This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3372620

Examining Limits of Small Multiples: Frame Quantity Impacts Judgments with Line Graphs

Helia Hosseinpour, Laura E. Matzen, Kristin M. Divis, Spencer C. Castro, and Lace Padilla

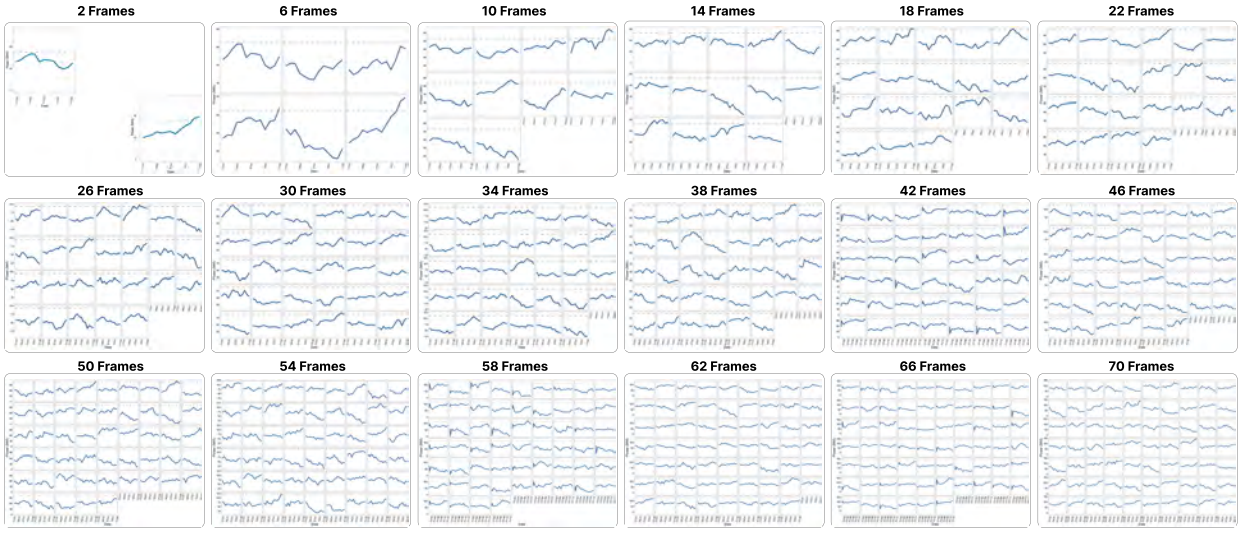

Fig. 1: Example of the small multiple stimuli used in Experiment 1 that varied in frame quantity from 2 to 70, incremented by four frames. The stimuli depicted power (in megawatts) over time (one year per frame).

***Abstract*—Small multiples are a popular visualization method, displaying different views of a dataset using multiple frames, often with the same scale and axes. However, there is a need to address their potential constraints, especially in the context of human cognitive capacity limits. These limits dictate the maximum information our mind can process at once. We explore the issue of capacity limitation by testing competing theories that describe how the number of frames shown in a display, the scale of the frames, and time constraints impact user performance with small multiples of line charts in an energy grid scenario. In two online studies (Experiment 1 n = 141 and Experiment 2 n = 360) and a follow-up eye-tracking analysis (n = 5), we found a linear decline in accuracy with increasing frames across seven tasks, which was not fully explained by differences in frame size, suggesting visual search challenges. Moreover, the studies demonstrate that highlighting specific frames can mitigate some visual search difficulties but, surprisingly, not eliminate them. This research offers insights into optimizing the utility of small multiples by aligning them with human limitations.**

***Index Terms*—Small multiples, time-series data, cognition**

I. Introduction

Small multiples have been a popular visualization technique since the late 1800s [1]. They present different views of a dataset through multiple small frames [2]. These frames maintain a consistent scale and axes and are typically arranged in a two-dimensional grid layout [3]–[5]. Visualizations from displays of election data to population demographic shifts utilize small multiples [6], [7]. Their popularity can likely be attributed to extensive research demonstrating their effectiveness in conveying complex data trends [8]–[12].

Although previous research has demonstrated small multiples' effectiveness [8], their potential constraints remain underinvestigated. Notably, human *cognitive capacity limits* and *visual search capabilities* could inform guidelines for designing small multiples for large datasets. Cognitive capacity limits refer to the maximum amount of information the mind can effectively process and retain at any given moment [13]. Visual search capabilities are how the visual system completes pattern recognition and identifies task-relevant items in an array [14]. Cognitive capacity limits and visual search capabilities are two theories that could inform when and how small multiples become less effective. However, no work has evaluated how cognitive processes drive the use of small multiples.

Missing guidance about the cognitive underpinning of small multiples becomes apparent in environments such as energy grid control rooms (see Figure 2). Energy grid control displays might feature over 100 small multiple frames, potentially challenging cognitive limits, and visual search capabilities. Without guidance on the impact of frame quantity on performance, designers may produce displays that cognitively overload analysts, potentially leading to errors, such as data misinterpretation or oversight. Understanding the relationship between cognitive mechanisms and small multiples is even more critical when analysts face time [15] or display constraints [16], which underscores the need for aligning small multiple design guidelines with human cognitive abilities.

We created small multiple line charts to explore capacity limits and visual search capabilities, and to provide cognitively informed guidance to small multiple designers. In the first experiment, we examined the relationship between frame quantity and the accuracy of responses across seven visualization tasks. In the second experiment, we controlled for

Fig. 2: Pennsylvania-New Jersey-Maryland (PJM) Interconnection control room, using numerous small multiples [17].

the scale of the frames and the presence of time constraints. Finally, in a small follow-up eye-tracking investigation, we explored participants' visual search strategies during the tasks.

Overview and Contributions: This work contributes the first empirical evidence that increasing the number of frames in small multiples results in a linear decline in accuracy. Although the notion that larger frame quantities are more difficult to use may seem obvious, this observation challenges the practice of using small multiples in vast datasets without adding appropriate cognitive supports. For example, we found that highlighting specific frames for comparison can mitigate the negative impact of increased frame quantities on accuracy. The second significant contribution of this work is that it provides converging evidence that visual search strategies are an essential driver of small multiples usage. The implications of this finding are that interventions that focus on making the visual search process easier may provide the most benefits for small multiples. Visual search interventions would allow users to explore specific frames selectively, such as zooming, panning, and filtering [18], reducing the reliance on visual search. This work also provides an example of using converging evidence to test competing cognitive theories, which is a practice that could strengthen theory development in visualization research [19]–[21].

II. RELATED WORK

A. Small Multiples

Small multiples can mitigate the risk of overlooking critical information patterns by providing a broader view of the data [12]. Their application includes multifaceted exploration [22], pattern exploration [23], and general data exploration [12]. Small multiples also facilitate direct visual comparisons, particularly for comparing, exploring, and analyzing time series data [8], [9], [24]–[27]. Furthermore, they facilitate data parameterization [28], [29].

Researchers have examined various tasks with small multiples, including topology-based, adjacency, and connectivity tasks [11]; trend comparison [8], [10]; maximum, slope, and discrimination commonly used for temporal visualizations [9]; and visual exploration such as identify, correlate, compare, and cluster tasks [12]. These tasks often involve questions with correct answers, allowing for accuracy assessment.

A study examining the effectiveness of trend animation in simulated presentation and analysis scenarios found that small multiples result in more accurate visualization analysis than trend animation [8]. Additionally, studies have found that small multiples lead to expedited task completion (although results can be task dependent), fewer errors, and improved accuracy when contrasted with trend animation [10], [11], [30]. Scholars recommend maintaining the same encoding across frames [31]–[33] and chart type arrangements [34]. Another study addressed small multiples' success with bar and line encodings across resolutions and numbers of displays [35].

Finally, previous research explored different quantities of frames. For instance, in the study comparing trend animation, a static image, and small multiples, researchers performed tests with small (i.e., 18) and large (i.e., 80) quantities of frames, finding that participants made fewer errors with the smaller dataset [8]. Across the spectrum of studies, frame quantities tested have included 2 [27], [34], 4 [9], 12 [36], 16 [10], with some studies capping the number at 25 [12], 48 [37], or even 1,178 frames [4]. However, questions remain regarding the performance of different task types over different quantities of frames, especially within the framework of cognitive capacity.

B. Cognitive Effort vs. Visual Search Theories

To investigate the relationship between the number of frames and task performance in small multiples, we conducted a series of studies designed to test two competing cognitive theories. These theories make distinct predictions about how accuracy varies with different frame quantities.

The first theory pertains to cognitive capacity limits [38]. It suggests that for tasks where users need to retain information mentally, there is a threshold beyond which they cannot or will not hold more information [39]. Previous research in cognitive science has shown that performance declines once users reach their capacity limit [40]. If the cognitive capacity limit theory describes performance with small multiples, there should be a point at which performance dramatically declines for tasks requiring users to store data points in their minds and compare across frames. This pattern would resemble an inverse sigmoidal curve as seen in past work (e.g., [40], [41]). Understanding the specific number of frames at which performance deteriorates would be crucial for designers to know and account for in their designs.

Another viable theory suggests that working with small multiples does not rely heavily on cognitive capacity. Instead, users might rely on their visual system for pattern recognition and identification of specific data points. Visual search tasks use cognitive effort, but the effort does not increase drastically with larger search arrays [14]. Based on the visual search theory, we would expect a linear decline in performance as the number of frames increases [14], [42]. Such a decline could imply that the main challenge of adding more frames is that it becomes more time-consuming and requires greater effort for users to identify the interesting frames. Determining which theory better describes performance will help designers create visual supports that cognitively assist their users.

III. METHODS AND AIMS

A. Experiments

In this research, Experiment 1 assessed the impact of increasing frame quantity in full-scale small multiples on

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3372620

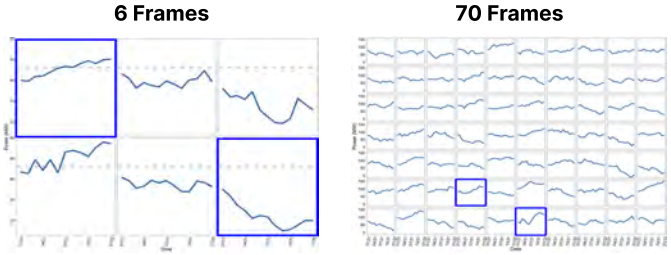

Fig. 3: Example stimuli used in the Compare tasks where two frames are highlighted in blue.

participants' accuracy. Experiment 2 introduced controlled conditions for frame scale and time constraints. Finally, we conducted a supplemental eye-tracking study to examine participants' visual search patterns.

Experiment 1: In this experiment, we evaluated how an increase in small multiple frames impacted the accuracy of participants' responses. We created a set of small multiples with full-scale frames that varied in frame quantity (see Figure 1). *Scale* refers to the size of the frames, and *full-scale* means that the frames were scaled to fill the screen.

Experiment 2: For the second experiment, we added two conditions to control the frames' scale and the time taken to complete the tasks. We created a set of fixed-scale stimuli for which all frame sizes remained equally sized regardless of frame quantity (Figure 4 shows examples). We also controlled for the time taken to complete each task. We presumed that with larger frame quantities, participants would take longer to complete the tasks, which would interact with accuracy. We included a condition in which participants had unlimited time and one where they had 40 seconds to complete each task. We chose 40 seconds by testing the study with our five research assistants and averaging their times.

Follow-up eye-tracking: Finally, we conducted an exploratory eye-tracking study to examine how different tasks and frame sizes impacted participants' visual search strategies.

B. Tasks, Design, and Procedures

To complete the tasks, the participants were asked to assume the role of an energy grid operator, and their job was to monitor the energy output. We used Brehmer and Munzner's [43] multilevel task typology to guide the selection of the following seven tasks under the **Identify**, **Compare**, and **Summarize** query phases.

- Identify 1: *Click on one graph with the highest peak power.*
- Identify 2: *Click on one graph with both the biggest change and the highest average power.*
- Identify 3: *Click on all graph(s) where the blue line goes above the dashed gray line.*
- Compare 1: *Of these two graphs highlighted in blue, click on one graph with the highest peak power.*
- Compare 2: *Of these two graphs highlighted in blue, click on one graph with both the biggest change and the highest average power.*
- Summarize 1: *Is the general trend in the graphs going up or down?*
- Summarize 2: *What is the average power across all plots?*

The Identify and Compare tasks required participants to click on the frame/s. The Summarize 1 task was a multiple-choice question, and Summarize 2 was an open-ended question for which the participants gave numerical answers. Each task had a correct answer, detailed in Section IV. The accuracy of participants' responses was evaluated using the correct answers, but they did not receive performance feedback.

We designed each task to examine different aspects of the two competing theories we were interested in testing in this work. We will describe our detailed predictions in the following section (III-C).

C. Hypothesis

To test the competing theories, we developed a series of tasks aimed at probing aspects of each theory. We designed the Identify 1 and 3 tasks to rely primarily on visual search, where participants could scan the full display and find the frame with the highest power or those that crossed a threshold.

For the Identify 2 task, participants had to hold two pieces of information in their minds (the biggest change and the highest average power) and update them as they compared the frames. Holding several pieces of information in the mind requires mental storage and, therefore, non-negligible amounts of cognitive capacity. Similarly, both Summarize tasks required participants to mentally average the data across all the frames. We expected this activity to be highly cognitively demanding and to become more demanding with more frames. We also did not want to test conditions that confirmed only our predictions. Therefore, we designed the Compare tasks not to require any serious visual search or cognitive demands as a control.

We designed the tasks to explore aspects of the various theories, but we did not know which theory was correct. Therefore, we preregistered hypotheses more generally about overall accuracy across the tasks on the Open Science Framework (OSF), which were:

- **H1A**: For the Identify and Summarize tasks, we hypothesized that accuracy would worsen with increased frame quantity.
- **H2A**: For the Compare tasks, we hypothesized that there would be no change in performance accuracy as frame quantity increased because participants were comparing two frames.

We conducted Experiment 2 to control for some possible confounds in Experiment 1. In particular, as the number of frames increased, the size of the frames became smaller. It was unclear from the findings of the first experiment if our results were due to the increasing frame quantity or the decreasing resolutions of each frame. The other possible confound in Experiment 1 was that participants could take longer to scan the small multiples with larger frame quantities. To control for both, we ran Experiment 2, in which we compared small multiples with fixed-sized frames to frames scaled to fill the screen. We also included a condition in which participants had unlimited time or time constraints. For the tasks that we designed to rely on visual search and cognitive capacity, we

hypothesized that introducing time constraints would magnify the effects in those conditions by not allowing participants to improve their accuracy by simply taking longer. We preregistered the following hypotheses regarding time pressure.

- **H1B**: For the Identify and Summarize tasks, we hypothesized that there would be differences in performance accuracy between participants with or without time constraints.
- **H2B**: For Compare tasks, we hypothesized that there would be no difference in performance accuracy between participants with or without time constraints.

Design. Experiment 1 used a 7 (Identify 1, Identify 2, Identify 3, Compare 1, Compare 2, Summarize 1, and Summarize 2 tasks) $\times$ 18 (2 to 70 frames, incremented by four) $\times$ 20 (randomly generated seeds) mixed study design. The tasks were between-subjects, and we randomly assigned participants to one of the task groups. Frame quantities (see Figure 1) and randomly generated seeds were within subjects. Frame quantities were blocked, so participants completed the 20 trials for each frame quantity together. We randomized the order in which they completed the blocks. We also presented the 20 trials within each block in a randomized order, which used a different seed to generate different plots of the same frame quantity. Each participant completed a total of 360 trials.

Experiment 1 was intentionally long and served the purpose of creating a normed dataset for determining accuracy across various stimuli, a common practice in psychological sciences (e.g., [44]). This practice informed our selection of representative stimuli for Experiment 2, ensuring that outliers due to data generation aspects were avoided.

Experiment 2 was also a mixed 4 (full-scale unlimited time, full-scale time-constrained, fixed-scale unlimited time, and fixed-scale time-constrained) $\times$ 7 (Identify 1, Identify 2, Identify 3, Compare 1, Compare 2, Summarize 1, and Summarize 2 tasks) $\times$ 5 (2, 6, 10, 30, and 70 frames) design. The scale and time constraints were between-subjects, and we randomly assigned participants to one of four between-subjects groups. Tasks and frame quantities were within subjects. Tasks were blocked, so participants completed the five trials for each task together. We randomized the order in which they completed the blocks. We presented participants with the five trials of varying frame quantities within each block in a randomized order, totaling 35 trials. We selected a subset of the frame quantities to be tested with the seven tasks because the results of Experiment 1 revealed a linear trend in accuracy. We, therefore, needed up to only five frame quantities to show the trend again. Additionally, we aimed to mitigate participant fatigue, which was a concern in the lengthy Experiment 1. Experiment 2 was preregistered on OSF (link).

In the exploratory eye-tracking study, participants completed nine trials using a small subset of the stimuli from the prior experiments. The participants completed each of the seven tasks once using a 70-frame stimulus. These trials assessed how the participants' patterns of eye movements changed for the different tasks, using the maximum set size to elucidate the differences between tasks. The other two trials used full-scale and fixed-scale versions of a six-frame stimulus, which participants used to complete the Identify 3 task. The

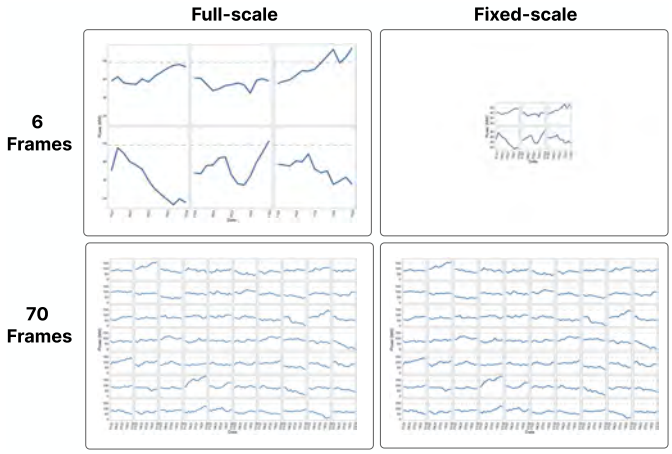

Fig. 4: Illustration of the scale condition in Experiment 2 for the small multiples with six (top) and 70 (bottom) frames. On the left is the full scale, and on the right is the fixed scale example.

Identify 3 task was selected because Experiment 2 showed that changing the frame size had a significant impact on accuracy for this task. We used the six-frame stimuli because they had enough frames to require visual search and also had a substantial size difference between the fixed- and full-size frames. The nine trials were presented in a different random order for each participant.

Procedures: The first two experiments were conducted using the online survey software Qualtrics [45]. Participants provided IRB-informed consent and were required to have a screen size of at least 9.4 $\times$ 6.6 inches. They received instructions about how to zoom their browser window to 100% and had to confirm that they completed the steps. Next, participants received instructions about the experiment context, how to read the visualizations, and how to do the tasks. Then, they practiced clicking on a frame, including instructions on deselecting it if they wanted to change their answer (full instructions on OSF). Following the instructions, participants completed a prescreening attention check, and only those who passed proceeded to the study. Demographic data were collected at the end of the experiments.

Participants in Experiment 2 with time constraints were given 40 seconds per question, and a practice question was included to ensure readiness. After the time limit, the screen auto-progressed with different messages based on the participants' performance. If they answered the questions in time, they were allowed to progress. If they did not answer the question in time, the system warned them and prompted them to click the next button when they were prepared.

During the eye-tracking study, participants completed nine trials while their eye movements were tracked with a Tobii Spectrum eye tracker recording at 1200 Hz. They were seated with their eyes approximately 60 cm from the computer monitor but could move their heads freely. The stimuli were presented on a monitor with 1920 $\times$ 1080 resolution and were 26.5 cm by 19.5 cm on the screen or approximately 25 by 18.5 degrees of visual angle. The individual frames subtended approximately 2.4 degrees of visual angle for the fixed-scale

stimuli and 7.6 degrees of visual angle for the full-scale stimuli. Prior to the task, the eye tracker underwent a nine-point calibration process. At the start of each trial, participants received task instructions on the screen and initiated the task by clicking the mouse. Next, a fixation cross was presented in the center of the screen for one second, followed by the stimulus, which remained on the screen until the participant finished the task for that trial.

D. Stimuli

For Experiment 1, we generated small multiple line charts with full-scale frames. These stimuli consisted of 18 frame quantities ranging from 2 to 70, shown in Figure 1, with each quantity having 20 versions. The small multiples depicted energy output, with wattage on the y-axis and time (one year per frame) on the x-axis across different regions (each frame represented one region). We simulated 20 datasets for each frame quantity using R [46] by generating time series data, allowing the trend lines' slope, variance, and intercept to vary. The datasets included columns for time, power, and names. We repeated the date sequence for the required quantity of frames and used random numbers (seeded) from a normal distribution to generate power values for each frame. We then combined the columns to create 20 datasets per frame quantity and stored them in a list (stimuli available on OSF).

The frame array and line charts for each small multiple were generated using the facet wrap function within the *ggplot2* package [47] in R. Additionally, we incorporated a dashed gray line into each frame, indicating a fictional critical energy threshold. We included the threshold lines primarily for our Identify 3 task, in which participants were asked to identify the frames in which the blue line exceeded the dashed gray line. We determined the threshold line value in a manner that guaranteed only two frames in each small multiple surpassed the threshold. Further, for Compare tasks, we highlighted the frames we wanted the participants to compare (see Figure 3). We independently selected the highlighted frames' locations for each quantity to ensure randomness. However, we maintained the distance between frames, ensuring that participants never had to compare widely spaced frames. The distance between the highlighted frames was determined based on the two-frame small multiple and consistently applied to all other small multiples. For each small multiple, the same two frames were highlighted for all participants.

In Experiment 2, we tested five quantities of the small multiples from Experiment 1 (2, 6, 10, 30, and 70) because the results of the first experiment revealed a linear relationship between accuracy and frame number. We selected 2, 6, and 10 as the first three intervals, 30 as the midpoint, and 70 as the last point. We tested both their full- and fixed-scale versions (also created in R) using the same data (see examples in Figure 4).

The follow-up eye-tracking study used one 70-frame stimulus from Experiment 1 and the fixed- and full-size versions of a six-frame stimulus from Experiment 2.

E. Participants

We conducted Experiments 1 and 2 online and recruited participants from Prolific [48] who were paid California minimum wage ($15 per hour). Experiment 1 took approximately 1.5 hours to complete, and Experiment 2 took approximately 15 minutes. All participants were prescreened to be at least 18 years old, residing in the United States, and limited to participation in only one of the two online experiments.

A total of 141 participants completed Experiment 1. There were 20 participants in each task condition, with the exception of the Summarize 1 task, which had 21 participants. This sample size was adequate due to the large number of trials (360 trials) in Experiment 1 [49], [50]. There were 70 male, 65 female, and 4 nonbinary/third gender participants. One participant did not indicate their gender. The mean age of participants was 33.48 (SD = 11.72). There were 360 participants in Experiment 2, with 90 in each of the four between-subject conditions. We determined the required sample size (90 participants per group) based on an anticipated effect size calculated from prior work ($f^2 = .09$) using the *pwr* [51] package in R [46]. The participant breakdown was as follows: 171 male, 174 female, 10 nonbinary/third gender, and 2 preferred not to indicate their gender. The mean age of participants was 25.64 (SD = 15.4).

The eye-tracking experiment was collected in person at Sandia National Laboratories. The five participants (two male and three female) were Sandia employees who were compensated for their time at their normal hourly rate. It took the participants approximately 8 minutes to complete the eye-tracking study, including the calibration phase.

IV. RESULTS

We analyzed the participants' accuracy for each task via multilevel logistic and linear regression models using R [46] with the *tidyverse* [52] and fitted mixed-effects models with the *lme4* [53] packages. We generated the visualizations using the *ggplot2* [47] and *ggdist* [54] packages. In six of the tasks, we implemented binomial logistic regressions to model accuracy with coded levels of *0 = incorrect* and *1 = correct*. Participants responded to the seventh (i.e., Summarize 2) task with an open-ended question, and we modeled the continuous absolute error with a linear regression equation. The data and analysis are in the supplemental materials (OSF link).

Calculations were made to determine the correct responses for each task. For **Identify 1**, correct responses were the frame with the highest value. **Identify 2's** correct responses were selected by multiplying the average power value by the range of power values for each frame and determining the largest number. The coded correct responses for **Identify 3** were all the frames for which the power value was larger than the horizontal line value. False positives (i.e., selecting an extra frame) and False negatives (i.e., failing to select a frame) both resulted in an incorrect response. For **Summarize 1**, if the starting power value was smaller than the ending power value, then the response *up* was coded as correct and vice versa. The **Summarize 2** task required participants to respond with a value they believed to be the average power across all frames. This value was compared to the true value to determine the

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3372620

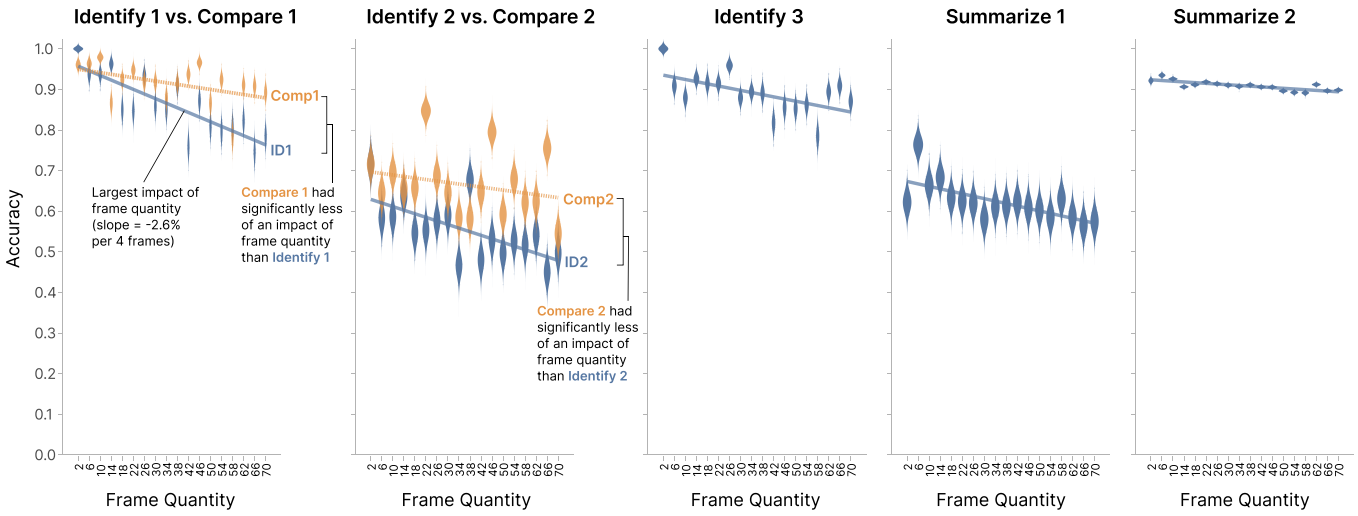

Fig. 5: Experiment 1 results showing the impact of frame quantity on accuracy for Identify, Compare, and Summarize tasks. Identify 1 and 2 (blue solid lines) are plotted in the same panel as Compare 1 and 2 (orange dashed lines) for direct comparisons.

absolute error, which was then scaled as a proportion of the possible values based on the range of power values displayed.

$$Accuracy = 1 - \frac{Absolute\ Error}{Maximum\ Power\ Value}$$

Of the 7,200 trials, two were not recorded by the Qualtrics online survey software [45]. Four of the 7,198 total responses were removed as they were above the maximum possible power value, resulting in 7,194 total observations.

A. Experiment 1

Experiment 1 examined the relationship between accuracy and frame quantity per task. We predicted that an increase in frame quantity would lead to a decrease in accuracy for the Identify and Summarize tasks (**H1A**) but not for the Compare tasks (**H1B**). To test H1A, we conducted six multilevel logistic regressions. We used frame quantity (2-70 incremented by four) as a predictor for accuracy (the outcome variable) in the Identify and Summarize 1 tasks, with random intercepts for the seeds (20 repeated trials of small multiples) and participants (also 20). Accuracy was modeled using this equation:

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot \text{Frame\_Quantity} + b_{\text{Trial}} + b_{\text{Participant}}$$

where $\text{logit}(p)$ represents the log-odds of the probability (p) of the binary outcome between 0 and 1. β_0 is the intercept coefficient, β_1·Frame_Quantity is the slope of frame quantity that is the main predictor, and b_{Trial} and $b_{\text{Participant}}$ are the random intercepts associated with the grouping variables *Trial* and *Participant*, respectively. We performed a similar analysis for Summarize 2 by fitting a linear model to the data to predict the continuous untransformed error. Significance was determined if the p-value (p) was below 0.05 and the confidence intervals (CI) did not include 0 [50]. For the Identify and Summarize tasks (shown in Figure 5 with blue solid regression lines), the **frame quantity significantly reduced accuracy and increased error in each task (H1A confirmed)**. The model output is shown in Table I.

TABLE I: Model output for the impact of frame quantity on accuracy in Experiment 1 (left), sorted based on the size of the slope. The right shows p-values for the direct comparisons between tasks. *** is p< .001, ** is p< .005, and * is p< .05.

	Impact of Frame Quantity					Task Comparisons					
Task	*b*	z, t	*M*	*SD*	*p*	I1	I3	C1	I2	S1	C2
I1	−.026	−14.5	.85	.36	**		***	***	***	***	***
I3	−.016	−7.8	.89	.31	**	***			*	***	***
C1	−.015	−6.8	.92	.28	**	***				**	***
I2	−.01	−8.1	.57	.50	**	***	*				**
S1	−.007	−6.4	.61	.49	**	***	***	**			
C2	−.005	−3.8	.67	.47	**	***	***	***	**		
S2	−.0004	−12.4	.91	.10	**	-	-	-	-	-	-

To test H1B, we conducted the same analysis on the Compare tasks and found that **frame quantity also significantly predicted accuracy in Compare 1 and 2 (H1B unconfirmed)**. It was surprising that accuracy still decreased when the two relevant frames were highlighted for the participants. Although this decline was significant for the Compare tasks, it was less pronounced than in the Identify tasks, which asked the same question but required participants to assess all the frames instead of focusing on two. We plotted the Compare tasks (orange) with the corresponding Identify tasks (blue) in Figure 5 (panels 1 and 2). As conveyed by the differences between the slopes, highlighting the relevant frames reduced the impact of frame quantity, but did not eliminate it.

Follow-up analysis: Identify vs. Compare. We conducted a follow-up analysis to compare the effect of frame quantity in Identify 1 to Compare 1 and Identify 2 to Compare 2. The goal was to determine if there was a significant difference between the slopes of the Identify and Compare tasks. We used the same modeling procedures as previously described but with accuracy predicted by task (Identify vs. Compare), frame quantity, and an interaction between the two. In both

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3372620

models (Identify 1 vs. Compare 1, and Identify 2 vs. Compare 2), the results revealed significant interactions between task and frame quantity, suggesting that the effect of frame quantity is significantly different for the two tasks. For the model comparing Identify 1 to Compare 1 the slope of Identify 1 (b = -.026) was significantly steeper than Compare 1 (b = -.015) (b = .01, $z(14{,}396) = 3.9$, $p < .001$, CI[.005, .017]). A significant interaction also indicated that Identify 2 (b = -.01) had a meaningfully steeper slope than Compare 2 (b = -.005) (b = .005, $z(14{,}396) = 3.08$, $p = .002$, CI[.002, .009]). These findings are annotated in Figure 5. Although H1B was not strictly confirmed, highlighting the frames can meaningfully improve accuracy in the tasks we tested.

Follow-up analysis: task comparison. We used the same multilevel logistic regression equation described above, with accuracy predicted by an interaction between frame quantity and task, while changing the referent task in the model to compute comparisons between all tasks (see Table I). We did not include Summarize 2 in this analysis because we did not feel that multiple-choice responses could fairly be compared to a continuous measure of accuracy. Our analysis revealed that the impact of frame quantity was significantly larger for Identify 1 (slope: -2.6% per 4 frames) than the five other tasks. These comparisons suggest that participants' ability to find the frame with the highest power (Identify 1) was the most difficult task for larger frame quantities in small multiple line charts in the context of the stimuli and tasks tested in this study.

Follow-up analysis: strategies for Identify 2 and Compare 2. In our analysis, we observed that Identify 2 and Compare 2 (second panel of Figure 5) had the highest variability. These tasks required participants to select a frame with *both* the greatest change and the highest average power. Surprisingly, even in the case of Compare 2, with two highlighted frames, accuracy was low (M = 67%, SD = 47%). This low accuracy might be attributed to participants not performing both parts of the tasks concurrently. They instead picked the frame with either the greatest change or the highest average power. In our original analysis, the correct frame had the highest mean × range value. Here, we separately identified the frame with the highest average wattage and the frame with the widest range to evaluate if the participants performed only one part of the task. Figure 6 illustrates the variations in accuracy calculations. For Identify 2 (top row of Figure 6), there is no clear indication that participants exclusively picked the frame with the greatest range or the highest mean. However, for Compare 2 (second row), it does appear that participants more commonly selected the frame with the highest mean.

Discussion: The primary finding from Experiment 1 was that small multiples with more frames were more challenging to use for all tasks we tested. The consistently significant effect of frame quantity was surprising for the Compare tasks. However, the results revealed that the effect of frame quantity was significantly smaller for them than for the corresponding Identify tasks. A limitation of Experiment 1 is that a confound between frame quantity and the size of the frames could impact these findings. In the stimuli for Experiment 1, we allowed the small multiples to fill up the screen, resulting in large sizes when fewer frames were presented and small sizes with

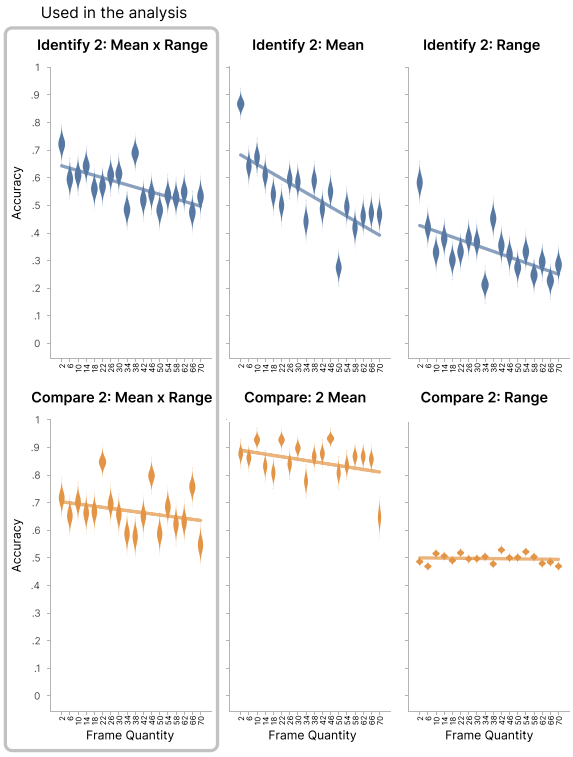

Fig. 6: Accuracy results in Experiment 1 for Identify 2 (top row) and Compare 2 (bottom row) tasks using various accuracy calculations. Columns represent different accuracy calculations: mean × range, the highest mean value, and the largest range (from left to right, respectively).

increased frames (see Figure 4). Another consideration with larger frame quantities is that participants might take longer searching through all the frames. To control for frame size and time to complete the tasks, we conducted Experiment 2.

B. Experiment 2

The goal of Experiment 2 was to determine the effects of scale and time to complete the tasks. We hypothesized that time constraints would impact accuracy in the Identify and Summarize tasks (**H1B**) but not the Compare tasks (**H2B**). To test these hypotheses, we conducted multilevel model analyses in which accuracy was predicted by frame quantity (2, 6, 10, 30, vs. 70 frames), time constraints (unlimited time vs. 40 seconds), and frame scale (full-scale vs. fixed-scale), with random slopes for each participant. In these models, the untimed and fixed-scale conditions were the referents.

This analysis revealed that **time constraints significantly reduced accuracy for Identify 1** (b = -0.56, $z(1{,}824)$ = -3, p = .003, CI[-.19, -.92]) and **Identify 3** (b = -1.37, $z(1{,}824)$ = -5.6, $p < .001$, CI[-.89, -1.85]). The meaningful differences in accuracy for Identify 1 and 3 are annotated in Figure 7. Although this finding supports H2B (no impact of time constraints for Compare tasks), it does not fully support H1B, which predicted time constraints to affect all Identify and Summarize tasks. This finding does, however, support the theory that the limiting factor of small multiples with larger frame quantities is the more extensive visual search

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3372620

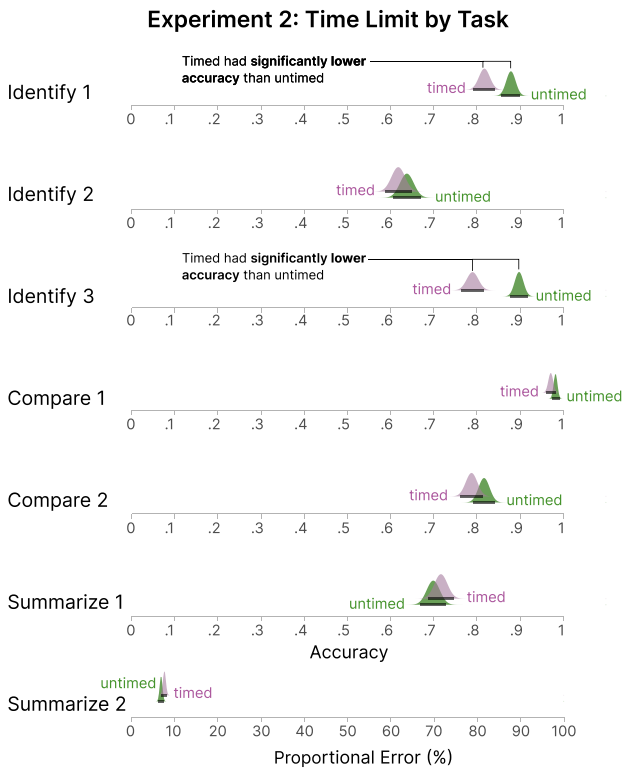

Fig. 7: Experiment 2 accuracy results for the timed condition in lavender and the untimed condition in green. Time constraints significantly reduced the accuracy of Identify 1 and 3 tasks.

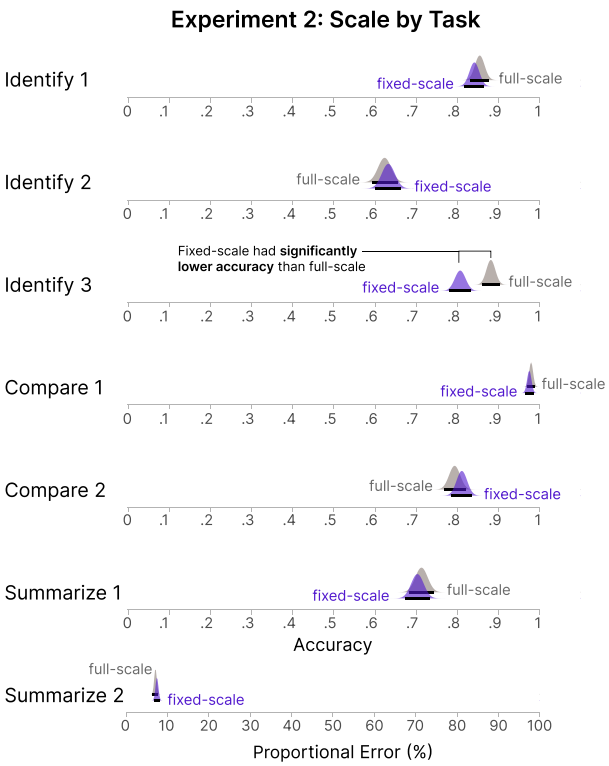

Fig. 8: Experiment 2 accuracy results for the fixed-scale condition in purple and the full-scale condition in gray. The fixed-scale graphs had significantly lower accuracy than the full-scale graphs for Identify 3.

they require rather than cognitive capacity limitations. The two tasks impacted by time constraints possibly involved a serial search process, in which participants scanned each frame to identify the highest point (Identify 1) or when data exceeded a threshold (Identify 3). Time constraints would hinder a thorough search and reduce accuracy in such cases. In contrast, Summarize tasks likely relied on *ensemble processing*, in which the visual system extracts the overall essence without an exhaustive search [55]. To explore participants' search strategies, we conducted an eye-tracking study to analyze gaze patterns for each task.

The other main finding is annotated in Figure 8 where for **Identify 3, the fixed-scale graphs** (M = 80.7%, SD = 39.5%) **had significantly lower accuracy than the full-scale graphs** (M = 88.1%, SD = 32.4%) (b = .83, z(1,824) = 3.6, p = .0003, CI[.38, 1.29]).

Discussion: In Experiment 2, we found that a time constraint of 40 seconds significantly reduced participants' accuracy of responses for the Identify 1 and 3 tasks. These two tasks likely required a serial search, driving the accuracy reduction. We further found that making all small multiple frames the same size regardless of the number of frames significantly reduced participants' accuracy of responses for the Identify 3 task. With smaller-sized frames, the resolution is reduced, making it challenging to detect when the blue line crosses the dotted threshold (shown in Figure 9). Interestingly, we did not find a significant impact of frame size for the other tasks. Despite this, we assume that the reduced resolution

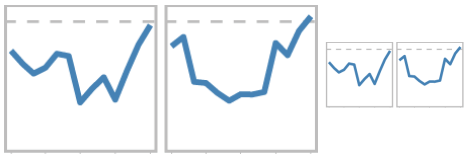

Fig. 9: Illustration of the different scales of frames.

of small frames will affect any task that requires fine-grain discrimination of visual features, as in Identify 3. Further, there is certainly a point at which the frame would be too small for any task, but we did not attempt to find the minimum frame size in this work. To dig deeper into these findings, we conducted a follow-up eye-tracking study.

C. Eye-Tracking: Visual Analysis

Fixations were calculated using the Tobii I-VT fixation filter [56] (for plots of the fixations, see Figure 10). We conducted a visual analysis of the eye-tracking heatmaps and found that, for the Identify tasks, participants visually scanned across the entire array. Participants completed the most thorough visual search for the Identify 3 (mean number of fixations = 75.8) and Summarize 2 tasks (mean number of fixations = 116.4) (see Table II). The Compare 1 and 2 tasks did not involve a thorough visual search (16.2 and 33.2 mean fixations, respectively), likely driving the reduced impact of increasing frame quantities on accuracy.

Participants generally oriented their foveae in the middle of the array in the Summarize 1 task. This gaze pattern could be

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3372620

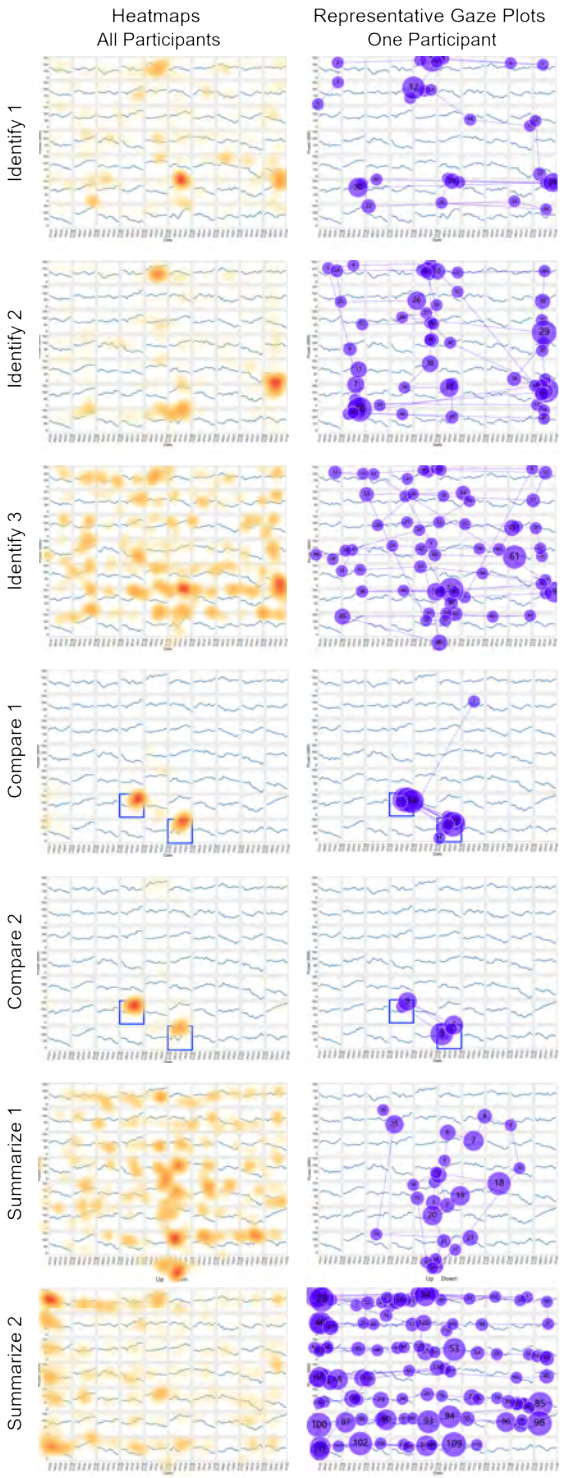

Fig. 10: Heat maps showing the combined fixation locations for all participants (left column) and gaze plots showing eye movement patterns from one representative participant (right column) for all seven tasks.

indicative of ensemble processing, in which the visual system computes summary statistics' gist from a set (for reviews, see Whitney et al. [55] and Szafir et al. [57]). Our accuracy results revealed a small overall impact of frame quantity during the Summarize 1 task (slope = -.007), which was significantly smaller than the effects observed in Identify 1, 3, and Compare 1 tasks. Lastly, Summarize 2 showed higher fixation counts on both the frames and the y-axis, reflecting additional mental calculations and more use of the y-axis.

TABLE II: Eye-tracking gaze fixation counts and standard deviations. The table includes all nine trials from the eye-tracking study. The Experiment 1 tasks used a 70-frame stimulus and the Experiment 2 tasks used a 6-frame stimulus.

Exp 1 Tasks	**Average # of Fixations**	**Total # of Frames Fixated**	**Average # of Fixations to the Y-Axis**
Identify 1	59.0 (39.16)	30.2 (17.21)	0.8 (1.30)
Identify 2	75.8 (27.52)	33.0 (8.43)	0.2 (0.45)
Identify 3	51.0 (17.18)	33.4 (8.73)	0.2 (0.45)
Compare 1	16.2 (5.31)	5.0 (1.58)	1.4 (1.52)
Compare 2	33.2 (23.21)	10.6 (12.58)	0.6 (0.89)
Summarize 1	54.2 (32.34)	32.0 (15.13)	1.0 (2.24)
Summarize 2	116.4 (32.35)	47.6 (11.91)	11.6 (8.65)
Exp 2 Tasks			
Identify 3, full-scale	25.8 (8.67)	5.8 (0.45)	0.2 (0.45)
Identify 3, fixed-scale	23.6 (4.67)	5.6 (0.89)	0.8 (1.79)

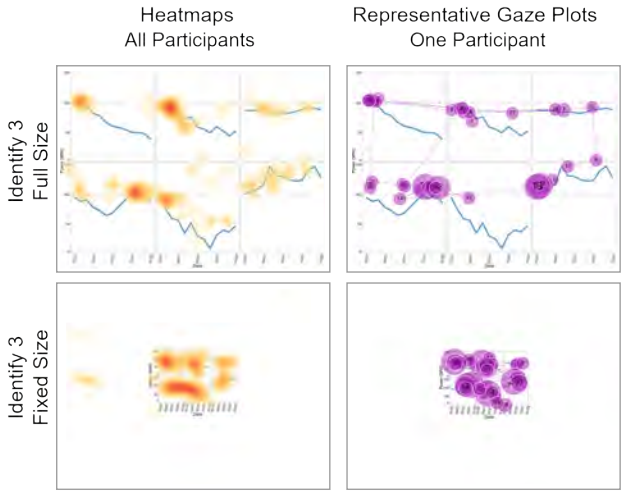

Fig. 11: Heat maps showing the combined fixation locations for all participants (left column) and gaze plots showing eye movement patterns from one representative participant (right column) for the full-scale and fixed-scale versions of the Identify 3 task.

Impact of frame size on fixations: When the small multiples had only six frames, the participants typically fixated on all frames regardless of their size (see Figure 11). On average, they had a higher number of fixations for the full-scale than for the fixed-scale frames in the Identify 3 task (see Table II). The larger frames required a higher number of fixations because the features of interest (like the area around the threshold line) were also larger, and participants' fixations were concentrated in those areas. The larger size of the relevant features also can explain why accuracy was higher for the full-scale small multiples in the Identify 3 task, as it was easier for the participants to see the details.

Discussion: The purpose of the eye-tracking analysis was to investigate the variations in individuals' eye movements

across different tasks. It is important to note that the study's small sample size does not allow for broad generalizations to the general population. However, when considering how individuals' visual search patterns change for each task, we see some preliminary patterns that may contextualize the behavioral findings in Experiments 1 and 2. For the small multiples with 70 frames, the participants often scanned across the full array, which is consistent with the visual search theory. The eye movement patterns for the Compare tasks showed that highlighting effectively focused the participants' attention on specific plots [58]. Regarding frame size, when the frames were larger, the participants fixated more on the parts of the frame that were important for the task, such as the area above the threshold line. On the full-scale small multiples, these essential features were larger and easier to see than in the fixed-scale small multiples. Although the findings of the eye-tracking inquiry are consistent with the behavioral findings, more work is needed to determine if they can be generalized to a more representative population.

V. GENERAL DISCUSSION

This study examined the effect of varying frame quantities in small multiple line charts on participants' performance in seven visualization tasks. We also explored the impact of scale, time, accuracy, and participant strategies. A key discovery was that as the number of frames in small multiples increased, accuracy declined in a linear fashion (Experiment 1 results). The decline in accuracy was not fully explained by a reduction in resolution in smaller frame sizes or differences in task completion times (Experiment 2 results). Such a linear decrease in accuracy points more toward a visual search challenge [14] rather than a cognitive limitation, which would have been revealed by an inverse sigmoid function of performance [40]. We found further support for the visual search theory by demonstrating that we can diminish the impacts of large frame quantities by highlighting specific frames, eliminating the need for a full array search. Additional corroboration for this theory is demonstrated in the eye-tracking analysis that shows viewers scanning large portions of the array, except when two frames were highlighted (Compare 1 and 2). This finding reinforces the effectiveness of small multiples, showing that there was not a threshold at which the number of frames exceeded users' cognitive capacity in our test context. Our results suggest that the main limitation of small multiples is not cognitive but practical, in which an excess of frames on a constrained display might render individual frames unreadable or require a great deal of visual searching.

The observed decrease in accuracy with increased frame quantity in small multiples carries significant implications for user support. Our findings point to extensive visual searching as the primary factor contributing to reduced accuracy. To mitigate this issue, modifications to the user interface can be considered. For instance, the interface could allow users to select specific frames, with the display subsequently updating to showcase only those chosen frames. Techniques such as *zooming*, *panning*, *filtering*, and *interactive piling* [59] offer ways for users to selectively explore specific frames [18]. Designers can leverage the insights from this study to determine the desired accuracy level and adjust the number of frames their system allows users to selectively display to achieve the desired accuracy. In this manner, this research can help optimize existing interaction methods to better align with human capabilities.

Experiment 2 demonstrated that the fixed-scale frames yielded poorer performance compared to the full-scale frames for the Identify 3 task, for which participants were required to pinpoint graphs in which the blue line surpassed the dashed gray line. The decline in accuracy with the smaller frames for this specific task is logical; as the frame size decreases, users' ability to discern instances in which the blue line barely exceeds the threshold becomes challenging. Intriguingly, we did not observe significant performance differences with the smaller frames for the other tasks. We theorize that this might be because the other tasks did not hinge as much on discerning fine details but rather on recognizing the overall trend of the line. It is crucial to note that our study did not encompass a broad spectrum of graph types, marks, or tasks. It is conceivable that specific designs are harder to interpret at reduced sizes, and numerous other tasks may also prove challenging when presented within smaller frames.

Experiment 2 also showed that time constraints reduced the accuracy of Identify 1 and 3 tasks. Identify 3 required participants to click on all graph(s) where the blue line exceeded the dashed gray line, the only task allowing for more than one frame selection. Identify 3 was also the only task for which participants did not know the correct number of frames. We ensured that there were only two correct frames, but participants were unaware of this. Even if the participants guessed that there were two correct frames, this task would be highly impacted by time pressure. Further, fixed-scale frames made it more challenging to discriminate small gaps between the threshold and power lines (shown in Figure 9). The eye-tracking data contextualized these results, indicating that when the frames were larger in the Identify 3 task, the participants fixated more on the relevant parts of the frames.

Limitations and Future Directions: An unexplained finding is that accuracy was significantly impacted by frame quantity for the Compare tasks. Work in cognitive psychology indicates that visual identification tasks are harder to complete when surrounded by distractors [60], [61]. More work is needed to determine if these findings can generalize to data visualizations or if there are other factors influencing the impact of frame quantity on Compare tasks.

Our study focused on line charts across seven specific tasks and resolutions, but further research is needed to validate our findings. Our observation of decreasing accuracy with increasing frame quantities may have broader applicability, but we acknowledge that our stimuli and tasks may not fully mirror the challenges faced by analysts, such as energy grid operators, in real-world settings. We prioritized experimental control over ecological validity, for example, by maintaining consistent frame distances in the Compare tasks. In these tasks, we chose to ensure that the only change in difficulty was from the number of frames, not the distance between the frames. However, this control may not represent the real-world use of

small multiples. Future studies should explore the impact of more naturalistic displays, tasks, and a wider range of chart types.

Lastly, we encountered several logistical limitations in online Experiments 1 and 2. Firstly, Experiment 1 was lengthy without scheduled breaks, potentially leading to participant fatigue. We performed a follow-up analysis and found no significant effects for order of stimuli upon accuracy, which is in the supplemental materials (OSF link), but Experiment 1 was still quite long. Secondly, we were restricted to US participants due to our institutional guidelines. Future work should consider more diverse populations. Of the 500 participants, all but one met the required screen size criteria. Each participant self-reported whether they had zoomed their browser window to 100%. We cannot verify complete adherence to our zooming guidelines, but we are confident in the participants' integrity, especially since only one individual inaccurately reported having the correct screen size. Still, future work should consider in-person studies to ensure consistent browser window settings among participants. Additionally, conducting in-person studies with real analysts familiar with small multiples, such as energy grid operators, could address participant unfamiliarity and provide more practical insights.

VI. CONCLUSIONS

The popularity of small multiples has led designers to utilize them in various applications, such as political forecasts and energy grid control rooms. This study demonstrates the effectiveness of small multiple line charts with large datasets for tasks that do not necessitate extensive visual search. On the other hand, we also present examples of tasks that pose significant challenges to users with large arrays of small multiples. Our findings highlight the importance of interaction methods such as zooming, panning, and filtering that can offload the need for extensive visual search of small multiples with larger numbers of frames. Like any visualization, employing small multiples requires careful design, particularly when dealing with large datasets, as there is no one-size-fits-all approach.

ACKNOWLEDGMENTS

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This work is also supported by NSF Grant #2238175.

REFERENCES

[1] E. Muybridge, "The horse in motion," https://www.loc.gov/item/97502309/, 1878, [Accessed 17-May-2023].

[2] E. R. Tufte, *The visual display of quantitative information*. Graphics press Cheshire, CT, 1983, vol. 2.

[3] W. Meulemans, J. Dykes, A. Slingsby, C. Turkay, and J. Wood, "Small multiples with gaps," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 381–390, Jan. 2017.

[4] M. Burch and D. Weiskopf, "A flip-book of edge-splatted small multiples for visualizing dynamic graphs," in *Proc. 7th Int. Symp. Vis. Inform. Commun. and Interact.*, 2014, pp. 29–38.

[5] W. Meulemans, M. Sondag, and B. Speckmann, "A simple pipeline for coherent grid maps," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 1236–1246, Feb. 2021.

[6] M. Chambers, "How to create a small-multiple tile map," tableau.com, accessed: Feb. 10, 2024. [Online]. Available: https://www.tableau.com/blog/how-create-small-multiple-tile-map-54303

[7] Z. Gemignani, "Better know a visualization: Small multiples," juiceanalytics.com, accessed: Feb. 10, 2024. [Online]. Available: https://www.juiceanalytics.com/writing/better-know-visualization-small-multiples

[8] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko, "Effectiveness of animation in trend visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 14, no. 6, pp. 1325–1332, Nov.-Dec. 2008.

[9] W. Javed, B. McDonnel, and N. Elmqvist, "Graphical perception of multiple time series," *IEEE Trans. Vis. Comput. Graphics*, vol. 16, no. 6, pp. 927–934, Nov.-Dec. 2010.

[10] M. Brehmer, B. Lee, P. Isenberg, and E. K. Choe, "A comparative evaluation of animation and small multiples for trend visualization on mobile phones," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 364–374, Jan. 2020.

[11] D. Archambault, H. Purchase, and B. Pinaud, "Animation, small multiples, and the effect of mental map preservation in dynamic graphs," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 4, pp. 539–552, Apr. 2011.

[12] S. van den Elzen and J. J. van Wijk, "Small multiples, large singles: A new approach for visual data exploration," in *Comput. Graph. Forum*, vol. 32, no. 3pt2. Wiley Online Library, 2013, pp. 191–200.

[13] R. Marois and J. Ivanoff, "Capacity limits of information processing in the brain," *Trends Cogn. Sci.*, vol. 9, no. 6, pp. 296–305, Jun. 2005.

[14] T. Drew, S. E. Boettcher, and J. M. Wolfe, "Searching while loaded: Visual working memory does not interfere with hybrid search efficiency but hybrid search uses working memory capacity," *Psychonomic Bull. & Rev.*, vol. 23, pp. 201–212, Feb. 2016.

[15] N. Ahituv, M. Igbaria, and A. V. Sella, "The effects of time pressure and completeness of information on decision making," *J. Manage. Inf. Syst.*, vol. 15, no. 2, pp. 153–172, Sept. 1998.

[16] A. Bezerianos and P. Isenberg, "Perception of visual variables on tiled wall-sized displays for information visualization applications," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2516–2525, Dec. 2012.

[17] pjminterconnection, "Who's who in the pjm control room," Mar 5, 2019, accessed Feb. 10, 2024. [Online]. Available: https://www.youtube.com/watch?v=MF0a8JKPA6A&t=8s

[18] J. S. Yi, Y. a. Kang, J. Stasko, and J. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 13, no. 6, pp. 1224–1231, Nov.-Dec. 2007.

[19] L. M. Padilla, "A case for cognitive models in visualization research: Position paper," in *2018 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV)*, Oct. 2018, pp. 69–77.

[20] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci, "Decision making with visualizations: a cognitive framework across disciplines," *Cogn. Res.: Princ. and Implications*, vol. 3, no. 1, pp. 1–25, July 2018.

[21] M. Bancilhon, L. Padilla, and A. Ottley, "Improving evaluation using visualization decision-making models: A practical guide," in *Visualization Psychology*. Cham: Springer International Publishing, 2023, pp. 85–107.

[22] L. Bavoil, S. P. Callahan, P. J. Crossno, J. Freire, C. E. Scheidegger, C. T. Silva, and H. T. Vo, "Vistrails: Enabling interactive multiple-view visualizations," in *VIS 05. IEEE Visualization, 2005.* IEEE, 2005, pp. 135–142.

[23] F. Lekschas, B. Bach, P. Kerpedjiev, N. Gehlenborg, and H. Pfister, "Hipiler: visual exploration of large genome interaction matrices with interactive small multiples," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 522–531, Jan. 2018.

[24] I. Boyandin, E. Bertini, and D. Lalanne, "A qualitative study on the exploration of temporal changes in flow maps with animation and small-multiples," in *Comput. Graph. Forum*, vol. 31, no. 3pt2. Wiley Online Library, 2012, pp. 1005–1014.

[25] B. Bach, N. Henry-Riche, T. Dwyer, T. Madhyastha, J.-D. Fekete, and T. Grabowski, "Small multipiles: Piling time to explore temporal patterns in dynamic networks," in *Comput. Graph. Forum*, vol. 34, no. 3. Wiley Online Library, 2015, pp. 31–40.

[26] M. Meyer, B. Wong, M. Styczynski, T. Munzner, and H. Pfister, "Pathline: A tool for comparative functional genomics," in *Comput. Graph. Forum*, vol. 29, no. 3. Wiley Online Library, 2010, pp. 1043–1052.

[27] 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 *Proc. SIGCHI Conf. Human Factors*, 2009, pp. 1303–1312.
[28] J. Marks, B. Andalman, P. A. Beardsley, W. Freeman, S. Gibson, J. Hodgins, T. Kang, B. Mirtich, H. Pfister, W. Ruml, K. Ryall, J. E. Seims, and S. M. Shieber, "Design galleries: A general approach to setting parameters for computer graphics and animation," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 73–84.
[29] T. Jankun-Kelly and K.-L. Ma, "Visualization exploration and encapsulation via a spreadsheet-like interface," *IEEE Trans. Vis. Comput. Graphics*, vol. 7, no. 3, pp. 275–287, July-Sept. 2001.
[30] C. Quijano-Chavez, L. Nedel, and C. M. Freitas, "Comparing scatterplot variants for temporal trends visualization in immersive virtual environments," in *2023 IEEE Conf. Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2023, pp. 669–679.
[31] E. R. Tufte, *Envisioning information*. Graphics Press Cheshire, CT, 1990.
[32] Z. Qu and J. Hullman, "Evaluating visualization sets: Trade-offs between local effectiveness and global consistency," in *Proc. 6th Workshop Beyond Time and Errors Novel Eval. Methods Vis.*, 2016, pp. 44–52.
[33] ——, "Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 468–477, Jan. 2018.
[34] B. Ondov, N. Jardine, N. Elmqvist, and S. Franconeri, "Face to face: Evaluating visual comparison," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 861–871, Jan. 2019.
[35] B. Yost and C. North, "The perceptual scalability of visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 12, no. 5, pp. 837–844, Sept.-Oct. 2006.
[36] J. Liu, A. Prouzeau, B. Ens, and T. Dwyer, "Design and evaluation of interactive small multiples data visualisation in immersive spaces," in *2020 IEEE Conf. Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2020, pp. 588–597.
[37] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg, "Evaluation of alternative glyph designs for time series data in a small multiple setting," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2013, pp. 3237–3246.
[38] N. Cowan, "Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system." *Psychol. Bull.*, vol. 104, no. 2, p. 163, Sept. 1988.
[39] A. Shenhav, M. M. Botvinick, and J. D. Cohen, "The expected value of control: an integrative theory of anterior cingulate cortex function," *Neuron*, vol. 79, no. 2, pp. 217–240, July 2013.
[40] K. Fukuda, E. Awh, and E. K. Vogel, "Discrete capacity limits in visual working memory," *Current Opinion Neurobiol.*, vol. 20, no. 2, pp. 177–182, Apr. 2010.
[41] S. J. Luck and E. K. Vogel, "The capacity of visual working memory for features and conjunctions," *Nature*, vol. 390, no. 6657, pp. 279–281, Nov. 1997.
[42] J. M. Wolfe, "Guided search 6.0: An updated model of visual search," *Psychonomic Bull. & Rev.*, vol. 28, no. 4, pp. 1060–1092, Aug. 2021.
[43] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2376–2385, Dec. 2013.
[44] M. G. Calvo, A. Gutiérrez-García, and M. Del Líbano, "What makes a smiling face look happy? visual saliency, distinctiveness, and affect," *Psychol. Res.*, vol. 82, pp. 296–309, Mar. 2018.
[45] "*Qualtrics*," (2014), Qualtrics, LLC. [Online]. Available: https://www.qualtrics.com.
[46] R Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing, 2018. [Online]. Available: https://www.R-project.org/
[47] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. New York, NY, USA: Springer-Verlag, 2016. [Online]. Available: https://ggplot2.tidyverse.org
[48] "Prolific," prolific.com, accessed: Feb. 10, 2024. [Online]. Available: https://www.prolific.com
[49] D. Oberfeld and T. Franke, "Evaluating the robustness of repeated measures analyses: The case of small sample sizes and nonnormal data," *Behav. Res. Methods*, vol. 45, pp. 792–812, Sept. 2013.
[50] A. Aron and E. N. Aron, *Statistics for Psychology*. Englewood Cliffs, NJ, USA: Prentice-Hall, Inc, 1999.
[51] S. Champely, *pwr: Basic Functions for Power Analysis*, 2020, r package version 1.3-0. [Online]. Available: https://CRAN.R-project.org/package=pwr
[52] 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," *J. Open Source Softw.*, vol. 4, no. 43, p. 1686, Nov. 2019.
[53] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.
[54] M. Kay, *ggdist: Visualizations of Distributions and Uncertainty*, 2022, r package version 3.2.0. [Online]. Available: https://mjskay.github.io/ggdist/
[55] D. Whitney, J. Haberman, and T. D. Sweeny, "From textures to crowds: multiple levels of summary statistical perception," *New Vis. Neurosci.*, pp. 695–710, Jan. 2014.
[56] A. Olsen, "The tobii i-vt fixation filter," *Tobii Technol.*, vol. 21, pp. 4–19, Mar. 2012.
[57] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri, "Four types of ensemble coding in data visualizations," *J. Vision*, vol. 16, no. 5, pp. 11–11, Mar. 2016.
[58] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cognitive Psychol.*, vol. 8, no. 4, pp. 441–480, Oct. 1976.
[59] F. Lekschas, X. Zhou, W. Chen, N. Gehlenborg, B. Bach, and H. Pfister, "A generic framework and library for exploration of small multiples through interactive piling," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 358–368, Feb. 2021.
[60] K. R. Cave and Z. Chen, "Identifying visual targets amongst interfering distractors: Sorting out the roles of perceptual load, dilution, and attentional zoom," *Attention, Perception, & Psychophys.*, vol. 78, pp. 1822–1838, Oct. 2016.
[61] M. P. Eckstein, "Visual search: A retrospective," *J. Vision*, vol. 11, no. 5, pp. 14–14, Dec. 2011.

VII. Biography Section

Helia Hosseinpour is a Ph.D. student at the University of California Merced in the Cognitive and Information Sciences department. Her research focuses on decision-making, goal-pursuit, and how cognitive processes activated outside of awareness may impact how we use and perceive data visualizations.

Laura E. Matzen is a Distinguished Member of the Technical Staff in the Cognitive Science and Systems department at Sandia National Laboratories. Her research focuses on how visual cues impact decision making and human-system interactions.

Kristin M. Divis is a Principal Member of the Technical Staff in the ISR Advanced Exploitation and Human Systems Integration department at Sandia National Laboratories. Her research interests include human-systems integration; interaction design; data visualization; and developing user-centered, next generation technology in support of national security.

Spencer Castro is an Assistant Professor at the University of California Merced in the Management of Complex Systems department. He studies attention and effort limitations' effect on performance in realistic settings, such as multitasking with mobile devices while driving.

Lace Padilla is an Assistant Professor at Northeastern University in the Computer Science department. She studies how to align data visualizations with human decision-making capabilities.