Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations

Jeffrey Heer[1], **Nicholas Kong**[2], **and Maneesh Agrawala**[2]

[1] Computer Science Department
Stanford University
Stanford, CA 94305 USA
jheer@cs.stanford.edu

[2] Computer Science Division
University of California, Berkeley
Berkeley, CA 94720-1776 USA
{nkong, maneesh}@cs.berkeley.edu

ABSTRACT
We investigate techniques for visualizing time series data and evaluate their effect in value comparison tasks. We compare line charts with *horizon graphs*—a space-efficient time series visualization technique—across a range of chart sizes, measuring the speed and accuracy of subjects' estimates of value differences between charts. We identify transition points at which reducing the chart height results in significantly differing drops in estimation accuracy across the compared chart types, and we find optimal positions in the speed-accuracy tradeoff curve at which viewers performed quickly without attendant drops in accuracy. Based on these results, we propose approaches for increasing data density that optimize graphical perception.

Author Keywords
Visualization, graphical perception, time series, line charts, horizon graphs.

ACM Classification Keywords
H.5.2. Information Interfaces: User Interfaces.

INTRODUCTION
Time series—sets of values changing over time—are one of the most common forms of recorded data. Time-varying phenomena are central to many areas of human endeavor and analysts often need to simultaneously compare a large number of time series. Examples occur in finance (e.g., stock prices, exchange rates), science (e.g., temperatures, pollution levels, electric potentials), and public policy (e.g., crime rates), to name just a few. Accordingly, visualizations that improve the speed and accuracy with which human analysts can compare and contrast time-varying data are of great practical benefit.

Effective presentation of multiple time series is an instance of a larger problem in visualization research: increasing the amount of data with which human analysts can effectively work. Toward this aim, researchers and designers have devised design guidelines and visualization techniques for making more effective use of display space. Tufte [27] advises designers to maximize data density (data marks per chart area) and researchers regularly promote visualization techniques (e.g., [12, 22, 25]) for their "space-filling" properties. Such approaches excel at increasing the amount of information that can be encoded within a display. However, increased data density does not necessarily imply improved graphical perception for visualization viewers.

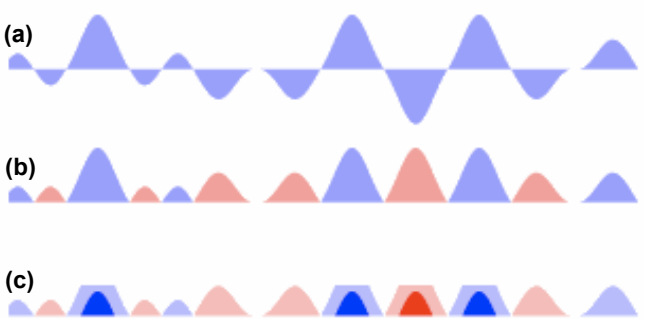

Figure 1. (a) Filled line chart. Area between data values on line and zero is filled in. **(b) "Mirrored" chart.** Negative values are flipped and colored red, cutting the chart height by half. **(c) 2-band horizon graph.** The chart is divided into bands and overlaid, again halving the height.

Consider the three time series charts in Figure 1. The first graph is a filled line chart — a line chart with the area between the data value on the line and zero filled in. The second graph "mirrors" negative values into the same region as positive values, and it relies on hue to differentiate between the two. The mirror chart doubles the data density compared to the line chart. The third chart, called a *horizon graph* [7], further reduces space use by dividing the chart into bands and layering the bands to create a nested form. With two layered bands the horizon graph doubles the data density yet again.

Such increases in data density enable designers to display more charts in a fixed area and thereby make it easier for viewers to compare data across multiple charts. Yet, mirroring negative values, dividing the series into bands, and layering the bands may also obscure patterns in the data

CHI 2009, April 4–9, 2009, Boston, MA, USA.

and reduce estimation accuracy. Few [7] argues that the benefits of increased data density in horizon graphs outweigh the drawback. However, it remains unclear how mirroring, dividing, and layering time series data affects the ability of analysts to quickly and reliably spot trends and compare values. Do viewers correctly interpret mirrored negative values? Does mental unstacking of layered charts interfere with estimation?

In this paper, we evaluate space-efficient techniques for visualizing time series data through a series of controlled experiments. We investigate the effects of chart height and layering on the speed and accuracy of value comparison tasks. We identify transition points at which smaller chart heights result in differing drops in estimation accuracy across chart types, and we provide guidelines indicating which charts work best at which scales. We also note an unexpected effect: estimation times decrease as charts get smaller, though estimation accuracy also decreases.

We begin by reviewing related work on both graphical perception studies and time series visualization techniques. Next, we present two graphical perception experiments of time series charts. The first investigates different variants of horizon graphs and the second examines both chart type and chart size. We then discuss the implications of our experimental results and propose guidelines for improving graphical perception of space-efficient time series charts.

GRAPHICAL PERCEPTION

A volume of prior research has investigated the degree to which visual encoding variables such as position, length, area, shape, and color facilitate comprehension of data sets. Following Cleveland [5], we use the term *graphical perception* to denote the ability of viewers to interpret such visual encodings and thereby decode information in graphs.

Bertin [2] provides the first systematic treatment of visual encodings, rank-ordering visual variables according to their effectiveness for encoding nominal, ordinal, and quantitative data. For example, Bertin posits that spatial position best facilitates graphical perception across all data types, while color hue ranks highly for nominal (category) data but poorly for quantitative data. Bertin bases his rankings on his experience as a graphic designer and cartographer.

Cleveland and McGill [5] place the ranking of visual encodings on a more rigorous scientific footing through perceptual experiments with human subjects. Subjects were shown charts and asked to compare the quantitative values of two marks by estimating what percentage the smaller value was of the larger. This accuracy measure is then used to test and refine the ranking of different visual variables.

Many other researchers have applied experimental methods to graphical perception tasks. Simkin and Hastie [23] test value discrimination and estimation for bar, divided bar, and pie charts. Spence and Lewandowsky [24] use a two-alternative discrimination task to investigate perception of percentages in bar charts, pie charts, and tables. Multiple projects [14, 26] investigate shape discrimination of scatter plot symbols. More recently, Wigdor et al. [30] apply the approach of Cleveland and McGill to measure how visual variable rankings vary due to perspective distortions that occur when seated at a table-top display.

Each of these studies measures how a visual encoding variable (i.e., position, size, hue, etc.) affects the accuracy and/or response time of estimating values of the underlying data. Additional work has built upon these studies to create cognitive performance models of graph decoding [8, 15, 23]. Establishing the effectiveness of visual encodings for graphical perception tasks is also vital for the design of effective visualizations and the development of automatic presentation software [16, 17].

Once a designer (or software) selects suitable visual encodings for the data variables of interest, one still needs to specify the aspect ratio [1, 6] and overall chart size. Size is of particular concern when analysts deal with many data sets and wish to make comparisons across them. The goal is to maximize the amount of data shown without hampering graphical perception. Despite a wealth of work on individual visual variables and (to a lesser extent) their interactions [18], there is relatively little research into the impact of chart size and density on graphical perception. Cleveland et al. [4] investigate scale effects on correlation perception in scatterplots, but vary axis ranges only, not display size. Woodruff et al. [31] present methods for promoting constant data density in semantic zooming applications, but without an empirical evaluation. Lam et al. [13] study the effects of low and high resolution displays on visual comparison and search tasks. They focus primarily on the cognitive costs of switching between display types. Their low- and high-res displays use different visual encoding variables (color vs. position), confounding analysis of the impact of display size. In this paper, we present studies of comparison tasks for time series data and measure both accuracy and time across various chart size and data density conditions.

TIME SERIES VISUALIZATION

Given the ubiquity of time series data, researchers have developed myriad approaches to time series visualization.

Line Charts

The most common form of time series visualization is the line chart, which uses position encodings for both time and value. Line charts often encode time as progressing from left to right along the horizontal axis, and encode time-varying values along the vertical axis. Line segments connect successive points and the slope of the line encodes the rate of change between samples.

Collections of time series can be overlaid on the same axes to facilitate comparison of values. However, placing multiple charts in the same space can produce overlapping curves that reduce the legibility of individual time-series. A popular alternative to overlaying multiple series is to use *small multiples* [27] showing each series in its own chart.

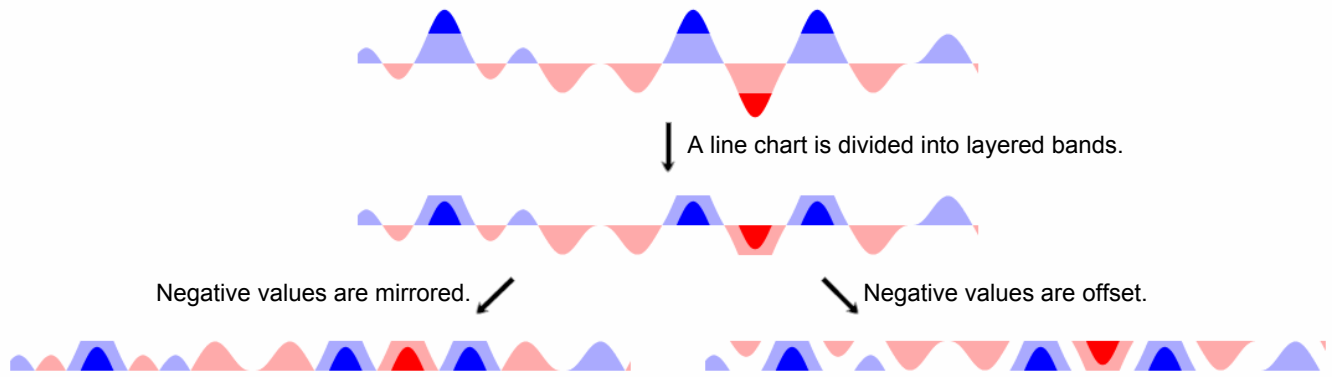

Figure 2. Horizon graph construction. A normal line chart is divided into bands defined by uniform value ranges. The bands are then layered to reduce the chart height. Negative values can be mirrored or offset into the same space as positive values.

Multiple charts are typically enumerated vertically and aligned horizontally to aid comparison of events and trends. Small multiples can be used with *sparklines* [28]—word-size data graphics—to form a data-dense display.

Optimizing Line Chart Aspect Ratios

Researchers have investigated ways to improve graphical perception by optimizing the display of line charts. In his book Visualizing Data, Cleveland [6] demonstrates how the aspect ratio of a line chart affects trend perception. He proposes using an aspect ratio at which the average absolute orientation of line segments in the chart is equal to 45 degrees. This technique, called *banking to 45°*, aims to maximize the discriminability of the orientations of the line segments in a chart. Heer and Agrawala [8] extend this approach by identifying trends at multiple data scales and computing a set of trend-specific aspect ratios. These techniques for banking to 45° leave one free size parameter: given a fixed height the aspect ratio will determine the width, and vice versa. A visualization designer must still choose either the height or width of the chart.

Stacked Time Series

Stacked graphs are an approach to time series visualization that simply stack time series on top of each other. The result is a visual summation of time series values that provides an aggregate view stratified by individual series. Projects such as NameVoyager [29] and sense.us [10] used animated stacked graphs to explore demographic data.

Though seemingly effective for aggregate patterns, stacked graphs are awkward for comparing individual series. Visual stacking is not an informative aggregation for many data types (e.g., temperature) or for negative values. Comparing values involves length (stack height) comparisons rather than more accurate position judgments [5]. Furthermore, viewers often misinterpret the space between curves [5], perceiving minimum rather than vertical distance. Byron and Wattenberg [3] suggest sorting the stacks to mitigate this problem. While sorting can improve perception, it cannot eliminate the issue. For these reasons stacked graphs are not ideal for comparing individual series and we remove them from consideration in the present work.

Animation

Directly animating values over time is another means of displaying time-series data. Examples include animating marks on a map to show time-varying geographic data and animating scatterplots to show trends (e.g., Gapminder [20]). Researchers have found that animating between time slices facilitates value change estimation better than static transitions between views [11], but that animation results in significantly lower accuracy in analytic tasks compared to small multiples of static charts [19]. Given these results, we restrict our focus to spatial representations of time.

Horizon Graphs

A *horizon graph* is a relatively new chart type that increases the density of time series graphs by dividing and layering filled line charts. Saito et al. [21] first developed the technique under the name "two-tone pseudo-coloring" and Panopticon [7] independently commercialized and branded the approach. As illustrated in Figure 2, one can construct a horizon graph by first segmenting a line chart along the vertical axis into uniformly-sized non-overlapping bands. The bands are then layered on top of each other and negative values are reflected around the zero point. Hue (blue or red) indicates positive or negative values, and saturation and/or intensity indicate the band level. Horizon graphs reduce the height of a line chart with positive and negative values by a factor of $2 \times \#\ bands$. We refer to this particular technique as a *mirrored graph* due to the reflection of negative values around the zero point.

We have devised an alternative approach which we call an *offset graph*, also shown in Figure 2. The construction is similar to mirrored graphs, except that rather than reflecting negative values, we offset the negative values such that the zero point for the negative values is at the top of the chart. In other words, we "slide up" the negative values. As a result, slopes for negative values are preserved, but the positive and negative values no longer share a common zero point.

Both mirror and offset horizon graphs show promise for increasing the amount of data that can be shown in a fixed display space. Both variants make use of a layered position encoding of values. Viewers can make position judgments to compare absolute differences between values in the same band. However, comparing differences across bands or making relative (proportional) judgments requires viewers to parse the band structure and mentally "unstack" the band ranges. In a set of graphical perception experiments, we explore how these additional cognitive operations affect the speed and accuracy of value estimation.

APPROACH AND METHODS

Our objective was to quantify the effects of chart sizing and layering on the speed and accuracy of graphical perception. To this end we ran two experiments. The goal of the first experiment was to determine the impact of band number and horizon graph variant (mirrored or offset) on value comparisons between horizon graphs. The goal of the second experiment was to compare line charts to horizon graphs and investigate the effect of chart height on both.

In both experiments, subjects completed discrimination and estimation tasks for points on time series graphs. Since the use case of horizon graphs is to compare data across several time series plots, we asked subjects to simultaneously view two separate graphs and compare a point on one graph to a point on the other, as shown in Figure 3. Subjects first reported which point represented the greater value and then estimated the absolute difference between the two. For each trial, we measured the *estimation error* as the absolute difference between a subject's estimation and the actual value difference between comparison points.

In order to reduce learning effects, we told subjects to take as many practice trials as they wished and instructed them to practice until they had reached a steady performance level. After each practice trial, the experimental software showed subjects the correct responses.

When analyzing the experimental data, we were concerned with the impact of outliers due to keying errors and extreme responses. Therefore we used 80% trimmed means, a more robust statistic, to analyze estimation time and accuracy. The statistic is the arithmetic mean of the middle 8 deciles of the data. In other words, we drop both the bottom and top 10% of the data. Cleveland et al. [4, 5] use a similar tactic in their work on graphical perception. In our analyses we used per-subject trimmed means for each experimental condition.

EXPERIMENT 1: HORIZON GRAPH COMPARISON

We designed our first experiment to address two questions:

(a) How does the choice of mirrored or offset horizon graph affect estimation time or accuracy?

(b) How does the number of bands in a horizon chart affect estimation time or accuracy?

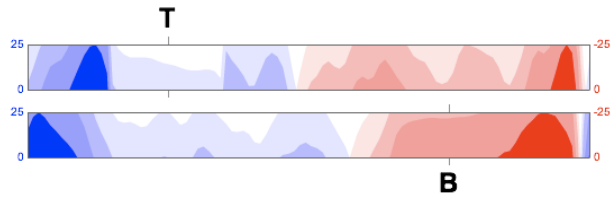

Figure 3. Example trial with a 4-band mirrored graph. Each band covers 25 values; the total range is [-100, 100]. Subjects reported if T or B was larger, and by how much.

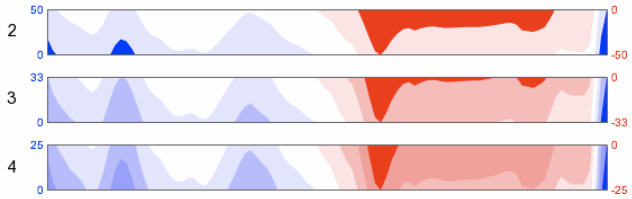

Figure 4. Offset horizon graphs with 2, 3, and 4 bands.

We hypothesized that offset graphs would result in faster, more accurate comparisons than mirror graphs, as offset graphs do not require mentally flipping negative values.

With respect to layering, we hypothesized that increasing the number of bands would increase estimation time and decrease accuracy across graph variants. We believe that increasing the bands increases the difficulty of the task by requiring additional perceptual discrimination to identify the bands and higher cognitive load to remember the band structure and perform mental arithmetic.

Method

In each trial, subjects viewed two charts, each with a position marked either T or B (Figure 3). Subjects first performed the discrimination task in which they reported whether position T or position B represented the greater value. Subjects then performed the estimation task in which they reported the absolute difference between the values at positions T and B. We asked subjects to answer as quickly as possible while trying to make estimates accurate to within 5 values. All charts were 500 pixels wide and 40 pixels tall. The y-axis of the time-series ranged from -100 to 100 values. We labeled the y-axis of each chart with the ranges for the first band (e.g., 0-50 or 0-33, see Figure 4).

We created the time-series by running a symmetric, discrete triangle smoothing filter over a random walk. We provide the details of our smoothing approach in Appendix A.

The experiment used a 2 (chart) × 3 (band) within-subjects design. We tested mirrored and offset horizon graphs with 2, 3, and 4 bands (Figure 4). A fully crossed design with 16 trials per condition resulted in 3×2×16 = 96 trials per subject. As we were interested in observing effects due to layering, each trial compared two values in different bands. We counterbalanced the trials to cover all pairs of bands.

To avoid confusion across conditions, we tested each cell of the experiment in a separate block. We preceded each block with practice trials in which we showed subjects the correct answers after they responded. We designed the experiment to test only for effects due to layering and kept the physical (pixel) height of the charts constant. We also fixed the horizontal location of the comparison points for every trial.

We deployed the experiment on the web as a Flash applet. Eighteen unpaid subjects (13 male, 5 female), participated in the study and were recruited through campus mailing lists. All were graduate or undergraduate engineering students. Each subject used their own machine and browser, so there was no control for screen resolution. Since we did not vary the chart size, effects due to resolution should be at least partially accounted for by the within-subjects design.

Results

For all conditions discrimination accuracy averaged 99% or higher, so we focus on the results of the estimation task. To test for significant effects, we first conducted a Repeated Measures MANOVA on the combined (error, time) results. The RM-MANOVA found a significant main effect for band count ($F(4,68) = 11.01$, $p < 0.001$), but did not find an effect for chart type ($F(2,16) = 0.367$, $p = 0.699$) nor any interaction ($F(4,68) = 0.211$, $p = 0.163$). We then performed univariate analysis of time and error for band counts.

Estimation Error Increases in 4-Band Condition

Univariate analysis of the estimation error found a significant main effect for band count ($F(2,34) = 58.27$, $p = 0.013$). Figure 5 shows the mean estimation errors by band count. Pair-wise comparison of the band counts found that estimation accuracy was not significantly different across the 2 and 3 band cases ($p = 0.768$), but that the 4 band case was less accurate than both the 2 band (mean difference of 1.52 units, $p = 0.042$) and 3 band (mean difference of 1.59 units, $p = 0.026$) cases.

Estimation Time Increases With Band Count

Univariate analysis of estimation times found a significant main effect for band count ($F(2,34) = 431.18$, $p < 0.001$). Figure 6 shows the mean estimation times by band count. Pair-wise comparison of the band counts found significant differences between all levels ($p < 0.001$ in all cases), with a mean increase of 2.89 seconds between 2 and 3 bands and an increase of 1.91 seconds between 3 and 4 bands.

Discussion

We found no significant difference in either estimation time or accuracy between chart types and reject our hypothesis that offset graphs would provide better performance than mirror graphs. Rather, the results suggest that mirrored and offset graphs are comparable for value comparison tasks.

However, the results confirm our hypothesis regarding the effects of band count on performance: both estimation time and error increased with more bands. Across graph types, using 2 or 3 bands had similar error levels ($M = 4.12$ and $M = 4.04$ units, respectively), while 4 bands resulted in significantly higher estimation error ($M = 5.64$ units). After the experiment, multiple subjects verbally reported that as the band count rose they experienced increased difficulty identifying and remembering which band contained a value and that performing mental math became fatiguing. Subjects also noted that working with ranges of 33 values in the 3-band condition was more difficult than working with the ranges in the 2 and 4 band conditions that were multiples of five. Though estimation time was slower with 3 bands than with 2, accuracy did not suffer similarly.

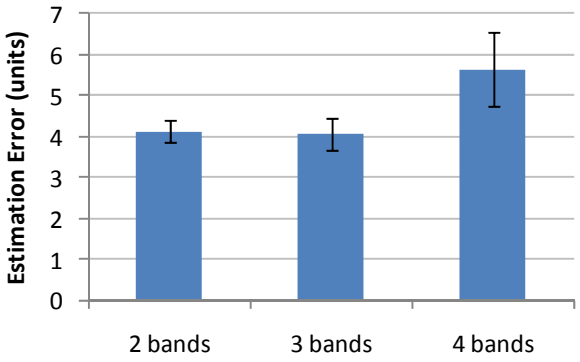

Figure 5. Estimation Error by Band Count. 4-band charts have significantly higher error than 2- or 3-band charts.

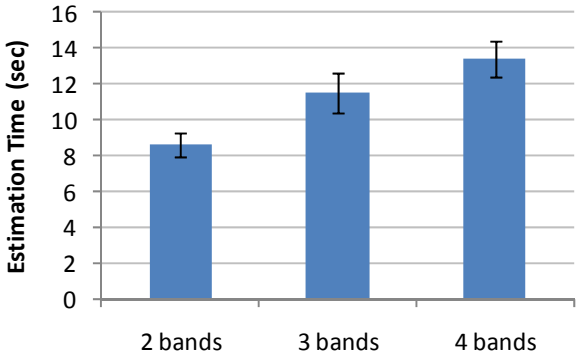

Figure 6. Estimation Time by Band Count. Estimation time increases significantly with each additional band.

EXPERIMENT 2: CHART SIZE AND LAYERING

We designed our next experiment to answer the questions:

(a) How do mirroring and layering affect estimation time and accuracy compared to line charts?

(b) How does chart size affect estimation time and accuracy?

In our first experiment we found that mirrored and offset graphs had comparable estimation times and accuracies. Mirrored graphs are also used in commercial products, and so we removed offset graphs from consideration in this experiment and focused on comparing mirrored graphs to filled line charts. The first experiment also found that 2- and 3-band charts had comparable accuracy, but that 3-band charts were significantly slower. Consequently, we limited the maximum band level to two. Thus, in this experiment,

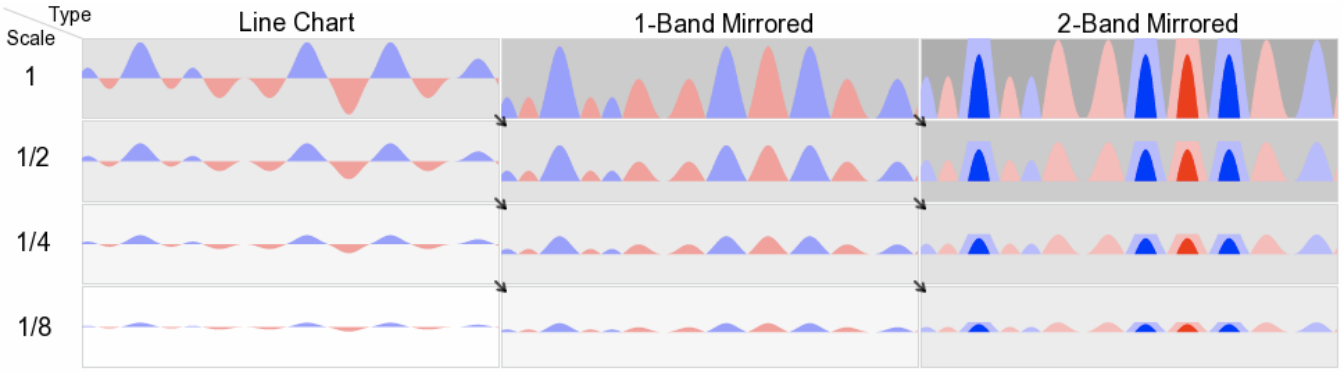

Figure 7. Chart Type and Scale Conditions in Experiment 2. We crossed 3 chart types and 4 chart heights. The diagonally adjacent cells indicated by arrows and shading have the same *virtual resolution*: the un-mirrored, un-layered size of the chart.

we compared line charts, mirrored charts without banding, and mirrored charts with two bands. We also varied the chart height for each type across four scales (Figure 7).

We hypothesized that at larger chart heights line charts would be faster and more accurate than mirror charts both with and without banding, and that mirror charts without banding would be faster and more accurate than those with banding. For the 2-band condition, we expected that mentally unstacking the charts would result in slowdowns akin to those seen in Experiment 1. In the mirroring-only condition, we expected comparisons across positive and negative ranges to be slower than comparisons made with non-mirrored line charts.

We also hypothesized that as chart heights decreased, error would increase monotonically, but would do so unevenly across chart types due to their differing data densities. We expected 2-band horizon graphs to result in better accuracy than the other chart types once the chart height fell under a threshold size, as the "unstacked" version of a horizon graph provides more pixels per unit value. Thus we predicted the presence of transition points in the height of the charts at which charts with higher data density result in higher accuracy. A primary goal of the experiment was to determine such transition points, should they exist.

Method

As in the prior experiment, in every trial subjects viewed two charts marked with comparison points and performed discrimination and estimation tasks. We instructed subjects to answer as quickly as possible while attempting to make estimates accurate to within 5 values. All charts were 500 pixels wide and we varied chart height as a factor.

The experiment used a 3 (chart) × 4 (size) within-subjects design. We tested 3 chart types (normal, 1-band mirrored, 2-band mirrored) and 4 scale factors (1, 1/2, 1/4, 1/8) where scale factor 1 corresponded to a height of 48 pixels. A fully crossed design with 10 trials per cell resulted in 4×3×10 = 120 trials per participant. We counterbalanced the trials for value differences between points. In each trial the comparison points were located in different bands. We counterbalanced the trials to cover all pairs of bands.

We recruited thirty paid subjects (17 male, 13 female) via a research participation pool. Subjects were undergraduate students from a variety of majors. All subjects performed the experiment on a 14.1" LCD monitor at 1024 × 768 pixel resolution. At scale factor 1, the physical chart size was 13.9 × 1.35 centimeters. Subjects sat normally at a desk and we did not constrain their movement.

We subsequently ran a follow-up experiment to further test performance at extremely small sizes and investigate accuracy transitions between the 1- and 2-band conditions. The follow-up used a 2 (chart) × 3 (size) within-subjects design, comparing the 1- and 2-band mirrored conditions and scale factors of (1/8, 1/12, 1/24). At the smallest scale, the chart height was only 48/24 = 2 pixels tall. We recruited eight paid subjects (6 male, 2 female) via campus e-mail lists. All subjects were graduate engineering students and used a 14.1" LCD monitor at 1024 × 768 pixel resolution. Six subjects had previously participated in Experiment 1.

Results

For all conditions, discrimination accuracy averaged 98% or higher for the main experiment and 96% or higher for the follow-up, so we focus on the results of estimation tasks. In the main experiment, a RM-MANOVA for (error, time) found significant effects for chart type ($F(4,116) = 11.086$, $p < 0.001$) and chart height ($F(6,174) = 7.099$, $p < 0.001$), but no interaction effect ($F(12,348) = 0.921$, $p = 0.526$). In the follow-up experiment, a RM-MANOVA similarly found significant effects for chart type ($F(2,6) = 21.630$, $p = 0.002$) and height ($F(4,28) = 5.555$, $p = 0.002$), but no interaction effect ($F(4,28) = 0.689$, $p = 0.605$).

Estimation Error Increases as Chart Height Decreases

Univariate analysis of estimation error found significant effects for both chart type ($F(2,58) = 7.550$, $p = 0.001$) and chart height ($F(3,87) = 12.369$, $p < 0.001$). Pair-wise comparisons showed a disadvantage for line charts against both 1- and 2-band mirror charts ($p < 0.001$ and $p = 0.015$,

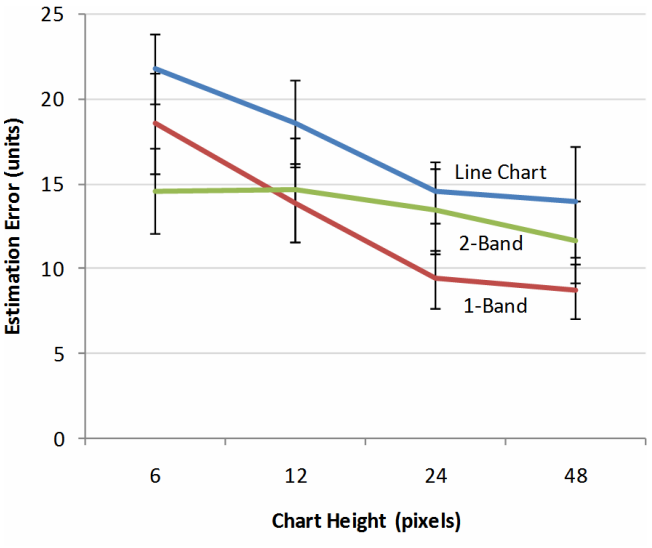

Figure 8. Estimation Error by Chart Type and Height. The 2-band mirror chart crosses the 1-band case at a chart height of 12 pixels (scale factor 1/4).

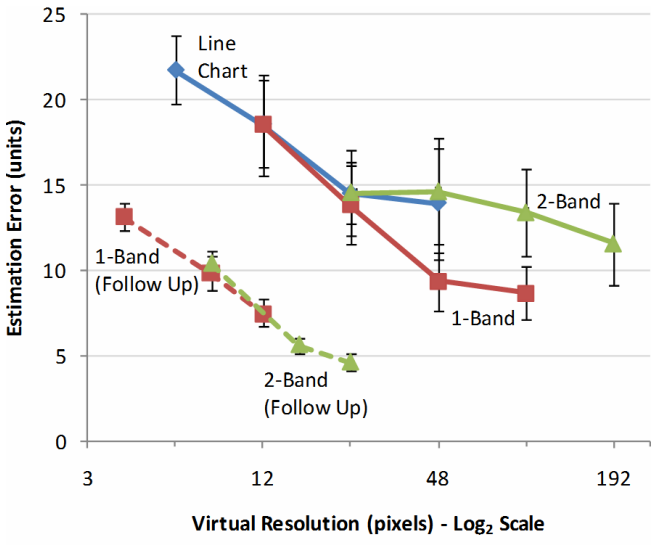

Figure 9. Estimation Error by Chart Type and Virtual Resolution. Error levels hold relatively stable at high virtual resolutions, but increase linearly at smaller resolutions.

respectively). One-band mirror charts had lower error than line charts at all scale factors (Figure 8). Our follow-up experiment found significant effects for chart type ($F(1,7) = 23.189$, $p = 0.002$) and height ($F(2,14) = 44.283$, $p < 0.001$), with error increasing as chart height decreases.

As shown in Figure 8, accuracy decreased at smaller chart heights. In the main experiment, this effect was most pronounced for line charts and 1-band mirror charts. Estimation error remained steady for scale factors of 1 and 1/2. At smaller sizes, both chart types had monotonically increasing error. Estimation error for 2-band mirror charts stayed relatively stable, equaling or beating the line and 1-band mirror charts at scales of 1/4 (12 pixels) and lower.

Estimation Error Increases with Virtual Resolution

The preceding analysis indicates a crossover point at which 2-band scale charts begin to outperform other chart types in terms of estimation accuracy. We hypothesized that increases in error are attributable to a chart's *virtual resolution*. We define virtual resolution as the un-mirrored, un-layered height of a chart. The virtual resolution for a line chart is simply its height. For a 1-band mirror chart it is twice the height. For a 2-band mirror chart it is four times the height.

Figure 9 plots the estimation accuracies of the chart types by their virtual resolutions. As we successively decreased chart height by a factor of two, we plotted virtual resolution on a base 2 logarithmic scale. For large virtual resolutions, the plot shows plateaus where the error level is stable. At lower resolutions, the error rate rises in a similar manner across charts. While the 2-band mirror chart has a greater baseline error rate, it also has a greater virtual resolution at a given chart height. It accordingly maintains the baseline error level for chart heights at which performance degrades in other chart types. At resolutions below 24 pixels, error appears to increase linearly as the virtual resolution halves. To test this observation, we ran a linear regression of error and virtual resolution at resolutions of 24 pixels and below. The regression fits with $R^2 = 0.986$ and a slope of -4.1 units / $\log_2$ pixel, indicating a linear increase.

We ran our eight subject follow-up experiment to see if our hypothesis would hold at smaller scales. We expected to find that the 2-band chart degrades in performance at the same virtual resolutions at which the other charts degrade. The results are shown in the bottom left corner of Figure 9. The baseline error rate was substantially less in our follow-up; we attribute the disparity to our different subject pools. (Follow-up subjects were engineering grad students, and many participated in Experiment 1. The base error rate in the follow-up is closer to that of Experiment 1.) We found that 1- and 2-band charts had nearly identical error levels at matching virtual resolution values. We also found that the errors increased at rates similar to the main experiment. Linear regression of error and virtual resolution fits with $R^2 = 0.980$ and a slope of -3.5 units / $\log_2$ pixel, again indicating a linear increase in error as chart heights halve.

Layering Increases Estimation Time, Mirroring Does Not

Univariate analysis of estimation time found a significant effect for chart type ($F(2,58) = 16.686$, $p < 0.001$). Two-band mirror charts were slower than normal time series by 2.05 sec on average ($p < 0.001$) and 1-band mirror charts by 1.91 on average sec ($p < 0.001$). The result is consistent with Experiment 1, where increasing the band count slowed estimation. We found no significant difference between 1-band mirror charts and line charts ($p = 0.632$). In our follow-up experiment, we found 1-band charts to be faster than 2-band charts by 0.85 sec ($F(1,7) = 10.911$, $p = 0.013$).

Estimation Time Decreases with Chart Height

Analysis of estimation times also found an effect for chart height ($F(3,87) = 5.139$, $p = 0.003$). As the chart height

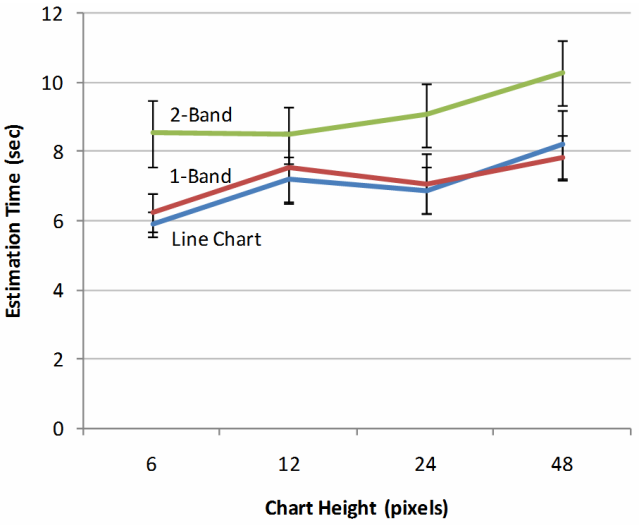

Figure 10. Estimation Time by Chart Type and Height. Line and 1-band mirror charts result in similar estimation times. Both are significantly faster than 2-band charts.

decreases, so does the estimation time. As plotted in Figure 10, estimation time is affected primarily by the chart height and *not* the virtual resolution of the graph, since an estimation time vs. virtual resolution plot would require the 2-band line to shift right two bins and the 1-band line to shift right one bin. At scale 1/2 (24 pixels), estimation times were faster than for larger charts by an average 1.1 sec. Interestingly, error increased less than 2 units across all chart types between scale 1 and scale 1/2. In our follow-up we found no effect on estimation time for the smaller scale charts ($F(2,14) = 1.525$, $p = 0.252$).

Discussion

Our first hypothesis was that at large chart sizes, line charts would outperform both mirror chart types, and that 1-band mirror charts would outperform the 2-band case. The hypothesis was only partially confirmed. At the two largest chart sizes, 1-band charts were faster and more accurate than 2-band charts. Contrary to our hypothesis, 1-band mirror charts exhibited equal or better speed and accuracy than normal line charts that were twice as tall.

We also hypothesized that estimation error would increase as chart size decreased, and would do so unevenly across chart types. This hypothesis was confirmed. We found that at scale factor 1/4 the error rate was comparable across charts and that 2-band mirror charts provided better accuracy at lower sizes. We found that virtual resolution is a good predictor of error for scale factor 1/4 and below. At the larger sizes, error appears to stabilize at a baseline rate, though more study may be needed to confirm the stabilization at even larger sizes.

Our follow-up experiment investigated chart heights as small as 2 pixels, at which point the information conveyed by position encoding is extremely coarse. Subjects reported relying on color to form estimates at this small size. Note that we rendered the charts using anti-aliasing, so each pixel could still encode a range of values. Thus, our results may characterize the transition from a positional encoding to a color encoding such as those used in pixel-oriented time series visualization techniques [12, 13].

We also found that subjects made estimates faster as chart size decreased. Interestingly, this result appears to depend on the physical chart height rather than virtual resolution. Two subjects verbally reported that they felt they could achieve more accurate results with the larger charts, and so spent more time to get that accurate result. It is possible that subjects form accuracy expectations based on the perceived chart size and allocate time accordingly.

The data also show that in some cases smaller charts led to faster estimation times but equivalent error levels. For all three chart types, scale factor 1/2 (24 pixels) resulted in faster but comparably accurate performance over charts twice as large. As detailed in the next section, this result suggests optimal points for setting a chart's default height, even when screen space is not under contention.

DESIGN IMPLICATIONS

Based on our experimental results, we offer the following design implications for optimizing time series visualizations.

Mirroring Does Not Hamper Graphical Perception

One unexpected result was that mirroring a chart—flipping the negative values around zero—neither slowed estimation time nor hurt estimation accuracy. As mirroring cuts the size of the chart in half without any observed downside, we advocate its use when space constraints warrant, so long as the viewer knows how to interpret the chart.

Layered Bands Are Beneficial As Chart Size Decreases

We found that dividing a chart into layered bands reliably increased estimation time and increased estimation error at constant chart heights. However, we also found that 2-band mirrored charts led to better estimation accuracies for chart heights less than 24 pixels (6.8 mm on our displays). For larger chart sizes, we advise scaling 1-band mirrored charts. For smaller sizes, we advise adding layered bands.

We discourage the use of 4 or more bands, as this resulted in increased time and error, and subjects complained that interpreting 4-band charts was difficult and tiring. The case for 3-band charts is less clear: at a chart height of 48 pixels estimation accuracy was comparable to the 2-band case, but estimation time was slower. Our virtual resolution model predicts benefits for 3-band charts at heights under six pixels, but more research is needed to verify the prediction. As a result, we recommend using 2-band charts for charts heights of 6 pixels (1.7 mm) or more.

Optimal Chart Sizing

Our results show that estimation error stayed stable at larger chart sizes, but that smaller sizes led to faster estimations. Therefore, for each chart type there is at least one size that minimized estimation time while preserving accuracy. For

both normal line charts and 1-band mirror charts, we found a chart height of 24 pixels (6.8 mm on our 14.1" 1024 × 768 pixel displays) to be optimal. For 2-band line charts, we found optima at 12 and 6 pixels (3.4 and 1.7 mm) – performance is about equal at both these sizes. Thus these sizes may be used to optimize graphical perception even when there are no space constraints. However, our subjects were instructed to make estimates accurate within 5 values. Future work is needed to ascertain if similar results occur under different target accuracies.

LIMITATIONS AND FUTURE WORK

One limitation of the present work is that we only measured the results of value comparison tasks. Graphical perception of time series typically involves observing rates of change in addition to comparing values. One reason we focused on value comparison is that graphical perception of rates of change has been studied previously [1, 5] and techniques for determining aspect ratios optimized to aid trend perception already exist [6, 8]. However, it is likely that value estimation is affected by local context within a chart, including line slopes. As we randomized the slope across all comparison points, we believe our results are robust to any contextual effects. Still, future work is needed to determine the nature and extent of any such effects.

Another limitation of our study is that we only varied chart heights and did not investigate the effects of chart width or of distance between comparison points. As time-varying data is encoded along the vertical dimension, we assumed that chart height would be the primary determinant of estimation performance. Furthermore, applying aspect ratio optimization [6, 8] to time series leaves only one free size parameter. Thus, determining an optimal aspect ratio and height will fix the total chart size. However, a large vertical or horizontal distance between points could adversely affect both estimation accuracy and time. We leave studies of the effects of distance between comparison points to future work. We also note that while we varied chart heights, we did not vary physical pixel sizes. Determining whether our results remain valid for higher resolution displays (i.e., smaller pixels) is also left to future work.

In our experiments we discovered that accuracy stabilized at the larger chart heights we investigated. However, we did not determine if those accuracy rates would hold at still larger chart sizes. Furthermore, for larger charts we would also expect additional axis labels, tick marks, and gridlines. We suspect that adding such marks reduces estimation error in larger charts. A potentially fruitful direction for future work is to evaluate if our optimal height results also imply an optimal physical spacing for tick marks and gridlines.

Another open question is where dividing and layering fits within the rank-ordering of visual variables for depicting quantitative data [2, 5, 16]. Virtual resolutions being equal, our results show that a pure position encoding is preferable to layering. More specifically, we found that unlayered charts are faster and at larger sizes more accurate than layered charts. Layered charts were more accurate than 2 pixel tall mirror charts that relied primarily on saturation to encode values. Thus, for encoding quantitative values, layering should be preferred over using a color encoding (c.f., [13]). Future investigation may determine how layering ranks against other visual variables. Our work shows promise for layering, at least for charts that can be layered without suffering from occlusion. For example, bar charts might be layered with similar results. Although other chart types such as scatter plots could also be layered, it is doubtful that such an approach would improve graphical perception.

Finally, while our results provide guidance for optimizing the display of time series data, we stop short of devising a perceptual and cognitive model that more fully explains our observations. Our results could be used to corroborate and extend existing cognitive models of graph comprehension [8, 15, 23]. Future work, including eye-tracking studies, might provide additional insight into both our own results and other issues in graphical perception.

ACKNOWLEDGMENTS

We thank our subjects for their participation and the X-Lab (http://xlab.berkeley.edu) for recruiting assistance. We also thank Stephen Few for drawing our attention to horizon graphs. The second author was funded by an NSERC Postgraduate Scholarship. This research was supported by NSF grant CCF-0643552.

REFERENCES

1. Beattie, V., Jones, M.J. The impact of graph slope on rate of change judgements in corporate reports. *ABACUS*, **38**(2):177-199, 2002.
2. Bertin, J. *Sémiologie Graphique*, Gauthier-Villars: Paris, 1967. English translation by W.J. Berg as *Semiology of Graphics*, University of Wisconsin Press: Madison, WI, 1983.
3. Byron, L., Wattenberg, M. Stacked Graphs — Geometry and Aesthetics. *IEEE Trans. on Visualization and Comp. Graphics*, **14**(6):1245-1252, Nov/Dec 2008.
4. Cleveland, W.S., Diaconis, P., McGill, R. Variables on Scatterplots Look More Highly Correlated When the Scales are Increased. *Science*, **216**(4550):1138-1141, Jun 1982.
5. Cleveland, W.S., McGill, R. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, **79**(387):531-554, Sep 1984.
6. Cleveland, W.S. *Visualizing Information*. Hobart Press. 1993.
7. Few, S. Time on the Horizon. *Visual Business Intelligence Newsletter*, Jun/Jul 2008. Online at http://www.perceptualedge.com/articles/visual_business_intelligence/time_on_the_horizon.pdf

8. Gillan, D.J., Callahan, A.B. A Componential Model of Human Interaction with Graphs: VI. Cognitive Engineering of Pie Graphs. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. **42**(4):566-591. Winter 2000.
9. Heer, J., Agrawala, M. Multi-Scale Banking to 45°. *IEEE Trans. on Visualization and Comp. Graphics*, **12**(5):701-708, Sep/Oct 2006.
10. Heer, J., Viégas, F., Wattenberg, M. Voyagers and Voyeurs: Supporting Asynchronous Collaborative Information Visualization. *Proc. ACM CHI*, pp. 1029-1038, Apr 2007.
11. Heer, J., Robertson, G. Animated Transitions in Statistical Data Graphics. *IEEE Trans. on Visualization and Comp. Graphics*, **13**(6):1240-1247, Nov/Dec 2007.
12. Keim, D.A. Designing Pixel-Oriented Visualization Techniques: Theory and Applications. *IEEE Trans. on Visualization and Comp. Graphics*, **6**(1):59-78, 2000.
13. Lam, H., Munzer, T., Kincaid, R. Overview Use in Multiple Visual Information Resolution Interfaces. *IEEE Trans. on Visualization and Comp. Graphics*, **13**(6):1278-1285, Nov/Dec 2007.
14. Lewandowsky, S., Spence, I. Discriminating Strata in Scatterplots. *Journal of the American Statistical Association*, **84**(407):682-688, Sep 1989.
15. Lohse, J. A. Cognitive Model for the Perception and Understanding of Graphs. *Proc. ACM CHI*, pp. 137-144, Apr/May 1991.
16. Mackinlay, J.D. Automating the Design of Graphical Presentations of Relational Information. *ACM Trans. on Graphics*, **5**(2):110-141, 1986.
17. Mackinlay, J.D., Hanrahan, P., Stolte, C. Show Me: Automatic Presentation for Visual Analysis. *IEEE Trans. on Visualization and Comp. Graphics.* **13**(6):1137-1144, Nov/Dec 2007.
18. Palmer, S. *Vision Science: Photons to Phenomenology*. MIT Press, 1999.
19. Robertson, G., Fernandez, R., Fisher, D., Lee, B., Stasko, J. Effectiveness of Animation in Trend Visualization. *IEEE Trans. on Visualization and Comp. Graphics*, **14**(6):1325-1332, Nov/Dec 2008.
20. Rosling, H. TED 2006, http://gapminder.org/video/talks/ted-2006-debunking-myth-about-the-third-world.html
21. Saito, T., Miyamura H.N., Yamamoto, M., Saito,H., Hoshiya, Y., Kaseda, T. Two-Tone Pseudo-Coloring: Compact Visualization for One-Dimensional Data. *Proc. IEEE InfoVis*, pp. 173-180, Oct 2005.
22. Shneiderman, B. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. on Graphics*, **11**(1):92-99, 1992.
23. Simkin, D., Hastie, R. An Information-Processing Analysis of Graph Perception. *Journal of the American Statistical Association*, **82**(398):454-465, Jun 1987.
24. Spence, I., Lewandowsky, S. Displaying proportions and percentages. *Applied Cognitive Psychology*, **5**:61-77, 1991.
25. Stasko, J., Zhang, E. Focus+Context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy Visualizations. *Proc. IEEE InfoVis*, pp. 57-65, 2000.
26. Tremmel, L. The Visual Separability of Plotting Symbols in Scatterplots. *Journal of Computational and Graphical Statistics*, **4**(2):101-112, Jun 1995.
27. Tufte, E. *The Visual Display of Quantitative Information*. Graphics Press, 1983.
28. Tufte, E. *Beautiful Evidence*. Graphics Press, 2006.
29. Wattenberg, M., Kriss, J. Designing for Social Data Analysis. *IEEE Trans. on Visualization and Comp. Graphics*. **12**(4):549-557, Jul/Aug 2005.
30. Wigdor, D., Shen, C., Forlines, C., Balakrishnan, R. Perception of Elementary Graphical Elements in Tabletop and Multi-Surface Environments. *Proc. ACM CHI*, pp. 473-482, Apr 2007.
31. Woodruff, A., Landay, J., Stonebraker, M. Constant Information Density Visualizations of Non-Uniform Distributions of Data. *Proc. ACM UIST*, pp. 19-28, 1998.

APPENDIX A: CHART GENERATION

In each trial of our experiment the subject had to estimate the magnitude difference between the y-coordinate of two query points, T and B (Figure 3). Given a signed offset distance *d* between the query points as input, we generated a pair of charts as follows. First, we randomly chose the y-value for T and added the offset *d* to it to set the y-value for B. The x-coordinates for T and B were set a priori and fixed for all charts in the experiment. Once the query point T or B was set, we used a random walk, with a step size of +/- 1 in x and y, to fill in the remaining values in the chart. To ensure that the chart was band-limited we then convolved the chart with a 5-tap triangle filter with parameters [0.11 0.22 0.33 0.22 0.11]. Because the smoothing process could shift the position of the query point, we translated the y-values in a neighborhood of 20 points about the query points to maintain the necessary offset distance between T and B. To further smooth the chart we repeated the convolution and translation process but using a symmetric 3-tap triangle filter with parameters [0.25 0.5 0.25]. The key features of our approach are that the offset distance between the query points were fixed, the charts appeared randomly different from trial to trial, and the charts did not contain high-frequencies because of the smoothing.