limited the participant pool to workers who are located in the United States with a minimum 95% approval rate and at least 100 approved tasks.

Participants carried out the task in four steps. They were first presented a description of the task with an option of accepting it. Once the task was accepted, participants completed a training session in which they were asked to estimate the proportion values of 25% and 75% in a sequence of two tasks. The order of the tasks was randomized for each participant. Participants were given feedback on the accuracy of their responses. After the training session, participants were randomly asked to estimate 25 randomly selected icon arrays from our 808 stimuli (a subset shown in Figure 4) in a sequence of tasks. They were paid for 0.75 USD per task.

For each icon array, participants were asked to estimate the proportion value represented, which was made difficult by the fact that it only appeared on the screen for a short amount of time. Similar to that in Experiment 1, they responded by clicking on an uninitialized slider. Participants could change their response by dragging the slider handle and the value selected on the slider was also shown under the slider. Before each stimulus was shown, a white cross centered in the stimulus area appeared for 2 seconds to drive the participant's attention followed by the presentation of the stimulus. We omitted the confidence measure as Experiment 1 has demonstrated that confidence tends to positively correlate with estimation accuracy.

5.2 Results

We filtered responses that were off by more than 15% (which is approximately two standard deviations away from the mean estimation error) and were left with 17,514 responses (87.57% of total responses). Overall, a within-subject ANOVA comparing the effect of arrangement and centrality scale (σ^2) on estimation error suggests there to be a significant main effect of both arrangement (F = 95.24, p < 0.001) and centrality scale (F = 5.16, p = 0.023). Participants tended to overestimate the objective probability when they viewed icon arrays with the central arrangement (M_{error} = 0.96, SE_{error} = 0.059), and underestimate the objective probability when they view icon arrays with the edge arrangement (M_{error} = -0.72, SE_{error} = 0.057), supporting results from Experiment 1. Overall, for the central arrangement, the overestimation seems to peak when the objective probability is around 60 to 70 percent. For the edge arrangement, the underestimation seems to peak around 30 to 50 percent.

For the main effect of centrality scale, we found that the higher the scale value, which means the more 'random' the arrangement appears, the lower the participants' estimation error. A post-hoc linear regression revealed that with a one unit increase in scale value, the estimation error decreases by 0.053 (SE = 0.02487, p = 0.033). We can also see this in Figure 5. From left to right, the vertical distance between 0 (no estimation error) and the lines representing estimation error decreases, signifying less error as the scale increases.

There is also a significant interaction between arrangement and centrality scale (F = 40.36, p < 0.001), as shown in Figure 5. With high centrality scale (σ^2 = 2.0), the central and edge arrangement looked more dissimilar and the patterns of participants' estimation

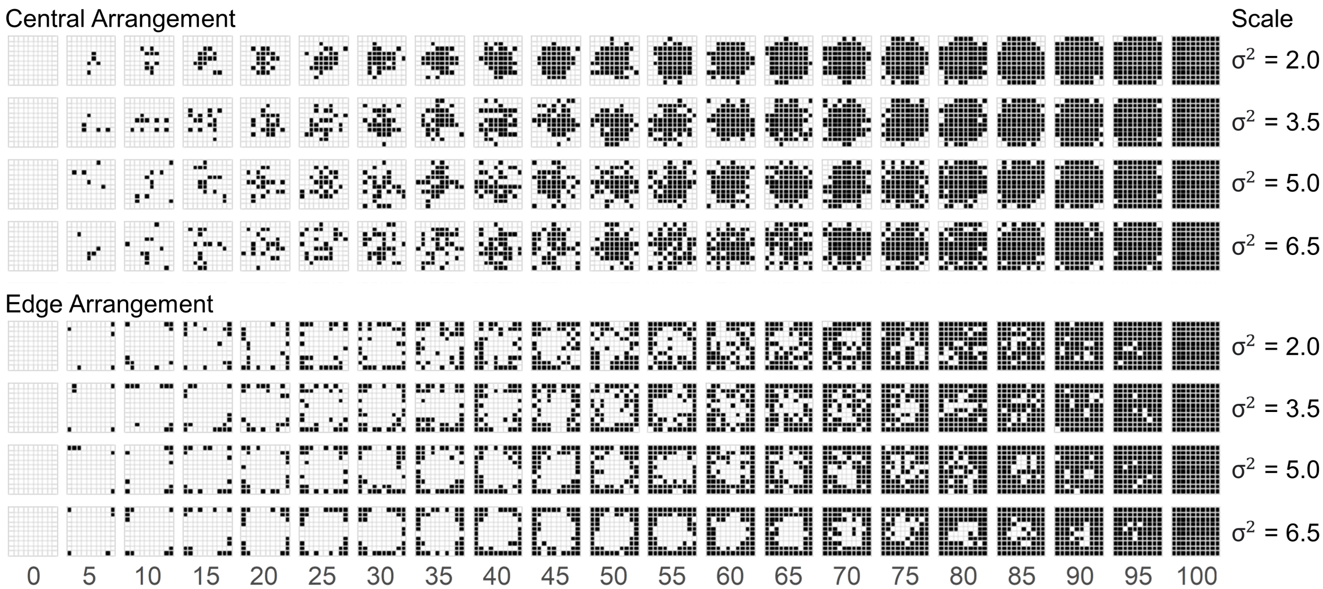

Figure 4: Icon arrays with varying centrality across a range of probability values. Each column shows icon arrays depicting a fixed probability value (0%, 5%, 10%, ...), with varying centrality in each row. Centrality is the highest on the top row and the lowest on the bottom row.

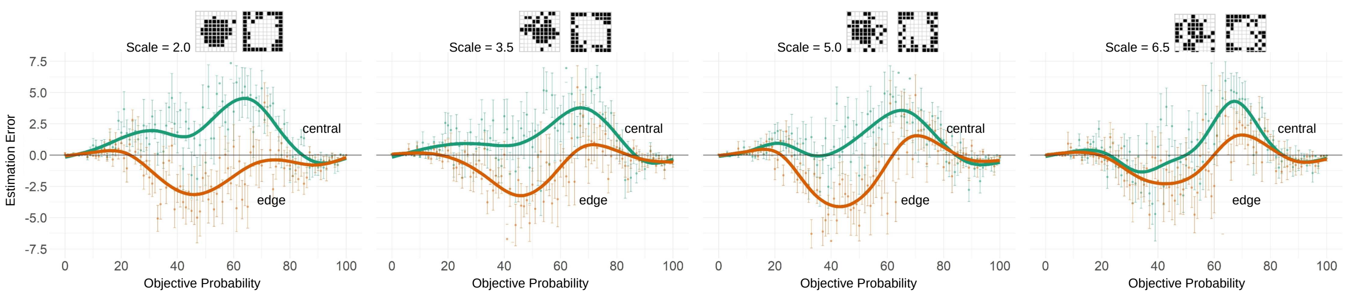

Figure 5: Average estimation error for central and edge arrangements at four levels of centrality (scale = 2.0, 3.5, 5.0, 6.5), across varying levels of probability values. Example icon arrays are shown on top at 45% for central and edge arrangements.

error appeared more distinct (see the left most panel in Figure 5). There is a consistent overestimation in the central arrangements and underestimation in the edge arrangements for most levels of objective probabilities. With low centrality scale (σ^2 = 6.5), both the central and edge arrangements start to look more similar to a random arrangement, and participants begin to exhibit similar bias patterns with that of viewing a random arrangement. The estimation error follows a more cyclical pattern, with underestimation in the range of 20 to 60 percent, overestimation in the range of 60 to 80 percent, and underestimation thereafter.

5.3 Discussion

Experiment 2 reproduced the biased pattern in probability estimations for the central and edge arrangement found in Experiment 1 and revealed that centrality plays an important role biasing our perception of probabilities in icon arrays. People overestimate probabilities in the central arrangement and underestimate probabilities in the edge arrangement, and these misestimations can be understood as centrality effects such that higher centrality scale (e.g., when σ^2 = 2.0) will lead to more overestimation and underestimation. We also observed that when centrality scale is low (e.g., when σ^2 = 6.5), the icon arrays in the central and edge arrangement begin to resemble the random arrangement, and participants' estimates begin to follow a similar cyclical pattern. This additionally supports our findings from Experiment 1, and is consistent with existing findings in the psychophysics literature [35, 58].

6 EXPERIMENT 3 RANDOM ARRANGEMENT

Following the discussion at the end of Experiment 1 and 2, we now conduct a systematic investigation of probability perception with icon arrays in the random arrangement to refine the perceptual bias pattern we observed in Experiment 1 and Experiment 2. This investigation also helps us identify potential strategies that participants use to make proportion estimations, such as whether they were using particular values as reference points [44]. As mentioned in the previous sections, according to Steven's Power Law [58], people tend to overestimate small values (e.g., probabilities close to 0) and underestimate large values (e.g., probabilities close to 100), creating a cyclical, biased pattern of perception. But if they are using additional values as reference points, the cyclical bias pattern will reset itself at the reference point, treating values slightly smaller than the reference point as 'large' values and values slightly bigger than the reference point as 'small' values. For example, if people are using the 50% mark as a reference point to make probability estimates in icon arrays, we should see them underestimating values slightly below 50, and overestimating values slightly above 50.

We additionally tested the robustness of this perceptual bias across various visual styles of icon arrays. This will help us better understand how much this effect generalizes depending on the particular visual settings we chose in our stimulus design. We tested four visual styles: large grids, medium grids, medium grids with light shading, and small grids, as shown in Figure 6.

6.1 Participants, Stimuli, and Procedure

We recruited 800 workers from Mechanical Turk following the same inclusion criteria as in Experiment 2, randomly assigning 200 participants to each of the four style conditions. Similar to that in Experiment 2, each worker was asked to estimate the proportion of grids filled in 25 randomly selected icon arrays. Again, as quality control, we eliminated responses that were off by more than 15%. For the 20000 responses we received, 17766 were retained (88.83%).

We generated icon arrays in the random arrangement and presented them on a 1024 by 600 pixel light gray background. Similar to that in the previous experiments, we generated icon arrays on a white background, divided into square cells with a regular grid of gray colored lines. The cells were filled by drawing colored squares centered in a cell to represent probabilities. The new manipulation for this experiment is the visual styles, which included large grids, medium grids, medium grids with light shading, and small grids, as shown in Figure 6. The objective proportion values shown with these icon arrays were uniformly sampled from 0% to 100%. In total, we generated 101 icon arrays for each visual style, one for each of the whole number percentage values from 0 to 100, totaling 404 stimuli used.

6.2 Results

Figure 6 shows participants' estimations across the four visual styles. While some variation existed across styles, the cyclical-pattern of bias seemed robust to the superficial changes we tested, suggesting that this perceptual bias is mostly driven by the perception of randomly arranged icon arrays themselves. On average, the estimation error participants made for the four visual styles is very similar: M_{large} = -0.086 (SE = 0.076), $M_{mediumDark}$ = 0.25 (0.080), $M_{mediumLight}$ = -0.31 (0.080), and M_{small} = 0.32 (0.084). A within-subject ANOVA comparing estimation error between these four visual style supports this observation. Although there exists a small main effect of visual style on estimation error ($F(3, 17754) = 8.39$, $p < 0.001$, $\eta^2 = 0.0014$), there is no evidence of an interaction effect between visual style and objective probability ($F(3, 17754) = 0.80$, $p = 0.49$). This suggests that the overall trends in estimation error varied slightly between different visual styles but is preserved across varying levels of objective probabilities.

As expected, in our ANOVA, we also observed a main effect of objective probability on estimation error ($F(1, 17754) = 112.31$, $p < 0.001$, $\eta^2 = 0.0063$). As shown in Figure 6, overall, there was a slight overestimation in the probability from 0 to 20 percent, transforming into a considerable amount of underestimation in the region between 20 and 60, returning back to overestimation around 60 to 90, and ending with a slight underestimation in the range 90 to 100.

6.3 Discussion

Based on Steven's Power Law, we know that when people tend to follow a cyclical bias pattern in their estimation errors [53, 55, 58]. However, recent work has demonstrated that the smaller and bigger values are relative. Depending on whether participants are using additional values as reference marks, this cyclical pattern could repeat itself around the reference marks [44]. In estimations of proportions in a stacked bar chart, for example, participants often use the 0% mark, the 50% mark, and the 100% mark as their reference points, so they overestimate values slightly above 0 and slightly above 50, and underestimate values slightly below 50 and slightly below 100, creating a two-cycle pattern [44].

Our results suggest that when participants estimate probabilities with icon arrays in the random arrangement, they tend to use 0% and around 60% as a reference point. They overestimate values slightly above 0 and slightly above 60, and underestimate values slightly below 60 and slightly below 100, following similar cyclical patterns identified in [44] and [35].

7 GENERAL DISCUSSION AND CONCLUSION

In this work, we report a bias in the perception of icon arrays that causes people to misestimate proportions. Both the visual arrangement of filled grids (e.g., top, row, central) and the objective probabilities (e.g., 5%, 20%, 85%) have an effect on people's estimation accuracy. Across all levels of objective probabilities, participants were fairly accurate in estimating proportions when viewing the top, row, and diagonal arrangements. However, they overestimated the proportions with the central arrangement and underestimated the proportions with the edge arrangement, with their overestimation peaking around 60% and their underestimation peaking around 40%. They were also biased to either overestimate or underestimate when viewing the random arrangement depending on the objective proportions, following a cyclical pattern consistent with existing findings in the psychophysics literature. More specifically, they underestimate proportions in the range of roughly 20% to 60% and 90% to 100%, and overestimate proportions in the range of roughly 0% to 20% and 60% to 90%. The bias seems related to, but not fully explained by, biases in proportion perception grounded in Steven's Law [35, 53].

Experiment 2 demonstrates that centrality is in fact a powerful lever in manipulating perceived probabilities across a wide range of values. So much so, in fact, that extreme centrality grids can reverse the perceptual effects seen in randomly populated grids.

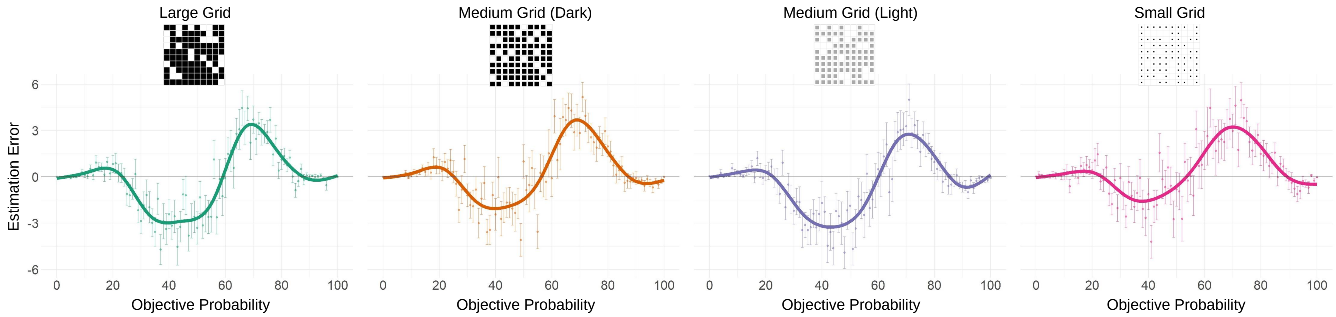

Figure 6: Average error as a function of the depicted probability values, for the four visual styles. All icon arrays tested are in the random arrangement.

For example, in Experiment 1 and Experiment 3, the random arrangement containing 45 filled grids is consistently underestimated, but in Experiment 2, the highly central grids with 45 black squares are consistently overestimated, though this effect becomes weaker and weaker as centrality decreases. Experiment 2 also shows that the relationship between perceptual bias and centrality is highly non-linear and complicated. The source of this non-linearity remains unclear. While it could arise from centrality effects alone, it is also possible that there are other visual and perceptual features coupled to centrality and proportion that drives perceptual bias.

8 LIMITATIONS AND FUTURE DIRECTIONS

We identify several limitations that provide promising future research directions based on our three experiments. Why did we observe different patterns of perceptual bias depending on the arrangements? We tested the specific effect of centrality and found that centrality plays a part in biasing people's probability estimations. However, we have yet to explore why the central and edge arrangements trigger completely opposite bias patterns such that people underestimate probabilities with one and overestimate with the other. Additionally, while the cyclical patterns we observe in probability estimations with the random arrangement seem consistent with existing psychophysics literature, the actual driver of this pattern is less well understood. Hollands and Dyre suggest it may arise from participants alternating between different perceptual estimation strategies (e.g., switching from using area to number encoding) from trial to trial, but admit that "[t]his explanation, is, of course, ad hoc" (p.518) [35]. Future work at the intersection of visualization and human perception research could further investigate what perceptual or cognitive mechanisms are in play to create this cyclical bias pattern in probability estimation. Furthermore, we also observed that participants were almost equally as accurate estimating probabilities in the top, row, and diagonal arrangements, despite their visual differences. What strategies were participants using when they estimated probabilities in these arrangements that made them more accurate? Future work can extend our investigation to explore these strategies people use and uncover the factors that lead to more accurate probability estimations.

Future work can also test a wider set of icon array arrangements beyond the six identified in our approach. In the real world, visualization designers can get very creative with how they arrange icons in an icon array to convey key statistics (for some examples, see [9]), so the effectiveness of less intuitive, less commonly seen arrangements should also be tested. Additionally, following our discussion in Section 2.1, we suspect other visual factors that can affect visual attention and visual grouping, such as color, shape, and spacing between icons, can also influence probability perception. Visualization designers in the real-world often leverage these visual features to make the key message stand out or the visualization more engaging, such as coloring icons a salient color or using a unique icon shape [1, 6, 54]. So it's critical for future work to also explore the effect of these other visual features on probability estimations in icon arrays.

We used a grid of 100 squares for all three studies to make our results more conservative. Participants may have an easier time estimating the probabilities when the total number of icons remained a fixed, prototypical number, such that each icon represents exactly 1%. We suspect people may be even more prone to perceptual biases when they estimate probabilities in icon arrays with a less prototypical number of total icons, such as 25 or 70 icons. Further, it is common for visualization designers to use an irregular number of icons to convey a key probability in the real world, so future work could test how these results might change with more or less-dense icon arrays.

Additionally, in our current experiment, we asked participants to report the percentage they perceived by moving a slider with number labels. Some participants might have mentally counted the number of filled icons and then reported that number in a straightforward fashion. Others might have relied on forming a mental representation of the icon array using surrogate visual features like area and density. These participants then have to translate that visual information into verbal information. Existing work in cognitive psychology has shown that a viewer who is forced away from thinking about area into thinking about numbers tends to misperceive that numeric quantity [7, 37]. Recent work has also shown that translating visual information to verbal information might introduce bias and inaccuracies in value reporting [12]. In addition to investigating participants' strategies when making probability estimations, researchers could also test out other experimental setups to record the participants' probability estimates in a way that doesn't force the participants into switching between thinking about numbers and area, such as asking the participants to compare two icon arrays and report which one portrays a higher probability.

See [17] for some alternative experimental setups future researchers can use.

Future work can also apply neural network techniques to train networks on a wide variety of icon arrays and proportion estimates. Researchers can use these trained networks to identify potential causes of perceptual biases by extracting systematic regularities between the icon arrays shown and people's proportion estimates, which enables them to generate icon arrays that should be perceived veridically. For example, we showed that manipulating centrality, which is the distance of the focal icons from the center of the grid, can change probability estimates. A neural network can help us quantify this effect by learning the features of an icon array that leads to similar and dissimilar human responses from viewing an icon array with varying centrality.

9 IMPLICATIONS FOR DESIGN

In this paper, we identified six icon array arrangements and systematically tested how they affected probability perception. It's clear that how filled grids are spatially arranged in an icon array significantly impacts viewer perception. One solution based on this work is to explicitly annotate the percentage depicted in the icon arrays to reduce perceptual biases. We also encourage visualization designers to use the data collected in our work as a reference tool to help them predict how a viewer might react to an icon array design. Below we discuss several recommendations for icon array designers to consider in their work.

Presenting icon arrays with top or row arrangements will help viewers confidently make the most accurate probability estimations. Presenting icon arrays with the diagonal arrangements will achieve a similar effect in terms of perceptual accuracy, but the viewer is likely much less confident about their estimations. In scenarios where the designer requires the viewer to make a prompt and accurate decision with some probabilistic information, regardless of the probability value, it's better to present them with icon arrays in the top or row arrangement.

On the other hand, presenting icon arrays with the central arrangement makes a viewer overestimate the probability, and presenting icon arrays with the edge arrangement makes a viewer underestimate the probability. For example, a patient looking at an icon array in the *central* arrangement depicting a 60% probability of having an allergic reaction might be more stressed than they would actually be, because they likely overestimated that 60% to be closer to 70%. A patient looking at an icon array in the *edge* arrangement depicting a 45% probability of risk might be more optimistic than they would want to be, because they likely underestimated that 45% to be closer to 40% risk.

As for the random arrangement, we demonstrate that using a risk theatre, where filled grids are randomly arranged to help people better grasp the concept of uncertainty [46, 50], may be at the expense of sacrificing perceptual accuracy and confidence, despite its other merits. People are biased to overestimate or underestimate the probability depicted in an icon array with random arrangement depending on the objective probability, perceiving it to be higher in the 0 to 20 and 60 to 90 percent range, and lower in the 20 to 60 and 90 to 100 percent range. So in the context of communicating a projected election outcome, say a Democratic candidate has a 70% chance of winning, a voter looking at the risk theatre depicting that 70% chance will perceive it to be higher, at around 75% chance. This difference in perceived probability and actual probability might mean the difference between going to vote or not going to vote.

We also discovered that when people make proportion estimates in icon arrays, they tend to use the 60% mark as their reference point. As a result, they are the most accurate when perceiving probabilities around 60%. Although further investigation with regards to why people might be using the 60% mark as reference is warranted, when designing an icon array to communicate a key probability value, it might be worthwhile to transform that key probability value to be around 60%. For example, in a competition between A and B, instead of presenting a prediction of the outcome using an icon array that depicts "a 40% chance of A winning", viewers are more likely to perceive the probability more accurately if the outcome is presented as "a 60% chance of B winning".

In summary, we recommend that icon array designers consider how their audience will perceive the probabilities depicted in their designs and weigh the trade-off between perceptual accuracy and other design goals.

REFERENCES

[1] Kiran Ajani, Elsie Lee, Cindy Xiong, Cole Nussbaumer Knaflic, William Kemper, and Steven Franconeri. 2021. Declutter and Focus: Empirically Evaluating Design Guidelines for Effective Data Communication. *IEEE Transactions on Visualization and Computer Graphics* (2021).
[2] Jüri Allik and Tiia Tuulmets. 1991. Occupancy model of perceived numerosity. *Perception & Psychophysics* 49, 4 (1991), 303–314.
[3] Jessica S Ancker, Yalini Senathirajah, Rita Kukafka, and Justin B Starren. 2006. Design features of graphs in health risk communication: a systematic review. *Journal of the American Medical Informatics Association* 13, 6 (2006), 608–618.
[4] Eduardo B Andrade. 2011. Excessive confidence in visually-based estimates. *Organizational Behavior and Human Decision Processes* 116, 2 (2011), 252–261.
[5] Dan Ariely. 2001. Seeing sets: Representation by statistical properties. *Psychological science* 12, 2 (2001), 157–162.
[6] Scott Bateman, Regan L Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. 2010. Useful junk? The effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2573–2582.
[7] Ty W Boyer and Susan C Levine. 2015. Prompting children to reason proportionally: Processing discrete units as continuous amounts. *Developmental psychology* 51, 5 (2015), 615.
[8] Jacquelyn Burkell. 2004. What are the chances? Evaluating risk and benefit information in consumer health materials. *Journal of the Medical Library association* 92, 2 (2004), 200.
[9] Alyxander Burns, Cindy Xiong, Steven Franconeri, Alberto Cairo, and Narges Mahyar. 2021. Designing with Pictographs: Envision Topics without Sacrificing Understanding. *IEEE Transactions on Visualization and Computer Graphics* (2021).
[10] David Burr and John Ross. 2008. Response: visual number. *Current Biology* 18, 18 (2008), R857–R858.
[11] Lisa Cantrell, Megumi Kuwabara, and Linda B Smith. 2015. Set size and culture influence children's attention to number. *Journal of Experimental Child Psychology* 131 (2015), 19–37.
[12] Cristina R Ceja and Cindy Xiong. 2021. Show or Tell? Visual and Verbal Representations Bias Position Recall. *arXiv preprint arXiv:2108.08407* (2021).
[13] Sang Chul Chong and Anne Treisman. 2003. Representation of statistical properties. *Vision research* 43, 4 (2003), 393–404.
[14] William S Cleveland and Robert McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association* 79, 387 (1984), 531–554.
[15] Stanislas Dehaene. 2011. *The number sense: How the mind creates mathematics.* OUP USA.
[16] Stanislas Dehaene and Jean-Pierre Changeux. 1993. Development of elementary numerical abilities: A neuronal model. *Journal of cognitive neuroscience* 5, 4 (1993), 390–407.
[17] Madison A Elliott, Christine Nothelfer, Cindy Xiong, and Danielle Albers Szafir. 2020. A Design Space of Vision Science Methods for Visualization Research. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1117–1127.

[18] Angela Fagerlin, Brian J Zikmund-Fisher, and Peter A Ubel. 2011. Helping patients decide: ten steps to better risk communication. *Journal of the National Cancer Institute* 103, 19 (2011), 1436–1443.

[19] Deb Feldman-Stewart, Nancy Kocovski, Beth A McConnell, Michael D Brundage, and William J Mackillop. 2000. Perception of quantitative information for treatment decisions. *Medical Decision Making* 20, 2 (2000), 228–238.

[20] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[21] Steven L Franconeri, Douglas K Bemis, and George Angelo Alvarez. 2009. Number estimation relies on a set of segmented objects. *Cognition* 113, 1 (2009), 1–13.

[22] Christopher D Frith and Uta Frit. 1972. The solitaire illusion: An illusion of numerosity. *Perception & Psychophysics* 11, 6 (1972), 409–410.

[23] Mirta Galesic, Rocio Garcia-Retamero, and Gerd Gigerenzer. 2009. Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychology* 28, 2 (2009), 210.

[24] Rocio Garcia-Retamero and Edward T Cokely. 2013. Communicating health risks with visual aids. *Current Directions in Psychological Science* 22, 5 (2013), 392–399.

[25] Titia Gebuis and Bert Reynvoet. 2012. Continuous visual properties explain neural responses to nonsymbolic number. *Psychophysiology* 49, 11 (2012), 1649–1659.

[26] Norman Ginsburg. 1976. Effect of item arrangement on perceived numerosity: Randomness vs regularity. *Perceptual and motor skills* 43, 2 (1976), 663–668.

[27] EJ Goldstein, RM Kahn, ML Alpert, BP Ginsberg, FL Greenway, and DM Citron. 1987. Ciprofloxacin versus cinoxacin in therapy of urinary tract infections. A randomized, double-blind trial. *The American journal of medicine* 82, 4A (1987), 284–287.

[28] Donna Gresh, Léa A Deleris, Luca Gasparini, and Dylan Evans. 2011. Visualizing risk. In *Proceedings of IEEE Information Visualization Conference*. 1–10.

[29] Justin H. Gross. 2016. *How to better communicate election forecasts — in one simple chart*. https://www.washingtonpost.com/news/monkey-cage/wp/2016/11/29/how-to-better-communicate-election-forecasts-in-one-simple-chart/?noredirect=on

[30] Justin Halberda, Sean F Sires, and Lisa Feigenson. 2006. Multiple spatially overlapping sets can be enumerated in parallel. *Psychological science* 17, 7 (2006), 572–576.

[31] Sarah T Hawley, Brian Zikmund-Fisher, Peter Ubel, Aleksandra Jancovic, Todd Lucas, and Angela Fagerlin. 2008. The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient education and counseling* 73, 3 (2008), 448–455.

[32] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 203–212.

[33] Rebecca Hess, Vivianne HM Visschers, and Michael Siegrist. 2011. Risk communication with pictographs: The role of numeracy and graph processing. *Judgment and Decision Making* 6, 3 (2011), 263.

[34] Ulrich Hoffrage, Samuel Lindsey, Ralph Hertwig, and Gerd Gigerenzer. 2000. Communicating statistical information.

[35] JG Hollands and Brian P Dyre. 2000. Bias in proportion judgments: the cyclical power model. *Psychological review* 107, 3 (2000), 500.

[36] Nicole Jardine, Brian D Ondov, Niklas Elmqvist, and Steven Franconeri. 2019. The perceptual proxies of visual comparison. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1012–1021.

[37] Yoonkyung Jeong, Susan C Levine, and Janellen Huttenlocher. 2007. The development of proportional reasoning: Effect of continuous versus discrete quantities. *Journal of Cognition and Development* 8, 2 (2007), 237–256.

[38] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 chi conference on human factors in computing systems*. 5092–5103.

[39] Helen Kennedy, Rosemary Lucy Hill, William Allen, and Andy Kirk. 2016. Engaging with (big) data visualizations: Factors that affect engagement and resulting new definitions of effectiveness. *First Monday* 21, 11 (2016).

[40] Sarah Kliff and Aatish Bhatia. 2022. *When They Warn of Rare Disorders, These Prenatal Tests Are Usually Wrong*. https://www.nytimes.com/2022/01/01/upshot/pregnancy-birth-genetic-testing.html?smtyp=cur&smid=fb-nytimes&fbclid=IwAR1yOiAgSAB79GLf8CAFWgdmFyLKToMG5-QnJWJp4ct80mIDGroJxvy3EJw

[41] Elke Kurz-Milcke, Gerd Gigerenzer, and Laura Martignon. 2008. Transparency in risk communication: graphical and analog tools. In *Annals of the New York Academy of Sciences*. Blackwell, 18–28.

[42] Laura Landro. 2017. *How to Get Patients to Take More Control of Their Medical Decisions*. https://wellnet.com/broker-resource/news/how-to-get-patients-to-take-more-control-of-their-medical-decisions/

[43] Tali Leibovich, Naama Katzin, Maayan Harel, and Avishai Henik. 2017. From "sense of number" to "sense of magnitude": The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences* 40 (2017).

[44] Caitlyn M McColeman, Lane Harrison, Mi Feng, and Steven Franconeri. 2020. No mark is an island: Precision and category repulsion biases in data reproductions. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1063–1072.

[45] Arthur L Miller and Robert A Baker. 1968. The effects of shape, size, heterogeneity, and instructional set on the judgment of visual number. *The American Journal of Psychology* 81, 1 (1968), 83–91.

[46] Clare Milliken. 2020. *Understanding Uncertainty*. https://magazine.northwestern.edu/exclusives/understanding-uncertainty?commentsstart2772=10

[47] Wendy Nelson, Valerie F Reyna, Angela Fagerlin, Isaac Lipkus, and Ellen Peters. 2008. Clinical implications of numeracy: theory and practice. *Annals of behavioral medicine* 35, 3 (2008), 261–274.

[48] Aude Oliva. 2005. Gist of the scene. In *Neurobiology of attention*. Elsevier, 251–256.

[49] Stefania Pighin, Lucia Savadori, Elisa Barilli, Laura Cremonesi, Maurizio Ferrari, and Jean-François Bonnefon. 2011. The 1-in-X effect on the subjective assessment of medical probabilities. *Medical Decision Making* 31, 5 (2011), 721–729.

[50] Xiaoying Pu and Matthew Kay. 2020. A probabilistic grammar of graphics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[51] I Qualtrics. 2013. Qualtrics. *Provo, UT, USA* (2013).

[52] Valerie F Reyna. 2008. A theory of medical decision making and health: fuzzy trace theory. *Medical decision making* 28, 6 (2008), 850–865.

[53] Emir H Shuford. 1961. Percentage estimation of proportion as a function of element type, exposure time, and task. *Journal of Experimental Psychology* 61, 5 (1961), 430.

[54] Mark Smiciklas. 2012. *The power of infographics: Using pictures to communicate and connect with your audiences*. Que Publishing.

[55] Ian Spence. 1990. Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance* 16, 4 (1990), 683.

[56] David Spiegelhalter, Mike Pearson, and Ian Short. 2011. Visualizing uncertainty about the future. *science* 333, 6048 (2011), 1393–1400.

[57] Jan Stellamanns, Dana Ruetters, Keshav Dahal, Zita Schillmoeller, and Jutta Huebner. 2017. Visualizing risks in cancer communication: a systematic review of computer-supported visual aids. *Patient education and counseling* 100, 8 (2017), 1421–1431.

[58] Stanley S Stevens. 1957. On the psychophysical law. *Psychological review* 64, 3 (1957), 153.

[59] Alan R Tait, Terri Voepel-Lewis, Brian J Zikmund-Fisher, and Angela Fagerlin. 2010. Presenting research risks and benefits to parents: does format matter? *Anesthesia and analgesia* 111, 3 (2010), 718.

[60] Alberto Testolin, Serena Dolfi, Mathijs Rochus, and Marco Zorzi. 2020. Visual sense of number vs. sense of magnitude in humans and machines. *Scientific reports* 10, 1 (2020), 1–13.

[61] Michael Theil. 2002. The role of translations of verbal into numerical probability expressions in risk management: a meta-analysis. *Journal of Risk Research* 5, 2 (2002), 177–186.

[62] Kevin E Tiede, Felicia Ripke, Nicole Degen, and Wolfgang Gaissmaier. 2020. When Does the Incremental Risk Format Aid Informed Medical Decisions? The Role of Learning, Feedback, and Number of Treatment Options. *Medical Decision Making* 40, 2 (2020), 212–221.

[63] Sithi Wangeamsermsuk and Arisara Jiamsanguanwong. 2018. The Comparison of Iconicity Level of Icon Arrays on Risk Perception. *Journal of Advances in Information Technology Vol* 9, 4 (2018).

[64] Evan A Wilhelms and Valerie F Reyna. 2013. Effective ways to communicate risk and benefit. *The virtual mentor: VM* 15, 1 (2013), 34.

[65] Lei Yuan, Steve Haroz, and Steven Franconeri. 2019. Perceptual proxies for extracting averages in data visualizations. *Psychonomic bulletin & review* 26, 2 (2019), 669–676.

[66] Brian J Zikmund-Fisher, Angela Fagerlin, and Peter A Ubel. 2008. Improving understanding of adjuvant therapy options by using simpler risk graphics. *Cancer* 113, 12 (2008), 3382–3390.

[67] Brian J Zikmund-Fisher, Peter A Ubel, Dylan M Smith, Holly A Derry, Jennifer B McClure, Azadeh Stark, Rosemarie K Pitsch, and Angela Fagerlin. 2008. Communicating side effect risks in a tamoxifen prophylaxis decision aid: the debiasing influence of pictographs. *Patient education and counseling* 73, 2 (2008), 209–214.