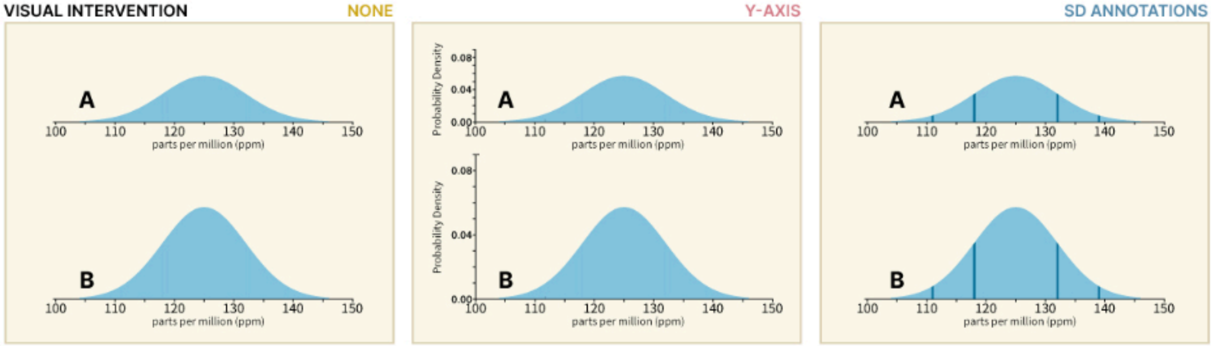
# Impact of Vertical Scaling on Normal Probability Density Function Plots

Racquel Fygenson and Lace Padilla



Which PDF shows a higher probability of being present at 130 ppm or more?

Fig. 1: Despite their difference in vertical height, PDFs A and B are all statistically identical.

**Abstract**— Probability density function (PDF) curves are among the few charts on a Cartesian coordinate system that are commonly presented without y-axes. This design decision may be due to the lack of relevance of vertical scaling in normal PDFs. In fact, as long as two normal PDFs have the same means and standard deviations (SDs), they can be scaled to occupy different amounts of vertical space while still remaining statistically identical. Because unfixed PDF height increases as SD decreases, visualization designers may find themselves tempted to vertically shrink low-SD PDFs to avoid occlusion or save white space in their figures. Although irregular vertical scaling has been explored in bar and line charts, the visualization community has yet to investigate how this visual manipulation may affect reader comparisons of PDFs. In this paper, we present two preregistered experiments ($n$ = 600, $n$ = 401) that systematically demonstrate that vertical scaling can lead to misinterpretations of PDFs. We also test visual interventions to mitigate misinterpretation. In some contexts, we find including a y-axis can help reduce this effect. Overall, we find that keeping vertical scaling consistent, and therefore maintaining equal pixel areas under PDF curves, results in the highest likelihood of accurate comparisons. Our findings provide insights into the impact of vertical scaling on PDFs, and reveal the complicated nature of proportional area comparisons.

**Index Terms**—visualization, probability density function, uncertainty, vertical scaling, perception, area chart

## 1 Introduction

Area encoding, which represents information through the size and shape of geometric regions, plays a pivotal role in many data visualizations [33]. In most approaches, areas are visually encoded through the proportional allocation of space. For example, bar charts typically represent larger amounts with taller bars, and treemaps represent larger percentages of a whole with bigger rectangles. This visual metaphor that equates a larger area with a larger amount of a plotted concept is informed by humans' learned experiences in the world [50]. Probability density functions (PDFs) are mathematical constructs that characterize a distribution of relative likelihoods for a range of possible outcomes. PDFs are ubiquitous in modern statistical education and employed outside of classrooms to communicate scientific results. Whereas some analytical results can also be encoded via other methods, results of Bayesian analyses often rely on PDF plots to convey prior and posterior beliefs [51]. In the case of communicating Bayesian results, misconstruing PDFs could drastically affect interpretation and lead to false conclusions that may propagate through research.

Typically, PDFs are visualized using an area chart, like those shown in Fig. 1. These PDF plots encode possible (continuous) outcomes along the x-axis, and probability density along the y-axis, such that the area under the curve represents 100% probability [46]. However, unlike other area charts, the y-axis of a PDF does not encode readily usable information. In Fig. 1, for example, the top and bottom PDFs in each column are statistically identical (normal distributions, $\mu = 125$, $SD = 7$), although one occupies much less pixel area than the other. To extract PDF plots' underlying probabilities, readers must estimate the proportions of the area under a plot's curve. Unfortunately, creators of PDFs may feel compelled to irregularly scale PDF plots such that each plot has the same height but different pixel areas (see Fig. 2, far right). The demand for this type of scaling is so common that statistical plotting packages, such as *bayesplot* [14] or *ggdist* [28], have presets to accommodate such a design decision.[1]

Vertical compression may be motivated by spatial constraints or aesthetic preferences. For example, consider a researcher who needs to present the results of their Bayesian analysis in a research paper. The researcher has nine different posterior distributions that they want to visualize using PDF plots, and a page limit constraint due to journal requirements. Initially, when plotting the results, the PDF plots overlap and occasionally occlude one another (see Fig. 2, a), making the figure hard to read and visually cluttered. The researcher contemplates spacing the plots to avoid overlap, as in Fig. 2 (b), but this design is space-inefficient. Ultimately, they decide to vertically compress the Bayesian posteriors to have the same height, as in Fig. 2 (c). Although this solution may satisfy the researcher, it introduces the risk of misguiding

• *Racquel Fygenson and Lace Padilla are with Northeastern University. E-mail: fygenson.r | l.padilla@northeastern.edu*

*Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx*

[1] `area_method = "equal_height"` in `mcmc_areas()` in bayesplot, `normalize = "xy"` in `stat_halfeye()` in ggdist

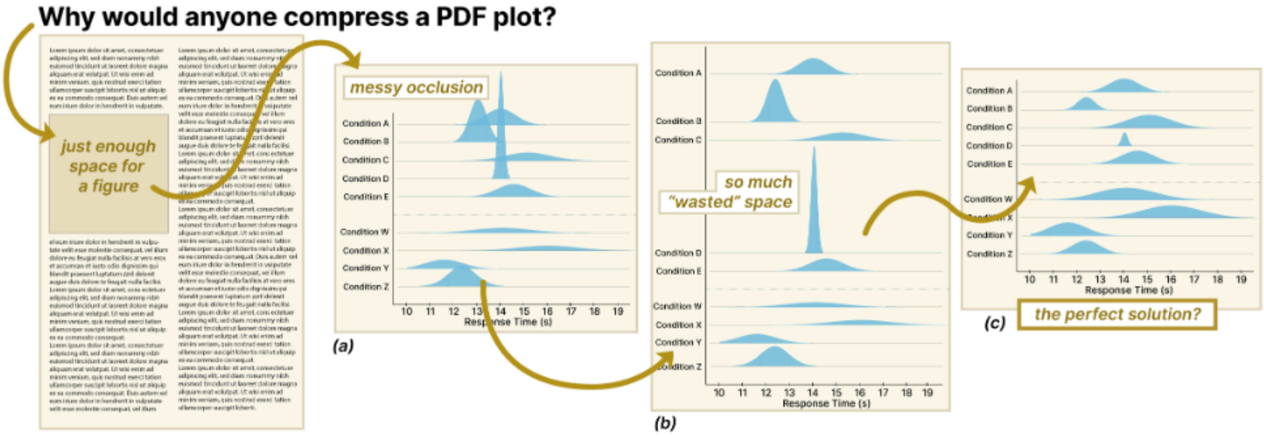


Fig. 2: A decision flow that may lead to visualization designers compressing some PDF plots at different rates. Findings from our experiments show that this "perfect solution" can lead to imperfect interpretation. Panel A shows equal-area PDFs with occlusion, B shows equal-area PDFs without occlusion, and C shows equal-height PDFs that have different areas.

readers because common visual encodings associate larger areas with greater quantities, potentially leading readers to incorrectly infer that shorter plots represent lower probabilities. In this paper, we investigate the impact of vertical scaling on readers' comparison of cumulative probabilities in PDFs.

We contribute two preregistered human-subjects experiments ($n$ = 600, $n$ = 401). Experiment 1 explores the effect of compressing the height and resulting area of PDF plots on readers' comparisons of cumulative probabilities (i.e., the probability of some, but not all, outcomes occurring – illustrated in the question text in Fig. 1). Experiment 1 investigates two levels of compression across statistically equivalent PDF plots and two visual interventions. Experiment 1 establishes that vertical compression of otherwise identical PDF plots can mislead readers. In practice, statistically identical PDF plots are not likely to be scaled differently because these plots already occupy the same height and pixel area. Experiment 2 prioritizes ecological validity by testing more realistic plot compression. We investigate the comparison of PDF plots with different standard deviations, scaled to equal heights—a method commonly preset in visualization software [14, 28]. Additionally, this experiment employs the same visual interventions as Experiment 1.

In our Discussion, we distill key design guidelines for practitioners who wish to present PDF plots that can be accurately compared to one another, particularly emphasizing the importance of avoiding irregular vertical compression altogether. Lastly, we discuss open questions surrounding PDF plot comprehension and outline potential future work.

These experiments mark important extensions of prior work. We expand past examinations of the risk of y-axis scaling from bar and line graphs to a more complex mark type. We also build upon perceptual studies that investigate the complicated nature of comparing irregularly-shaped areas, and provide experimental evidence to illuminate current theory and systems contributions within probability communication.

## 2 Background and Motivation

### 2.1 Area Judgments in Visualization

Perceptual comprehension of PDF plots has not been extensively studied; however, there is a strong body of research on human perception of other two-dimensional areas. For example, research has examined the comparison of rectangular areas [12, 32, 34, 47], irregular polygons [1] and ovals [12, 34, 60]. Notably, polygonal area estimation has been shown to be significantly more accurate when viewed in comparison to another equally-scaled, polygon than when viewed one-by-one [1]. Recent perceptual work has also distinguished that perceived area of shapes can differ from the actual pixel area of those shapes [59, 60], and that general dimensions of a set of shapes (i.e., "additive areas") may be used by readers as proxies for area judgment [59]. Other vision research has led to the hypothesis that perimeter length is used as a proxy for visually estimating the area of rectangles [32], and that reader judgment of circles' area and perimeter is significantly less accurate than their judgment of radius length [44], suggesting that readers might be prone to comparing straight-line distances, such as the height of PDF plots, over the area of nonrectilinear shapes.

In light of these findings, our investigation into the perceptual comprehension of PDF plots aims to bridge the gap between existing research on area measurement and the unique characteristics of PDF plot comparisons. The established reliance on perimeter length [32] and straight-line distances, coupled with the variability in perceived versus actual pixel area [59, 60], underscores the complexity of accurate area judgment in nonrectilinear shapes such as PDF plots.

### 2.2 Vertical Scaling in Visualization

Representing increasing quantities in graphs via upward vertical space reflects a universal convention observed across diverse human groups and historical periods [50]. A wealth of research supports the cognitive association between higher vertical positions and larger values [16, 31, 57]. This "More Is Up" concept suggests that numerical increases and vertical elevation align in intrinsic human perception and that representing growth of a quantity through increasing height can enhance reader clarity and understanding of graphical data [39, 45, 50].

Thus, visualizations commonly use difference in height of two visual objects to represent which object encodes larger values. Because the heights of visual elements can intrinsically represent quantities, literature on visualization best practices often advises against varying the vertical scaling of individual visual objects [19, 49]. Past research has explored how rescaling and truncating y-axes in line and bar charts can lead to misperceptions of encoded values and worse reader accuracy [6, 9, 37, 48]. Research also suggests that superimposing variables in a single graph with multiple y-axes (i.e., dual axis charts) should be avoided [21]. This body of work on vertical space in visualizations provides the basis for our hypothesis that inconsistent vertical scaling may decrease readers' ability to successfully compare PDF plots.

### 2.3 Vertical Stacking In Visualizations

Vertically stacking individual graphs is a commonly used layout to facilitate easy comparison of visualizations. However, vertically stacking PDF plots can lead to occlusion and significant space consumption, often prompting irregular compression of these plots, as discussed in Sec. 1. This problem is not unique to PDF plots; past research has investigated the impact of vertically stacking time series (i.e., line or area charts that encode a temporal variable along their x-axes [15, 17, 23, 40]), and stacking multiple line charts with a wide range of different aspect ratios [22]. In particular, research on time series visualizations has proposed solutions to enhance the compactness of multiple charts that require visual comparison. These solutions include horizon graphs, which cut large peaks and align them flush with the x-axis on top of their bases. Another option is braided graphs, which interweave

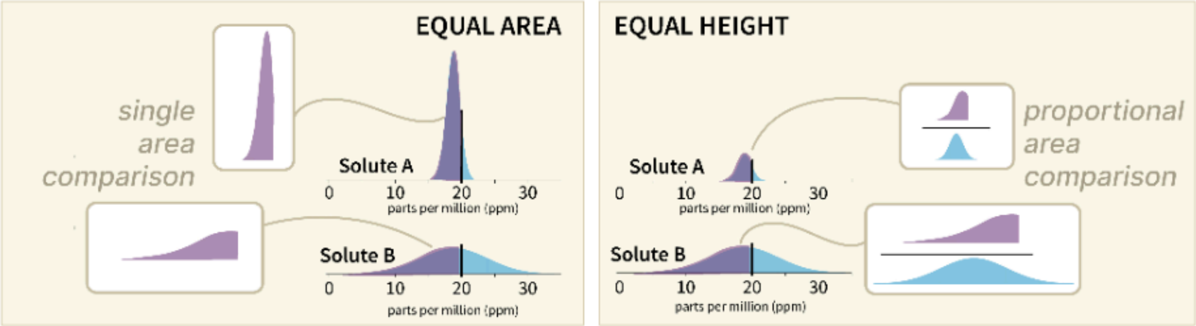


Fig. 3: Correct visual strategies for comparing cumulative probabilities of two PDFs. Left: Equal-area PDF plots can be compared via a single area comparison. Right: Equal-height PDF plots must be compared via a proportional area comparison.

multiple time series by dynamically ordering their segments, so they appear from the largest value at the back to the smallest value at the front of the composite graph [17, 23, 40]. These solutions save vertical space by truncating and moving lines such that the area between the lines and x-axis is minimized, which could make them less ideal for area-encoded charts, like PDF plots.

Vertical stacking of area charts that encode distributions has been explored in the form of ridgeline (a.k.a "joy" [30]) plots, which stack area charts such that their contours resemble the ridges of mountains [55]. Individually, ridgeline plots overlap with one another to save vertical space and facilitate comparison of "relative heights across groups" [55]. Thus, accurate comprehension of these plots is contingent on equal vertical scaling across all ridgelines; unlike PDF plots, irregular vertical scaling of ridgeline plots is canonically (and mathematically) incorrect. Ridgeline plots' overlapping can lead to occlusion in cases where distributions with a wide range of heights occupy similar horizontal positions. Although we explore only the impact of scaling PDF plots in this paper, further work comparing PDF plots and ridgeline plots may shed more light on best practices for communicating probability densities.

### 2.4 Probability Distribution Visualizations

Communicating probability distributions is commonly required to adequately relay experimental results [36]. There are several methods for doing so, each of which has different implications (for review, see [38]).

Confidence intervals (with and without indication of a distributional mean), box plots, and standard deviation intervals show only summarizing statistical moments of distributions to reduce visual complexity and facilitate easier comparisons. These methods are still used after many decades, taught in statistics curricula (e.g., [46]), and popular in scientific communication to experts and the general public [52]. These visualizations, however, do not depict distribution shape and reduce the statistical resolution of the information they communicate [10]. Confidence intervals and box plots also have been shown to fall prey to the *deterministic construal error*, in which readers disproportionately attribute more probability to values that lie inside delineated visual encodings than to those that lie right outside those encodings, essentially discounting the desired conveyance of uncertainty [10, 24].

PDF plots are another classic method for visualizing probability distributions by using area to encode distributional data. This method is part of a broader array of techniques that employ area to articulate distributional characteristics, including violin [18], ridgeline [56], and raincloud plots [3]. These plots maintain high statistical resolution and are less known to incur deterministic construal errors because of their continuous visual nature [10]. However, as we exhibit in this paper, many readers, especially those without strong levels of graph literacy, may not intuitively grasp important properties of PDF plots.

For scientists interested in making probability densities more accessible to an audience without strong statistical training, *frequency framing* may be of interest. Natural frequency framing (i.e., presenting probabilities as "6 times out of 100" instead of "6%") has been shown to be a more intuitive method for conveying uncertainty in textual contexts [11]. In recent years, visualization researchers have used this theory to inspire new visual encodings in which probability is conveyed discretely, such as quantile dot plots (QDPs) and hypothetical outcome plots (HOPs) [20, 29]. Both of these solutions are more likely to be correctly interpreted by members of the general public in some contexts [13, 20, 25, 26, 29]. At the same time, HOPs is an animated solution, making it incompatible with nondigital formats and requiring a larger amount of viewing time for more precise readings [29]. QDPs do not suffer from these drawbacks but sacrifice statistical detail in favor of discrete dots. These drawbacks are especially relevant for low (20)-quantile QDPs, which past evaluations have found to be more effective than higher (100)-quantile versions [29].

Although each of the aforementioned methods has its own advantages and disadvantages, PDF plots remain widely used for presenting experimental outcomes, particularly in Bayesian analysis, which yields posterior distributions as its statistical result. Moreover, past work details how individuals, including statistical experts, can struggle to read skewed PDF plots correctly [41]. We contribute to this research by investigating an unstudied, yet common, method of visually manipulating PDF plots and offering subsequent design recommendations.

### 2.5 Interpreting PDF Plots

Unlike standard Cartesian-coordinate-system plots, in which value can be derived from the y-axis positions of visual objects [53], correctly comparing PDF plots does not require referencing the y-axis [46]. In fact, attempting to estimate y-axis values along a PDF plot's curve can lead readers to mathematically incorrect conclusions. PDF plots convey probability via the area beneath their curves [46]. A distinct characteristic of PDF plots is that their total under-curve area must sum to one (i.e., unity) [46]. We hypothesize this assumption of unity can be de-emphasized, and sometimes not communicated at all, when the heights of neighboring PDF plots are rescaled to different degrees.

We are not the first to consider how the visual representation of density plots can mislead viewers. Pu and Kay define a "correct" probabilistic visualization as one in which the "proportions of visual elements (such as counts or areas) and their spatial placement reflect the underlying probability distribution, including any... part-to-whole relationships." [42]. Although individual PDF plots meet this requirement for correctness regardless of their scaling, multiple PDF plots that are scaled incongruently may not accurately reflect underlying probability distributions in a *perceptually* relative manner. Pu and Kay illustrate several examples where density estimates are plotted in a single frame without statistical corrections to account for the part-to-whole perceptual relationship that a singular frame with multiple interior pieces communicates. Pu and Kay posit that these are incorrect probability visualizations because they are likely to cause viewers to misinterpret the densities in their true context, and suggest that independently scaling y-axes by sample size is a necessary correction for appropriate visual comparison of part-to-whole density plots [42]. We propose an extension of Pu and Kay's notion of correctness to emphasize the importance of scaling in comparative probabilistic visualizations.

We describe two PDF plots that cover an equal amount of physical or pixel space beneath their curves as *equal-area* plots. In equal-area plots, the probability of some but not all outcomes occurring (i.e., the “cumulative probability" as shown in purple in Fig. 3) can be compared by contrasting the portions of areas that lie between outcomes of interest under each plot’s curve [46]. For example, on the left side of Fig. 3, the probability that Solute A presents at 20 ppm or less versus that of Solute B can be measured by visually determining which purple area is larger. On the right side of Fig. 3, where PDF plots have equal heights but not equal areas, to compare the same probabilities readers must visually determine which two-part area proportion is larger. Because of the multiple visual calculations it requires, the latter comparison may be more cognitively complex. If readers are comparing equal-height PDF plots, or other PDF plots without equal area, then comparing height, length, or total area can lead to false conclusions.

Other strategies may exist as well. For example, when comparing normal PDF plots, regardless of their area or height, readers familiar with the correlation of standard deviations and well-known probabilities in normal PDFs ($\mu \pm$ SD covers $\approx 68\%$ of the PDF, $\mu \pm 2$ SD covers $\approx 95\%$) can use these mathematical concepts to estimate general likelihoods of outcomes [46]. When comparing normal PDF plots, readers could also contrast the horizontal position of the plots’ means and SDs to comparatively estimate their cumulative probabilities.

PDF plots convey more information than just cumulative probabilities. The shape of PDFs can be used to convey probability distributions (e.g., large area concentrated around plot tails convey skewed probabilities). However, non-normal PDFs are drastically harder for people, even statistical experts, to interpret [41]. For this reason, we restrict our investigation to the comparison of normal PDF plots. However, in the Discussion section we note that future work should exploring accurate comparison of non-normal PDFs, so as to shed light on probability communication under a wide range of circumstances.

## 3 Experiment 1

We investigate the effects of vertically compressing PDF plots through two preregistered experiments. The first experiment[2] investigates readers’ comparisons of two statistically identical PDF plots that are scaled to different heights, and uses re-test conditions to examine potential misinterpretation. This experiment also explores two visual interventions and their impact in reducing possible misinterpretations.

### 3.1 Materials and Methods

#### 3.1.1 Investigative Questions and Hypotheses

We hypothesized that participants will be more likely to report that taller PDF plots (e.g., Fig. 1, bottom row) show a higher cumulative probability than their shorter counterparts and that we can mitigate this misconception with design interventions. Specifically, we hypothesized that **(H1)** adding a y-axis (e.g., Fig. 1, middle) and **(H2)** adding vertical lines to indicate standard deviations from the mean (SD annotations) (e.g., Fig. 1, right) will decrease the likelihood of incorrect comparisons. Lastly, we hypothesized that **(H3)** SD annotations will be more likely to reduce incorrectness than y-axes, but not to a large extent.

#### 3.1.2 Stimuli

We tested a range of paired normal PDF plots. Each pair either had both plots with 5, 7, or 9 SDs and the same mean, as shown in Fig. 4’s top row. We generated the plots using Python and edited them in Adobe Illustrator. We compressed each plot using Illustrator’s transform tool, generating versions that are 50% and 75% of the original height (Fig. 4, middle row). We explore the effect of two visual interventions: a y-axis that scales along with its PDF (Fig. 1, center) and lines demarcating $\pm 1$ and $\pm 2$ standard deviations (SDs) from the mean (Fig. 1, right). We hypothesized that the y-axis condition would further indicate vertical compression, possibly making the compression more salient and allowing readers to notice that the presented plots are statistically identical. Because the SD lines do not move as the heights of their plots are compressed, we hypothesized that these *SD annotations* may make it easier to notice the x-axis positioning of plots are statistically identical. To control for effects related to the visual location of the plots, we counterbalanced the composition (horizontal vs vertical stacking) and the position of the scaled plots as shown in Fig. 4’s bottom row.

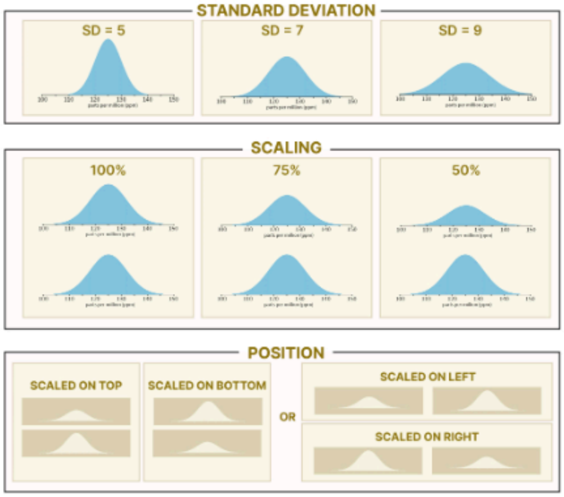


Fig. 4: Within subjects conditions in Experiment 1. Top row: Normal PDF plots with SDs of 5, 7, and 9. Middle row: scaling of 100% (equal area and height), 75%, and 50%. Bottom row: positions of compressed PDF plot on the top and bottom or left and right, depending on stacking.

#### 3.1.3 Experimental Design

We utilized a 3 (*Scaling*: 100% vs 75% vs 50%) x 3 (*Visual Intervention*: No Intervention vs Y-axis vs SDAnnotations) x 2 (*Stacking*: Horizontal vs Vertical) x 2 (*Position* of compressed PDF: Top/Left vs Bottom/Right) x 3 (*SD*: 5, 7, 9) mixed-subjects design. This design results in 15 graph combinations (3 with equally scaled plots and 12 with differently scaled PDF plots). The between-subject variables were *Stacking*—whether compared plots were horizontally or vertically faceted—and *Visual Intervention*—whether plots had a scaled y-axis, SD annotations, or no visual intervention—making six participant groups. The within-subject variable of interest was vertical *Scaling*, which varied the difference in height between compared PDF plots, either asking participants to compare two plots with equal scaling (100%), one plot that was scaled to 75% of the height of the other, or one plot that was scaled to 50% of the height of the other.

We included *Position* as a within-subject manipulation control for where the scaled plot was located. Thus, we ensured that the scaled graph occurred in all four locations (top, bottom, left, right). We included three standard deviations (*SD*=5, *SD*=7, and *SD*=9) to increase the number of trials and ensure test-retest reliability. We did not have predictions for *Position* or *SD* and considered these covariates in our analysis. Prior to this experiment, we ran a series of pilot studies to confirm question legibility and inform a power analysis to determine sample size. The pilot data inform our preregistration.

#### 3.1.4 Participants

In our pilot data, we observed a large effect in which participants were more likely to incorrectly compare PDF plots with unequal pixel areas. However, we anticipated a more conservative effect size for the visualization interventions, falling within the range of small to medium, according to Cohen’s guidelines [7]. We utilized the *pwr* package in R to determine the required sample size by specifying six degrees of freedom, a significance level of 0.05, a desired power of 0.8, and *f2* = 0.135 [5, 43]. This analysis indicated that we need a sample size of approximately 100 participants per between-subject group.

We recruited participants via the online platform Prolific.com. Participants were all above the age of 18, currently residing in the U.S., self-reported as fluent in English, had an approval rate $\geq 98\%$ on Prolific, had not participated in any of our pilot studies, and used desktop displays to complete the study.

We crowdsourced our participants so we could examine whether the general public has difficulty interpreting PDF plots with different

[2] https://osf.io/eu2th

modifications. This study primarily involves perceptual judgment, which is why we felt it was appropriate to use participants from the general public to examine this principle. Further, in our experiment we employ a scenario from [20], which crowdsourced its participants. It is possible that more educated participants may not exhibit the same biases, and future research should examine the relationship between statistical knowledge and the results observed in this study.

## 3.2 Procedure

Participants completed Experiment 1 online via Qualtrics [8] on their personal machines. After giving IRB-approved consent to participate in the study, if participants successfully completed an attention check, they were asked to "please make [their] screen window as large as possible." Next, participants were shown the definition of a solute and provided with a scenario in which scientists measured the concentration of samples of sea water and generated corresponding plots. We adopted this scenario from previous research on how a general audience interprets probability distributions through HOPs [20]. Participants were then shown an example PDF plot with basic "how-to-read" instructions. They then answered 15 multiple-choice questions, one about each pair of charts shown, which were presented in a randomized order. Each of these multiple choice questions asked participants, "Which solute, if either, has a higher probability of being present at X or more ppm in the sampled sea water?". The threshold X ppm was in the same position across all PDF plots (at 130ppm as shown in Fig. 1) but varied in actual number depending on the x-axis labeling of the stimuli. All pairs of PDFs were statistically identical, so the correct answer to all multiple-choice questions was that **Solute A and B have the same probability**. We selected this task to investigate if a scaling-related bias occurs; the task requires readers to consider the peaks of PDF plots, which are affected by vertical scaling. This task is not representative of all PDF plot use cases, nor is it a quintessential use case.

Following the experiment, participants completed a short graph literacy test [35], and finally answered demographic questions. The full survey is available in our Supplemental Materials.

### 3.2.1 Analysis

Our preregistered analysis[3] consists of several binomial Bayesian models. Two of the preregistered models investigate participants' strategy–whether participants were more likely to report taller or shorter PDF plots as showing a higher cumulative probability. Across all conditions, these models show that participants were much more likely to indicate that taller plots depicted higher cumulative probabilities. Although useful for considering perceptual tactics, we focus this paper on the more complex results stemming from our model of binary accuracy. We include the strategy models in our Supplemental Materials.

Below, we focus on the analysis for evaluating Experiment 1's accuracy across variables. To do so, we utilize the R packages *tidyverse* v. 2.0.0 for data processing [54], *brms* v. 2.20.4 for Bayesian modeling [4], and *tidybayes* v. 3.0.6 for data processing and visualization [27, 43].

We assess the amount of variance in binary correctness explained by the interaction between *Scaling* and *Visual Interventions*, as well as their lower-order terms (levels of each variable described in Sec. 3.1.3). Additionally, we account for the variance explained by *Stacking*, *Position*, *Graph Literacy*, and *SD* in the following model:

$$\begin{aligned} BinaryCorrectness \sim\ & Scaling \times Visual\,Interventions \\ & + Stacking + Position \\ & + Graph\,Literacy + SD + (1|ID). \end{aligned} \tag{1}$$

We evaluate correctness with a response of "A and B have the same probability" coded as 1 and all other responses coded as 0. Our model specified an interaction term between Scaling and Intervention so that we could test all of Experiment 1's hypotheses (H1 - H3). We include *Stacking*, *Position*, *Graph Literacy*, and *SD* as covariates to account for their potential meaningful effect on our effects of interest.

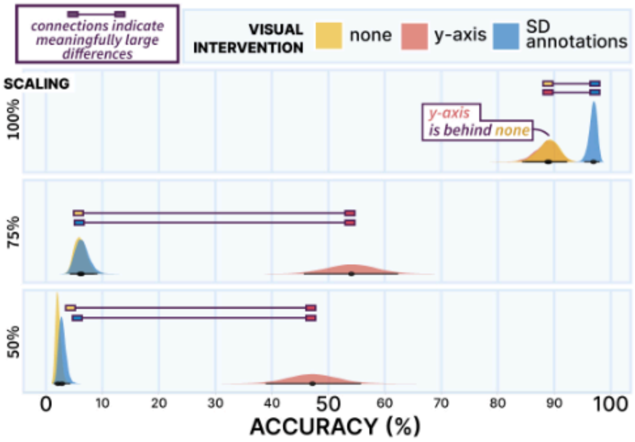


Fig. 5: Bayesian posteriors of accuracy as determined by scaling of PDF plots (rows) and visual interventions (colors).

For all models, we include random intercepts for participants. Our model specifications include uninformative priors centered at 0, with a standard deviation of 2.5. Our preregistered model originally included priors center at .5, which we corrected here. To assess the impact of an effect, we utilize 95% credible intervals, considering predictors with credible intervals excluding zero as having a reliable effect on participants' judgments. Additionally, for all models, we analyze results including and excluding participants who fail to pass the attention check. We preregistered a possible exclusion of these participants, if a sensitivity analysis indicates they significantly skew results.

## 3.3 Results

### 3.3.1 Participants

We collected a total of 600 participants ($n$ = 100 per between-subject condition). 297 were female, 290 male, 11 nonbinary, and 2 opted not to say. The median age was 38.5 years (*mean* = 40.4, *SD* = 12.9), and the median short graph literacy score was 2 out of 4 (*mean* = 2.3, *SD* = 1.1). The median survey completion time was 8 minutes and 46 seconds, making the average compensation $13.58/hr.

### 3.3.2 Binary Correctness

**Examining H1.** To examine if including a **y-axis** meaningfully decreases the number of participants that incorrectly interpret differently scaled PDF plots, we analyzed the accuracy of participant responses using Eq. (1). We computed this model with and without participants who failed an attention check and found no marked differences between the groups. Below we report the results of the entire sample for a more conservative statistical analysis.

Our model for Binary Correctness reveals interactions between all of the *Scaling* and *Visual Intervention* conditions. See Fig. 5 for a visual description of posteriors. Overall, the results reveal that irregular vertical scaling drastically decreases binary accuracy, which can be mitigated by the **y-axis** condition.

To investigate the model's interactions, we change its referents to each of the *Scaling* and *Visual Intervention* conditions, as is recommended to examine interaction effects [2]. These analyses reveal that the interactions are driven by two distinctly different relationships between *Visual Interventions* and *Scaling*. The model outputs for these comparisons at each level of *Scaling* are shown in Tab. 1. Firstly, as depicted in the top row of Fig. 5, each visualization condition has a high level of accuracy when comparing two distributions of equal area (100% scaling). Further, there is no evidence for a difference between the **no intervention** condition and the **y-axis** scale (see Tab. 1 first row). However, there is a meaningful difference between the **SD annotation** and **y-axis**, and **SD annotation** and **no intervention** conditions (second and third rows of Tab. 1). Fig. 5 is annotated with these results. In contrast, the response patterns when the distributions were not equally scaled (75% and 50%) are drastically different. For both the 75% and 50%

[3] https://osf.io/eu2th

| Scaling | Intervention Comparison | Est. | l-95% CI | u-95% CI |
|---|---|---|---|---|
| 100% | Y-axis x None | 0.02 | -0.53 | 0.56 |
| 100% | Y-axis x SD Annot. | 1.34 | 0.75 | 1.92 |
| 100% | SD Annot. x None | -1.19 | -1.76 | -0.61 |
| 75% | Y-axis x None | -2.89 | -3.40 | -2.40 |
| 75% | Y-axis x SD Annot. | -2.81 | -3.33 | -2.31 |
| 75% | SD Annot. x None | -0.13 | -0.66 | 0.40 |
| 50% | Y-axis x None | -3.67 | -4.23 | -3.12 |
| 50% | Y-axis x SD Annot. | -3.34 | -3.89 | -2.80 |
| 50% | SD Annot. x None | -0.42 | -1.03 | 0.16 |

Table 1: Breakdown of interaction effects in Model 1 (Binary Correctness) across scaling levels. Rows compare visual intervention conditions at three scaling levels and show 95% credible intervals in log odds. Darker cell backgrounds indicate stronger effects.

*Scaling* conditions, the **y-axis** intervention's accuracy was substantially higher than that of the other two visual interventions. Moreover, no evidence for differences was detected between the **SD annotations** and the **no intervention** conditions within the 75% or 50% scaling contexts. We detail the direct comparisons in Tab. 1 and annotate them in Fig. 5. The improvement in accuracy among participants reviewing PDF plots with a **y-axis** in the 75% and 50% conditions **provides evidence for H1**. This enhancement in performance validates the utility of y-axes in PDF plots under specific scaling conditions.

**Examining H2.** To examine if adding vertical lines to indicate standard deviations would improve comparison of differently scaled PDF plots, we can look at the direct comparisons in Tab. 1 (rows 6 and 9). For both 75% and 50% conditions, we find no meaningful difference between the **SD annotation** condition and **no intervention**. Interestingly, for the 100% condition, accuracy is higher for the **SD annotation** than the other two conditions. However, overall, we **do not find sufficient evidence for H2**, finding that **SD annotations** do not consistently improve accuracy. It is worth reiterating **SD annotation's** positive impact on correctness (top-right corner of Fig. 5), indicating a potential use for **SD annotations** when PDF plots are visually similar.

**Examining H3.** To investigate whether **SD annotations** improve accuracy more than **y-axes** when comparing differently scaled plots, we can look to the direct comparisons in Tab. 1 (rows 5 and 8), which shows evidence for the effect that **y-axes** meaningfully improve accuracy over **SD annotations** for the relevant 75% and 50% scaling conditions. Therefore, we **document no support for H3**.

**Covariates.** We observe no evidence for effects of the covariates *Stacking* or *SD* on participant accuracy in this model. We note the covariates *Position* and *Graph Literacy* account for a meaningful proportion of variance in the model's outcomes. All the effects we report control for the effect of these covariates. We describe some of the effects of *Graph Literacy* in Sec. 5.1, and the full model output detailing all the effects is in the supplemental materials.

# 4 Experiment 2

## 4.1 Methods

The results from Experiment 1 reveal that, at a minimum, roughly 50% of participants incorrectly deduced that vertically compressing PDF plots decreases the probability shown. Experiment 1 varies only vertical height, which in turn varies area; thus, its experimental design confounds pixel area and height. Additionally, Experiment 1 tests identical PDFs, which are unlikely to be irregularly scaled in practice. To explore the effects of compressing PDF plots in a more realistic layout, we preregistered Experiment 2[4].

In this experiment, we asked participants to compare the cumulative probabilities of PDF plots with different SDs, but which have been scaled to occupy the same pixel height (i.e., *equal-height* PDF plots, as shown in Fig. 6). We also test this comparison against a control in which participants compare uncompressed, "equal-area" PDF plots with different SDs, and, thus, different heights. Experiment 2's comparison investigates a design decision that may occur when chart creators are keen to save space and do not want to allocate a great deal of vertical white space to charts with small SDs, as illustrated in Fig. 2 (c).

### 4.1.1 Investigative Questions & Hypotheses

Motivated by our results from Experiment 1, we investigated if participants may be conflating pixel area with probability. To do so, we varied the SDs of compared PDFs (1 SD & 5 SD, 2 SD & 5 SD, 3SD & 5 SD, 4 SD & 5 SD, 5 SD & 5 SD – see Fig. 6) and hypothesized that **(H4)** participants would be more likely to incorrectly compare cumulative probabilities of two PDFs when their plots had a large difference in SDs. We reasoned that PDF plots with larger differences in SDs would have larger differences in pixel area, potentially misleading more participants to incorrectly choose the visually larger plots.

We tested the same visual interventions as in Experiment 1, which we again hypothesized **(H5)** would mitigate some of the inaccuracy from vertically compressing PDF plots to have equal-heights (Fig. 6, right columns). Although we saw only the y-axis intervention increase accuracy in Experiment 1, we reasoned that perhaps the SD annotations would prove more useful when comparing PDF plots of varying SDs.

Lastly, we included a control *equal-area* condition in which we asked participants to compare PDF plots with different SDs that were proportionally scaled to occupy the same pixel area, as is traditional practice when visualizing PDFs (Fig. 6, *equal area*). We hypothesized that **(H6)** participants would be more likely to accurately compare cumulative probabilities when PDF plots occupied equal areas.

### 4.1.2 Stimuli

Experiment 2 compares PDF plots with varying SDs. We created all plots using similar methods and tools to Experiment 1. We detail the reasoning for each stimulus in Sec. 4.1.1.

### 4.1.3 Experimental Design

We utilized a 4 (*Visual Intervention*: Equal-area, Equal-height, Y-axis, SD Annotations) x 5 (*SD Pairs*: 1-5 SD, 2-5 SD, 3-5 SD, 4-5 SD, 5-5 SD ) x 2 (*Position* of smaller-SD PDF: Top, Bottom) mixed-subjects design. *Visual Intervention* was Experiment 2's only between-subjects condition. *SD Pairs* and *Position* were within-subjects conditions. In Experiment 1, we found no meaningful effect from stacking plots vertically or horizontally, so in Experiment 2, we tested only vertically stacked plots. Vertical stacking is more likely to result in occlusion from tall, low-SD PDFs overlapping (see Fig. 2), and thus is more likely to motivate designers to compress plot heights. This design resulted in 9 graph combinations (4 *SD Pairs* with *Position* = Top, 4 *SD Pairs* with *Position* = Bottom, and *1 SD* Pair = 5-5 SD with Position = N/A).

### 4.1.4 Participants

We used the effect size and power analysis from Experiment 1 to inform our preregistered sample size of 100 participants per between-subjects group. We recruited participants using the same criteria as in Sec. 3.1.4.

### 4.1.5 Procedure

Experiment 2 mimicked the procedure of Experiment 1. We provided all the same instructional information and added additional labeling to the SD annotation condition to indicate its statistical implications. To accommodate the PDF plots in Experiment 2, we also asked a slightly different cumulative probability question. We asked participants, "Which solute, if either, has a higher probability of being present at X or less ppm in the sampled seawater?" As in Experiment 1, the threshold X was set at the same position across all plots–at $x = 20$ in Fig. 6 plots–but varied in number depending on x-axis labeling.

Participants were again instructed to select one of three possible multiple choices: "Solute A has a higher probability", "Solute B has a higher probability", or "Solute A and B have the same probability". In almost all pairs of plots, the PDF with the lower SD was the correct answer. This distribution had more probability concentrated around its mode. For the stimulus with the *SD Pair* 5-5 SD, the correct answer was that the plots show the same probability. Experiment 2 participants

[4] https://osf.io/uxb7n

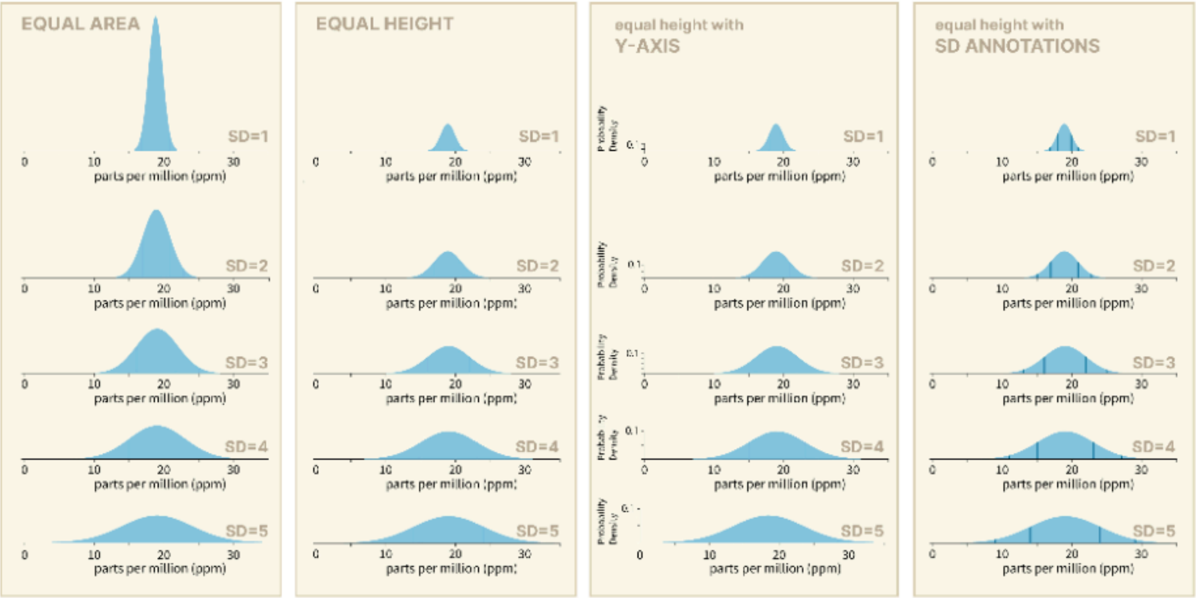


Fig. 6: Visual Intervention conditions shown between subjects. SDs vary from 1 at the top of each column to 5 at the bottom, and each pair of SDs is compared within subjects.

| Interaction | Est | l-CI | u-CI |
|---|---|---|---|
| 5/5 v 4/5 x Equal-area v Equal-height | -2.27 | -3.20 | -1.36 |
| 5/5 v 3/5 x Equal-area v Equal-height | -1.65 | -2.57 | -0.75 |
| 5/5 v 2/5 x Equal-area v Equal-height | -1.61 | -2.51 | -0.70 |
| 5/5 v 1/5 x Equal-area v Equal-height | -1.62 | -2.51 | -0.73 |
| 5/5 v 4/5 x Equal-area v Y-axis | -1.17 | -2.07 | -0.29 |
| 5/5 v 3/5 x Equal-area v Y-axis | -0.72 | -1.61 | 0.14 |
| 5/5 v 2/5 x Equal-area v Y-axis | -0.75 | -1.63 | 0.11 |
| 5/5 v 1/5 x Equal-area v Y-axis | -1.15 | -2.06 | -0.27 |
| 5/5 v 4/5 x Equal-area v SD Lines | -1.97 | -2.89 | -1.05 |
| 5/5 v 3/5 x Equal-area v SD Lines | -1.85 | -2.77 | -0.95 |
| 5/5 v 2/5 x Equal-area v SD Lines | -1.94 | -2.86 | -1.05 |
| 5/5 v 1/5 x Equal-area v SD Lines | -1.75 | -2.66 | -0.85 |

Table 2: Interactions for referents 5-5 SD Pair & Equal Area from Model 2. Rows show 95% credible intervals in log odds. Darker cell backgrounds indicate stronger effects.

filled out the same postexperiment information as those in Experiment 1. The full survey is available in our Supplemental Materials.

### 4.1.6 Analysis

We preregistered a binomial Bayesian model to test our hypotheses in Experiment 2[5]. Using the same packages and logic as in Experiment 1, we assess Binary Correctness as follows:

$$\begin{aligned} Binary\,Correctness \sim Visual\ Interventions \times SD\ Pairs \\ + Position + Graph\,Literacy + (1|ID). \end{aligned} \tag{2}$$

This model specifies an interaction term between *Visual Intervention* and *SD* Pairs so that we could test all of Experiment 2's hypotheses (H4 - H6). We include *Position* and *Graph Literacy* as covariates to control for their potential meaningful effect on our effects of interest. We code correct answers (as described in Sec. 4.1.5) as 1 and incorrect answers as 0. We also include participants as random intercepts and specify the same priors and credible intervals as in Experiment 1. We evaluate this model with the full sample of participants and with a population that excludes participants who failed a simple attention check, again preregistering a possible exclusion criteria if participants who fail the attention check meaningfully impact results.

## 4.2 Results

### 4.2.1 Participants

We collected a total of 401 participants ($n$ = 101 for SD Annotation intervention, $n$ = 100 for each other between-subject conditions). 195 were female, 197 male, 7 nonbinary, and 2 opted not to say. The median age was 37.0 years (*mean* = 38.8 years, *SD* = 13.8 years), and the median short graph literacy score was 3 out of 4 (*mean* = 2.4, *SD* = 1.1). The median survey completion time was 8 minutes and 31 seconds, making the average compensation $10.32/hr.

### 4.2.2 Binary Correctness

We investigate our hypotheses using our Bayesian model, setting **equal area** as the referent for *Visual Interventions* and 5-5 SD as the referent for *SD pair*. We see very few differences between the model results using data from the entire tested population and those using data from just participants who pass the attention check. In this paper, we discuss only the more statistically conservative results of the entire population, but include the second model in our Supplemental Materials.

**Examining H4.** To evaluate whether participants' accuracy declines as SD disparity increases between compared plots, we employ Eq. (2). This analysis unveils numerous meaningful interactions between visual interventions and specific SD pairs, as detailed in Tab. 2. To decode these interactions, we systematically adjusted the model's referents to each visual intervention and SD pair condition combination. These adjustments allowed us to examine participants' accuracy by assessing the influence of varying SD pairs across visual interventions.

Fig. 7 illustrates a uniform increase in accuracy when transitioning from SD pairs 4-5 to 1-5. Our analysis reveals that this trend is consistently observed across all visual interventions–specifically, a meaningful difference in accuracy occurs between comparisons separated by two levels. For instance, the accuracy for SD pairs 4-5 is smaller than that of 2-5 and 1-5. Similarly, the accuracy of 1-5 was meaningfully larger than that of 3-5, with the sole exception of comparisons with **y-axes**. The comparisons between the SD pairs for each visual intervention are documented in Tab. 3, and those with credible intervals that do not include zero are annotated at the bottom of Fig. 7.

These empirical findings substantiated a meaningful effect, albeit in a direction opposite to that which we originally hypothesized, thereby **providing no evidence for H4**. Interestingly, we consistently find that participants' accuracy improves when there are larger differences between the standard deviations of the two PDF plots they are compar-

[5] https://osf.io/uxb7n

ing. Future work stands to benefit from investigating the relationship between PDF's SDs and their accurate comparison.

**Examining H5.** To examine if the visual interventions we tested improve accuracy, we look at the comparisons in our model between the visualization types that we revealed in the previous interaction examination. Firstly, we observe no evidence for a main effect between visual interventions at the level of our referent *SD Pair* (5-5 SD). As the top row of Fig. 7 shows, accuracy is relatively high for this *SD pair* condition, with only marginal differences between visual interventions.

Secondly, we observed a noteworthy pattern in our analysis of SD pairings that are not equivalent (4-5 SD through 1-5 SD): the **equal-area** plots consistently outperformed most other visual interventions. Specifically, for the SD pairs (1-5 SD) and (4-5 SD), the **equal-area** plots yielded superior accuracy compared to all other visual interventions. However, for SD pairs 2-5 SD and 3-5 SD, only negligible differences between the **equal-area** visualization and the **y-axis** intervention were detected. We annotate these effects in Fig. 7 and show the meaningful comparisons in Tab. 4.

Further examination of Fig. 7 reveals a consistent ranking among the visualizations for SD pairs ranging from 1-5 SD to 4-5 SD. Here, the **equal-height** plots typically underperformed or were comparable to **SD annotations**, followed by the **y-axis** condition, and finally the **equal-area** visualizations. Focusing on the **equal-height** PDF plots, we find that at the level of 4-5 SD, it performs meaningfully worse than the **y-axis** and **equal-area** interventions. For 3-5 SD, 2-5 SD, and 1-5 SD, the **equal-height** plots perform only markedly worse than the **equal-area** plots. Consequently, we find **some support for H5**, only within the context of **y-axis** performance outperforming the **equal-height** plot in some cases. In other cases, the **y-axis** condition performs as well as the **equal-area** plot. We also **do not find sufficient evidence for H5 in the context of SD annotations**, which did not show sizable improvement from **equal-height** plots.

**Examining H6.** We can use the previously described analysis to also examine if **equal-area** PDF plots increase accuracy over **equal-height** PDF plots when the *SD pair*s are not equal (i.e., not 5-5 SD). We find that the **equal-area** visual intervention meaningfully improves the likelihood of accuracy compared to all of the other visual interventions for *SD pair*s 1-5 SD and 4-5 SD. Further, we find that **equal-area** plots meaningfully improve the likelihood of accuracy over **equal-height** and **SD annotations** for 2-5 SD and 3-5 SD. These conclusions **provide evidence for H6**, finding that **equal-area** versions of PDF plots produce the best performance out of the visual interventions we tested.

**Covariates.** We observe no evidence for effects of the covariates *Position* or *Graph Literacy* in Model 2.

## 5 Discussion

In this paper, we provide empirical evidence of the impact of irregular vertical compression on PDF comparisons. Although nonuniform vertical scaling of visualizations is generally thought to mislead readers, previous work on vertical scaling has examined only line and bar charts [6, 9, 37, 48]. Similarly, previous work on area comparisons is largely reserved to ovals and polygons [1, 12, 32, 34, 47, 60]. We explore a manipulation that is not uncommon in PDF visualizations–as can be seen by its implementation into presets in popular statistical plotting packages [14, 28]–and highlight its potential to mislead readers, along with some mitigating tactics.

The two experiments we present in this paper vary the width and height of PDF plots, not only providing general design guidance, but also generating insight about the implications of mapping probability to area. Our findings, although specific to compressing PDF plots, can shed light on the mental strategies that lay audiences use to make sense of area-encoded probability visualizations. For example, we present strong evidence from both experiments that PDF plots with equal areas are more consistently correctly interpreted than PDF plots with equal heights or other vertical compression. These patterns could indicate that, even when informed of the unity of each PDF plot, individuals rely on single area comparisons to make judgments about probabilities instead of the mathematically correct strategy of proportional area judgments, as depicted in Fig. 3, right. Future research should investigate whether this finding is attributable to readers' failure to recognize PDF plots as continuous part-to-whole visualizations, to the increased difficulty of single area judgments compared to two-part proportional area judgments, or to a combination of both factors. Regardless of underlying reasons, this work provides key evidence of the complicated nature of interpreting and comparing PDF plots.

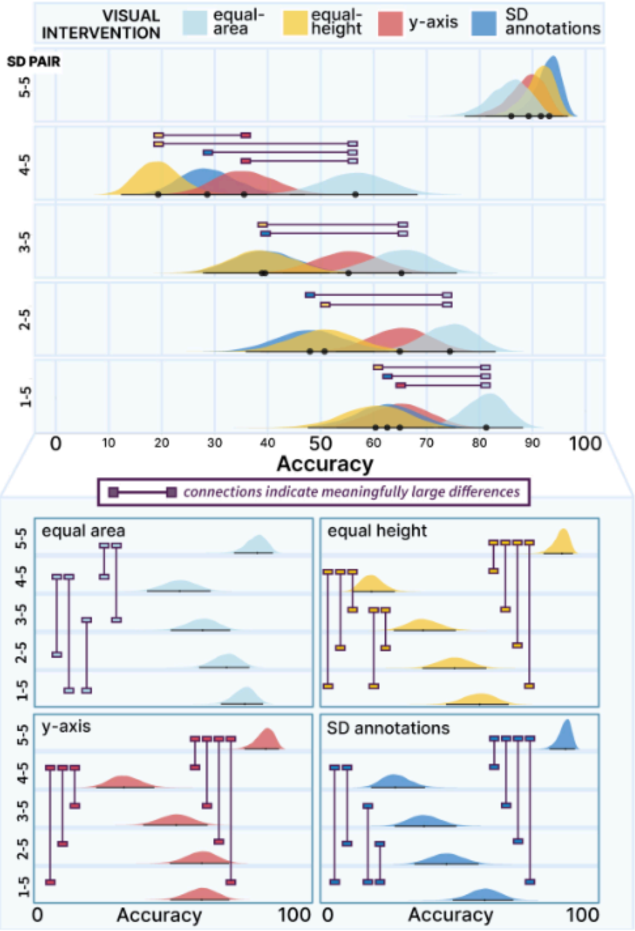


Fig. 7: Bayesian posteriors of accuracy as determined by *SD Pairs*, and *Visual Interventions* (colors). Rectangular connectors indicate comparisons where the credible interval do not include zero within and across *SD Pairs* and *Visual Interventions*.

### 5.1 Implications for Visualization Design

In practice, probability densities that need to be visualized will most likely not share identical means and standard deviations. Experiment 2 demonstrates how **equal-height PDF plots are susceptible to incorrect comparison** and to what degree y-axes and SD annotations can improve comparison. We find that, regardless of visual intervention, **equal-area PDF plots lead to more accurate probability comparisons** than equal-height PDF plots. Frustratingly, this recommendation can require large amounts of dedicated space, or can result in visual occlusion from PDF plot overlap. In some cases, the overlap may not seem inhibiting, although the range of effects from slight to extreme PDF plot overlap, like that in ridgeline plots, is yet to be explored.

If visualization designers are compelled to compress PDF plot heights, which our experiments indicate is inadvisable, adding a y-axis could mitigate some miscomprehension. This signal was especially strong in Experiment 1, in which we found an effect of graph literacy. To understand the impact of graph literacy on visual interventions, we conducted an exploratory analysis by adding an interaction term between visual intervention and graph literacy in Experiment 1. This analysis revealed that the influence of graph literacy becomes espe-

| SD Pair Comparison | Intervention | Est. | l-95%CI | u-95%CI |
|---|---|---|---|---|
| 1-5 v 3-5 | Equal-area | -0.77 | -1.29 | -0.28 |
| 1-5 v 4-5 | Equal-area | -1.14 | -1.64 | -0.64 |
| 2-5 v 4-5 | Equal-area | -0.77 | -1.27 | -0.27 |
| 3-5 v 5-5 | Equal-area | 1.02 | 0.39 | 1.66 |
| 4-5 v 5-5 | Equal-area | 1.42 | 0.80 | 2.06 |
| 1-5 v 3-5 | Equal-height | -0.87 | -1.36 | -0.39 |
| 1-5 v 4-5 | Equal-height | -1.85 | -2.35 | -1.35 |
| 1-5 v 5-5 | Equal-height | 2.02 | 1.34 | 2.73 |
| 2-5 v 3-5 | Equal-height | -0.50 | -0.98 | -0.01 |
| 2-5 v 4-5 | Equal-height | -1.48 | -1.98 | -0.98 |
| 2-5 v 5-5 | Equal-height | 2.38 | 1.71 | 3.08 |
| 3-5 v 4-5 | Equal-height | -1.01 | -1.52 | -0.49 |
| 3-5 v 5-5 | Equal-height | 2.85 | 2.17 | 3.58 |
| 4-5 v 5-5 | Equal-height | 3.72 | 3.02 | 4.43 |
| 1-5 v 4-5 | Y-axis | -1.24 | -1.72 | -0.77 |
| 1-5 v 5-5 | Y-axis | 1.51 | 0.86 | 2.18 |
| 2-5 v 4-5 | Y-axis | -1.18 | -1.67 | -0.70 |
| 2-5 v 5-5 | Y-axis | 1.58 | 0.93 | 2.26 |
| 3-5 v 4-5 | Y-axis | -0.79 | -1.26 | -0.31 |
| 3-5 v 5-5 | Y-axis | 1.97 | 1.31 | 2.66 |
| 4-5 v 5-5 | Y-axis | 2.72 | 2.04 | 3.41 |
| 1-5 v 2-5 | SD Annot. | -0.60 | -1.08 | -0.13 |
| 1-5 v 3-5 | SD Annot. | -0.94 | -1.42 | -0.47 |
| 1-5 v 4-5 | SD Annot. | -1.44 | -1.91 | -0.95 |
| 1-5 [illegible] | SD Annot. | 2.15 | 1.44 | 2.86 |
| 2-5 [illegible] | SD Annot. | -0.88 | -1.37 | -0.39 |
| 2-5 [illegible] | SD Annot. | 2.68 | 1.97 | 3.43 |
| 3-5 [illegible] | SD Annot. | -0.53 | -1.02 | -0.05 |
| 3-5 [illegible] | SD Annot. | 3.04 | 2.34 | 3.80 |
| 4-5 v 5-5 | SD Annot. | 3.52 | 2.80 | 4.28 |

Table 3: SD comparisons with 95% credible intervals that do not include zero from testing H4, broken down by Visual Intervention. The first SD condition listed in each row is the referent. Units are in log odds and darker cells indicate stronger effects.

cially impactful for participants when a y-axis is present, as shown in Fig. 8. This finding implies that for audiences with a high level of graph literacy, adding y-axes to compressed PDF plots could mitigate misunderstandings to some degree. However, considering the more minimal impact of the y-axis intervention in Experiment 2, our results about its usefulness for a general audience are mixed. Also, a background grid may function similarly to y-axes by making visual compression more obvious, but the effects of other visual interventions still need to be explored. Finally, we find **little evidence that standard deviation annotations are useful.**

Beyond the visual interventions that we discuss in this paper, designers could also encode densities with a different mark type. For example, distributions with large SDs may be too short to be visible when uniformly scaled along with other PDF plots. In these cases, designers may consider visualizations that do not encode probability with height, such as gradient plots [10] or dual histogram intervals [58]. Although mark types that do not encode cumulative probability with area are likely unaffected by the misconceptions highlighted in this paper, we caution readers that they may exhibit other unknown misconceptions.

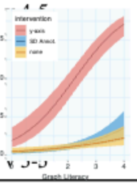


Fig. 8: Conditional effects for graph literacy in Exp. 1. Darker lines show means, color bands show 95% CIs.

| Visual Intervention Comparison | SD Pair | Est. | l-95%CI | u-95%CI |
|---|---|---|---|---|
| Equal-area v Equal-height | 1-5 | -0.97 | -1.68 | -0.25 |
| Equal-area v Equal-height | 2-5 | -0.99 | -1.70 | -0.28 |
| Equal-area v Equal-height | 3-5 | -1.05 | -1.77 | -0.36 |
| Equal-area v Equal-height | 4-5 | -1.62 | -2.33 | -0.94 |
| Equal-area v Y-axis | 1-5 | -0.76 | -1.44 | -0.06 |
| Equal-area v Y-axis | 4-5 | -0.80 | -1.49 | -0.13 |
| Equal-area v SD Annot. | 1-5 | -0.87 | -1.59 | -0.17 |
| Equal-area v SD Annot. | 2-5 | -1.09 | -1.79 | -0.41 |
| Equal-area v SD Annot. | 3-5 | -1.03 | -1.74 | -0.34 |
| Equal-area v SD Annot. | 4-5 | -1.11 | -1.81 | -0.44 |

Table 4: Breakdown of meaningful interaction effects in Model 2 by SD Pair. Rows compare Visual Interventions per SD Pair and show 95% credible intervals in log odds. Darker cells indicate stronger effects.

## 5.2 Limitations and Future Work

We limit the scope of this investigation to normal probability distributions, which is not the entire set of PDFs that science communicators might need to present. We do this in part because non-normal PDFs can be challenging to read [41], perhaps because they do not adhere to symmetry or well-known percentage-to-SD ratios. In the future, it would be worthwhile to investigate reader comparison of non-normal and nonsymmetric PDF plots. However, given that these plots are even more challenging to read [41], it is possible that additional errors may arise. The results we present here act as preliminary motivation for further investigating how manipulations of distributional area plots affect reader comprehension. The studies we present indicate a difference in accuracy dependent on vertical scaling, but do not advise on the mental strategies that produce this difference. Future work is needed to investigate the mechanisms behind this observed loss of accuracy.

Additionally, the two experiments we present hold many variables constant, leading to cleaner signals but reduced ecological validity. It would be useful to evaluate how readers compare more than two PDF plots, especially those that are aligned along a single x-axis, like we use to communicate our results in the top of Fig. 7. Investigating vertically compressed PDFs across multipage reports could also lend interesting, ecologically valid findings to this body of research.

We also have yet to explore how PDF plots' height-to-width ratio impacts the perception of them individually. Future work could ask, 'How does perceived certainty change as the height-to-width ratio of a singular PDF plot shifts?' There may be an optimal height-to-width ratio for PDF curves that has yet to be uncovered. Lastly, future work could build on our experimental design by evaluating analogous manipulations to raincloud, ridgeline, violin, and quantile dot plots, as well as other visualizations that encode probability via area.

## 6 Conclusion

In this paper, we contribute evidence of the impact of compressing PDF curves on reader comprehension. Specifically, we find that equal-area PDF plots consistently result in more accurate comparisons than their equal-height counterparts. We also test potential visual interventions to improve the accuracy of comparing differently compressed PDF plots. In some cases, we find adding y-axes can improve the accuracy of comparisons of compressed plots. In most cases, we find adding standard deviation annotations impacts the accuracy of comparisons very little. Our experimental data also informs base standards for the accuracy of readers' comparisons of cumulative probabilities in PDF plots. We find that when two PDF plots are visually and statistically identical, a general audience (graph literacy *mean* = 2.4 of 4, *SD* = 1.1) can accurately compare cumulative probabilities around 80% of the time. This number drops when PDF plots have different SDs or vertically scaling. Our findings inform best practices for visualizing PDFs and provide motivation for future work exploring PDF comprehension.

## Supplemental Materials

All supplemental materials are available on OSF at https://osf.io/7k5un/, released under a CC BY 4.0 license. In particular, they include
(1) Exp1 > Exp_1_Survey: Qualtrics .QSF and .PDF versions of Experiment 1's survey
(2) Exp2 > Exp_1_Survey: Qualtrics .QSF and .PDF versions of Experiment 2's survey
(3) Exp1 > Exp_1_Analysis:Anonymous data, analysis files, and models for Experiment 1 (.pdf, html, and .Rproj files)
(4) Exp2 > Exp_2_Analysis: Anonymous data, analysis files, and models for Experiment 2 (.pdf, html, and .Rproj files)
(5) stimuli.zip: all stimuli

## Acknowledgments

The authors wish to thank the members of the Khoury Visualization lab and Connor Wilson in particular for experiment design and stimuli feedback.

This work was funded by NSF Grant #2238175

## References

[1] P. Adamic, V. Babiy, R. Janicki, T. Kakiashvili, W. W. Koczkodaj, and R. Tadeusiewicz. Pairwise comparisons and visual perceptions of equal area polygons. *Percept Mot Skills*, 108(1):37–42, Feb 2009. doi: 10.2466/PMS.108.1.37-42 2, 8

[2] L. S. Aiken, S. G. West, and R. R. Reno. *Multiple regression: Testing and interpreting interactions*. Sage, 1991. doi: 10.1177/109821409301400208 5

[3] M. Allen, D. Poggiali, K. Whitaker, T. R. Marshall, J. van Langen, and R. A. Kievit. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res*, 4:63, 2019. doi: 10.12688/wellcomeopenres.15191.2 3

[4] P.-C. Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01 5

[5] S. Champely. pwr: Basic functions for power analysis, 2020. R package version 1.3-0, https://CRAN.R-project.org/package=pwr. 4

[6] W. S. Cleveland, M. E. McGill, and R. McGill. The shape parameter of a two-variable graph. *Journal of the American Statistical Association*, 83(402):289–300, 1988. doi: 10.1080/01621459.1988.10478598 2, 8

[7] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2 ed., 1988. doi: 10.4324/9780203771587 4

[8] Q. Company. Qualtrics, March 2024. https://www.qualtrics.com. 5

[9] M. Correll, E. Bertini, and S. Franconeri. Truncating the y-axis: Threat or menace? In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376222 2, 8

[10] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics*, 20(12):2142–2151, 2014. doi: 10.1109/TVCG.2014.2346298 3, 9

[11] L. Cosmides and J. Tooby. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58(1):1–73, 1996. doi: 10.1016/0010-0277(95)00664-8 3

[12] V. Di Maio and P. Lánský. Area perception in simple geometrical figures. *Percept Mot Skills*, 71(2):459–466, Oct 1990. doi: 10.2466/pms.1990.71.2.459 2, 8

[13] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3173718 3

[14] J. Gabry and T. Mahr. bayesplot: Plotting for bayesian models, 2024. R package version 1.11.0, https://mc-stan.org/bayesplot/. 1, 2, 8

[15] A. Gogolou, T. Tsandilas, T. Palpanas, and A. Bezerianos. Comparing similarity perception in time series visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):523–533, 2019. doi: 10.1109/TVCG.2018.2865077 2

[16] M. Hartmann, V. Gashaj, A. Stahnke, and F. W. Mast. There is more than "more is up": Hand and foot responses reverse the vertical association of number magnitudes. *Journal of Experimental Psychology. Human Perception and Performance*, 40(4):1401–1414, 2014. doi: 10.1037/a0036686 2

[17] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, p. 1303–1312. Association for Computing Machinery, New York, NY, USA, 2009. doi: 10.1145/1518701.1518897 2, 3

[18] J. L. Hintze and R. D. Nelson. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998. doi: 10.1080/00031305.1998.10480559 3

[19] D. Huff. *How to lie with statistics*. WW Norton, New York, NY, Oct. 1993. 2

[20] J. Hullman, P. Resnick, and E. Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLoS ONE*, 10, 2015. doi: 10.1371/journal.pone.0142444 3, 5

[21] P. Isenberg, A. Bezerianos, P. Dragicevic, and J.-D. Fekete. A study on dual-scale data charts. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2469–2478, 2011. doi: 10.1109/TVCG.2011.160 2

[22] W. Javed and N. Elmqvist. Stack zooming for multifocus interaction in skewed-aspect visual spaces. *IEEE Transactions on Visualization and Computer Graphics*, 19(8):1362–1374, 2013. doi: 10.1109/TVCG.2012.323 2

[23] W. Javed, B. McDonnel, and N. Elmqvist. Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):927–934, 2010. doi: 10.1109/TVCG.2010.162 2, 3

[24] S. Joslyn and S. Savelli. Visualizing uncertainty for non-expert end users: The challenge of the deterministic construal error. *Frontiers in Computer Science*, 2:58, 2021. doi: 10.3389/fcomp.2020.590232 3

[25] A. Kale, M. Kay, and J. Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 10 2020. doi: 10.1109/TVCG.2020.3030335 3

[26] A. Kale, F. Nguyen, M. Kay, and J. Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):892–902, 2019. doi: 10.1109/TVCG.2018.2864909 3

[27] M. Kay. *tidybayes: Tidy Data and Geoms for Bayesian Models*, 2023. R package version 3.0.6. doi: 10.5281/zenodo.1308151 5

[28] M. Kay. *ggdist: Visualizations of Distributions and Uncertainty*, 2024. R package version 3.3.2. doi: 10.5281/zenodo.3879620 1, 2, 8

[29] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. 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*, CHI '16, p. 5092–5103. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2858036.2858558 3

[30] R. Kosara. Joy plots, Jul 2017. https://eagereyes.org/blog/2017/joy-plots. 3

[31] T. Loetscher, C. J. Bockisch, M. E. R. Nicholls, and P. Brugger. Eye position predicts what number you have in mind. *Current biology: CB*, 20(6):R264–265, 2010. doi: 10.1016/j.cub.2010.01.015 2

[32] J. Mates, V. Di Maio, and P. Lánský. A model of the perception of area. *Spat Vis*, 6(2):101–116, 1992. doi: 10.1163/156856892x00172 2, 8

[33] T. Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press, 2015. doi: 10.1201/b17511 1

[34] J. Nachmias. Judging spatial properties of simple figures. *Vision Res*, 48(11):1290–1296, May 2008. doi: 10.1016/j.visres.2008.02.024 2, 8

[35] Y. Okan, E. Janssen, M. Galesic, and E. A. Waters. Using the short graph literacy scale to predict precursors of health behavior change. *Medical Decision Making*, 39(3):183–195, 2019. PMID: 30845893. doi: 10.1177/0272989X19829728 5

[36] L. Padilla. Know your experimental uncertainty. *Interactions*, 29(6):21–23, nov 2022. doi: 10.1145/3564022 3

[37] L. Padilla, H. Hosseinpour, R. Fygenson, J. Howell, R. Chunara, and E. Bertini. Impact of covid-19 forecast visualizations on pandemic risk perceptions. *Scientific reports*, 12(1):2014, 2022. doi: 10.1038/s41598-022-05353-1 2, 8

[38] L. Padilla, M. Kay, and J. Hullman. *Uncertainty Visualization*, pp. 1–18. Wiley StatsRef: Statistics Reference Online, 2021. doi: 10.1002/9781118445112.stat08296 3

[39] P. Parsons. Conceptual metaphor theory as a foundation for communicative visualization design. In *VisComm 2018 Schedule*. IEEE VIS Workshop on

Visualization for Communication (VisComm), Berlin, Germany, 2018. 2

[40] C. Perin, F. Vernier, and J.-D. Fekete. Interactive horizon graphs: improving the compact visualization of multiple time series. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, p. 3217–3226. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2470654.2466441 2, 3

[41] C. Peterson and A. Miller. Mode, median, and mean as optimal strategies. *Journal of Experimental Psychology*, 68(4):363, 1964. doi: 10.1037/h0040387 3, 4, 9

[42] X. Pu and M. Kay. A probabilistic grammar of graphics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376466 3

[43] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. 4, 5

[44] A. Raidvee, M. Toom, K. Averin, and J. Allik. Perception of means, sums, and areas. *Attention, Perception, & Psychophysics*, 82(2):865–876, 2020. doi: 10.3758/s13414-019-01938-7 2

[45] J. S. Risch. On the role of metaphor in information visualization. *arXiv:0809.0884 [cs]*, 2008. doi: 10.48550/arXiv.0809.0884 2

[46] S. M. Ross. Chapter 6 - normal random variables. In S. M. Ross, ed., *Introductory Statistics (Third Edition)*, pp. 261–296. Academic Press, Boston, third edition ed., 2010. doi: 10.1016/B978-0-12-374388-6.00006-5 1, 3, 4

[47] B. Schneider and R. Bissett. "ratio" and "difference" judgments for length, area, and volume: are there two classes of sensory continua? *J Exp Psychol Hum Percept Perform*, 14(3):503–512, Aug 1988. doi: 10.1037//0096-1523.14.3.503 2, 8

[48] R. R. Seva, K. K. Chinjen, N. Estoista, and J. A. Wu. Indicator distance and color effects in comprehension of multiple time series graph. In *10th Annual International Conference on Industrial Engineering and Operations Management*, 2020. doi: 10.46254/AN10.20200097 2, 8

[49] E. R. Tufte. *The visual display of quantitative information*. Graphics Press, Cheshire, CT, 2 ed., Jan. 2001. doi: doi/10.5555/33404 2

[50] B. Tversky. Visualizing thought. *Topics in Cognitive Science*, 3(3):499–535, 2011. doi: 10.1111/j.1756-8765.2010.01113.x 1, 2

[51] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, and C. Yau. Bayesian statistics and modelling. *Nat. Rev. Methods Primers*, 1(1), Jan. 2021. doi: 10.1038/s43586-020-00001-2 1

[52] A. M. Van Der Bles, S. Van Der Linden, A. L. Freeman, J. Mitchell, A. B. Galvao, L. Zaval, and D. J. Spiegelhalter. Communicating uncertainty about facts, numbers and science. *Royal Society open science*, 6(5):181870, 2019. doi: 10.1098/rsos.181870 3

[53] J. Vince. *Cartesian Coordinates*, pp. 27–34. Springer London, London, 2010. doi: 10.1007/978-1-84996-023-6_5 3

[54] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686 5

[55] C. O. Wilke. *Visualizing Many Distributions at Once*, chap. 9. O'Reilly Media, Inc., 1 ed., 2019. 3

[56] C. O. Wilke. *ggridges: Ridgeline Plots in 'ggplot2'*, 2024. R package version 0.5.6. 3

[57] B. Winter and T. Matlock. More is up… and right: Random number generation along two axes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 35, 2013. 2

[58] F. Yang, M. Cai, C. Mortenson, H. Fakhari, A. D. Lokmanoglu, J. Hullman, S. Franconeri, N. Diakopoulos, E. C. Nisbet, and M. Kay. Swaying the public? impacts of election forecast visualizations on emotion, trust, and intention in the 2022 us midterms. *IEEE Transactions on Visualization and Computer Graphics*, 2023. doi: 10.1109/TVCG.2023.3327356 9

[59] S. R. Yousif and F. C. Keil. The additive-area heuristic: An efficient but illusory means of visual area approximation. *Psychol Sci*, 30(4):495–503, Apr 2019. doi: 10.1177/0956797619831617 2

[60] S. R. Yousif and F. C. Keil. Area, not number, dominates estimates of visual quantities. *Scientific Reports*, 10(1):13407, 2020. doi: 10.1038/s41598-020-68593-z 2, 8