Declutter and Focus: Empirically Evaluating Design Guidelines for Effective Data Communication

Kiran Ajani, Elsie Lee, Cindy Xiong, Cole Nussbaumer Knaflic, William Kemper, and Steven Franconeri, *Member, IEEE*

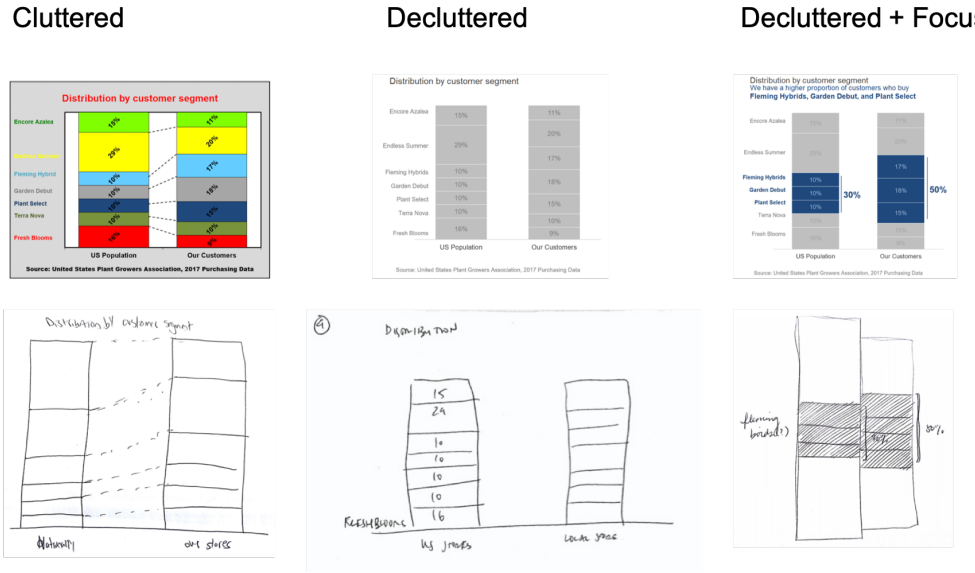

Fig. 1. We empirically evaluated the effects of two visualization design themes frequently prescribed by practitioner guides: *declutter* and *focus*. Decluttered designs showed small advantages for subjective ratings, and adding focus to the designs showed additional subjective rating advantages, along with a strong influence on what data pattern was remembered by viewers.

Abstract—Data visualization design has a powerful effect on which patterns we see as salient and how quickly we see them. The visualization practitioner community prescribes two popular guidelines for creating clear and efficient visualizations: declutter and focus. The *declutter* guidelines suggest removing non-critical gridlines, excessive labeling of data values, and color variability to improve aesthetics and to maximize the emphasis on the data relative to the design itself. The *focus* guidelines for explanatory communication recommend including a clear headline that describes the relevant data pattern, highlighting a subset of relevant data values with a unique color, and connecting those values to written annotations that contextualize them in a broader argument. We evaluated how these recommendations impact recall of the depicted information across cluttered, decluttered, and decluttered+focused designs of six graph topics. Participants were asked to redraw previously seen visualizations, to recall their topics and main conclusions, and to rate the varied designs on aesthetics, clarity, professionalism, and trustworthiness. Decluttering designs led to higher ratings on professionalism, and adding focus to the design led to higher ratings on aesthetics and clarity, and better memory for the highlighted pattern in the data, as reflected both by redrawings of the original visualization and typed free-response conclusions. The results largely empirically validate the intuitions of visualization designers and practitioners.

Index Terms—data visualization, data communication, data storytelling, empirical evaluation, visualization aesthetics

Each day, across organizations, research labs, journalism outlets, and classrooms, tens of millions of people attempt to communicate specific patterns in data using visualizations. One estimate from Microsoft, albeit 20 years old, put the number of PowerPoint presentations alone at 30 million per day [1]. Given the ubiquity of visual data communication, it is vital that visualizations transmit intended patterns to audiences quickly and clearly. In contrast, dozens of best-selling practitioner guides (Table 1) argue that business-as-usual visualizations are ineffective, confusing, or even misleading [2–38]. These books prescribe multiple tactics for improving graphical communication, but two themes stand out as common across many of them. The first guideline is to *'declutter'* a visualization (contrast the first and second examples in Figure 1), by removing unnecessary elements like gridlines, marks, legends, and colors. The second is to *'focus'* visualizations (contrast the second and third examples in Figure 1), by providing annotation and highlighting that lead a viewer to focus on a given pattern in the data.

Dozens of books intended for practitioners prescribe these designs because they note that real world visualizations and presentations tend to violate these guidelines across organizations. While we know of no empirically driven estimates of their prevalence, paper author C.N.K. estimates, based on training more than 25,000 people across various

- *Kiran Ajani is with Case Western Reserve University School of Medicine. E-mail: kiran.ajani@case.edu.*
- *Elsie Lee is with University of Michigan School of Information. E-mail: elsielee@umich.edu. *corresponding author*
- *Cole Nussbaumer Knaflic is with storytelling with data.*
- *Cindy Xiong, William Kemper, and Steven Franconeri are with Northwestern University.*

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

Table 1. Data Visualization Practitioner Guides

Book Title	Author	Declutter	Focus
Info We Trust	Andrews		•
Good Charts	Berinato	•	•
The Functional Art	Cairo	•	•
The Truthful Art	Cairo	•	•
Data At Work	Camoes	•	•
Trees, Maps, and Theorems	Doumont	•	•
DataStory	Duarte		•
slide:ology	Duarte	•	•
Effective Data Storytelling	Dykes	•	•
Effective Data Visualization	Evergreen	•	•
Presenting Data Effectively	Evergreen	•	•
Now You See It	Few	•	•
Information Dashboard Design	Few	•	•
Speaking PowerPoint	Gabrielle	•	•
Storytelling with Graphs	Gabrielle	•	•
Avoiding Data Pitfalls	Jones	•	•
Communicating Data With Tableau	Jones	•	•
Data Visualisation	Kirk	•	•
Storytelling With Data	Knaflic	•	•
Storytelling With Data: Let's Practice	Knaflic	•	•
#MakeoverMonday	Kriebel et al.	•	•
The Book of Trees	Lima		•
Design for Information	Meirelles		•
Visualization Analysis & Design	Munzner		•
Tableau Your Data!	Murray	•	•
Better Presentations	Schwabish	•	•
Elevate the Debate	Schwabish	•	•
Visual Explanations	Tufte	•	•
Visual Display of Quantitative Information	Tufte	•	•
Envisioning Information	Tufte	•	•
The Power of Data Storytelling	Vora	•	•
Information Visualization: Perception for Design	Ware		•
Visual Thinking for Design	Ware		•
Big Book of Dashboards	Wexler et al.	•	•
WSJ Guide to Information Graphics	Wong	•	•
Data Points	Yau	•	•

organizations and industries, that the overwhelming majority of visualizations intended for explanatory purposes include graphical clutter and do not focus attention. Note that this sample may be an overestimate, because these organizations had self-selected for visualization design training. But it is consistent with the fact that dozens of best-selling books have taken the time to make these prescriptions – that would be unlikely if those practices were already common in the real world.

While these two guidelines are argued to improve the comprehension and clarity of data communication, given their prevalence as prescriptions and their potential for impact in daily life, these recommendations have not been sufficiently empirically evaluated. The declutter guideline has some less controversial elements, such as replacing legends with direct labels, which are highly likely to improve performance [3–7, 9–14, 17–22, 26–32, 34–38]. However, other aspects of this process present potential tradeoffs. In the example in Figure 1, the overall aesthetic of the visualization might improve by removing dotted lines connecting the segments (contrast the first and second graphs), but that could also decrease the precision of comparing values between the two stacks. Removing the variety of colors for the nominal categories (as shown between the first and second graphs) might lead to higher aesthetic appeal, but it may also render the different categories harder to match. Are these 'clutter' elements really such an impediment for a powerful visual system that takes up almost half of the human brain [39], or might they perhaps lead to a level of minimalism that the viewer finds boring? Likewise, the focus guideline points viewers toward a particular pattern in the data, with one intention of having a viewer better remember the shape of the data. But again, given such a powerful visual system, is this step really needed? Might the viewer instead feel that a single story is being 'pushed' on them?

We empirically tested the impacts of these design guidelines, across ratings from undergraduate students (aesthetics, clarity, professionalism, and trustworthiness) and memory recall (via drawings and typed responses) for cluttered, decluttered, and decluttered+focused versions of visualized datasets. In collaboration with a practitioner guide author (paper author C.N.K.), we generated six example visualization topics in each of the three design styles shown in Figure 1. We find that decluttered visualizations are generally rated more positively on professionalism, but larger benefits appeared for decluttered+focused visualizations. These were rated more positively on aesthetics and clarity, and led to improved memory for focus-relevant data patterns, as measured by drawings and typed conclusions from memory. However, the focus manipulation did not have an additional effect beyond the declutter manipulation on ratings of professionalism, and neither decluttering alone nor the addition of focusing showed strong evidence for improving trustworthiness of a graph.

1 Related Work

1.1 Decluttering a Visualization

One form of argument for minimalist design in a visualization is Tufte's 'data-to-ink ratio' [30]. While the definitions of 'data' vs. 'ink' can be vague and subject to context [40], the general prescription is to remove any unnecessary elements in a visualization. In many cases, this rule is invoked as a reason to omit 'chartjunk,' pictorial ornamentation and metaphors such as arranging increasingly large bars in a graph as a monster's teeth, or putting a memorable dinosaur in the background of a graph. Some past work shows that these pictorial embellishments can lead to either better or worse performance on immediate reports depending on the details of task and context [41, 42], and found no significant effect of the presence of chartjunk on decision-making [43] or memory for the data pattern [44]. Chartjunk can lead to higher engagement and aesthetics ratings [42], as well as better short- and long-term memory for whether the visualization was previously seen [45] and what the data content or message was [46–48].

Our present work focuses not on such pictorial 'chartjunk,' but with 'clutter,' "...conventional graphical paraphernalia routines added to every display that passes by: over-busy grid lines and excess ticks, ... the debris of computer plotting..." [30]. A critique of the default settings of Microsoft Excel 2007 shows that the software created visualizations that nudge users toward redundant gridlines, excessive labels, excessive color, and 3D effects [33, 34, 49–51]. Authors of data visualization books also suggest reducing the number of colors present in a display to as few as possible, eliminating separate colors to indicate nominal categories in the data (e.g., using different colors for marks in a line or bar graph), or other areas of a visualization such as the background [3, 21, 49, 52].

Some existing studies have sought to determine whether decluttering would lead to objectively better performance on graphical perception and memory tasks. One study found that within a simulated monitoring dashboard for nine metrics, removing unnecessary elements (tick marks, verbose scale number labeling, redundant readouts, colored backgrounds highlighting relevant thresholds, etc.) did not improve response times or situation awareness accuracy [53]. A more extreme manipulation stripped away the graphical display entirely so that values were only displayed as text digits, and this did improve response times [53]. But this more drastic change might improve performance not because it omits clutter, but because it contained larger versions of text digits and a more precise data representation that was likely more

suitable for the participant's high-precision monitoring task. Another study asked participants to compute simple means, differences, and comparisons on bar graphs with only two bars, while manipulating the presence of individual bits of 'clutter,' finding that including axis tick marks slightly increase response times, but completely removing the x- and y-axes can slightly slow responses, with both effects occurring for bar but not line graphs [41]. Finally, another study showed that graphs with 'data redundancy' – symbolic numbers placed on or near visual marks that already show those numbers visually – actually showed *higher* quality memory reports for their content [47].

Other studies evaluate the impact of decluttering on preference ratings. When they are important for a precise estimation task, viewers prefer lower-contrast gridlines compared to heavier lines that can obscure the underlying data, and the authors even provide quantitative alpha values for the preferred range [54], and show how this range varies by the color of the gridlines [55]. Another study evaluated ratings of beauty, clarity, effectiveness, and simplicity of visualizations with high vs. low 'data-ink' ratios [56]. Surprisingly, visualizations with lower data-ink ratios (more 'clutter') were rated more positively, potentially because the more minimalistic style was unfamiliar, given that the cluttered designs are encountered more frequently in tools like Microsoft Excel. Another study found similar results, especially for extremely minimalistic designs [57], and again argued that the lower level of familiarity might drive a distaste for overly decluttered designs.

1.2 Focusing a Visualization

While broad statistics about the data values, such as distributions and outliers, are available quickly [58], picking out one of the many – even dozens or hundreds – of potential patterns, trends, and relations of interest within it is an inefficient perceptual process [59], requiring seconds or minutes of processing to unpack the 'paragraphs-worth' of information implicit in a single visualization [60]. One study showed that once a pattern is seen within a visualization, a 'curse of expertise' biases people to tend to think that others will focus on that pattern as well, even when they don't. Viewers saw background information that focused on a particular data pattern (i.e., relationships between two lines from a four-line graph) in a visualization. They then asked participants to forget that story, and predict which of multiple possible patterns would be most salient to a viewer who had not heard that story. Despite reminders that other viewers had not heard the story, participants still incorrectly predicted that others would see what they saw in the visualization [61].

This problem motivates a frequent practitioner guideline to focus a visualization: if there is a single pattern that a viewer should extract and remember, then the designer should state the pattern clearly with direct annotation, and highlight the key data values that create that pattern [2–38]. One study found that visualizations with a title showed higher quality memory reports for their content, especially when the title included 'message redundancy,' an additional section of text that focused the viewer on a key pattern in the visualization [47]. That study also found that titles were robustly fixated (especially when placed at the top as opposed to the bottom of the visualization), and that later descriptions from memory tended to reflect a rewording of the content expressed in the title. Other work focuses not just on focusing a single view on a single visualization, but instead on the broader practice of 'storytelling': creating a sequence of views that follows an argument or other narrative as a rhetorical tool to guide a viewer through a more complex set of data patterns over time, with some considering the added complexities of user interactivity, path choice, and drilldown [62–65]. Some of these studies pick out real-world examples of highlighting relevant data values in a visualization [63].

1.3 Contributions of the present work

Relatively more existing work has tested the declutter guideline than the focus guideline, but this work has not converged on a clear answer. Some of this work shows an advantage to decluttering [54], some shows no difference [53], some shows a disadvantage [47], and some shows a mix [41]. For preference ratings, at least two recent studies actually show a preference for more cluttered designs, perhaps because they are more familiar [56, 57], though these studies use graphs that are already highly minimalistic and abstract (e.g. an Excel bar graph of 'Sales' across four regions 'North,' 'South,' etc.). Other work that finds little impediment of clutter on objective performance uses very simple displays, such as 2-value graphs [41], or relies on a specific dashboard monitoring task [53]. While well-controlled, these highly abstract graphs and specific tasks may not reflect the preference and performance effects generated by the types of clutter seen frequently by participants. To test this idea, we use realistic examples of both graph topics and clutter types drawn from a guidebook derived from many years of experience in organizational presentation settings [21].

Little work has tested the effects of the focus guideline. In the closest work, titles (particularly those that focused on a single message) led to better objective memory of the content in the visualization [47]. While highly suggestive, these data are correlational because presence of a title was not randomly assigned, so it is possible that a third variable contributed to that advantage. For example, a particularly salient data pattern could drive both the original visualization author and experimental participant to notice that more memorable pattern. We therefore test the same visualizations across three designs, in a between-subject counterbalanced manipulation. Furthermore, that study was concerned with measuring 'objective' memory for correct vs. incorrect information in the visualization, while our focus is on measuring how strongly a focus manipulation could subjectively emphasize one possible pattern in the data over others. This previous work also measured relatively long-term memory for visualizations after seeing 100 total visualizations for 10 seconds each, simulating a longer-term exposure to many visualizations. In contrast, our goal was to test an immediate understanding after a 10-second viewing period, in order to simulate the experience of being shown data on a handout or presentation slide in a meeting, conference, or discussion. Note that other tasks might be less memory-dependent, such as having an unlimited time to inspect a slide during a long meeting. Future work would be needed to assess whether the conclusions of this study generalize beyond our memory-based tasks. In addition to memory reports via typed text, we add a novel method of collecting drawings from memory, to see if the focus manipulation would affect not only what they say, but what they remember seeing. Finally, we know of no existing work that measures preference ratings (aesthetics, etc.) across the increasingly prevalent practice of implementing the focus guideline in visualization designs.

2 Experiment Overview

Figure 1 depicts an example of three design variations – cluttered, decluttered, and decluttered+focused – for one of the visualization topics used in our experiments. We take representative examples from a popular practitioner guide [21], 100,000+ copies sold according to storytellingwithdata.com [66], and the alternate designs were created in collaboration with that book's author.

We measured several metrics of communication effectiveness across these three design variations, including whether viewers would be more likely to recall the intended message of each visualization, as measured by qualitative coding of visualizations drawn from memory and from typed free-responses. We focused on recall, rather than descriptions made while participants actively viewed the stimuli, for two reasons. First, recall has been used most frequently in related work (e.g., [47]). Second, we wanted to simulate the experience of integrating information across a sequence of data patterns during a presentation, which is critical for leading viewers to a conclusion or decision at the presentation's end. We therefore sought to identify the most 'sticky' message from each visualization. We additionally measured whether they would rate the design variations differently across quantitative scales of aesthetics, clarity, professionalism, and trustworthiness, as well as qualitative explanations of those ratings.

2.1 Participants

An omnibus power analysis from the quantitative data collected from a pilot experiment (see Supplementary Materials) suggested a target sample of 24 participants would give us more than 95% power to

detect an overall effect of graph version (cluttered, decluttered, decluttered+focused) on quantitative ratings of aesthetics, clarity, professionalism, and trustworthiness. All 24 participants whose data were used in the final analysis are either students or community members at Northwestern University (18 female, age range 18 to 26, average age 19.5, all normal or corrected-to-normal vision). We replaced a subset of our initial 24 participants to resolve a condition counterbalancing error and this replacement was performed blind to participant results, only governed by which conditions they were shown. They participated in return for $10/hour or for course credit. Undergraduate students are familiar with graphs and other data visualizations, and this population is most likely to later move on to become the audience addressed by the practitioner guides in Table 1 – employees in businesses and other organizations. However, future work should test those populations more directly.

3 Materials and Procedure

3.1 Stimuli

The stimuli for this experiment (Figure 2) consisted of cluttered, decluttered, and decluttered+focused (which will be shortened to 'focused' from here forward) versions each of six different example graphs: concerns about an automotive design split by concern category ('Car'), holiday shopping frequencies over time split by gender ('Holiday'), news sources over time split by medium ('News'), the distribution of customer-preferred flower seeds split by customer category ('Plants'), US prisoner offenses split by category ('Prison'), and retail prices of tires over time split by manufacturer ('Tires'). The graphs consisted of vertical bar charts, horizontal bar charts, stacked vertical bar charts, and line graphs. We created the three different visualization designs of the same graph, using examples adapted from Nussbaumer Knaflic [21].

The cluttered graphs contained a set of features listed as 'clutter' across the practitioner guides listed in Table 1: a more diverse color palette, low-legibility fonts, gridlines, heavy borders, background shading, axis tick marks, data markers, redundant numeric labeling, diagonally rotated text, overuse of bolded text, and 3D shading cues. While not all real-world graphs include all of these cluttered elements, we included a large number of them for each cluttered design to maximize the chances of these additional elements affecting participant performance.

Following the guidelines of the sources listed in Table 1, the decluttered graphs removed backgrounds, all colors, chart borders, gridlines, tick marks, and even axis lines in some cases, but kept the graph's title, data values, and axes and their labels. Text, including axis and tick-mark labels were oriented horizontally instead of diagonally. Text was spatially aligned to other elements in the graph; for example, instead of being centered, titles were left-aligned with the y-axis. White space was added between major elements (e.g. between the title and graph, or graph and footnote). Legends were converted into direct labels. For example, for the News graph, the decluttered version removed the excessive data points and gridlines present in the cluttered version, and added white space.

Again taking inspiration from the sources in Table 1, the focused graphs added a single highlight color (e.g. red) to the grayscale graph, intended to focus the viewer on a given pattern. In the Holiday graph topic, intensity of color was manipulated to focus the viewer particularly to a pair of data values. A one-sentence annotation described a conclusion that could be drawn from that data pattern, with key words in the same font color as the highlighted data pattern. For example, in the News graph, the focused version adds contrast between the blue 'Internet' line and the other news source lines to draw attention to the pattern of increasing usage. The specific changes made between the cluttered, decluttered, and focused designs for each graph topic are outlined in Figure 2.

3.2 Procedure

3.2.1 Part 1

We presented each of the six graph topics to 24 participants in a Latin square counterbalanced design that balanced presentation order across examples. Participants were placed into one of six counterbalancing groups, each of which had a predetermined balance of cluttered, decluttered, and focused versions such that each participant viewed two examples of each of the three visualization designs for a total of six graphs. The combination of which visualization design was seen with which graph topic was balanced across the counterbalancing groups, resulting in an equal presentation of each visualization design and graph topic across all participants.

Redraw Task. Participants saw the following prompt: "You are about to be shown a graph for 10 seconds. After the graph is displayed, you will be asked to redraw as much as you can remember about the graph. Please do not draw anything on the paper in front of you until after the graph has disappeared from the screen." Participants were additionally presented with a scenario that provided context for the graph they were about to see. For example, before seeing any version of the Plants graph, they read the following scenario: "A local botanical garden sells seven different brands of flower seeds to the community. These different seed companies also sell all over the United States. Our local botanical garden wants to know how these brands are selling in our own store, compared to how they sell around the United States in general." All of the scenarios are available in Supplementary Materials. After the 10-second exposure, participants saw the prompt: "Please take 1-2 minutes to redraw whatever you can remember from the graph that you just saw on the piece of paper in front of you. Advance to the next page once you have redrawn the graph." Participants then redrew as much of the graph as possible from memory. All drawings are available in the Supplementary Materials.

Free Response Conclusions. Participants wrote about the subject matter and conclusions that could be drawn from the graph. On the next page after redrawing the graph, they typed out the subject matter ("What was the subject matter of the graph?") and their perceived conclusion of the data ("What conclusions can be drawn from the graph?") in a free response text box. The subject matter responses were collected but are not reported because they were largely redundant with more detailed responses provided for the conclusion. For example, in response to the Plants graph, the subject matter written by Subject 1 was: "how seed brands are selling in a local botanical store compared to the US," while the conclusion response was: "which seeds are most popular locally, which seeds are most popular nationally, what seeds have poor sales locally and/or nationally." The subject matter and conclusion data is available in the Supplementary Materials.

3.2.2 Part 2

Quantitative Evaluation. We next presented all three visualization designs of each graph topic simultaneously to each participant, and asked them to rate the three designs according to four Likert scales (1-5 range). This process was repeated for each of the six topics for a total of 18 graphs (3 designs x 6 topics) in a counterbalanced order such that each topic appeared an equal number of times in each spot in the ordering. The four scales were:

- **Aesthetics**: "Overall, is this a visually appealing image?", rated from one ("very hideous") to five ("very beautiful").
- **Clarity**: "Is it clear what information is being presented and why?", rated from one ("I am utterly confused") to five ("makes perfect sense").
- **Professionalism**: "Does this graph look like something you would see in a professional environment?", rated from one ("very unprofessional") to five ("very professional").
- **Trustworthiness**: "Based on the presentation, how trustworthy is the person who made this graph?", rated from one ("very untrustworthy") to five ("very trustworthy").

Qualitative Evaluation. Finally, we asked an open-ended question: "Please explain your reasoning for choosing these ratings. (1-2 sentences)" for each of the four scales for all 18 graphs.

4 Pilot Experiment

We first piloted this experiment (17 participants, all received course credit at Northwestern University) in order to measure statistical power

Fig. 2. We created three versions (cluttered, decluttered, and focused) of six different graph topics following popular data visualization guidelines. The figure rows also show all changes for each topic across the three versions.

for our quantitative measures and to refine our stimuli. The stimuli and procedure were similar to those described above, with the following exceptions. For stimuli, we realized that some graph topics in the pilot condition appeared to have insufficient contrast in the focus of their conclusions between the decluttered and focused designs (e.g. the Prison topic), and made edits to the stimuli to address this for the experiment. Focused designs initially omitted source information, which trustworthiness rating explanations revealed to be important to participants, so these were added for all graph topics and designs after the pilot. The Holiday example had confusingly inconsistent temporal bins sizes on the x-axis (e.g., some bars represented 2-week ranges, some months), so for the experiment we changed these labels to consistently present one month intervals. Finally, the pilot originally used abstract data labels on the Plant and Tires topics, which were subsequently changed from "Segment 1, Segment 2, Segment 3, etc." to "Fresh Blooms, Terra Nova, Plant Select, etc." and from "Product 1, Product 2, Product 3, etc." to "BF Goodrich, Bridgestone, Continental, etc.," respectively.

In the pilot, when participants rated graphs along the four scales, they saw each of the 18 total graphs (3 designs x 6 topics) individually. We later decided these sequential ratings would make it more difficult for participants to directly compare the different designs. Therefore, in the experiment we showed all three designs for each topic simultaneously. We did not perform a formal analysis on the qualitative pilot data, as they were only intended to provide inspiration for refining the stimuli.

4.1 Pilot Results

An initial MANOVA examined the effect of graph version (cluttered, decluttered, and focused) on aesthetics, clarity, professionalism, and trustworthiness ratings of the visualizations. This test revealed a significant multivariate effect (Pillai's value = 0.34) of design, after accounting for within-subject error. We conducted 999 simulations using Bonferroni adjustment methods using the R pairwise.perm.manova function [67] to obtain post-hoc pairwise comparisons between the three levels of design with corrections for multiple testing. The test revealed that overall, across all four dimensions, focused designs were rated significantly higher across all metrics than decluttered designs ($p = 0.003$), which were rated significantly higher than cluttered designs ($p = 0.003$).

We entered the Pillai's value into G*Power [68] and conducted a power analysis. In this experimental design with 3 groups of visualization design (cluttered, decluttered, and focused) and 4 response variables (aesthetics, clarity, professionalism, and trustworthiness), the pilot experiment achieved 90.4% power at the alpha level of 0.05. With an increased sample size of 20 participants (with 3 measures for each response variable), we would be able to obtain 95.29% power at an alpha level of 0.05. We only analyzed the rating data from the pilot in order to conduct a power analysis for Experiment 1.

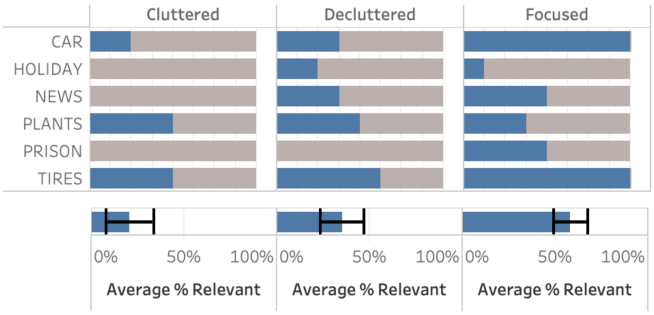

Fig. 3. Whether the redrawn visualization is coded as containing elements that are relevant (blue) or irrelevant (grey) to the pattern of data highlighted by the focused design.

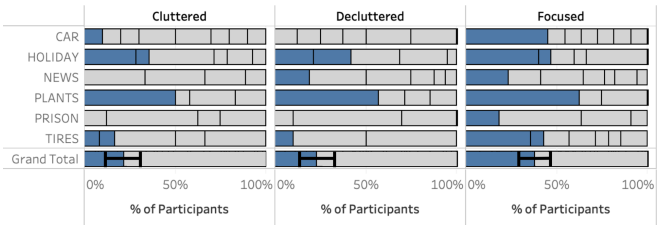

Fig. 4. Participants were more likely to recall the relevant conclusions from focused visualizations compared to cluttered and decluttered designs. Blue rectangles represent typed topics that were relevant to the focused design's highlighted pattern, grey rectangles represent irrelevant ones. The rectangle width depicts the percentage of participants naming that topic. Across topics, focused designs led to more focus-relevant conclusions

Table 2. Does the redrawn graph show the relevant conclusion?

Topic	Relevant
Car	Does the graph include the word "noise" (or a variation)?
Holiday	Is the first two months higher for women than men AND is the last two months lower for women than men?
News	Does the graph show "Internet" as increasing?
Tires	Do the lines converge on the right side of the graph?
Prison	Does the graph have both minor and major drug offenses labeled OR include a sentence about minor and major drug offenses?
Plants	Is the physical size OR percentage for the third, fourth, and fifth segments of US Population smaller than those of Our Customers?

5 Experiment Results

Based on pilot rating effect sizes, we conducted the experiment with 24 new participants in the same Latin-Square design such that all participants saw an equal amount of all three groups of design (cluttered, decluttered, focused). Because paper author C.N.K. is an author of practitioner books that prescribe the methods tested here, she could be perceived as having a conflict of interest in the outcome of this study. Therefore, while she offered advice in the design of the stimuli and offered abstract contextualization advice on the design of qualitative coding schemes, she was purposely excluded from both the data analysis stage and initial drafts of the manuscript.

5.1 Redraw Task Results

Our results for the redraw task are shown in Figure 3, which compares the percentage of relevant to irrelevant redraws for all three visualization designs (cluttered, decluttered, and focused) of all six graph topics. Qualitative coding of the contents of the redrawn graphs was performed by the first author and a second coder who were blind to both the study design and the condition manipulations. Any discrepancies were resolved by a tiebreaker vote from the second author. Ratings were based on a rubric of questions, as displayed in Table 2. To calculate inter-rater reliability (IRR), we used Cohen's Kappa for 2 Raters. The IRR kappa value for the redrawn graphs was strong [69]: Kappa = 0.814 (z = 16, $p < 0.001$).

We explored whether the redrawn visualization was more likely to reflect the graph author's intended message, as determined by the message that was featured most saliently in the focused graph condition, and operationalized as shown in Table 2. We created a binary 'relevancy score' based on whether or not the participants' redrawn graphs contained at least one key element that displayed the main trend. For example, the relevancy score for the Car topic was only based on whether or not the graph included the word "noise." The Holiday topic was the only topic that required two features to be present: "Is the first two months higher for women than men AND is the last two months lower for women than men?" This was because the annotation on the focused visualization concentrated on a higher proportion of women shopping earlier in the year compared to men. To determine whether participants actually understood this conclusion, their graph had to demonstrate women shopping more in earlier months and men shopping more in later months. If participants did show the intended trend in their redrawn graph of a topic, the trial would be coded as a 1, otherwise it would be coded as a 0. We hypothesized that the redraw visualizations would more successfully replicate main messages illustrated in the focused visualizations, compared to decluttered and cluttered visualizations.

Overall, we found that **participants included the most relevant information in their redraws when they viewed visualizations with the focused design**. We used a mixed-effect linear model to fit the relevance scores [70] under the three visualization designs. For fixed effects, in addition to visualization design, we also included graph topic and its interaction with graph design as predictors. We used a random intercept term accounting for individual differences as random effects to fit the Latin-Square design of the experiment. The regression model indicated a moderate effect of visualization design, $\chi^2 = 21.20$, $\eta^2 partial = 0.15$, $p < 0.001$. Post-hoc analysis with Tukey's adjustments suggests that, overall, participants redrew visualizations with the most relevant information when they viewed visualizations with the focused design compared to cluttered ($Est = 0.75$, $p < 0.001$) and decluttered ($Est = 0.62$, $p = 0.007$). There is no significant difference between relevance scores for the cluttered and decluttered designs ($Est = 0.13$, $p = 0.81$). There is also an effect of topic, χ^2=38.31, $\eta^2 partial = 0.17$, $p < 0.001$, suggesting that participants more readily included critical information for some visualization topics than they did others. However, the multiple comparisons of means using Tukey contrast did not reveal significant differences between visualizations of different topics. There is no interaction between design and topic $\chi^2 = 16.05$, $p = 0.10$.

5.2 Free Response Conclusion Results

Figure 4 shows a comparison of the relevant and irrelevant conclusions written by participants for each of the three visualization designs (cluttered, decluttered, and focused). The free response conclusions were categorized similarly to the redrawn graphs. The coders devised a set of categories from the entire set of responses using open coding based on grounded theory [71], such that a new category was created for each response that did not fit into previous categories. These categories were created in order to ascertain the number and variety of conclusions participants could potentially draw from the graphs. All conclusions were placed into one or more categories based on certain keywords. For example, any conclusion that mentioned steering as a concern in the Car graph was categorized as "Steering"; if the conclusion additionally mentioned engine noise as a top concern, the conclusion would be categorized as "Noise, Steering." For analysis of participants who mentioned more than one conclusion, if they said a relevant conclusion, their response was analyzed as "relevant" even if they wrote additional irrelevant conclusions as well. After all of the categories were created,

Table 3. Relevant Conclusion Categories

Topic	Conclusion	Criteria
Car	Noise	Mentions engine noise as the top / one of the top concern(s)
Holiday	Men	Mentions that men shop more in later months (November and December)
Holiday	Women	Mentions that women shop more / earlier than men
News	Internet	Mentions that internet usage is increasing
Plant	Local / National	Mentions a difference between local and national sales
Prison	Minor / Major	Mentions both major and minor drug offenses
Tires	Competition	Mentions that tire prices are becoming more competitive
Tires	Convergence	Mentions that tire prices converge

the coders determined which ones were "relevant" to the conclusions of the focused graphs based on the annotation on the graphs. The full set of categories had both relevant and irrelevant conclusions, though we only used the relevant conclusions for our final analysis. Table 3 shows a subset of only the relevant categories. The full table of conclusion categories as well as the categorization key are available in Supplementary Materials. The qualitative coding for the recalled conclusions was also done by the first author, who was blind to each trial's conditions, and a second coder who additionally was blind to the study design. Discrepancies were resolved for this task without a tiebreaker. To calculate inter-rater reliability (IRR), we used Cohen's Kappa for 2 Raters. The IRR kappa value for the recalled conclusions was strong [69]: Kappa = 0.88 (z = 32.5, $p < 0.001$).

We conducted a logistic general linear regression predicting the likelihood that participants would write a certain type of conclusion using: the conclusion type (relevant or irrelevant; whether the conclusion was relevant to the message of the original visualization shown to them), visualization design (cluttered, decluttered, and focused), and their interaction. Follow-up analysis of deviance using a Chi-square based ANOVA suggests that there is no significant main effect of design ($p = 0.82$), such that participants were equally likely to draw conclusions for each design, but a significant main effect of conclusion type ($p = 5.66e-08$), such that **participants overall were more likely to mention relevant information than irrelevant information in their conclusions**, and a significant interaction ($p < 0.001$). Note that Figure 4 shows more grey (irrelevant) than blue (relevant) – this is because each conclusion could have been placed into both relevant and irrelevant categories depending on its content. Although more irrelevant categories were mentioned overall, for each conclusion, participants more likely mentioned at least something that was coded as relevant. Close examination of the relevancy and design interaction in the model reveals that it was driven by **participants being more likely to write visualization-story-relevant conclusions in their free responses when the original visualization was a focused design**. Compared to showing participants a cluttered visualization, showing them a focused visualization would increase the likelihood of them writing a story-relevant conclusion afterwards during free-recall by 2.96 fold ($Est = 1.08$, $p = 0.006$). Showing participants a focused visualization would also significantly increase the likelihood of writing a story-relevant conclusion compared to showing them a decluttered visualization by 2.49 fold ($Est = 0.91$, $p = 0.017$). There is no significant difference in the likelihood of writing a story-relevant conclusion during free-recall between participants who viewed the cluttered and decluttered visualizations ($Est = 0.17$, $OR = 1.19$, $p = 0.67$).

Table 4. Correlation and Variance Inflation Factors.

	Aesthetics	Clarity	Prof.	Trust.	VIF
Aesthetics	*1*	0.60*	0.73*	0.59*	2.37
Clarity		*1*	0.61*	0.52*	1.76
Professionalism			*1*	0.69*	2.93
Trustworthiness				*1*	2.00

5.3 Quantitative Ratings on Aesthetics, Clarity, Professionalism, and Trustworthiness

The average quantitative ratings for the three visualization designs across these four variables are depicted in Figure 5. We conducted a MANOVA analysis examining the effect of visualization design on aesthetics, clarity, professionalism, and trustworthiness ratings of the visualizations. We additionally considered possible effects of visualization topic and an interaction between design and topic. One participant's data was not analyzed because they did not complete this component due to a computer power failure, for a total of $N = 23$. After adjusting for within-subject error, the analysis reveals a significant multivariate effect of design (Pillai's value = 0.57), a trending effect of topic (Pillai's value = 0.08), and a significant interaction between design and topic (Pillai's value = 0.28).

Overall, focused designs are rated as more aesthetically appealing, professional, and clear. Post-hoc analysis with Tukey's adjustment using the lme4 package in R studio [72] suggests that, comparing focused and cluttered designs, the focused designs are rated significantly higher on visual aesthetic appeal ($Est = 0.86$, $p = 0.005$), trendingly higher on clarity ($Est = 0.60$, $p = 0.08$), significantly higher on professionalism ($Est = 1.10$, $p < 0.001$) and not significantly different on trustworthiness ($Est = 0.22$, $p = 0.64$). Comparing focused and decluttered designs, the focused design was rated trendingly higher on aesthetics ($Est = 0.53$, $p = 0.06$), trendingly higher on clarity ($Est = 0.55$, $p = 0.06$), and not significantly different on professionalism ($Est = -0.01$, $p = 0.99$) or trustworthiness ($Est = -0.31$, $p = 0.31$). Comparing decluttered and cluttered designs, they are not significantly different on aesthetics ($Est = 0.33$, $p = 0.43$) or clarity ($Est = 0.05$, $p = 0.98$), but decluttered is rated as significantly more professional ($Est = 1.11$, $p < 0.001$), and trendingly more trustworthy ($Est = 0.53$, $p = 0.06$). Overall, visualization design seems to have a relatively large effect size on all four ratings dimensions (aesthetics, clarity, professionalism, and trustworthiness) above and beyond other factors such as topics, with a partial η^2 ranging from 0.326 to 0.475. Details are available in the Supplementary Materials.

The trending main effect of topic and significant interaction of topic and design were driven by the Prison visualization being rated higher than the Plants visualization on clarity ($Est = -0.72$, $p = 0.043$) and professionalism ($Est = 0.63$, $p = 0.046$), as well as the Tires story being rated higher on professionalism than the Car ($Est = 0.90$, $p < 0.001$), Holiday ($Est = 0.82$, $p = 0.002$), and Plants ($Est = 1.05$, $p < 0.001$) visualizations. Overall, the main effect of design persisted in the same pattern despite these interactions.

However, our four rating measures are not fully independent. For example, increasing the professionalism of a visualization may go hand-in-hand with making it more aesthetically pleasing. We therefore conducted collinearity diagnostics on these dependent variables in terms of variance inflation factors (VIF). A VIF greater than 1 would suggest some correlation between them. For example, aesthetics has a VIF of 2.37, which is 1.37 times bigger than what would be expected if it had no correlation with the other three dimensions. Table 4 shows the VIF and correlation between all four rating dimensions. All VIF values are below four, suggesting a small to moderate correlation between them. Professionalism seems to be the most highly correlated with the other three dimensions.

5.4 Qualitative Rating Results

In addition to the Likert scale ratings on the four dimensions above (aesthetics, clarity, professionalism, and trustworthiness), we also qualitatively coded the participants' open-ended comments about the reasons for their ratings (their entry in a single text box could contain explana-

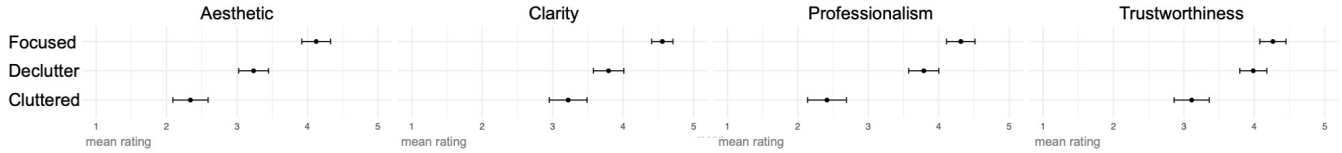

Fig. 5. Quantitative ratings for the three visualization design on a scale from 1 to 5 on aesthetics, clarity, professionalism, and trustworthiness.

Fig. 6. Qualitative ratings for the three visualization designs, where the largest bubble represents 88% of participants coded as matching that theme. Each bubble represents a different graph topic, allowing the viewer to gauge the consistency of comment frequency across topics (mappings from topics to bubbles are available in the full dataset in the Supplementary Materials)

tions of their ratings for one or all of the graph designs). The coding used the same four metrics as the quantitative ratings (Figure 6). Each one of these metrics had several specific categories to evaluate the metric in more detail, including positive or negative sentiment (e.g. "too much color"). As above, one participant's data was not analyzed because they did not complete this component due to a computer power failure, for a total of $N = 23$.

The open-ended comments were qualitatively coded by the second author. Similar to the free response conclusions, the comments were open coded, letting the codes arise from the open-ended comments. Codes were created as they were seen in the data, with new codes being created for concepts that were commonly being mentioned. Based on these codes, four prominent categories of comments were decided: color, aesthetics, information, and emphasis. For each category, there were codes that fell into either positive or negative comments about the category. For example, a common comment was that the participant thought that there were too many colors. This would have been coded as "too many colors" as well as the more general category of "dislikes color." Another participant commented, "Left graph uses too many colors which is distracting and messy," which was coded as "disliked color, too many colors, disliked aesthetics, cluttered." After all of the codes had been defined, the second author made sure that all of the comments were coded according to all of the codes.

For reliability, a second coder who was blind to both the study design and the condition manipulations coded the open-ended comments based on the codes decided by the second author. Any discrepancies between the second author and the blind coder were discussed and both coders agreed on a final decision. Inter-rater reliability (IRR) was calculated the same way as the previous coding. For the open-ended comments for the ratings qualitative coding, our kappa was 0.811 ($z = 83.5$, $p < 0.001$), indicating strong agreement [69].

In order to have the qualitative rating responses be alignable to the quantitative rating responses, we categorized each of the codes into the categories from the quantitative rating: aesthetics, clarity, professionalism, and trustworthiness. Several new codes were added (e.g. emphasis is untrustworthy) to address these new categories. The first author and the blind coder coded these new codes. Because of the combination of the quantitative rating categories and the qualitative open-ended comments, some of the categories have fewer instances of the code showing up in the data, such as the trustworthy category. This may suggest that participants did not consider trustworthiness as important as the other aspects of the graphs, since they mentioned it fewer times.

We found that overall **participants were more likely to associate the focused and decluttered design with positive sentiments, and the cluttered design with negative sentiments** via a regression analysis using a logistic general linear model with visualization design (cluttered, decluttered, focused) and sentiment (positive and negative), and their interactions, to predict the likelihood that participants would produce such a response, for each of the four dimensions (aesthetics, clarity, professionalism, and trustworthiness). Analysis of variance using Chi-square comparisons revealed an overall main effect of visualization design across all four dimensions: aesthetics ($\chi^2 = 48.71$, $p < 0.001$), clarity ($\chi^2 = 57.61$, $p < 0.001$), professionalism ($\chi^2 = 41.31$, $p < 0.001$), and trustworthiness ($\chi^2 = 28.89$, $p < 0.001$). For aesthetics, clarity, and professionalism ratings, there is also a significant main effect of sentiment and interaction of sentiment and design (details can be found in the Supplementary Materials).

Post-hoc analysis with the logistic model reveals that for aesthetics, participants were significantly less likely to associate the cluttered design with a positive sentiment ($Est = -1.15$, $p < 0.001$, OR = 0.39). They were also less likely to associate decluttered ($Est = -1.16$, $p < 0.001$, OR = 0.31) and focused designs ($Est = -1.16$, $p < 0.001$, OR = 0.18) with negative sentiments. Many participants claimed that the cluttered graphs were "disorganized" and "visually unappealing." They were significantly more likely to associate decluttered and focused designs with positive sentiment (Decluttered: $Est = 1.39$, $p < 0.001$; Focused: $Est = 3.00$, $p < 0.001$), with cluttered design as the reference. For clarity, compared to the cluttered design, participants were significantly more likely to associate the focused design ($Est = 1.61$, $p < 0.001$, OR = 5.01) with positive sentiment. One participant noted this difference for clarity and mentioned that the "[cluttered] one is unclear and looks messy with all the labels and colors. The [focused] one is the best as it emphasizes the data for internet, which conveys the message clearly." For professionalism, participants were significantly less likely to associate decluttered ($Est = -2.93$, $p < 0.001$, OR = 0.05) and focused designs ($Est = -1.65$, $p < 0.001$, OR = 0.19) with negative sentiments and more likely to associate them with positive sentiments ($Est = 1.83$, $p = 0.024$, OR = 6.26; $Est = 1.11$, $p = 0.034$,

OR = 3.03), compared to cluttered designs. Interestingly, some participants claimed that the cluttered graphs seemed to be made by children: "The middle [cluttered] graph looks very childish like something a middle schooler would make for a school project." Very few participants mentioned trustworthiness-related attributes in their free responses, and thus whether participants associate different designs with positive or negative sentiments regarding trustworthiness remains inconclusive. Of note, some participants expressed that the focused graphs seemed suspicious in the way that they pushed a single interpretation of the data: "The [focused] graph made the point of emphasis most clear, though took a hit in trustworthiness because the extra text made it seem like it had an agenda." We found these unanticipated responses to be particularly interesting. Detailed statistical analysis can be found in the Supplementary Materials.

6 Discussion

Participants were 2.5-3x more likely to recall a focus-relevant conclusion for focused designs relative to the non-focused designs, and were more likely to redraw the focused elements of the original visualization. They also found those designs more aesthetically appealing and clear, associating more positive sentiment with these visualizations and more negative sentiment with the cluttered visualizations. Across trustworthiness and professionalism, participants preferred decluttered or focused graphs to cluttered ones.

6.1 Redrawing and Free Response Discussion

Coding of the free response conclusions reveals how many different patterns a viewer might focus on when that focus is not guided by the visualization's designer. Across all of the relevant and irrelevant categorized conclusions, there were 12 total conclusion categories that participants entered for the Car graph, 9 for the Holiday graph, 9 for the News graph, 10 for the Tires graph, 10 for the Prison graph, and 7 for the Plants graph (these categories are partially represented by the rectangles within the bars of Figure 4, though that visualization allows the viewer to see which categories are the same across designs; full data are available in the Supplementary Materials). This variety further validates the practitioner focus guideline, showing that different viewers will otherwise focus on many possible patterns of data.

Across redrawn graphs and free responses, there were no statistically significant interactions among design and topic, which is to be expected given that these combinations were manipulated between-subjects (a single participant only saw one design for each topic). Yet, we pause for a moment to speculate on why some topics showed bigger focus effects than others.

For the Car topic, while 'noise' issues were the topic of focus, they were also present across three of the top largest value bars. This pattern is therefore consistent not only with the strong focus on that topic across redrawings and conclusions, but also in the redrawings (less so the conclusions) of the cluttered and decluttered designs.

For the Holiday topic, there was only a small effect of focusing in the redrawings, possibly because of the complexity of the topic. It required the participant to notice (and draw or describe) women and men showing opposite patterns in some months compared to others.

The News topic showed only a small effect of the focus manipulation across both the redrawings and written conclusions. But this was curiously matched by the decluttered design, even though the decluttered design did not focus on the key pattern.

The Plants topic also showed only a small increase in focus on the focus-relevant topic, perhaps because even in the cluttered design, a salient set of diverging lines showed the same trend that was highlighted in the focus design.

The Prison topic showed a strong focus effect, with half of participants who viewed the focused design drawing the focused trend, and none doing so for drawings or free response conclusions in the non-focus designs. This may be due to the focused pattern (picking out two bars that both referred to drug offenses) being less intrinsically salient, leaving more room for the highlighting to guide viewers. In the absence of that guidance, most participants instead focused on 'Immigration' as the largest bar, and around half of the conclusions focused on Immigration being the most prominent cause of incarceration.

We also informally examined whether redrawn graphs included other features unrelated to the focus manipulation. We used a binary (Y/N) code for whether various features were present for the redrawn graphs to find the features that were most likely to be missing or present across each combination of design (cluttered, decluttered, and focused) and graph topic. For example, participants who saw the cluttered Car graph drew axis labels and arranged bars from largest to smallest more than participants who saw either the decluttered or the focused version. Figure S1 in Supplementary Materials displays each of the 18 graphs paired with a handpicked drawing that most reflects the features that tended to be present or missing for that combination. Each stimulus and associated redrawn graph is supplemented with annotations identifying the main changes and features. We intend this analysis to be illustrative and to be used as inspiration, and do not have sufficient data to make firm claims about these choices or trends.

6.2 Quantitative and Qualitative Ratings Discussion

Aesthetics and clarity ratings were trended toward being higher for focused graphs compared to decluttered. Open-ended responses were consistent with these ratings, with many describing the decluttered graphs as "boring" or "vague." Decluttered designs were rated higher on aesthetics and clarity compared to cluttered designs. The cluttered graphs often generated negative sentiments towards the excessive use of color in the cluttered graphs. Professionalism ratings were similar for decluttered and focused graphs, and both were higher compared to cluttered. For trustworthiness, there was no significant difference across designs, though some qualitative comments indicated that the cluttered graphs seemed to be made by children or that the focused graphs had an agenda.

6.3 Limitations and Future Directions

One limitation of the present study is that the quantitative ratings scales for aesthetics, clarity, professionalism, and trustworthiness were somewhat correlated to each other. Our scales were inspired by past work (e.g., aesthetics and clarity) [56,57], with the addition of two new scales that we thought would be important to measure (professionalism and trustworthiness). All of the metrics correlated with each other at R values of 0.5-0.7, suggesting that these concepts are not fully dissociable. Particularly strong relationships that merit future consideration include professionalism's correlation with both aesthetics and trustworthiness.

We relied on a recall task (drawings and typed conclusions) to simulate the experience of seeing many individual views of data across an article or live presentation. We acknowledge that the results of this recall task may not extrapolate for longer viewing scenarios. Future work could also explore the differing impacts of the three graph designs using online measures of task accuracy, response time, or even talk-aloud procedures.

Another limitation is the scope of the tested population. The guidelines tested here are intended for visualizations placed in front of a wide variety of audiences, including people making presentations in organizations, educational settings, and the press. While our undergraduate participant serves as an initial model that should generalize to some degree, important differences might arise for other audiences. For example, a student might be more accustomed to being asked to browse through data for patterns they find interesting, while a busy executive might have less patience with a non-focused graph, demanding the 'so what' immediately. Similarly, decluttered graphs led to some polarization, with some finding them boring, and with a few people finding the cluttered graphs to be 'professional' in a positive way, perhaps because they see those cluttered designs as more familiar (see [56,57] for similar findings).

Another limitation of the present work is that the declutter design manipulation included a large set of design changes that are typically prescribed in contemporary practitioner guides. While the results show no downside of decluttering, and a mild improvement in ratings, there are likely cases where our 'clutter' could be beneficial. For example, one recent blog post from a practitioner book author decried blanket

rules against gridlines, which can be critical for helping people read values far from an axis [73]. A solution to this might be to empirically evaluate the effect of *every* declutter guideline element separately (gridlines, color variety, white space…) for every conceivable task or measure (speed to pick a value, accuracy to pick a value, memory for a trend, …), presenting an imposingly large space of empirical tests. Given the existing work, we are now largely satisfied with the state of the existing empirical literature on the declutter guideline, and trust the designer community's intuitions for guidance on such contextually-dependent decisions – but stand ready to offer evaluation should they fail to reach a clear consensus.

Finally, the focused designs included not only color highlighting of the most relevant data points, but also an annotation sentence that verbally pointed the viewer to a particular pattern. We purposely used both of these additions to generate a maximal contrast to the decluttered designs, which had neither. Future work should tease apart the independent contributions of these two techniques, perhaps starting by adding the relevant annotation (without pointing to the data values) on the decluttered designs.

7 Conclusions

Compared to the focus guideline, we found weaker but generally positive evidence for the benefits of the declutter guideline, with only a handful of undergraduate students finding decluttered designs too simple, or diverging from the typical cluttered style that is familiar to them. In a recall task, we found strong evidence for benefits of the combination of decluttered and focused designs, for both preference ratings and for its ability to point the students toward a single critical pattern. Without this manipulation, there was an impressive diversity of other patterns that participants found more salient. Note again we cannot be sure that these results would extrapolate beyond our memory-based task to more 'online' understanding tasks.

We were surprised to find little existing empirical evidence in support of the power of the focus guideline. One might ask: Why bother testing its power – doesn't everyone know that highlighting and naming a pattern is important, because it will lead people to remember it? In contrast, despite the massive number of practitioner guides making this prescription, few people in the real world tend to do it. For the most common reader of this article – an academic – think of the rate at which people explicitly guide you to the patterns in their data across talks, posters, and papers; we suspect that your own value will be somewhere around 5%. The rest of the world – businesses, nonprofits, classrooms, and labs – would likely report a similarly low rate. One cause of this is likely to be the *curse of expertise*, where visualization designers assume that others see what they see in their data [61]. The designer may assume that the audience knows a similar amount, or even more, about a topic than they do, and will intuitively know what patterns to focus on. However, the designer is typically the person who has studied the topic and its context most deeply, putting them in a unique position to guide their audience.

There are also contexts where one might avoid the focus manipulation. The designer may not know what patterns are most important, and may want to show data to gain the audience's perspective, or to initiate a group brainstorm. The technique may also feel heavy handed at times – even some of our participants reported lower trust in the focused design because they found them 'pushy.' There may be too many patterns to highlight, especially in a static document. We do not use the technique for Figures 3 and 4 in this paper simply because there are too many trends to highlight without creating clutter for the paper's reader. We are intrigued by the technique showcased recently by Kale et al (2020) [74] who use creative annotation techniques to clearly connect multiple key patterns in data to the text descriptions of those patterns.

A final reason to test the power of the focus guideline is to provide empirical support for the guideline prescriptions of the community of book authors and workshop instructors who directly or indirectly train hundreds of thousands of people in data visualization design. This community is often asked for 'proof' that their guidelines work, especially from people with a 'curse of knowledge' that their visualizations are already perfect [61]. This proof is needed not just for the readers or attendees of sessions – an already receptive audience – but also for the colleagues and supervisors that they will need to convince the following week. We hope that the present results provide that evidence for the efficacy of the focus guideline for improving the audience's reception to and memorability of the data they want to communicate, the point they want to make, and the action they hope to inspire.

Acknowledgments

The authors wish to thank Evan Anderson for assisting with blind qualitative coding, Evan Lee for digitizing redrawn graphs, and Caitlyn McColeman and Miriam Novack for assisting on the inter-rater reliability analysis. Thanks also to Nicole Jardine and Cristina Ceja for their helpful comments on this paper. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1901485.

References

[1] Ian Parker. Absolute powerpoint. *The New Yorker*, 28:76–87, 2001.
[2] RJ Andrews. *Info We Trust: How to Inspire the World with Data*. John Wiley & Sons, 2019.
[3] Nicolas P Rougier, Michael Droettboom, and Philip E Bourne. Ten simple rules for better figures. *PLoS Comput Biol*, 10(9):e1003833, 2014.
[4] Scott Berinato. *Good charts: The HBR guide to making smarter, more persuasive data visualizations*. Harvard Business Review Press, 2016.
[5] Alberto Cairo. *The Functional Art: An introduction to information graphics and visualization*. New Riders, 2012.
[6] Alberto Cairo. *The truthful art: data, charts, and maps for communication*. New Riders, 2016.
[7] Jorge Camões. *Data at work: Best practices for creating effective charts and information graphics in Microsoft Excel*. New Riders, 2016.
[8] Nancy Duarte. *Data story: explain data and inspire action through story*. Ideapress Publishing, 2019.
[9] Nancy Duarte. *Slide: ology: The art and science of creating great presentations*, volume 1. O'Reilly Media Sebastopol, CA, 2008.
[10] Brent Dykes. *Effective data storytelling: how to drive change with data, narrative and visuals*. John Wiley and Sons, Inc., 2020.
[11] Stephanie DH Evergreen. *Effective data visualization: The right chart for the right data*. Sage Publications, 2019.
[12] Stephanie DH Evergreen. *Presenting data effectively: Communicating your findings for maximum impact*. Sage Publications, 2017.
[13] Stephen Few. *Now you see it: simple visualization techniques for quantitative analysis*. Analytics Press, 2009.
[14] Stephen Few. *Information dashboard design: The effective visual communication of data*. O'Reilly Media, Inc., 2006.
[15] Bruce R Gabrielle. *Speaking PowerPoint: The new language of business*. Insights Publishing, 2010.
[16] Bruce R. Gabrielle. *Storytelling with graphs: a new approach beyond data visualization*. Insights Pub., 2018.
[17] Ben Jones. *Avoiding data pitfalls: how to steer clear of common blunders when working with data and presenting analysis and visualizations*. Wiley, 2020.
[18] Ben Jones. *Communicating Data with Tableau: Designing, Developing, and Delivering Data Visualizations*. " O'Reilly Media, Inc.", 2014.
[19] Andy Kirk. *Data visualisation: a handbook for data driven design*. Sage, 2016.
[20] Cole Nussbaumer Knaflic and Catherine Madden. *Storytelling with data: lets practice!* Wiley, 2020.
[21] Cole Nussbaumer Knaflic. *Storytelling with data: A data visualization guide for business professionals*. John Wiley & Sons, 2015.
[22] Andy Kriebel and Eva Murray. *# MakeoverMonday: Improving how We Visualize and Analyze Data, One Chart at a Time*. John Wiley & Sons, 2018.
[23] Manuel Lima. *The book of trees: visualizing branches of knowledge*. Princeton Architectural Press, 2014.
[24] Isabel Meirelles. *Design for information: an introduction to the histories, theories, and best practices behind effective information visualizations*. Rockport publishers, 2013.
[25] Tamara Munzner. *Visualization analysis and design*. CRC press, 2014.
[26] Daniel G Murray. *Tableau your data!: fast and easy visual analysis with tableau software*. John Wiley & Sons, 2013.

[27] Jonathan Schwabish. *Better presentations: a guide for scholars, researchers, and wonks*. Columbia University Press, 2016.

[28] Jonathan A Schwabish. *Elevate the debate: A multilayered approach to communicating your research*. John Wiley & Sons, 2020.

[29] Edward R Tufte and David Robins. *Visual explanations*. Graphics Cheshire, CT, 1997.

[30] Edward R Tufte. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 2001.

[31] Edward R Tufte, Nora Hillman Goeler, and Richard Benson. *Envisioning information*, volume 126. Graphics press Cheshire, CT, 1990.

[32] Sejal Vora. *The Power of Data Storytelling*. SAGE Publications India, 2019.

[33] Colin Ware. *Information visualization: perception for design*. Elsevier, 2012.

[34] Colin Ware. *Visual thinking: For design*. Elsevier, 2010.

[35] Steve Wexler, Jeffrey Shaffer, and Andy Cotgreave. *The big book of dashboards: visualizing your data using real-world business scenarios*. John Wiley & Sons, 2017.

[36] Dona M Wong. *The Wall Street Journal guide to information graphics: The dos and don'ts of presenting data, facts, and figures*. New York: WW Norton & Company, 2010.

[37] Nathan Yau. *Data points: Visualization that means something*. John Wiley & Sons, 2013.

[38] Jean-Luc Doumont. Trees, maps, and theorems. *Brussels: Principiae*, 2009.

[39] David C Van Essen, Charles H Anderson, and Daniel J Felleman. Information processing in the primate visual system: an integrated systems perspective. *Science*, 255(5043):419–423, 1992.

[40] Michael Correll and Michael Gleicher. Bad for data, good for the brain: Knowledge-first axioms for visualization design. In *IEEE VIS 2014*, 2014.

[41] Douglas J Gillan and Edward H Richman. Minimalism and the syntax of graphs. *Human Factors*, 36(4):619–644, 1994.

[42] Huiyang Li and Nadine Moacdieh. Is "chart junk" useful? an extended examination of visual embellishment. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 58, pages 1516–1520. SAGE Publications Sage CA: Los Angeles, CA, 2014.

[43] Robert D Helgeson and Robert A Moriarty. The effect of fill patterns on graphical interpretation and decision making. Technical report, AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH, 1993.

[44] James D Kelly. The data-ink ratio and accuracy of newspaper graphs. *Journalism Quarterly*, 66(3):632–639, 1989.

[45] Michelle A Borkin, Azalea A Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, 2013.

[46] Steve Haroz, Robert Kosara, and Steven L Franconeri. Isotype visualization: Working memory, performance, and engagement with pictographs. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1191–1200, 2015.

[47] Michelle A Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. Beyond memorability: Visualization recognition and recall. *IEEE transactions on visualization and computer graphics*, 22(1):519–528, 2015.

[48] Scott Bateman, Regan L Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. 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*, pages 2573–2582, 2010.

[49] Yu-Sung Su. It's easy to produce chartjunk using microsoft® excel 2007 but hard to make good graphs. *Computational Statistics & Data Analysis*, 52(10):4594–4601, 2008.

[50] Stephen Kosslyn. *Better PowerPoint (R): Quick Fixes Based On How Your Audience Thinks*. Oxford University Press, 2010.

[51] Andy Kirk. *Data Visualization: a successful design process*. Packt Publishing Ltd, 2012.

[52] Stephen Few. Show me the numbers. *Analytics Pres*, 2004.

[53] Anthony J Blasio and Ann M Bisantz. A comparison of the effects of data–ink ratio on performance with dynamic displays in a monitoring task. *International journal of industrial ergonomics*, 30(2):89–101, 2002.

[54] Lyn Bartram and Maureen C Stone. Whisper, don't scream: Grids and transparency. *IEEE Transactions on Visualization and Computer Graphics*, 17(10):1444–1458, 2010.

[55] Lyn Bartram, Billy Cheung, and Maureen Stone. The effect of colour and transparency on the perception of overlaid grids. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):1942–1948, 2011.

[56] Stephen Hill, Barry Wray, and Christopher Sibona. Minimalism in data visualization: Perceptions of beauty, clarity, effectiveness, and simplicity. *Journal of Information Systems Applied Research*, 11(1):34, 2018.

[57] Ohad Inbar, Noam Tractinsky, and Joachim Meyer. Minimalism in information visualization: attitudes towards maximizing the data-ink ratio. In *Proceedings of the 14th European conference on Cognitive ergonomics: invent! explore!*, pages 185–188, 2007.

[58] Danielle Albers Szafir, Steve Haroz, Michael Gleicher, and Steven Franconeri. Four types of ensemble coding in data visualizations. *Journal of vision*, 16(5):11–11, 2016.

[59] Christine Nothelfer and Steven Franconeri. Measures of the benefit of direct encoding of data deltas for data pair relation perception. *IEEE transactions on visualization and computer graphics*, 26(1):311–320, 2019.

[60] Priti Shah and Eric G Freedman. Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in cognitive science*, 3(3):560–578, 2011.

[61] Cindy Xiong, Lisanne van Weelden, and Steven Franconeri. The curse of knowledge in visual data communication. *IEEE transactions on visualization and computer graphics*, 2019.

[62] Robert Kosara and Jock Mackinlay. Storytelling: The next step for visualization. *Computer*, 46(5):44–50, 2013.

[63] Edward Segel and Jeffrey Heer. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148, 2010.

[64] Jeremy Boy, Francoise Detienne, and Jean-Daniel Fekete. Storytelling in information visualizations: Does it engage users to explore data? In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1449–1458, 2015.

[65] Jessica Hullman and Nick Diakopoulos. Visualization rhetoric: Framing effects in narrative visualization. *IEEE transactions on visualization and computer graphics*, 17(12):2231–2240, 2011.

[66] storytelling with data. `http://www.storytellingwithdata.com/`. Accessed: 2020-04-21.

[67] RVAideMemoire. `https://www.rdocumentation.org/packages/RVAideMemoire/versions/0.9-73/topics/pairwise.perm.manova`. Accessed: 2020-04-15.

[68] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191, 2007.

[69] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012.

[70] Douglas Bates. Fitting linear mixed models in r. *R news*, 5(1):27–30, 2005.

[71] Anselm Strauss and Juliet Corbin. Grounded theory methodology. *Handbook of qualitative research*, 17:273–85, 1994.

[72] Douglas M Bates. lme4: Mixed-effects modeling with r, 2010.

[73] You probably need gridlines. `https://stephanieevergreen.com/you-probably-need-gridlines/`. Accessed: 2020-04-21.

[74] Alex Kale, Matthew Kay, and Jessica Hullman. Visual reasoning strategies and satisficing: How uncertainty visualization design impacts effect size judgments and decisions. *arXiv preprint arXiv:2007.14516*, 2020.