Better sensitivity to linear and nonlinear trends with position than with color

Jessica K. Witt Colorado State University, Fort Collins, CO, USA

Amelia C. Warden Colorado State University, Fort Collins, CO, USA

Useful data visualizations have the potential to leverage the visual system's natural abilities to process and summarize simple and complex information. Here, we tested whether the design recommendations made for pairwise comparisons generalize to the detection of trends. We created two different types of graphs: line graphs and stripplots. These graphs were created from identical datasets that simulated temperature changes across time. These datasets varied in the type of trend (linear and exponential). Human observers performed a trend detection task for which they judged whether the trend in temperature over time was increasing or decreasing. Participants were more sensitive to trend direction with line graphs compared to stripplots. Participants also demonstrated a systematic bias to respond that the trend was increasing for line graphs. However, this bias decreased with increasing sensitivity. Despite the better sensitivity to line graphs, more than half of the participants found the stripplots more appealing and liked them more than the line graphs. In conclusion, our results indicate that, for trend detection, depicting data with position (line graphs) leads to better performance compared to depicting graphs with color (stripplots). Yet, graphs with color (stripplots) were preferred over the line graphs, suggesting that there may be a tradeoff between the aesthetic design of the graphs and the precision in communicating the information.

Introduction

Starting in mid-June of 2019, a trend on Twitter was to #*ShowYourStripes* (see Figure 1). The stripes correspond to warming stripes or stripplots that depicted differences in temperature across nearly two centuries. Red corresponds to a higher relative temperature within the time frame, and blue corresponds to a lower relative temperature. The plots depict trends at the global, continent, country, or state level. The plots have been downloaded nearly one million times, have graced the cover of the *Economist*, and have been applied to personal items from neckties and shirts to car paint (source: Wikipedia). The plots are aesthetically pleasing, but are they good at depicting trends in climate data?

According to classic work on graph design, plotting data in terms of relative position along a common scale leads to the best precision, and using color leads to the worst precision (Cleveland & McGill, 1985). In these experiments, participants were asked to specify the ratio between two data points. Ratio judgments were most accurate for graphs that specified the value of data with relative position on a common scale. These kinds of graphs include dot, bar, and line graphs. Ratio judgments were least accurate for graphs specifying the value of data with color.

If these recommendations with pairwise comparisons generalize to different kinds of graphs such as those that require visual averaging or those that show trends, line graphs, rather than stripplots, should be best for depicting data. The research showing this generalization is mixed. Some studies show better performance with scatterplots than stripplots. For example, participants were more accurate at selecting which of two distributions of dots in a scatterplot had the higher mean than selecting which of two stripplots had the higher mean (Legge, Gu, & Luebker , 1989). They were also more accurate at selecting which were sampled from a distribution with a higher variance. Similarly, other research showed better accuracy at finding the graph with the highest decrease among many small multiples for line graphs than for stripplots (Fuchs, Fischer, Mansmann, Bertini, & Isenberg , 2013). Together, these results showed that relative position leads to better accuracy than brightness or color.

In contrast, other research demonstrated that Cleveland and McGill's (1985) recommendations do not generalize. Several studies have shown that visual averaging is more accurate with stripplots than with line graphs (Albers, Correll, & Gleicher, 2014; Correll et al., 2012). In one study, the visual averaging task was to estimate the mean value of a stock based on a graph depicting the value of the stock across one year (Albers et al., 2014). In another study, the visual averaging

Citation: Witt, J. K., & Warden, A. C. (2021). Better sensitivity to linear and nonlinear trends with position than with color. *Journal of Vision*, *21*(5):12, 1–13, https://doi.org/10.1167/jov.21.5.12.

https://doi.org/10.1167/jov.21.5.12 Received December 28, 2019; published May 13, 2021 ISSN 1534-7362

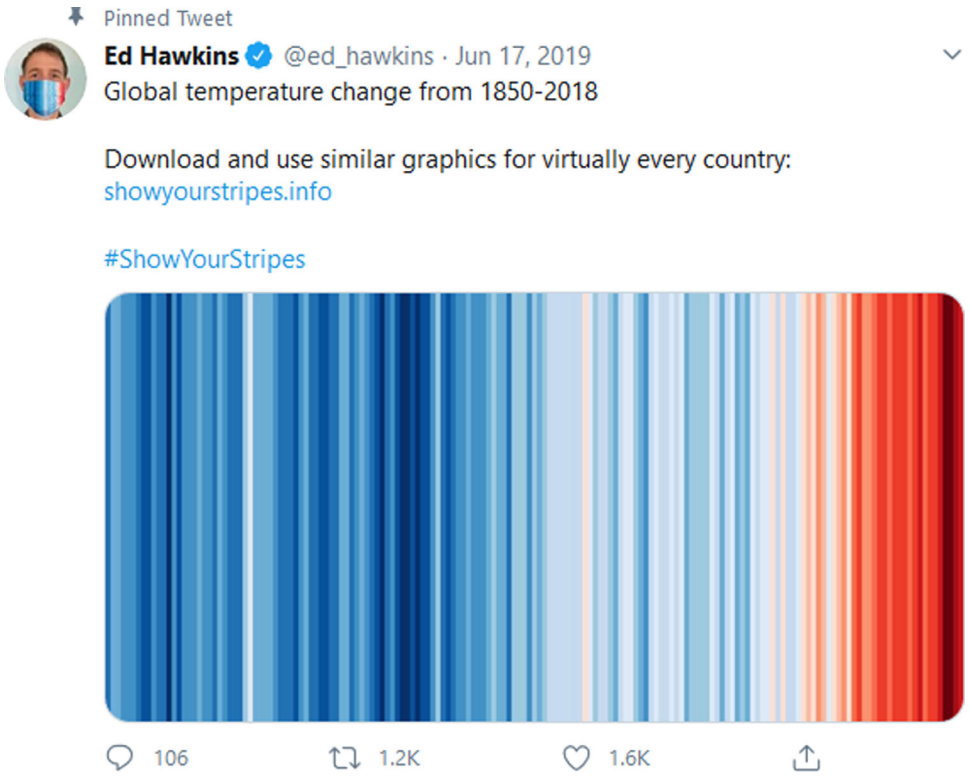

Figure 1. Tweet by Ed Hawkins showing global temperature change using a stripplot (Hawkins, June 17, 2019). Screenshot of tweet taken on September 3, 2020.

task was to determine which month had the highest average sales, which required averaging data within each month and comparing across months (Correll, Albers, Franconeri, & Gleicher, 2012). Both studies showed superior performance in terms of the number of correct answers (accuracy) with stripplots than with line graphs. These results inspired the recommendation that color rather than position should be used when presenting data for summary tasks (Szafir, Haroz, Gleicher, & Franconeri, 2016).

The seminal work by Cleveland and McGill (1985) clearly shows better pairwise precision for relative position than for color. So why would stripplots, which depict data using the less precise visual feature of color, lead to superior performance compared to line graphs, which use the more precise visual feature of relative position? Szafir and Colleagues (2016) raised the possibility that less precision for individual data points could actually promote performance at the level of the ensemble or group of data points. Specifically, they speculated that less precise representations would lead to less individuation of specific data points, which would enable more precise representations at the group level.

The visual system can extract information at the level of an individual object, but it can also extract summary information about a group of objects. This ability, termed *ensemble perception*, has been most frequently researched with respect to the mean of a group. For example, observers might see an array of circles and estimate their color, size, or position (Whitney & Yamanashi Leib, 2018). Ensemble perception has focused on whether observers can detect the mean across a variety of stimuli from low-level visual features such as color and orientation to high-level features such as facial expressions.

Despite the growing literature on ensemble perception, the field has not focused on the direct comparison of performance across features and is therefore silent with respect to whether color versus position leads to the best performance in a visualization. This is not surprising given that it is hard to equate a range of stimuli in one dimension (such as vertical position) to a range of stimuli in another dimension (such as color). However, research on ensemble perception suggests that better visual precision for individual objects is related to better visual precision for ensembles. Individual differences have shown a positive correlation between precision for individuals and for ensembles (Haberman, Brady, & Whitney, 2015). If this correlation is driven by a causal link, this suggests that summary tasks should be better when data are represented by position instead of color because precision for individual data points is better for position than color.

In contrast to this speculated causal link, developmental research has shown that children who have poor precision at the level of the individual item can still perform summary tasks (Sweeny, Wurnitsch, Gopnik, & Whitney, 2015). Similarly, poor precision at the individual level due to visual crowding does not prevent ensemble perception (Fischer & Whitney, 2011; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001). In addition, there can be changes in items at the individual level that go undetected, but ensemble coding is still correctly perceived (Haberman & Whitney, 2011).

The mixed literature gives rise to several questions. From the perspective of visualizations, one question is whether trend detection is better served by representing the data using color or vertical position. From the perspective of ensemble perception, the question is whether the visual feature that gives rise to the best precision for comparing individual objects also gives rise to the best precision for extracting summary statistics. As discussed by Rensink (2017), trend detection likely recruits ensemble perception processes. The summary statistic would be a probability distribution of how one feature (vertical position or color) varies with another feature (horizontal position). Thus research on trend detection has implications for understanding ensemble processes as well. Our research explored these questions as well as whether the patterns that hold for linear relationships generalize to non-linear relationships.

Furthermore, we will expand on previous literature using a signal detection theory framework to separate the effects of sensitivity versus bias (Macmillan & Creelman, 2008; Tanner & Swets, 1954; Wickens, 2002). When applying this framework to an experiment on detecting whether a trend depicts an increase or a decrease over time, sensitivity refers to the observers' abilities to discriminate between increasing and decreasing trends, and bias refers to the observers' tendencies to respond that a trend is increasing versus decreasing. By separating the effects of sensitivity and bias, a signal detection approach provides a more nuanced understanding of observers' performance than using the proportion of correct responses.

A signal detection approach could explain discrepancies in the literature about visualizations. For example, performance on detecting linear trends was better for line graphs than stripplots when the trends were decreasing (Fuchs et al., 2013). However, the performance was similar or possibly even better for stripplots than scatterplots when the trends were increasing (Rensink, 2015). These seemingly inconsistent patterns could easily be explained by bias. A bias to estimate trends as increasing when data are presented with color compared to position would lead to better performance with stripplots for increasing trends but not for decreasing trends.

We hypothesized that sensitivity for detecting trends in time series data would be higher with line graphs that represent data via position than with stripplots that represent data via color. This hypothesis is consistent with the idea that ensemble perception builds from the perception of individual elements. If the data instead show better sensitivity with stripplots, this would suggest that color, rather than position, is better for summary tasks including detecting averages and trends (Szafir et al., 2016). This result would also suggest that worse resolution at the pairwise level (cf., Cleveland & McGill, 1985) leads to better precision at detecting information in ensembles. Such a result would have important implications for the mechanisms driving ensemble perception. We also hypothesize differences in bias such that observers will be more biased to respond that the trends are increasing for stripplots versus line graphs. Such biases have not yet been explored. The current approach investigates what kinds of biases exist for the different kinds of plots and ensemble tasks.

Method

Participants

Fifty-seven participants were recruited from Amazon's Mechanical Turk ($n = 22$) and the Psychology Research Participant Pool at Colorado State University ($n = 35$).

Stimuli and apparatus

Stimuli were constructed in R (R Core Team, 2017). Data were simulated to create 189 datasets. Each dataset simulated temperature across 100 time points. Each dataset was constructed by simulating a trend with an intercept, slope, and exponent. Random noise was added to each dataset, and cyclical noise was added to some datasets. The equation for constructing the datasets is shown in Equation 1. Time was a vector ranging from 1 to 100. All other values were systematically manipulated parameters.

$$Temperature = 50 + slope * time^{exponent} + randomNoise + cyclicalNoise \quad (1)$$

There were three trend exponents: 1, 2, and 4. Each was constructed to be positive and negative by manipulating the sign of the slope. For the linear trend, the slope was −0.2 or 0.2 and the exponent was 1. For the exponential-2 trend, the slope was −0.002 or 0.002, and the exponent was 2. For the exponential-4 trend, the slope was −0.00002 or 0.00002, and the exponent was 4. There were 3 levels of random noise. The random noise was generated using the *rnorm* function in R with

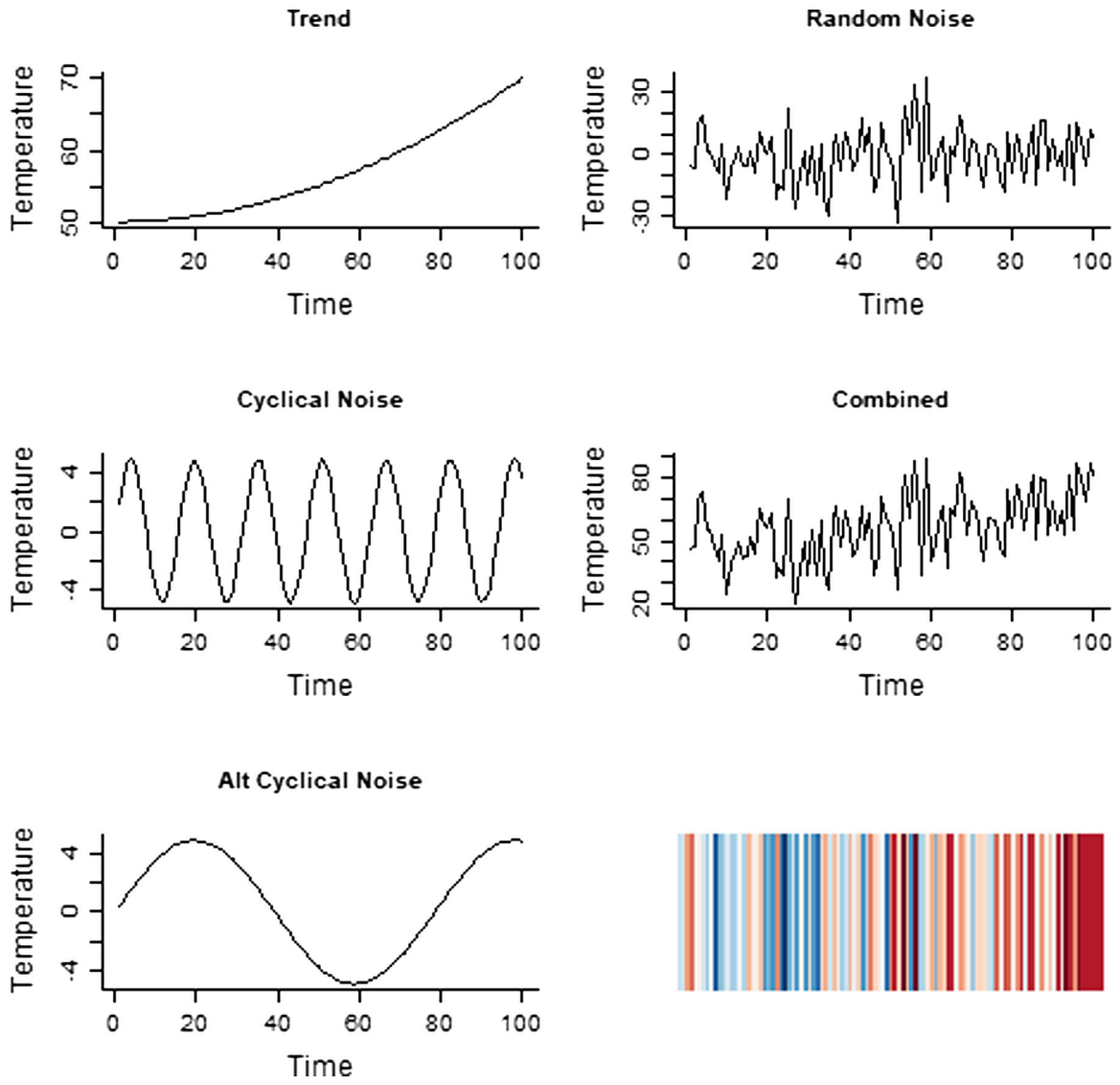

Figure 2. Illustration of the various components of each dataset. In this example, the trend is positive, exponential-2 with the medium level of random noise and the short cyclical noise. The combined plot shows how the data look after the three components are combined, and the stripplot shows how the data were depicted for the stripplot. The alt cyclical noise panel shows an alternative option that was the long cyclical noise.

a mean of 0 and a standard deviation (*SD*) of 8, 12, or 16. There were also three levels of cyclical noise. One level was to have no cyclical noise, meaning all the noise was random. The other two levels were a short and a long sine wave. This noise was calculated as five times the sine of the product of time (the vector from 1 to 100) multiplied by 0.4 or 0.08 for the short and long waves, respectively. An example of the different components that were combined to create the final dataset are shown in Figure 2. Examples of different combinations are shown in Figure 3. There was an additional trend type for which the slope was 0, and the exponent was 0. Any trends apparent in the data were due to spurious, unintended patterns in the noise. In hindsight, these stimuli should not have been included in the experiment, and data from these trials were excluded from the analyses. The seven trend types × 3 levels of random noise × 3 levels of cyclical noise × 3 repetitions of each produced 189 unique datasets.

Each dataset was used to create one line graph and one stripplot (see Figure 2). The range of the y-axis was left to the default values in R, which are ± 4% of the data range. For the stripplots, the data were presented as vertical lines colored as a function of relative temperature within each dataset. Numeric labels and axes were not presented in the stripplots. We used a diverging red-blue color ramp created from ColorBrewer. Many designer color tools, such as ColorBrewer, generate color ramps that transverse nonlinear paths through the perceptual color space

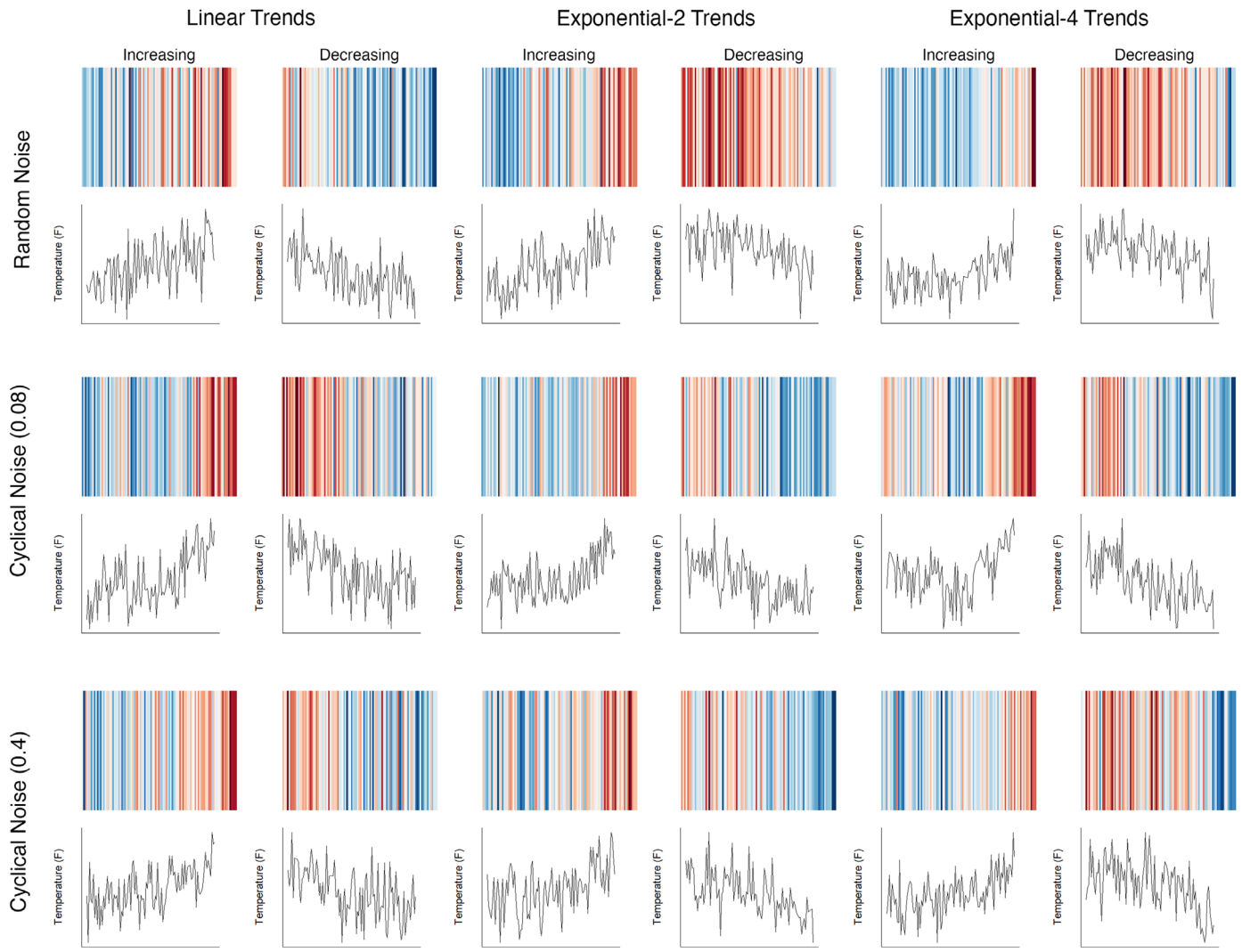

Figure 3. Examples of stripplots and line graphs for each trend exponent for each level of cyclical noise (levels 0, 0.08, and 0.4). Cyclical noise of 0 means the simulated data contained only random noise.

(Smart et al., 2020). The diverging red-blue color palette used here is symmetrical, meaning that there is an equal number of colors on either side of the midpoint. The color ramp has an orderly lightness, with more lightness change at the ends of the color ramp. The color ramp has more hue changes in the middle range and varied saturation throughout the ramp (Brewer, 1999a).

The stimuli were presented in Qualtrics. The trials consisted of two blocks (one stripplots block and one lines block) followed by three questions about the types of graphs used. The stripplot and line blocks were counterbalanced and consisted of 189 trials that were randomly presented to the participant. Once both blocks were completed, participants were asked questions about the visual appearance and likability of the line graph and the stripplot.

Procedure

The stimuli were presented through a Qualtrics survey. Instructions at the beginning indicated that the participants would see graphs depicting average annual temperature across time. Their task was to determine whether the overall trend in average temperature was increasing or decreasing. For clarity, participants were informed not to look at the last item or compare the first and last items when making their decisions. Instead, they were told to assess the entire time course of the graph to decide whether the average temperature was increasing or decreasing over time.

Before each test trial block, participants were given a description of the type of graph they would see for that block. The stripplots were described as presenting how global temperatures change over time

by using color. Dark red stripes indicated the highest temperatures occurring in a given year, and dark blue stripes indicated the lowest temperature occurring in a given year. The line graphs were described as presenting how global temperatures change over time using lines to represent trends. The height of the lines corresponded to the temperature occurring in a given year, with the highest values on the y-axis representing the higher temperature for that year, and the lowest values on the y-axis representing the lowest temperature for that year.

On each trial, a single graph was presented. The dimensions were set to be 400 pixels wide × 300 pixels tall. Participants responded by clicking either the box that said "Decreasing" or the box that said "Increasing." For one block of trials, all the graphs were the line graphs, and for the other block, all the graphs were stripplots. The starting order of graph type was counterbalanced across participants. Image presentation within a block was randomized.

Participants were asked three questions after completing both blocks: (1) "Which graph do you find more visually appealing?"; (2) "Which graph would you be more likely to share (e.g., via social media, email, text)?"; and (3) "Which graph do you like more?" Each question required a response of selecting either the "Line Graph" or the "Stripes Color Plot."

Results

Trend direction was determined by whether the correlation between temperature and time was positive or negative. Trend direction coincided with the intended trend direction based on the sign of the slope used to simulate the data. However, one of the linear graphs had a very low correlation ($r < .20$) because of noise and random sampling and was removed from the analysis.

Accuracy

Participants' responses were coded as accurate (1) if their response matched the direction of the trend and incorrect (0) otherwise. These data were analyzed with a general linear mixed model. The dependent measure was accuracy. The independent measures were trend direction (positive, negative), graph type (line graph, stripplot), trend type (linear, exponential-2, exponential-4), and all interactions. The random effects for participant were included for the intercepts and slopes for each main effect. Trend direction and graph type were both coded as 0.5 and −0.5; trend type was coded as 0, 1, 2.

The main effect of graph type was significant, $z = 3.30$, $p < .001$, estimate = 0.73, $SE = 0.22$. Participants were more accurate with the line graphs than the stripplots. However, this accuracy varied depending on the other aspects of the graphs, as revealed by significant interactions. All two-way interactions were significant, ps $< .03$. The three-way interaction was also significant, $z = -4.03$, $p < .001$. For stripplots, the difference in accuracy between increasing and decreasing trends varied as a function of the exponent, $z = -5.00$, $p < .001$, estimate = −0.37, $SE = 0.07$ (see Figure 4). In contrast, for the line graphs, the difference in accuracy between increasing and decreasing trends did not vary significantly as the exponent increased, $z = 1.15$, $p = .25$, estimate = 0.11, $SE = 0.09$.

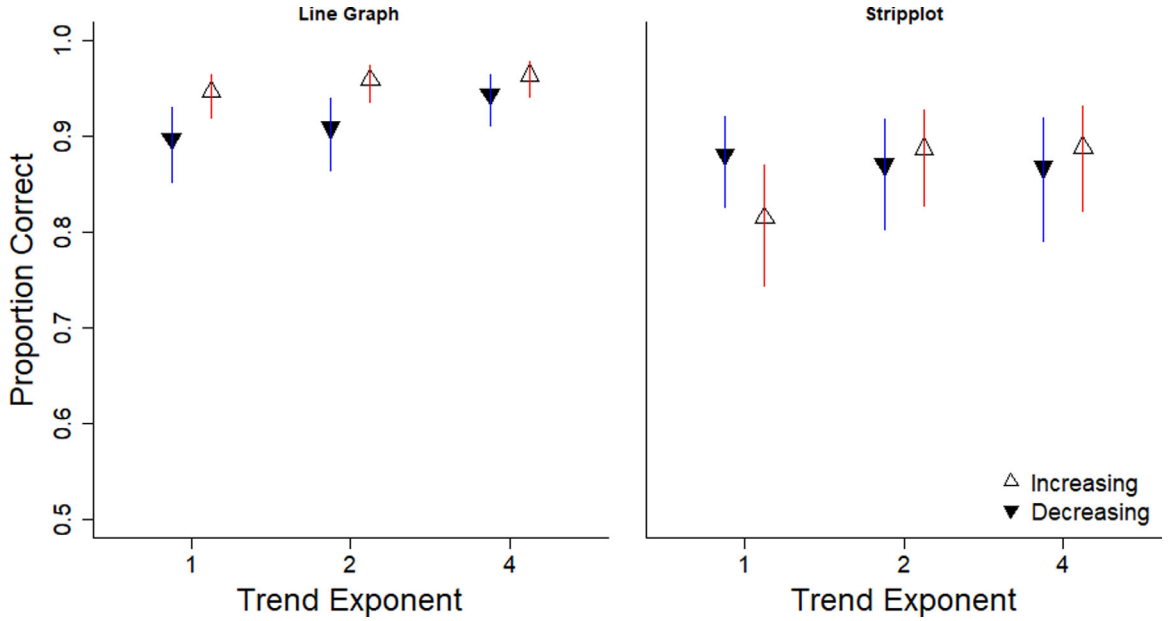

Figure 4. Proportion of correct responses is plotted as a function of trend exponent, trend direction, and graph type. Error bars are asymptotic 95% confidence intervals calculated from the model.

	Response	
Trend direction	"Increasing"	"Decreasing"
Increasing	Hit	Miss
Decreasing	False alarm	Correct rejection

Table 1. Classification of responses for signal detection analysis

Signal detection measures

Although accuracy may differ for line graphs and stripplots as a function of trend type and trend direction, a more parsimonious explanation is that the differences across trend directions reflect attenuated sensitivity coupled with specific biases. To explore this possibility, we analyzed the data using the signal detection measures of d' and c, which measure sensitivity and bias, respectively.

The hit rate was calculated as the proportion of trials for which the trend was increasing, and participants responded that it was increasing. The false alarm rate was calculated as the proportion of trials for which the trend was decreasing, but participants responded that it was increasing (see Table 1). Sensitivity was measured using d', which was calculated as the z-score of the hit rate minus the z-score of the false alarm rate. Bias was measured using c, which was calculated as -1 times the sum of the z-scores of the hit and false alarm rates divided by 2. Negative c scores indicate a bias to respond that the trend is increasing, and positive c scores indicate a bias to respond that the trend is decreasing. Both measures were calculated for each participant for each graph type and each trend type. With signal detection analysis, trend direction is collapsed within each category to be able to calculate both hits and false alarms.

Sensitivity

To assess sensitivity, we conducted a linear mixed model with d' as the dependent factor. The fixed effects were trend type (linear, exponential-2, exponential-4, coded as 0, 1, and 2, respectively), graph type (coded as $-.5$ and .5 for stripplots and lines graphs, respectively), and their interaction as the within-subjects factors. Random effects were included for each participant, including intercepts and slopes for graph type because the model was singular with random effect slopes for trend type.

Graph type significantly affected sensitivity, $t = 3.53$, $p < .001$, estimate $= 0.75$, $SE = 0.21$. Participants were more sensitive to the direction of the trend (higher d' values) when viewing line graphs compared to stripplots (see Figure 5). This advantage for the line graphs over the stripplots increased as the exponent of the trend

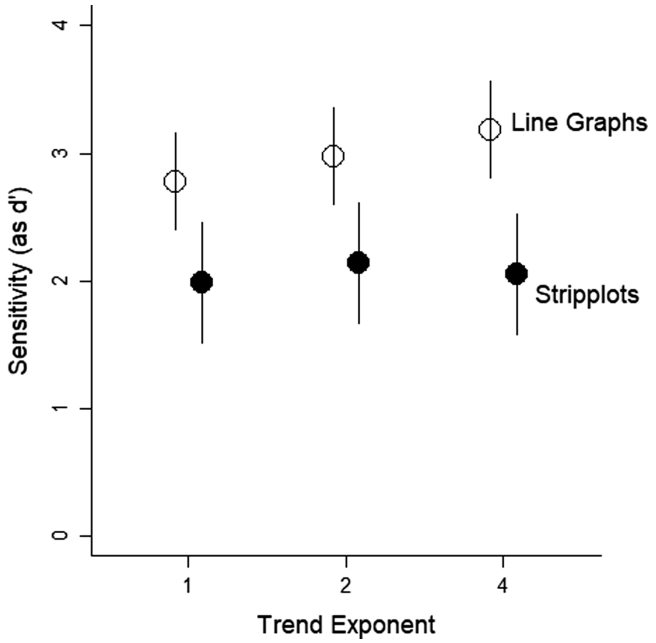

Figure 5. Sensitivity (measured as d') is plotted as a function of trend exponent and graph type. Higher d' values indicate better sensitivity. Error bars are 95% confidence intervals calculated from the model.

increased, $t = 2.71$, $p = .007$, estimate $= 0.17$, $SE = 0.06$. A multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016) revealed a similar pattern regardless of outlier exclusion (see Supplementary materials at https://osf.io/x372y/).

Why is sensitivity better with the line graphs than the stripplots? One possibility is that observers allocated their attention to different parts of the displays for the two graph types. To evaluate this possibility, we constructed four predictor variables on the basis of the simulated data used to create the stimuli. All four predictors were the difference scores between the means of two subsets of the simulated data. One predictor was the difference in means for the second half minus the first half. We calculated the mean temperature for the second half of the simulated data and the mean temperature for the first half of the simulated data, then calculated the difference between them (halves 1 and 2 or *halves1–2*). We also calculated the difference in means between the first and last quarter (*quarters1–4*) and the first and last tenths (*tenths1–10*). In addition, we calculated the difference in means between the third and fourth quarter (*quarters3–4*). If participants only attended to the first and last items, the tenths1–10 should best predict participants' responses. If participants only attended to the right half of the graphs, the quarters3–4 should best predict their responses.

We conducted a general linear mixed model for each predictor variable for each graph type. The dependent

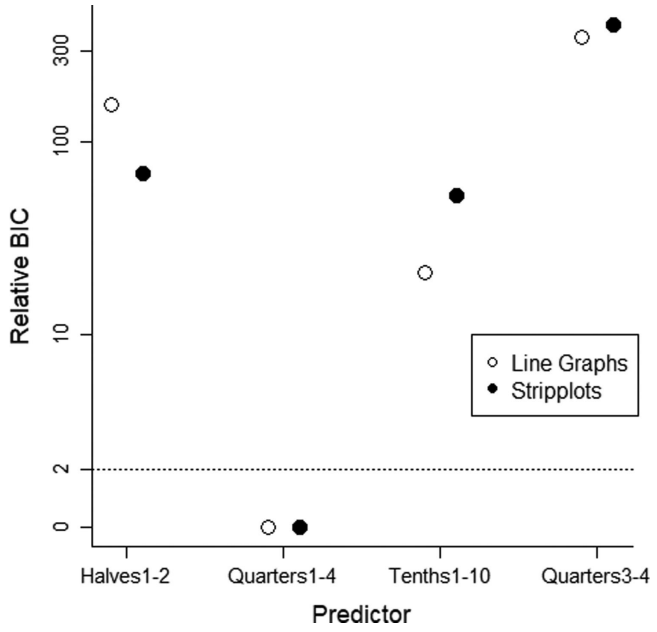

Figure 6. Relative BIC scores for each predictor for each graph type. Relative BIC was calculated by subtracting the minimum BIC score within each graph type from the model's BIC. A difference greater than 2 is evidence that the model with the lower score is a better model. The dotted line is at two. Note the nonlinear y-axis. See text for details on how the predictors were calculated.

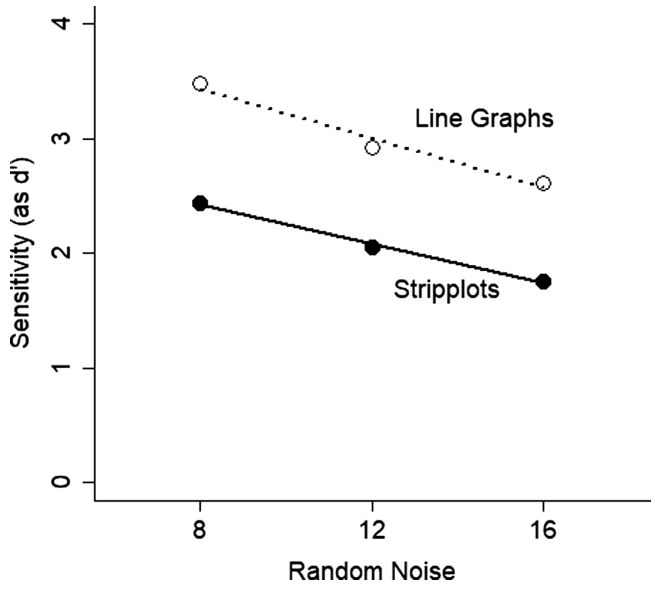

Figure 7. Sensitivity (measured as d') as a function of the random noise in the stimulus and graph type. Points correspond to mean d' scores, and lines correspond to linear regressions estimated from the linear mixed model.

variable was the participants' response, coded as 1 for increasing and 0 for decreasing. The fixed effect was the predictor. The random effect was participant including intercepts and slopes for the predictor. We compared Bayesian information criterion (BICs) for each model. Lower BIC indicate better model fit. For both line graphs and stripplots, the model with the lowest (best) BIC was the model that included quarter1–4 (see Figure 6). The next best models were tenths1-10, halves1–2, then quarters3–4, respectively (note that BIC can only be compared within graph type and not across graph types). Quarter1–4 better predicted participants' responses than tenths1–10, which shows that participants did not just focus on the first and last items. Critically, the fact that the same models emerged as the best predictors for both graph types suggests similar allocation of attention for the two graph types. This suggests that the ensemble processes for summarizing trends is more precise for information depicted using vertical position than using color. In other words, the reason the line graphs led to better sensitivity is not because participants knew how to better allocate their attention but rather because they were able to process the information more precisely or more completely (e.g., taking more of the information into account). Of course, we did not test all possible predictor variables, so it is possible that people allocated their attention in different ways across the two graph types and our measures did not capture this difference.

If trend detection relies on ensemble processes, we would expect sensitivity to get worse as the noise increased. Just as perceiving the mean is impaired by increasing the variability (Ariely, 2001), we would expect perceiving the trend to also be impaired by increasing the variability due to random noise. We calculated d' scores for each participant for each graph type for each level of random noise. We analyzed the d' scores with a linear mixed model. The fixed effects were level of random noise (coded as 0, 1, and 2 for low, medium, and high), graph type, and their interaction. The main effect for random noise was significant, $t = -4.75$, $p < .001$, estimate $= -0.43$, $SE = 0.09$. As random noise increased, d' decreased. The main effect for graph type was significant, $t = -6.10$, $p < .001$, estimate $= -1.01$, $SE = 0.17$. Participants were more sensitive to trend direction with line graphs than stripplots. The interaction between random noise and graph type was not significant, $t = 0.69$, $p = .49$, estimate $= 0.09$, $SE = 0.13$. Thus the impact of random noise on sensitivity was similar for both the line graphs and the stripplots (see Figure 7). That sensitivity to the information was similarly impacted by random noise in both conditions is consistent with the idea of a common mechanism underling the processing of both graph types.

To more directly test the possibility of a common mechanism for trend detection from both line graphs and stripplots, we calculated the mean d' scores for each participant for each graph type and conducted a correlation across the two graph types (see Figure 8).

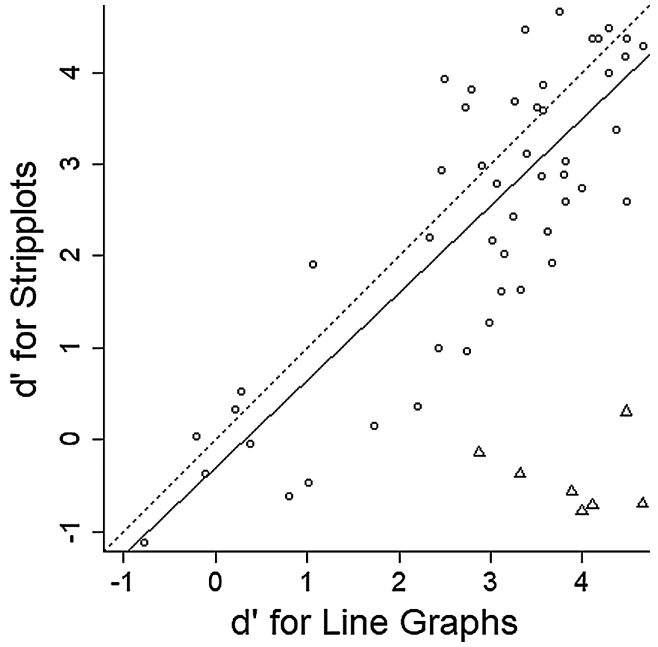

Figure 8. Mean d' scores for the stripplots as a function of d' scores for the line graphs. Each point corresponds to one participant. The triangles correspond to participants with a difference score greater than 3. The solid line represents the linear regression for participants with difference scores less than 3 (circles). The dotted line represents unity.

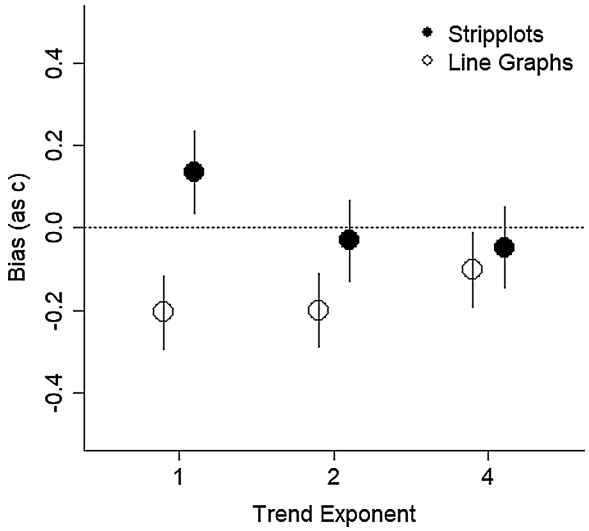

Figure 9. Bias (measured as c) is plotted as a function of trend exponent and graph type. Zero, shown with a horizontal dotted line, indicates no bias. Positive values indicate a bias to respond "decreasing," and negative values indicate a bias to respond "increasing." Error bars are 95% confidence intervals calculated from the model.

With all participants, the correlation was significant, $r = .54$, $df = 55$, $p < .001$. Excluding participants with a difference score greater than 3 (see triangles in Figure 8), the correlation was even greater, $r = .83$, $df = 48$, $p < .001$. These values are consistent with previous work on individual differences in the precision of ensemble perception (Haberman et al., 2015). This previous work showed moderate-to-high correlations between precision of perceiving the means of ensembles with elements varying across low-level features including orientation and color, but no correlation between low-level and high-level features such as perceiving mean facial expression. This was taken as evidence for common mechanisms underlying ensemble processes for low-level features. The current data replicate and extend this finding to a different kind of ensemble process, namely extracting the trend in the data.

Bias

Signal detection theory provides measures of sensitivity and of bias. We analyzed the measure of bias, c scores, with a linear mixed model. The fixed effects were graph type, exponent, and their interaction. We included random effects for participant including the intercepts and slopes for graph type. The main effect of graph type was significant, $t = -5.40$, $p < .001$, estimate $= -0.33$, $SE = 0.06$. However, this effect was modulated by trend exponent, $t = 4.77$, $p < .001$, estimate $= 0.14$, $SE = 0.03$ (see Figure 9). When the trend was linear (exponent = 1), participants were biased when viewing line graphs to respond that the trend was increasing, as revealed by negative c values, $t = -4.57$, $p < .001$, estimate $= -0.20$, $SE = .04$. When viewing stripplots, they were biased to respond that the trend was decreasing, as revealed by positive c values, $t = 2.71$, $p = .008$, estimate $= 0.13$, $SE = 0.05$. This difference was significant, $p < .001$. For the exponential-2 graphs, there was a similar bias to respond increasing for the line graphs, $p < .001$, but no bias for the stripplot, $p = .53$. This difference in bias between the two conditions was significant, $p = .009$. For the exponential-4 graphs, the difference between the two graph types was not significant, $p = .40$. When viewing the line graphs, the bias to respond that the trend was increasing was approximately half that found with the other trend types, $t = -2.26$, $p = .026$, estimate $= -0.10$, $SE = 0.05$.

For both graph types, it is clear that the bias reduces as the exponent increases. It is unclear to us why the bias for the stripplots was a bias to respond decreasing for the linear trend. Because red is associated with warming (Or & Wang, 2014; Tham, Sowden, Grandison, Franklin, Lee, Ng, Park, Pang, & Zhao, 2019) and threat (Elliot & Maier, 2007; Moller, Elliot, & Maier, 2009), we had expected a bias for the stripplots to respond that the temperature was increasing. However,

the stripplots also contained blue, which is associated with cooling (Tham et al., 2019), so perhaps the combination of having both colors in the plot resulted in interactive effects.

To further understand the bias, we compared these bias scores to the stimuli for which a flat trend was simulated, and thus any signal was spurious. We divided the flat trends into those for which random chance meant the simulated data exhibited a positive correlation and those that exhibited a negative correlation. We used these classifications to calculate d' and c scores. As expected, d' scores were low ($M = 0.86$, $SD = 0.65$). We expected low d' scores because the correlations were low ($rs < 0.20$), and prior research showed poor detection of low correlations (Rensink, 2017). For both the line graphs and the stripplots, there was a significant bias to respond that the trend was increasing, $ps < .001$ (line graph mean $= -0.37$, $SE = 0.06$; stripplots mean $= -0.25$, $SE = 0.06$). Thus it seems that there is a general bias for responding that the graphs show that temperature is increasing over time. For line graphs, the bias lessened as the exponent increased. For stripplots, the bias reversed for the linear graphs then lessened as the exponent increased.

Preferences

In the experiment, we also asked people to rate their preference for line graphs versus stripplots. The majority (60%) indicated they found the stripplots more visually appealing, though there was no difference in their ratings of which they liked more or which they were more likely to share (49%). These preferences are contaminated by the fact that they had already completed the rest of the experiment. If they found it easier to detect the trends depicted in the line graphs, that could have swayed their preference judgments. Somewhat consistent with this idea, participants who showed greater sensitivity with the line graphs than the stripplots only showed a 44% preference for the stripplots, whereas those who showed similar sensitivity for the two plots had a 62% preference for the stripplots. Participants were divided based on a median split of the mean difference in sensitivity, and the preference score was the mean of the three preference judgments. The difference in preference judgments between the two groups was not significant, $t = 1.59$, $p = .119$, $df = 54.86$, 95% CI [−0.05, 0.39], although, results show a slight pattern. Nevertheless, the fact that the preference was not overwhelmingly in favor of the line graphs shows a dissociation between visual sensitivity (for which the line graph was the clear winner) and preference (for which the line graph was not the clear winner).

Discussion

Data visualizations can leverage the visual system's vast capacity to process and summarize information. In some cases, data visualizations require a simple visual comparison of two data points. In other cases, data visualizations require visual summarizing of multiple data points. The question for the current research was whether the kinds of visualizations that best serve pairwise comparisons are also the kinds of visualizations that best serve tasks requiring trend detection. In other words, we can question the extent to which visualization design recommendations generalize across various tasks.

For pairwise comparisons, research clearly points to graphs like dot, line, and bar graphs as being superior because visual precision is highest when data are represented as position along a common scale (Cleveland & McGill, 1985). Pairwise comparisons require visually evaluating two data points to determine the extent to which one is greater than the other. When the task is to determine how much higher in the plot one data point is relative to another, performance is better compared with when the task is to determine the relative color between two data points. In other words, relative position leads to greater visual precision than color.

For summarizing tasks, observers must evaluate multiple data points to extract a summary statistic. Examples of summary statistics include the mean value across the entire timeline, multiple mean values across subsets of the data, or trends in the data such as whether the probability distribution along one dimension varies across another dimension. It is known that the visual system can perform the summary statistics tasks in general (Whitney & Yamanashi Leib, 2018) and summary statistics related to trend detection (Rensink, 2012). Extracting trends from scatterplots and stripplots both likely require a probability distribution of how one visual feature (vertical position or color) varies as a function of another visual feature (horizontal position). Correlation perception for color stripplots has shown to be somewhat more accurate than correlation perception for scatterplots, suggesting that basic visual features (e.g., color) can effectively convey correlation information (Rensink, 2015).

What is unknown is whether better precision at the level of pairwise comparisons coincides with better precision at the level of the ensemble of data points. Correlational research points to a relationship between visual precision at the level of the individual and at the level of the ensemble, but cannot assert a causal link (Haberman et al., 2015). In contrast, research on visualizations have documented several instances for which performance on summary tasks are better when the data are represented with color

rather than with position (Albers et al., 2014; Correll et al., 2012; Rensink, 2012, but see Fuchs et al., 2013). If performance were better with color than with vertical position, this would suggest that there is not a causal link between precision for individual elements and precision for ensembles.

Our data showed that graphs depicting trends with vertical position are superior to graphs depicting trends with color. Sensitivity to whether a trend increased or decreased was greater with the line graphs than with the stripplots. With respect to mechanisms underlying ensemble perception, the data are consistent with the idea that better visual precision for individual objects leads to better visual precision for ensembles of objects. This claim rests on the assumption that the results from Cleveland and McGill (1985) would generalize to the specific features tested here because we did not test precision for pairwise comparisons in our stimuli. The results do not prove a causal relationship between precision at the level of individual objects and precision at the level of ensembles. However, had the results showed the opposite pattern with better sensitivity for the stripplots than the line graphs, the data would have had different implications for mechanisms underlying ensemble perception. Specifically, the precision for pairwise comparisons would not generalize to the precision for ensemble perceptions. Note that perception of the individual objects need not be conscious, as demonstrated by work on visual crowding and change detection (Fischer & Whitney, 2011; Haberman & Whitney, 2011; Parkes et al., 2001).

In addition, the research speaks to common mechanisms underlying trend extraction for both kinds of graphs. Participants' sensitivity to the direction of the trend in line graphs was highly correlated with their sensitivity to the direction of the trend in stripplots. This correlation is indicative of a common underlying process involved in perceiving both types of graphs. That a correlation was found across the displays is consistent with prior research on extracting the mean from a group of objects. In those studies, participants estimated the mean orientation of triangles and the mean color of triangles (Haberman et al., 2015). Those who were more precise at estimating mean orientation were also more precise at estimating mean color. The authors proposed a low-level ensemble processor. Our data replicate this finding and extend it to trend detection. However, future research is needed to determine whether the low-level ensemble processor is involved in all ensemble judgments, including mean and trends, rather than separate low-level processes for each type of ensemble judgment.

There are several limitations to our data. One is that we used line graphs rather than point graphs (or scatterplots). The research on pairwise comparisons compared two points, rather than points implied within line graphs (Cleveland & McGill, 1985). While we can make claims about sensitivity due to relative vertical position, we are limited in making any claims about precision at the level of individual elements (which we did not measure) and how they relate to precision for ensembles. A causal link between precision at the level of the individual and of the ensemble was not assessed. Another limitation is that we only used one color palette for the stripplots. Perhaps a different choice in color palettes would have led to better sensitivity and perhaps even sensitivity similar to that found with the line graphs. In particular, ensemble perception is more precise when averaging across fewer colors than more colors (Maule & Franklin, 2015). If this result extends to trend detection, using a single-hue palette could shrink or eliminate the differences in sensitivity between line graphs and stripplots.

An additional limitation is that our task concerned a major global issue, namely global warming. Because we did not compare performance across different types of contexts, it is unclear the extent to which the biases revealed in the current data are due to the visualizations themselves or to the topic depicted in the graphs. In other words, it is unclear whether line graphs lead to a general bias to report that trends are increasing or if the bias is specific to the context of global warming.

Concerning data visualization, the results raise the issue of a one-size-fits-all design recommendation. Our data on trend detection revealed different design recommendations compared with prior results on other kinds of summary tasks, such as detecting the mean value for which the stripplots were superior (Albers et al., 2014). Even within our results, we found that bias differed depending on the magnitude of the trend's exponent. These conflicting results highlight the need for a systematic evaluation of different kinds of summary tasks. Understanding ensemble mechanisms involved in various types of tasks and how mapping data values to visual features and spatial positions contribute to ensemble processes could better inform design guidelines of visualizations and maximize the viewer's ability to process patterns in the underlying data. The various kinds of ensemble perception tasks may differ in the kind of visualization that would best promote sensitivity to the information. A systematic evaluation would benefit from using the signal detection measures used here to separate the effects of design features on sensitivity from those on bias. Making this distinction can help build an overarching framework of design recommendations for visualizing time series and trend data.

The current data expand on prior research on visualizations by exploring nonlinear time series trends. For line graphs, we found that higher exponents lead to better sensitivity to whether the trend was increasing or decreasing. This is sensible given that an observer can focus on just the last part of the display as the exponent

increased. However, this clustering of the most relevant information as the exponent increased did not improve performance for the stripplots. For stripplots, sensitivity was similar regardless of whether the trend was linear or nonlinear.

Concerning visualization design, the data showed that line graphs were superior to stripplots for detecting linear and nonlinear trends. Of course, the stripplots were rated as more visually appealing than the line graphs, so for design purposes, there was a tradeoff between better precision in communicating the information and design preference. Although a stripplot became a Twitter trend called *#ShowYourStripes,* graced the cover of the *Economist,* and was used to decorate various items from ties to cars, the line graphs are unlikely to be as sensationalized. The current data show tradeoffs: the cost of using the more aesthetically pleasing stripplots means an 18% reduction in sensitivity.

Keywords: visualizations, graphs, sensitivity

Acknowledgments

Supported by a Grant from the National Science Foundation (BCS-1632222) and a fellowship from the Colorado State University School of Global Environmental Sustainability to JKW. Stimulus materials, data, and analysis scripts can be found at https://osf.io/x372y/.

Commercial relationships: none.
Corresponding author: Jessica Witt.
Email: jessica.witt@colostate.edu.
Address: Department of Psychology, Colorado State University, Fort Collins, CO 80523, USA.

References

Albers, D., Correll, M., & Gleicher, M. (2014). Task-Driven Evaluation of Aggregation in Time Series Visualization. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI Conference,* 2014, 551–560, https://doi.org/10.1145/2556288.2557200.

Ariely, D. (2001). Seeing Sets: Representation by Statistical Properties. *Psychological Science, 12*(2), 157–162, https://doi.org/10.1111/1467-9280.00327.

Brewer, C. A. (1999a). Color Use Guidelines for Data Representation. Proceedings of the Section on Statistical Graphics, American Statistical Association, Alexandria VA., 55–56.

Cleveland, W. S., & McGill, R. (1985). Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science, 229*(4716), 828–833, https://doi.org/10.1126/science.229.4716.828.

Correll, M., Albers, D., Franconeri, S., & Gleicher, M. (2012). Comparing Averages in Time Series Data. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 1095–1104, https://doi.org/10.1145/2207676.2208556.

Elliot, A. J., & Maier, M. A. (2007). Color and psychological functioning. *Current Directions in Psychological Science, 16*(5), 250–254.

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology, 106*(3), 1389–1398.

Fuchs, J., Fischer, F., Mansmann, F., Bertini, E., & Isenberg, P. (2013). Evaluation of alternative glyph designs for time series data in a small multiple setting. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 3237–3246.

Haberman, J., Brady, T. F., & Whitney, D. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General, 144*(2), 432–446.

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review, 18*(5), 855.

Hawkins, E. [@ed_hawkins]. (2019, June 17). Global temperature change from 1850-2018. [Tweet with image attached]. *Twitter*, https.://twitter.com/ed_hawkins/status/1140772720508047360.

Legge, G. E., Gu, Y., & Luebker, A. (1989). Efficiency of graphical perception. *Perception & Psychophysics, 46*(4), 365–374.

Macmillan, N. A., & Creelman, C. D. (2008). *Detection Theory: A User's Guide* (Second Edition). New York: Psychology Press.

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision, 15*(4), 6, https://doi.org/10.1167/15.4.6.

Moller, A. C., Elliot, A. J., & Maier, M. A. (2009). Basic hue-meaning associations. *Emotion, 9*(6), 898.

Or, C. K., & Wang, H. H. (2014). Color–concept associations: A cross-occupational and-cultural study and comparison. *Color Research & Application, 39*(6), 630–635.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging

of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739.

Rensink, R. (2015). Visual Features as Carriers of Information. *Journal of Vision, 15*(12), 890–890, https://doi.org/10.1167/15.12.890.

Rensink, R. A. (2012). Invariance of correlation perception. *Journal of Vision, 12*(9), 433.

Rensink, R. A. (2017). The nature of correlation perception in scatterplots. *Psychonomic Bulletin & Review, 24*(3), 776–797.

Smart, S., Wu, K., & Szafir, D. A. (2020). Color Crafting: Automating the Construction of Designer Quality Color Ramps. IEEE Transactions on Visualization and Computer Graphics, *26*(1), 1215–1225, https://doi.org/10.1109/TVCG.2019.2934284.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science, 11*(5), 702–712, https://doi.org/10.1177/1745691616658637.

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4–5-year-old children. *Developmental Science, 18*(4), 556–568.

Szafir, D. A., Haroz, S., Gleicher, M., & Franconeri, S. (2016). Four types of ensemble coding in data visualizations. *Journal of Vision, 16*(5), 11, https://doi.org/10.1167/16.5.11.

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*(6), 401–409.

Team, R. C. (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Tham, D. S. Y., Sowden, P. T., Grandison, A., Franklin, A., Lee, A. K. W., Ng, M., Park, J., Pang, W., … Zhao, J. (2019). A systematic investigation of conceptual color associations. *Journal of Experimental Psychology: General, 149*, 1311–1332.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology, 69*, 105–129.

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. Oxford: Oxford University Press.