Measures of the Benefit of Direct Encoding of Data Deltas for Data Pair Relation Perception

Christine Nothelfer and Steven Franconeri

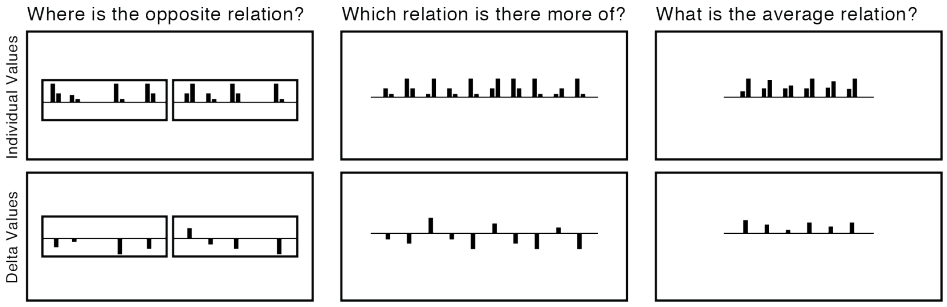


Fig. 1. Summary of tasks: participants are far more efficient in searching for a target relation (left), discriminating proportions of relations (center), and estimating the average relational magnitude (right) when values are encoded as deltas (bottom row) than when encoded as individual values (top row).

Abstract—The power of data visualization is not to convey absolute values of individual data points, but to allow the exploration of relations (increases or decreases in a data value) among them. One approach to highlighting these relations is to explicitly encode the numeric differences (deltas) between data values. Because this approach removes the context of the individual data values, it is important to measure *how much* of a performance improvement it actually offers, especially across differences in encodings and tasks, to ensure that it is worth adding to a visualization design. Across 3 different tasks, we measured the increase in visual processing efficiency for judging the relations between pairs of data values, from when only the values were shown, to when the deltas between the values were explicitly encoded, across position and length visual feature encodings (and slope encodings in Experiments 1 & 2). In Experiment 1, the participant's task was to locate a pair of data values with a given relation (e.g., Find the 'small bar to the left of a tall bar' pair) among pairs of the opposite relation, and we measured processing efficiency from the increase in response times as the number of pairs increased. In Experiment 2, the task was to judge which of two relation types was more prevalent in a briefly presented display of 10 data pairs (e.g., Are there more 'small bar to the left of a tall bar' pairs or more 'tall bar to the left of a small bar' pairs?). In the final experiment, the task was to estimate the average delta within a briefly presented display of 6 data pairs (e.g., What is the average bar height difference across all 'small bar to the left of a tall bar' pairs?). Across all three experiments, visual processing of relations between data value pairs was significantly better when directly encoded as deltas rather than implicitly between individual data points, and varied substantially depending on the task (improvement ranged from 25% to 95%). Considering the ubiquity of bar charts and dot plots, relation perception for individual data values is highly inefficient, and confirms the need for alternative designs that provide not only absolute values, but also direct encoding of critical relationships between those values.

Index Terms—Information visualization, marks, perception, attention, visual comparison, visual search, aggregation

Introduction

In their seminal work investigating graphical perception across various data encodings, Cleveland & McGill [3] stated that "the power of a graph is its ability to enable one to take in the quantitative information, organize it, and see *patterns and structure* not readily revealed by other means of studying the data." Data visualization design must consider not only which encodings provide an accurate percept of individual values, but also which encodings allow the human visual system to efficiently perceive and explore the relations and patterns among those values. Consider a bar graph depicting student test scores before and after an intervention program (Figure 2A). The absolute values of the test scores are not the critical information – it is the increase or decrease (i.e., ***relations***) in scores that supports the efficacy of the program. Figure 2B illustrates the helpful design technique of explicitly encoding differences (***deltas***) between those data values. Here we quantify this advantage to understand *how much* value this approach provides, and assess that advantage across 3 different comparison tasks to understand *when* it is most advantageous.

- *Christine Nothelfer is with Northwestern University. E-Mail: cnothelfer@gmail.com.*
- *Steven Franconeri is with Northwestern University. E-Mail: franconeri@northwestern.edu.*

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

1 Background and Related Work

There is surprisingly little empirical work exploring the most *efficient* way to visually forage through relations (among data values). Existing work tends to focus on how *precisely* a provided pair of data values can be compared. Cleveland & McGill [3,4] evaluated the precision of comparisons (e.g., What is the ratio of value X1 as a percentage of value X2?) between two provided values, or two values highlighted among others, across several encoding types. Those experiments resulted in a ranking of encodings according to precision for that particular task, e.g., position and length encodings afford more precision for ratio extractions, compared to line slopes or figure areas. Mackinlay [18] hypothesized rankings of visual feature encodings for

relations across more diverse conditions (e.g., whether data are quantitative, ordinal, or nominal), though those rankings were not empirically evaluated. While past work focuses on tasks that measure precision between two provided values, we argue that it is *at least* equally important to test general perception of data patterns (i.e., how we perceive spatial relationships between objects in order to judge magnitude relations, such as increases and decreases between values), as well as use tasks that measure the efficiency with which an observer can forage through large sets of such comparisons, rather than only one comparison at a time [3].

Others have examined how different visual designs impact graphical perception while visually comparing data values [12, 13, 19, 21]. For example, one study [21] examined how different bar chart designs (including stacked bar charts and simple bar charts) impact accuracy in comparing bar heights, including the negative effect of bar spacing. However, none of these studies address how visual designs impact foraging across large sets of data value comparisons. Such work would encourage the design and evaluation of alternative formats for efficient relation perception within data visualization.

The data visualization design technique of directly encoding deltas (Figure 2B) is intended to improve the efficiency of viewing and understanding the relations between data values. According to Gleicher et al.'s [7] taxonomy for categorizing visual displays for visual comparison, this design technique is considered an *explicit encoding*. However, this design technique does not come without costs – it removes the context of the original individual data values, which can be important for a variety of visual decisions. While removing the context of the original data points is an advantage if one only cares about the differences, it is a disadvantage if one needs to connect deltas to the original values (e.g., 'What is the average delta for only values past a certain threshold?'). Gleicher et al. additionally mention that insights tied to the individual data points, beyond the delta relationships, will not be visible and thus less likely to be discovered by the viewer. This design technique also takes up precious real estate, such as in data visualization dashboards – with each new data series added for potential comparison, there is a combinatorial explosion of the number of possible delta encodings among the possible pairings.

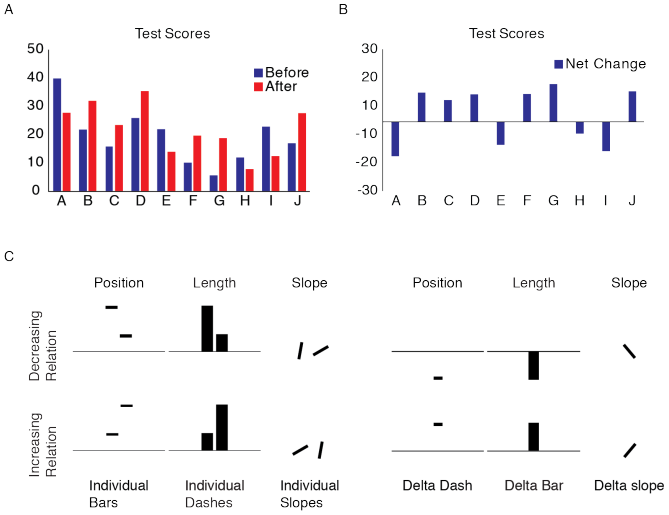


Fig. 2. (A-B) Two ways of encoding the same data set. How many students performed better on the second (red values) test? Is it easier to judge improvements when the data is encoded by bars depicting individual data points (2A) or differences (deltas) between data pairs (2B). (C) Encoding Types. Six encodings were used as the stimuli in Experiments 1, 2, and 3, though absolute values varied. Encodings on the left half graphically represents two data values (*individual value encodings*), while encodings on the right half represent the delta between those two values (*delta value encodings*). Each column shows values encoded by one of three visual feature encodings (position, length, or slope). Encodings on the bottom row (*increasing* relations) depict the opposite relation as that on the top row (*decreasing* relations).

If one is comparing data values to understand their relations, it seems expected that explicitly encoding these differences as deltas should improve relation perception. However, *how much better is relation perception when using deltas versus individual data values? Is this improvement large enough given the costs of displaying only the deltas? For which tasks is this design technique most useful?*

A recent study [19] explored a similar idea, by showing participants displays that included grouped bar charts (several bar pairs), grouped bar charts with delta overlays encoded as dashes, and only deltas encoded as dashes. They found that the latter two chart types outperformed the grouped bar charts, for response time and accuracy, when participants identified which category has the largest absolute change, but were better than grouped bar charts (for response time only) when participants identified the absolute change for a particular category, suggesting the design technique's benefit may vary by task. However, only 2 of the 6 visual comparison tasks required judging aspects of the *differences* between data value pairs, because the remaining 4 tasks required the participant to focus only one of the two data series (e.g., "Click on the month with the minimum value for 2016"). We build on this work by testing delta value encodings across 3 additional tasks, to further assess when this design technique is more useful. To preview our results, we find a wide range in improvement due to delta encodings (i.e., 25-95% better performance).

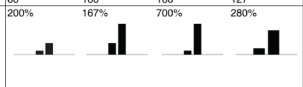
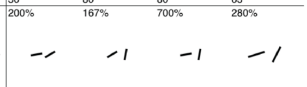
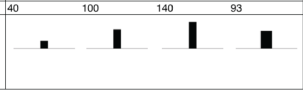
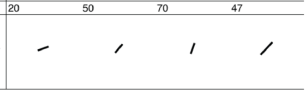
Beyond using 3 new tasks, we also build on Srinivasan et al. [19] by exploring a different type of relational judgment. While Srinivasan et al. focused on the magnitude of individual deltas (i.e., What's the largest difference? What's the difference size for this data pair?), we add a task requiring judgment of relational direction across multiple data pairs (Exp. 1: Where is the only increase?; Exp. 2: Do more pairs increase or decrease?). This is an important task because identifying the relation direction is the first step to many tasks. For example, if one is trying to estimate the greatest increase magnitude difference, one must first identify which data pairs show an increase in value. We also extend Srinivasan et al.'s tasks by requiring the viewer to summarize across *multiple* deltas across multiple value pairs (Exp. 3: What's the average delta?). Together, these are important patterns in datasets: an analyst might want to find a month with the biggest sales increase relative to last year (a task similar to that of Srinivasan et al. [19]), but also might want to identify which months dropped, whether there were months with a drop or increase, or what the average drop is across months, as our work tests.

1.1 Visual Search

One well-studied measure for processing efficiency is the visual search task, which operationalizes processing efficiency as the added time needed to find a target item among each additionally introduced distractor item (i.e., search rate) [11, 22, 24], assuming constant accuracy (which is typically required to be near ceiling). Finding targets that are unique in the typical data visualization encoding dimensions (e.g. red vs. blue color, long vs. short length), can produce search rates as efficient as 0 ms/item. But one of the robustly hardest visual search tasks is for relations. One study asked participants to search for a specific relation (e.g., dash above a plus) among distractors with the opposite relation (e.g., plus above a dash), and found that response times strongly increased with the number of relations within a display [17]. The size of this increase suggests that perceiving the relation between two items is a serial – or close to serial – process. Similarly, another study found that search for T's among L's (different spatial relationships of the same two line segments) yielded steep search rates [14]. In fact, across a number of visual search studies, search rates are far higher for spatial configurations of items than for simple items [23]. This serial processing may be due to the need to isolate each item within a relation individually with the attentional 'spotlight', in order to extract its location, independent of the other item [6]. Based on these results, one can imagine how poorly this process would scale to large data sets with complex patterns.

Importantly, to our knowledge, all of these studies ask viewers to find relations between qualitative visual identities (object X and object Y), and *no existing work has tested the efficiency of search for*

Table 1. Stimuli Relation Values. The 6 encodings depicted values from sets A-C in Experiments 1 and 2. Encodings representing individual data values depicted one low value and one high value within a value set, while encodings representing deltas depicted the difference between low and high values. Position and length encodings (left half) were scaled to fit the display monitor prior to being converted to pixels. While these absolute values are arbitrary given that the stimuli did not include axes, they are provided here for reproducibility.

	Values for Position and Length Encodings				Values for Slope Encodings			
	Set A	Set B	Set C	Preview	Set A	Set B	Set C	Preview
Low Value	20	60	20	33	10	30	10	17
High Value	60	160	160	127	30	80	80	63
% Difference	200%	167%	700%	280%	200%	167%	700%	280%
Example								
Delta Value	40	100	140	93	20	50	70	47
Example								

quantitative relations (a tall object left of a short object, among other objects that are mirror-reversals of that pattern). These two possible configurations differ in the shape of the envelope that surrounds them both (when you 'squint', small-large pairings create a triangle with a 'forward-slash' diagonal line at the top; large-small pairings create the mirror-reversal of that shape). The visual system excels at using such global shape-perception heuristics to avoid relying on relations between segmented objects, which is a far more computationally intensive process [6, 15]. And a simpler version of this shape contrast – searching for a forward-slash among backslashes – produces perfectly efficient search [26].

Given that visual search is highly efficient for simple visual feature encodings and inefficient for relations, it is likely that visual search for a specific data relation (e.g., Name a student whose test scores did not improve in Figure 2A) would be significantly more efficient when the *differences* among data value pairs are directly encoded with single visual feature encodings (Figure 2B) than when the same visual feature encoding is used to encode each individual data point (Figure 2A). Yet there are also reasons to believe that it may not offer a strong advantage, and to our knowledge this advantage has never been formally empirically quantified.

1.2 Ensemble Coding

The task of detecting the presence of a single anomalous relation between value pairs is a relatively simplistic case study for evaluating processing efficiency for relations in real-world displays. More realistic tasks likely require the collection of information from broader sets of objects at once. Another well-studied measure for this type of processing efficiency is an 'ensemble' snapshot of the visual statistics, such as a mean of the locations, orientations, or luminances of a set of objects [9]. Such summary measures may subserve many common data visualization tasks, such as distributional information about encoded values, pattern and motif recognition, or segmentation of values according to their place in the distribution [20], positioning ensemble coding as a critical tool in understanding data.

We might predict that any same processing bottleneck that we uncover in Experiment 1's search task could also constrain our ability to extract such 'ensemble' statistics. Given that observers can extract the mean position of glyphs in a scatterplot [8], mean size of a set of circles [1], and the mean orientation of a set of angled line segments [5], Experiments 2 & 3 will quantify how strongly relation perception within similar ensemble tasks (e.g., Did students generally improve after the intervention in Figure 2A? What is the average improvement?) benefits when the differences among data value pairs are directly encoded with simple visual feature encodings (Figure 2B).

1.3 The Current Study

In summary, there is limited research investigating how we perceive *relations* across a large set of objects of the type typically used in a data visualization display, and no existing research investigating how we perceive *relational direction* in these displays, or aggregate over multiple deltas across many value pairs. The design technique of explicitly displaying deltas is a promising solution, but requires empirical testing to evaluate *how strongly* and *when* it is effective. The overall aim of this study is to investigate how we can best perceive relations from multiple **data pairs** (i.e., 2 data values – a 'high' value and a 'low' value). We evaluate whether data should be depicted as individual data values – leaving the visual system to extract relations between object pairs – or as single visual feature encodings representing deltas for efficient relation perception by examining the magnitude of improvement when using delta value encodings across multiple comparison tasks.

We tested two formats for visually depicting simple magnitude relations (smaller/larger vs. larger/smaller) – either displaying individual data points or directly encoding the difference (delta) between pairs of data points – with three basic visual feature encodings (position, length, and slope). Participants searched for a particular relation in a display (e.g., Find a 'tall bar to the right of a short bar' among 'short bar to the right of a tall bar' pairs) in Experiment 1, discriminated proportions of relations (e.g., Which relation type was more prevalent, short bar to the left of a tall bar, or vice-versa?) in Experiment 2, and determined the average delta among relations within a display (e.g., What is the average bar-height difference across bar pairs?) in Experiment 3.

These three tasks provide a cross section of common visual comparisons – we include 2 tasks where only relation direction is relevant (Exp. 1 and Exp. 2), rather than absolute values (Exp. 3), and one task where the goal is to single out one particular data pair (Exp. 1) rather than summarize the entire set (Exp. 2 and Exp. 3). All three experiments revealed a powerful improvement in relational processing efficiency for direct encodings of the difference between data pairs. This delta advantage is most pronounced in our task involving locating a single particular relation (Exp. 1), and less pronounced in our tasks that involve ensemble coding (Exp. 2 and 3).

Contributions: This study contributes design guidelines built on empirical findings, new tasks that simulate realistic data visualization comparison tasks, and perceptual-psychology-inspired experimental methods to quantify performance differences. While, in all cases, we expect direct encoding of deltas to perform better compared to relational processing of absolute values, when we ask 'How do we know that?' (what empirical data actually underlie that gut response? [16]), the existing literature relies on importantly *different* stimuli and tasks.

Findings: Our results demonstrate that in most cases, it is staggeringly inefficient to perceive relations among data value pairs, but highly dependent on the viewer's task: the delta encoding advantage ranged widely from 25% (as in Exp. 3) to 95% (as in Exp. 1). We were also surprised to see that orientation encodings showed only moderate advantages (20-50% improvement) for direct delta encodings, due perhaps to surprisingly poor performance overall.

Tasks: This study demonstrates 3 distinct visual comparison tasks. These tasks expand upon the ratio comparison task typically used [3, 4], and the two ensemble coding tasks (Exp. 2 and 3) provide an approach for assessing how viewers perceive large sets of data points.

Methods: We provide perceptual psychology methods (psychophysics) that will allow researchers to assess the quantitative effect of a visual design and measure the advantage. Critically, this approach enables researchers to compare data visualization design techniques across different tasks and dependent measures. For example, the delta value encoding advantage in this study clearly depends on the viewer's task, which highlights a need for assessing a new chart type or visual feature encoding across a variety of visual tasks.

2 Encoding Types

Experiments 1 and 2 tested visual processing efficiency of relations across 6 encodings (see Figure 2C), while Experiment 3 used only 4

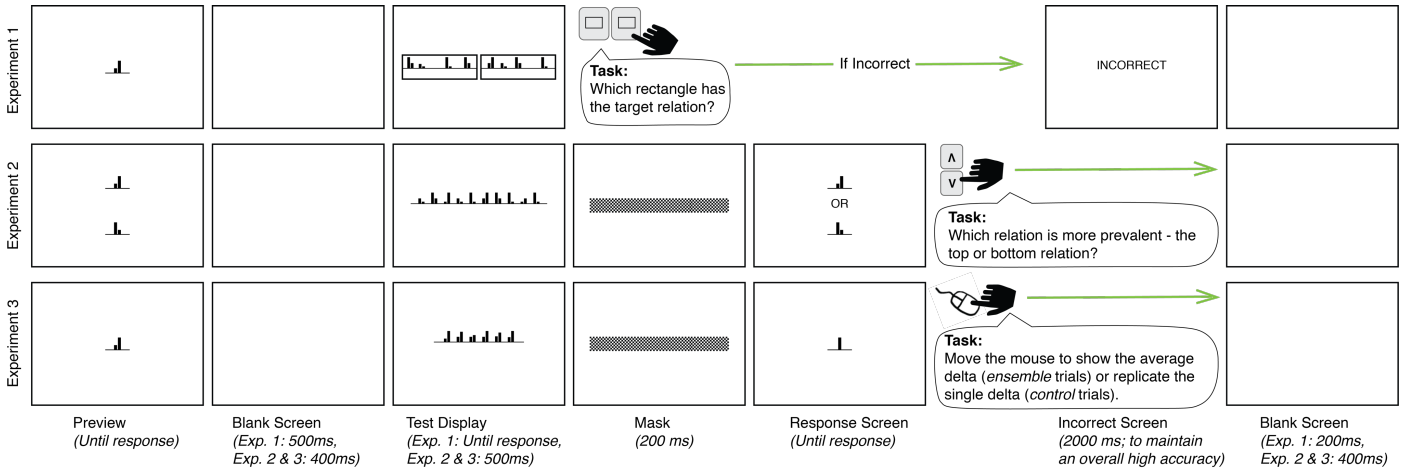


Fig. 3. Experiments 1, 2, and 3's designs. Stimuli shown here are not drawn to scale. Participants viewed a *preview* screen containing a target relation(s) to search for (Experiment 1) or judge (Experiments 2 and 3), followed by a *blank* screen, and then a *test* display. The *test* display remained until participants indicated which rectangle contained the target relation in Experiment 1, followed by an *incorrect* screen if they responded incorrectly. In Experiments 2 and 3, the *test* display was presented briefly, followed by a *mask* screen to prevent an after-image. In the final *response* screen, participants indicated which relation was more prevalent (Experiment 2) or moved the mouse to show the perceived average delta (Experiment 3). All trials concluded with a *blank* screen. Relations were located randomly, but always aligned to a bottom baseline.

encodings (*position* and *length* encodings). All experiments used the encodings to depict a 'high' value and a 'low' value of a data pair, though absolute values varied.

Individual Value Encodings: The **position** (i.e., distance) from baseline (*individual dashes*), the bar **length** (i.e., height) (*individual bars*), and **slope** (i.e., orientation in degrees, tilting upward from a horizontal angle (0 degrees)) (*individual slopes*) represent each value in the data pair. With the *individual slopes* encoding, all data values are scaled to be less than 90 so that lines are always positively sloped between 0 and 90 degrees.

Delta Value Encodings: The **position** (i.e., distance) from baseline (*delta dash*), the bar **length** (i.e., height) (*delta bar*), and **slope** (i.e., orientation in degrees) (*delta slope*) represent the difference value of the data pair. The *delta dash* and *delta bar* are above or below the baseline when representing an *increasing* relation (positive delta) or *decreasing* relation (negative delta), respectively. The *delta slope* tilts upward or downward from a horizontal angle (0 degrees) when representing an *increasing* relation (positive delta) or *decreasing* relation (negative delta), respectively. All data values (prior to calculating delta value) are scaled to be less than 90 for the *delta slope* encoding so that lines are always positively sloped when representing positive delta values and negatively sloped when representing negative delta values.

All stimuli were intended to simulate typical graph types: dot plots (horizontal dashes), bar charts, and slope graphs, respectively. Some of these encodings differ slightly from real world use cases (e.g., Does the *individual slopes* encoding truly simulate a slope graph when the lines are arranged horizontally rather than superimposed?), but this difference was intentional, in order to explore a more controlled space of visual feature encodings. For example, comparing the efficiency of *individual slopes* to *individual bars* reveals how the visual system extracts relational information from the visual feature encodings of slope versus length. Likewise, comparing the efficiency of *individual dashes* to *individual bars* allows exploration of the importance of the 'tops' of the bar values, compared to the area represented underneath.

Note that we refer to *individual bars* and *delta bars* encodings as visual representations in which values are encoded by length (height of the bars) for simplicity, but that length, position (distance from baseline to opposite end of the bar), and area (the filled region of the rectangle) all redundantly encode each value.

3 General Methods for Experiments 1-3

Thirty-nine university students participated in this study (13 participants per experiment, a typical sample size in perceptual psychology studies) in exchange for course credit or payment. All experiments lasted a total of 30-45 minutes, including the informed consent process, understanding experiment instructions, doing practice trials, and post-experiment debriefing.

Conducting this study in a lab environment allows a study facilitator to monitor participant attentiveness during the testing session to ensure high data quality (e.g., that participants are not multitasking while completing an experiment that measures split-second response times). Our set of experiments is a long duration within-subject design requiring participants to remain attentive for the full duration of the testing session, making speeded judgments in Experiment 1, and viewing rapidly-disappearing single-glance displays in Experiments 2 and 3.

Experiment 1-3's trials followed the same general procedure, summarized in Figure 3 (see *Supplemental Material* for more details). Because the stimuli in all three experiments do not display axes or data group labels, participants must rely on the shapes of the data encodings alone to provide their responses. That is, in Experiment 1 participants indicate which rectangle contains the opposite relation instead of the data group label associated with that pair of data points, in Experiment 2 participants indicate which relation (increasing or decreasing) they see more of instead of whether one data group is overall doing better or worse than the other data group, and in Experiment 3 participants adjust the height of a rectangle or dash to indicate the average delta value rather than stating this value numerically. Providing axes and data group labels could introduce a new source of error (e.g., participants not locating a data group label quickly enough, mixing up the two data group labels, or the degree to which one can translate visual shapes to numeric values), which would obfuscate the effects of our experimental manipulations alone. The goal of this study is to understand how our perception of data values is impacted by different encoding types, which is the first step before finding the corresponding numeric value or data group label.

4 Experiment 1: Visual Search for Relations

The goal of this experiment was to evaluate to what degree using simple visual feature encodings representing deltas (as opposed to encodings representing individual data values) can lead to efficient visual processing of the relations between the data values when an observer searches for a single known relation. A realistic example of this is identifying which states experienced an increase in healthcare enrollment.

We measured visual processing efficiency from the increase in response times as the number of pairs increased in the display for each

encoding. Response times that are slower with larger set sizes indicate more serial processing, and this metric can reflect relative processing efficiency across the six tested encodings.

4.1 Methods and Procedure

Stimuli

The stimuli in this experiment were 4 sets of data pairs (see *Sets A*, *B*, *C*, and *Preview* in Table 1; each data pair contains one *low value* and one *high value*). Data pair values were depicted by each of the 6 encodings (see *Encoding Types*) by either converting values to pixels (position and length values; left half of Table 1) or degrees (slope values; right half of Table 1) to create image files (see *Supplemental Material* for more details).

Half of the encoding types depicted individual data points (i.e., displayed both the *low* and *high value* within each data pair; left half in Figure 2C), while the other half depicted delta values (i.e., the *difference* between each data pair's values; right half in Figure 2C). Three of these data pair sets (*Set A*, *Set B*, and *Set C*) were used in the *test* displays, while the fourth set (*Preview*) was only used in the *preview* displays. There were 2 possible relations for each data pair (bottom row in Figure 2C depicts the *increasing* relation, while the top row depicts the *decreasing* relation), yielding a total of 48 unique stimuli. All image files were then scaled down by a factor of 0.62 to fit within the display, with each encoding spanning 1.05-2.39 visual degrees wide and 0.27-4.24 visual degrees tall.

There were two primary screen types:

Preview Screens*:* These screens featured the target relation for the given trial.

Test Screens: In the *test* displays, data pairs were arranged within two side-by-side black rectangle outlines arranged horizontally across a white screen (Figure 3). The rectangles contained 1, 2, 4, or 5 data pairs each (always the same number of data pairs per rectangle for any given trial), for a total set size of 2, 4, 8 or 10 data pairs. Each display contained one data pair of a target relation (e.g., small bar to the left of a tall bar), and the remaining (distractor) data pairs were of the opposite relation (e.g., tall bar to the left of a small bar). The target relation (assigned to a specific rectangle for each trial) and distractor relations locations were randomized within each rectangle along 5 evenly-spaced possible positions within each rectangle. Each data pair's absolute values were randomly selected from the three possible sets (*Sets A*, *B*, and *C* in Table 1).

Procedure

Task: Participants were asked to quickly indicate which rectangle contained the target relation.

Trial Procedure: The general procedure is described in Section 3 above. The *test* display remained on the screen until the participant quickly pressed the left or right rectangle key (the 'b' or 'n' key, respectively, each covered with a sticker showing a small rectangle to represent the corresponding screen rectangle – e.g., the left rectangle ('b') key represented the left rectangle on the screen) to indicate their response.

Design: Factors in the full factorial design included: 2 **data depictions** (individual value encodings, delta value encodings) x 3 **visual feature encodings** (position, length, slope) x 4 **set sizes** (2, 4, 8 or 10 data pairs in the search display) x 2 **target relations** (increasing (Figure 2C, bottom row), decreasing (Figure 2C, top row)) x 2 **target locations** (the target relation appears in the left or right rectangle) x 4 **repetitions** – yielding a total of 384 test trials. While participants were provided with feedback via an Incorrect Screen when responding incorrect, they were given only one attempt per trial.

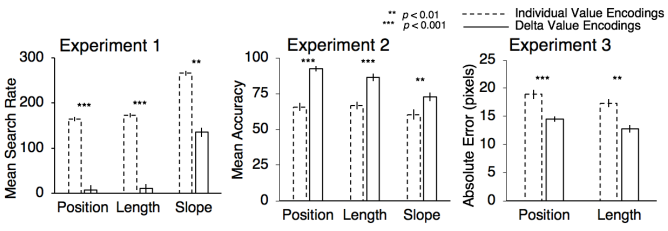


Fig. 4. Summary of Expriments 1-3 Results. Delta value encodings lead to faster search rates (left) when searching for an opposing relation, higher accuracies (center) for distinguishing which relation there is more of, and lower error (right) when perceiving the average delta value in the ensemble task. Error bars indicate within-subject standard error of the mean.

Table 2. Search Rates for Experiment 1. Descriptive statistics, ANOVA results, and follow-up t-test results shown for significant ($p < 0.05$) comparisons. Search rates represent visual processing efficiency, in that they summarize how response times improve or degrade with an increase in set size; the greater the search rate, the worse the impact of increasing the set size, and therefore the worse the visual processing efficiency.

Factor	*M*	*SE*	Statistical Test
Data Depiction			$F(1,12) = 72.98$, $p < 0.001$, $\eta p^2 = 0.86$
Delta Value Encodings	0.05 s	0.00 s	$t(12) = -8.72$, $p < 0.001$, $d = -2.42$
Individual Value Encodings	0.20 s	0.02 s	
Visual Feature Encoding			$F(1.20,14.36) = 23.78$, $p < 0.001$, $\eta p^2 = 0.66$
Slope Encodings	0.20 s	0.02 s	$t(12) = -4.95$, $p < 0.001$, $d = -1.37$
Length Encodings	0.09 s	0.01 s	
Position Encodings	0.09 s	0.01 s	$t(12) = -5.36$, $p < 0.001$, $d = -1.49$

Trial Order: Participants first completed 10 practice trials. Trials were randomly ordered within each of 4 test blocks, one block for each repetition.

4.2 Results and Discussion

Median response times to correct trials were calculated after grouping response times across trial repetitions. We ran a factorial, repeated measures ANOVA on the factors data depiction, visual feature encoding, set size, target relation, and target location on median response times. Degrees of freedom were Greenhouse-Geisser corrected for sphericity violations. Significant effects were followed up by two-tailed paired t-tests. Participants maintained a high number of trials (88% or greater) after incorrect and slow trial removal.

Search rates were also calculated to quantify visual processing efficiency, which indicates how much time is spent during visual search with each addition of another data pair. Highly efficient visual processing should not be affected too much by an increase in set size. Therefore, the greater the search rate, the worse the impact of increasing the set size, and the worse the visual processing efficiency. The search rate calculation allows for predictions of how performance differences should scale for data displays containing fewer or more relations. The median response times to the smallest set size (2) were subtracted from median response times to the largest set size (10), and divided by the difference in set sizes (8) to obtain search rates. Because search rates are much more indicative of visual processing efficiency, we ran a factorial, repeated measures ANOVA on the factors data depiction, visual feature encoding, target relation, and target location on search rates. Table 2 and Figure 4 (left) shows search rate results for Experiment 1.

Only the results for our primary experimental factors are listed here (see *Supplemental Material* for other analyses, including analyses on response times, and an analysis confirming no speed/accuracy trade-off).

Data Depiction: If the visual depiction of representing individual data values or their deltas impacts visual processing efficiency, then search rates for each encoding should be different depending on the data depiction. Indeed, search rates were impacted by data depiction, $F(1,12) = 72.98$, $p < 0.001$, $\eta p^2 = 0.86$. That is, **search rates were significantly slower for *individual value encoding* trials than for *delta value encoding* trials.** This means that *response times get slower with the addition of each data pair* when encoded by individual values in a data display, and at a much *slower rate* than when encoded by delta values.

Visual Feature Encoding: The particular visual feature encoding for the data pairs, regardless of depiction type (i.e., encoding individual values or delta values) impacted search rates, $F(1.20,14.36) = 23.78$, $p < 0.001$, $\eta p^2 = 0.66$. **Search rates were significantly slower for *slope* encodings than both *length* encodings and *position***

encodings. Search rates for *length* trials and *position* trials were not statistically different. This result is mostly consistent with the visual feature encoding ranking proposed by [3], which outlines that the visual precision for comparing two values is better when those values are encoded by position (dots in a dot plot) and by length (unaligned subsets within a stacked bar graph), than when encoded by slope (angled line segments in a line chart). However, this work also shows that values encoded by position is better than values encoded by length, which we failed to find here.

These results indicate that while response times are overall moderately worse for *length* trials than *position* trials, performance degrades similarly with the addition of data value pairs. On the other hand, performance is the worst for *slope* encodings.

Set Size: As expected, response times were slower as set size increased, $F(3,33) = 82.80$, $p < 0.001$, $\eta p^2 = 0.88$.

In sum, the data encoding greatly impacted the visual processing efficiency of relations between data values: relations between data pairs are much more efficient to visually extract when directly represented as deltas rather than individual data points. Consistent with prior work [3], the particular visual feature encodings of the values impact relation processing as well. What is most striking about the present results is that **plotting individual data values is staggeringly inefficient – *two to sixteen times worse* than delta value encodings** (each additional data pair adds 164-266 ms when represented as individual values, but only 8-135 ms when represented as delta values, resulting in a **massive 49-95% improvement**) – despite that these are perhaps the most ubiquitous encodings for data values.

5 Experiment 2: Ensemble Coding of Proportions

Experiment 1 revealed that the choice of data depiction led to enormous differences in processing efficiency for relations between values for the visual search of a known target – participants were significantly faster to search for a particular relation when its delta was encoded rather than its individual values. Experiment 2 explored whether this result generalizes to other tasks that require a viewer to compute visual relational information across a broader set of value pairs. The goal of this experiment was to emulate situations where an observer judges the proportions of relations. A realistic example of this is judging which gender is earning a higher salary the most often across a range of job titles.

Table 3. Accuracies for Experiment 2. Descriptive statistics, ANOVA results, and follow-up t-test results shown for significant ($p < 0.05$) comparisons for accuracies, except where otherwse noted (i.e., data depiction differences). Data depiction differences were calculated to further expolore how accurcies for each visual feature encoding and each target relation interacted with the two data depictions. The larger the data depiction difference, the greater the advantage offered by delta value encodings.

Factor	*M*	*SE*	Statistical Test
Data Depiction			$F(1,12) = 116.82$, $p < 0.001$, $\eta p^2 = 0.91$
Delta Value Encodings	84%	2%	$t(12) = -10.81$, $p < 0.001$, $d = -3.00$
Individual Value Encodings	64%	2%	
Visual Feature Encoding			$F(2,24) = 28.09$, $p < 0.001$, $\eta p^2 = 0.70$
Slope Encodings	67%	2%	$t(12) = 4.70$, $p = 0.001$, $d = 1.30$
Length Encodings	77%	3%	
Position Encodings	79%	2%	$t(12) = 7.00$, $p < 0.001$, $d = 1.94$
Data Depiction x Visual Feature Encoding			$F(2,24) = 17.38$, $p < 0.001$, $\eta p^2 = 0.59$
Visual Feature Encoding (*data depiction differences**)			
* *individual value encoding* accuracy - *delta value encoding* accuracy			
Slope Encodings	12%	3%	$t(12) = 2.37$, $p = 0.036$, $d = 0.66$
Length Encodings	20%	2%	$t(12) = -3.51$, $p = 0.004$, $d = -0.97$
Position Encodings	27%	2%	$t(12) = 7.04$, $p < 0.001$, $d = 1.95$
Difficulty Ratio			$F(2,24) = 43.80$, $p < 0.001$, $\eta p^2 = 0.78$
1:9	83%	3%	
3:7	73%	2%	
4:6	66%	2%	
Target Relation x Data Depiction			$F(1,12) = 5.87$, $p = 0.032$, $\eta p^2 = 0.33$
Target Relation (*data depiction differences**)			
Decreasing Relation Target	24%	2%	$t(12) = 2.42$, $p = 0.032$, $d = 0.67$
Increasing Relation Target	16%	2%	

We manipulated difficulty by adjusting the ratio of relations – it should be easier to perceive proportions of relations the more lopsided the ratio (i.e., 1:9), while it should be more difficult the more even the ratio (i.e., 4:6). If visual processing is more efficient when relations are represented by delta value encodings than by individual data value encodings, then performance on trials with individual data value encodings will drop as ratio difficulty increases (e.g., accuracy is worse with a display containing a 4:6 relations ratio than a 1:9 relations ratio), or at least more so than with delta value encodings. It is also possible that accuracy for individual data value encodings may instead (or additionally) be overall worse than accuracy for delta value encodings.

5.1 Materials and Procedure

Stimuli

The stimuli in this experiment used the same 6 encodings (see *Encoding Types*) depicting the same data pair values as Experiment 1. As before, data pair values were either converted to pixels (length and position values; left half of Table 1) or degrees (orientation values; right half of Table 1) to create image files. All image files were then scaled down by a factor of 0.5 to fit within display; this resulted in length/position values that were the exact same numerically as the orientation values, though they were the same relative values as in Experiment 1.

There were 3 primary screen types:

Preview Screens: These screens featured the two possible relations for the given trial, one above and one below the screen's center, showing the type of relation participants were to judge in the test display.

Test Screens: Each *test* display always contained 10 data pairs, comprised of both possible relations (e.g., 3 positively sloped lines and 7 negatively sloped lines) of a single encoding (see Figure 3). The specific amount of each relation was one of three possible difficulty ratios (1:9 (easiest ratio), 3:7 (medium difficulty), or 4:6 (most difficult ratio). All data pairs were randomized across 10 evenly-spaced positions, all aligned vertically to a bottom baseline at the center of a white screen.

Response Screens: These screens were identical to *preview* screens, except "OR" was written in the center of the screen between the two possible answers.

Procedure

Task: Participants were asked to quickly indicate which of two relation types was more prevalent in the *test* display.

Trial Procedure: The general procedure is described in Section 3 above. The *response* screen remained until the participant responded with the answer by pressing the 'T' for the top relation, 'G' for the bottom relation (keyboard keys were covered with stickers showing a "/\" and "\/", respectively).

Design: Factors in the full factorial design included: 2 **data depictions** (individual value encodings, delta value encodings) x 3 **visual feature encodings** (position, length, slope) x 3 **difficulty ratio** (1:9, 3:7, 4:6) x 2 **target relations** (relation composing the majority of the test display's relations (i.e., the correct answer): increasing, decreasing) x 8 **repetitions** – yielding a total of 288 test trials.

Trial Order: Participants first completed 10 practice trials. Trials were randomly ordered within each of 8 test blocks, one block for each repetition.

5.2 Results and Discussion

Accuracies (percent correct) were calculated after grouping correct/incorrect responses across trial repetitions. We ran a factorial, repeated measures ANOVA on the factors data depiction, visual feature encoding, difficulty ratio, and target relation on the accuracies. Significant effects were followed up by two-tailed paired t-tests. Figure 4 (center) and Table 3 show accuracy results for Experiment 2. Only the results for our primary experimental factors are listed here (see *Supplemental Material* for other analyses).

Data Depiction: If the visual depiction of representing individual data values or their deltas impacts visual processing efficiency, then participants' accuracies should be higher or lower depending on the type of data depiction. Indeed, accuracies were impacted by data depiction, $F(1,12) = 116.82$, $p < 0.001$, $\eta p^2 = 0.91$, such that **accuracies for *delta value encoding* trials were significantly higher than those for *individual value encoding* trials**.

Visual Feature Encoding: The particular visual feature encoding of the data value pairs impacted accuracies, $F(2,24) = 28.09$, $p < 0.001$, $\eta p^2 = 0.70$. **Accuracies were significantly lower for *slope* encodings** (i.e., *individual slopes* and *delta slope* encodings combined) **than both *length* encodings** (i.e., *individual bars* and *delta bar* encodings combined) **and *position* encodings** (i.e., *individual dashes* and *delta dash* encodings combined). Participants' accuracies for *length* and *position* trials were not statistically different. Considering the latter result was approaching significance, these results are roughly consistent with the visual feature encoding ranking proposed by Cleveland & McGill [3], as in Experiment 1.

Accuracies for each visual feature encoding also interacted with data depiction, $F(2,24) = 17.38$, $p < 0.001$, $\eta p^2 = 0.59$. To investigate this effect, data depiction differences (*delta value encoding* accuracy - *individual value encoding* accuracy) were calculated for each visual feature encoding for each participant. The larger the data depiction difference, the greater the advantage offered by delta value encodings. **The data depiction differences for *position* trials were significantly larger than for both *length* trials and *slope* trials, suggesting that the greatest *delta value encoding* advantage can be found when using *position* encodings.** Data depiction differences for *length* trials were marginally significantly larger than for *slope* trials. This pattern of results is also consistent with the Cleveland & McGill [3] visual feature encoding ranking. The worse the visual feature encoding, the lower the accuracies are to chance performance (50% accuracy), and the less possible difference there can be between trials of both data depictions for that visual feature encoding.

In sum, the data encoding greatly impacted the visual processing efficiency of relations between data values in the same manner as in Experiment 1: **relations between data pairs are much more efficient to visually extract when directly represented as deltas rather than individual data points**. Plotting individual data values is again incredibly inefficient – **accuracy improves by 30% when data values are plotted as delta values instead**. This, however, is a very conservative calculation, given that purely guessing yields an accuracy of 50% (our results indicate a delta advantage of 143% if examining accuracy beyond 50%). The effectiveness of the particular visual feature encodings of the values impact processing as well, and in such a way that is mostly consistent with prior work [3].

Participants' low performance on slope trials in both Experiments 1 and 2 was surprising – while slope is a lower-ranked visual feature encoding [3], it is still a basic visual feature encoding that tends to be easy to parse in visual search (though this depends on the particular slope angles in the display, [25]) and ensemble coding [5] tasks. We believe this discrepancy stems from the difference in display arrangements, given that encodings are arranged in a row in our tasks, instead of scattered across the screen as in other studies. Randomly-arranged slopes tend to form an overall texture that can be efficiently segmented [10]. Slopes arranged in a row across the screen tend to be grouped as $\wedge$ and $\vee$ shapes, which could make it harder to view the encodings as individual slopes to accomplish the tasks.

6 Experiment 3: Ensemble Coding of Magnitudes

Experiment 2 revealed that participants were significantly more accurate when discriminating proportions of value pairs when they were encoded as deltas rather than as individual values. While participants attended to the relation direction, they ignored the *degree* to which values differed within a value pair. Experiment 3 explored whether this result extends to cases when one is assessing the magnitude of difference between two values, across a set of data pairs. The goal of this experiment was to emulate situations where an observer computes the average difference between two values across multiples instances. A realistic example of this is judging the average difference in employment across multiple population segments before and after an important policy change.

We measured error while participants indicated the average data value difference (delta) across 6 data pairs. Deltas needed to be extracted from individual value encodings prior to averaging, but were simply averaged if delta value encodings were presented. If visual processing is more efficient when relations are represented by delta value encodings than by individual value encodings, then error should be *lower* on trials with delta value encodings than with individual value encodings.

Participants responded with the average delta (e.g., What's the average difference?) instead of average relation ratio (e.g., Here is the average low value – please draw the average high value relative to this average low value) because the latter would only apply to individual value encodings. That is, if participants were to respond with the average relation ratio among 6 individual value encodings, participants could be shown one value (e.g., a bar representing a high value) and be asked to indicate the appropriate other value (e.g., a bar representing a low value) such that both values are *proportional* to the average low and average high value of the set. There is no equivalent scenario for *delta value encoding* trials because participants are providing only one value.

6.1 Materials and Procedure

Stimuli

We suspected this task would require more effort and be more time consuming than Experiments 1 and 2 since there is a much wider range of responses, so we decided to use only the two best performing visual feature encodings (*length* and *position*; see *Encoding Types*). These two visual feature encodings are also much more commonly used than slope in visualization.

Randomly selecting a set of values from Table 1 on each trial would have resulted in the same correct response for each trial, on average across trials. Therefore, unlike Experiments 1 and 2, data pair values were generated for each experiment to provide a range of responses across trials. As before, data pair values were converted to pixels. Encodings were all drawn (via Matlab) on the screen rather than presented as image files (as in Experiments 1 and 2), since image values were created at the start of each participant's testing session using our stimulus value algorithm (see *Supplemental Material*).

The dashes (in *position* trials) and bars (in *length* trials) were always each 0.63 visual degrees wide, while the thicknesses of the dashes were always 0.13 visual degrees. *Individual value encoding* trials contained 0.47 visual degrees of empty space between the bar/dash pairs which represented each data pair. The data pair shown in the preview displays for *individual value encoding* trials represented the lowest possible value (19; 0.50 visual degrees) and 38 (1.01 visual degrees) as the higher value, while *delta value encoding* trial depicted a delta of 19, so that both types of *preview* displays depicted the exact same delta (19).

A set of 6 data pairs were generated for each *ensemble* trial according to an algorithm described in the *Supplemental Material*. This process was repeated for each visual feature encoding (*length*, *position*), each relation (*increasing*, *decreasing*), each of 3 delta distributions (from which each trial's deltas were selected), and repeated 4 times for a sufficient amount of trials. The same values were used for both *individual value encoding* trials and *delta value encodings* trials so that performance could be correlated. The same algorithm was used to create *control* trials, except only one data pair was created for each trial. This process was repeated for each visual feature encoding (*length*, *position*), each relation (*increasing*, *decreasing*), each of 3 delta distributions, and repeated 2 times for a sufficient amount of trials. The same values were used for both *individual value encoding* trials and *delta value encodings* trials so that performance could be correlated.

There were 3 primary screen types:

Preview Screens: These displays featured the general type of data pair for the given trial (i.e., the particular combination of data depiction, visual feature encoding, and relation – regardless of whether it is an *ensemble* or *control* trial; e.g., a single bar above the baseline during a *length delta value encoding* trial featuring *increasing* relations) that the participant was to judge in the subsequent *test* display. Each data pair present in the *preview* (as well as the *test*) displays depicted each data value during *individual value encoding* trials, but only their deltas during *delta value encoding* trials.

Test Screens: During *ensemble* trials, each *test* display always contained 6 data pairs comprised of the same relation within each trial (e.g., 6 pairs of a small bar to the left of a taller bar) of a single encoding (see Figure 3). All data pairs were centered across 6 evenly-spaced positions (4.20 visual degrees apart), all aligned vertically to a bottom baseline at the center of a white screen. During *control* trials, each *test* display contained 1 data pair located in the 2nd data pair position from the left (i.e., 6.29 visual degrees left of the screen's center). The data value pair always appeared in this position during *control* trials so that its location was consistent, because the *ensemble* trials containing 6 data value pairs at always the same 6 locations.

Response Screens: These displays featured a single bar (during *length* trials) or single dash (during *position* trials) arranged above or below the baseline depending on whether it was an *increasing* or *decreasing* relation trial, respectively. The data value represented by the bar or dash in these screens (i.e., the height of the bar, and the distance between the baseline and the top of the dash) was randomly selected from the range of possible delta values (19 to 126) so that any bias from the bar/dash's starting position would average out across trials. The height of the bar (the top of the bar if it was above the baseline (*increasing* relation trials) or the bottom of the bar if it was below the baseline (*decreasing* relation trials)) or position of the dash adjusted as the participant moved the computer mouse up or down, but was restricted to the range of possible delta values.

Procedure

Task: Participants were asked to determine the average delta across the 6 data pairs (i.e., the average difference between low and high values during *individual value encoding* trials, and simply the average value during *delta value encoding* trials) during *ensemble* trials, and to replicate the delta of the single data pair (i.e., replicate the difference between the low and high value during *individual value encoding* trials, and replicate the value displayed during *delta value encoding* trials) during *control* trials, depending on whether the *test* display contained 6 data pairs or only 1. Participants otherwise did not know prior to the *test* display whether they were about to view an *ensemble* or *control* trial since those trials' *preview* screens were identical for any given combination of data depiction, visual feature encoding, and relation.

Trial Procedure: The general procedure is described in Section 3 above. The *response* screen remained until the participant responded with the answer by moving the mouse up and down to adjust the bar/dash height, and then left-clicked to submit their response.

Design: Factors in the full factorial design included: 2 **tasks** (ensemble, control) x 2 **data depictions** (individual value encoding, delta value encoding) x 2 **visual feature encodings** (length, position) x 2 **display relations** (increasing, decreasing) x 3 **delta distributions** (distributions from which each trial's deltas were selected; means: 45.75, 72.50, 99.25) x 4 or 2 **repetitions** (ensemble trials and control trials had 4 and 2 repetitions, respectively) – yielding a total of 144 test trials (96 ensemble trials and 48 control trials).

Trial Order: Participants first completed 16 'slow' practice trials during which the *test* display remained on screen for twice as long (1000 ms) because this task is challenging and we suspected participants would need some extra time to fully understand the task. This was followed by 16 practice trials during which the *test* display remained on screen for the experiment trial duration (500 ms). Each block of practice trials contained trials for every combination of task, data depiction, and visual feature encoding; delta difference distribution was randomly selected for each practice trials. Trials were randomly ordered within each block (slow practice, practice, test trials).

Table 4. Errors for Experiment 3. Descriptive statistics, ANOVA results, and follow-up t-test results shown for significant ($p < 0.05$) comparisons for absolute errors (pixels between participant's response and the true mean delta).

Factor		*M*	*SE*	Statistical Test
Task				$F(1,12) = 64.59$, $p < 0.001$, $\eta p^2 = 0.84$
	Ensemble Task	15.9	1.08	$t(12) = -8.04$, $p < 0.001$, $d = -2.23$
	Control Task	10.24	1.12	
Data Depiction				$F(1,12) = 46.73$, $p < 0.001$, $\eta p^2 = 0.80$
	Delta Value Encodings	11.21	1.15	$t(12) = 6.72$, $p < 0.001$, $d = 1.86$
	Individual Value Encodings	14.94	1.00	
Visual Feature Encoding				$F(1,12) = 7.75$, $p = 0.017$, $\eta p^2 = 0.39$
	Length Encodings	12.25	0.87	$t(12) = -2.80$, $p = 0.016$, $d = -0.78$
	Position Encodings	13.89	1.26	

6.2 Results and Discussion

Mean absolute errors (pixels between participant's response and the true mean delta) were calculated after grouping absolute errors across trial repetitions. We ran a factorial, repeated measures ANOVA on the factors task, data depiction, visual feature encoding, display relation, and delta distribution on the mean absolute errors. Significant effects were followed up by two-tailed paired t-tests. Figure 4 (right) and Table 4 show error results for Experiment 3. Only the results for our primary experimental factors are listed here (see *Supplemental Material* for other analyses).

Task: A control task (i.e., replicate the delta from the single value pair displayed) was included to assess any error that may stem from simply replicating a delta with this experiment's response procedure. We expected error to be greater from *ensemble* trials than *control* trials because *ensemble* trials require the additional step of extracting the mean delta from all 6 value pairs (i.e., the process of ensemble coding). Indeed, task impacted errors, $F(1,12) = 64.59$, $p < 0.001$, $\eta p^2 = 0.84$, such that **errors from *ensemble* trials were greater than those from *control* trials**.

Identical deltas were displayed between *individual value encoding* and *delta value encoding* trials for each type of visual feature encoding. Participants' errors to *individual value encodings* and *delta value encodings* (for each visual feature encoding) may correlate to reveal whether there is a common mechanism between the two types of visual processes: a high correlation resulting from performance fluctuating as one replicates (*control* trials) and averages (*ensemble* trials) delta(s) within the *test* display suggests a common underlying process to perceiving deltas from pairs of values and perceiving directly drawn single deltas, while a low correlation suggests error unique to each data depiction type. Errors were correlated between *individual value encoding* and *delta value encoding* trials for each visual feature encoding within each task and averaged across all participants. Correlation values were overall quite low, though a little higher for *control* trials (*length* trials: $M = 0.32$, $SE = 0.36$; *position* trials: $M = 0.20$, $SE = 0.34$) than *ensemble* trials (*length* trials: $M = 0.18$, $SE = 0.30$; *position* trials: $M = 0.08$, $SE = 0.26$), suggesting either little commonality between the underlying mechanisms, or significant additional error resulting from the additional step of extracting a delta from a value pair prior to replicating (*control* trials) or averaging (*ensemble* trials) the delta(s).

Data Depiction: Critically, if the visual depiction of representing individual data values or their deltas impacts visual processing efficiency, then participants' errors should be higher or lower depending on the type of data depiction. Indeed, errors were impacted by data depiction, $F(1,12) = 46.73$, $p < 0.001$, $\eta p^2 = 0.80$, such that **errors for *delta value encoding* trials were significantly lower than those for *individual value encoding* trials**.

Task x Data Depiction: Surprisingly, the pattern of errors for each data depiction (lower errors for *delta value encodings* than for *individual value encodings*) did not vary by task – there was no interaction between task and data depiction, $F(1,12) = 0.73$, $p = 0.41$, $\eta p^2 = 0.057$.

Visual Feature Encoding: The particular visual feature encoding of the data value pairs impacted errors, $F(1,12) = 7.75$, $p = 0.017$, ηp^2

= 0.39. **Errors were significantly lower for *length* encodings than for *position* encodings.** While this result runs contrary to Experiment 1 and Experiment 2's visual feature encoding ranking, the encodings' means differ only minimally.

Error Direction: Participants tended to systematically underestimate their responses in more than half the conditions, notably in almost all *length* conditions and in almost all delta value encoding conditions (see *Supplemental Material* for analyses).

In sum, the data encoding significantly impacted the precision of the visual processing of relations between data values in the same manner as in Experiments 1 and 2: **relations between data pairs are visually extracted and averaged much more precisely when directly represented as deltas rather than individual data points**. Plotting individual data values is inefficient again – **error decreases by 25% when data values are plotted as delta values instead during *ensemble* trials**. However, directly-represented deltas tended to bias responses to **underestimate**, rather than under- and over-estimate.

Given that this experiment's task was particularly complex, we were surprised by the degree to which delta value encodings improved accuracy, and were expecting a massive result similar to Experiment 1. We suspect this is likely due to a perceptual heuristic which participants employed, and the fact that all test screens always contained value pairs with the same relation direction (i.e., all increasing relations). Since participants did not need to selectively 'filter' for a particular relation, it is possible that (in the *individual value encoding* trials) they averaged only the 'thin', top portion of each bar pair in the *length* trials, or averaged only the distance between the higher dash and its neighbouring lower dash. Neither of these approaches would work alone if the task was to estimate the average delta for *only one* of two types of relations in a display. Indeed, a more likely scenario would include value pairs of the opposite relation which need to be filtered out. Can we accurately *filter out* irrelevant relations to average the specific relations of interest? Given that opposing relations are significantly harder to distinguish when encoded by individual value encodings rather than delta value encodings (i.e., Experiment 2's results), it is likely that performance in this type of scenario would be far worse than it already is when values are represented individually.

One limitation is that participants responded by adjusting the height of a single bar or dash, which is most similar to the delta encoding itself. It is possible that method could bias responses in favour of delta value encodings. If so, this would be important to examine further given that it would provide a task-specific instance in which delta value encodings provide less (or even no) benefit.

7 Experiments 1-3: Results Summary

The primary results are summarized in Figure 4. The key take-aways across the three experiments are as follows:

A) **Delta value encodings consistently yield better relation perception than individual encodings.** Participants searched faster for the opposite relation (Exp. 1), were more accurate in perceiving which relation there was more of (Exp. 2), and had lower error rates when perceiving the average delta (Exp. 3) when deltas were explicitly encoded, than when individual data points were encoded. While [19] showed delta encoding response time benefits, the present data add that responses times *continue* to slow down with the addition of each individually encoded data pair. Experiment 2 uniquely shows the benefit of delta encodings in perceiving proportions of relations. Lastly, we show that delta encodings improve perception of the average delta, adding to the finding of Srinivasan et al. [19] that accuracy for measuring a specific, single delta was comparable across chart types.

B) Further, delta value encodings were far more efficient than individual value encodings for processing, but **highly depended on the task**-- delta encodings accelerated search rates by 49-95% in Exp. 1, improved accuracy by 30% in Exp. 2, and lowered error rates by 25% in Exp. 3. While it may be tempting to generalize a guideline across these tasks, testing delta encodings further across a larger swath of tasks is necessary in order to conclude which types of visual decisions benefit most from delta encodings, beyond our three specific tasks. The fact that we find great variability in the delta advantage points to the need for this exact type of additional testing.

C) Lastly, **delta value encodings tend to bias people to underestimate the average delta** (Exp. 3). While this pattern is intriguing, its perceptual root is unclear.

8 Guidelines

Show deltas, but only when necessary: Across our three tasks, delta encodings consistently yielded the best performance over individually depicted data values: a 25-95% improvement depending on the task. Unfortunately, depicting deltas requires more space in webpages, slides, or dashboards, and they typically cannot simply replace the visualization of the base values, which provide the context for those differences, so they should be used judiciously. If perceiving data value pair differences is central to the task (e.g., identifying whether an intervention improved the majority of test scores in Figure 1A, unlike identifying which student had the single highest test score), encoding differences should offer far better performance, even more so when the viewer is identifying data pairs with a particular relation (like in Experiment 1). The difference overlay technique tested by [19] presents a potentially powerful way to show both absolute values as well as deltas between pairs of values, in a way that adds visual processing power, with little evidence of drawbacks.

Use position and length encoding: Consistent with prior work [3], position (e.g., dot plots) and length (e.g. bar charts) encodings led to far better performance compared to slope encodings, at least for the present displays and tasks.

Require less relational extraction: Extracting relations (i.e., perceiving both the *difference* between a data value pair and the direction of that difference – whether it is an *increase* or a *decrease*) is a slow process, and *continues to slow down* with the addition of *each* data value pair (Exp. 1 – response times increase as set size increases). Therefore, when it is not possible to explicitly show delta (e.g., for data triplets instead of pairs), one might consider decreasing the resolution of included categories, to decrease the amount of relational processing demanded of the viewer.

9 Limitations and Future Work

Loss of context: As already stated, naturally much context is lost with the loss of the original individual data values by displaying only deltas. In light of this, **our results highlight the need for visualizations that both display original data values *and* showcase relational differences**. In fact, grouped bar charts with deltas values explicitly overlaid were preferred by viewers [19]. Further, this approach is highly valuable in situations of data exploration, in which the viewer *does not know ahead of time* which aspects are important or relevant to compare.

How do these results scale? The present experiments tested relations between two data values as a starting point. How do these results scale when the number of data points, and thus relations, increases (e.g., a bar chart containing multiple bars per data group)? Relatedly, our displays always contained multiple relations, but our participants never made comparisons between any relations (i.e., relations between relations).

Number of objects, or density? Density is confounded with increased set size in Experiments 1, though equally so across encodings. Because both can sometimes lead to response time increases [24], future work should measure their relative contributions to search inefficiency, which may lead to new concrete guidelines (e.g., more data may not slow viewers down as much, as long as it is well-spaced).

REFERENCES

[1] D. Ariely. Seeing sets: Representation by statistical properties. *Psychological science*, 12(2), 157-162, 2001.

[2] D. H. Brainard. The Psychophysics Toolbox. *Spatial Vision*, 10, 433–436, 1997.

[3] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387), 531-554, 1984.

[4] W. S. Cleveland and R. McGill. Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716), 828-833, 1985.

[5] S. C. Dakin and R. J. Watt. The computation of orientation statistics from visual texture. *Vision Research*, 37(22), 3181-3192, 1997.

[6] S. L. Franconeri, J. M. Scimeca, J. C. Roth, S. A. Helseth and L. E. Kahn. Flexible visual processing of spatial relationships. *Cognition*, 122(2), 210-227, 2012.

[7] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4), 289-309, 2011.

[8] M. Gleicher, M. Correll, C. Nothelfer and S. Franconeri. Perception of average value in multiclass scatterplots. *IEEE transactions on visualization and computer graphics*, 19(12), 2316-2325, 2013.

[9] J. Haberman and D. Whitney. Ensemble perception: Summarizing the scene and broadening the limits of visual processing. *From perception to consciousness: Searching with Anne Treisman*, 339-349, 2012.

[10] C. G. Healey, K. S. Booth, and J. T. Enns. High-speed visual estimation using preattentive processing. *ACM Transactions on Computer-Human Interaction* (TOCHI), 3(2), 107-135, 1996.

[11] C. G. Healey and J. T. Enns. Attention and visual memory in visualization and computer graphics. *IEEE transactions on visualization and computer graphics*, 18(7) 1170-1188. 2012.

[12] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. *Proceedings of the ACM SIGCHI conference on human factors in computing systems*, 203-212, 2010.

[13] J. Heer, N. Kong and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. *Proceedings of the ACM SIGCHI conference on human factors in computing systems*, 1303-1312, 2009.

[14] L. Huang and H. Pashler. Attention capacity and task difficulty in visual search. *Cognition*, 94(3), B101-B111, 2005.

[15] J. E. Hummel. Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual cognition*, 8(3-5), 489-517, 2001.

[16] R. Kosara. An empire built on sand: reexamining what we think we know about visualization. Proceedings of the ACM Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization,162-168, 2016.

[17] G. D. Logan. Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, 20(5), 1015, 1994.

[18] J. Mackinlay. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)*, 5(2), 110-141, 1986.

[19] A. Srinivasan, M. Brehmer, B. Lee and S. M. Drucker. What's the Difference?: Evaluating Variations of Multi-Series Bar Charts for Visual Comparison Tasks. *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 304, 2018.

[20] D. A. Szafir, S. Haroz, M. Gleicher and S. Franconeri. Four types of ensemble coding in data visualizations. *Journal of vision*, 16(5), 11-11, 2016.

[21] J. Talbot, V. Setlur and A. Anand. Four experiments on the perception of bar charts. *IEEE transactions on visualization and computer graphics*, 20(12), 2152-2160, 2014.

[22] A. Treisman and S. Gormican. Feature analysis in early vision: Evidence from search asymmetries. *Psychological review*, 95(1), 15, 1988.

[23] J. M. Wolfe. What can 1 million trials tell us about visual search? *Psychological Science*, 9(1), 33-39, 1998.

[24] J. M. Wolfe, K. R. Cave and S. L. Franzel. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance*, 15(3), 419, 1989.

[25] J. M. Wolfe, S. R. Friedman-Hill, M. I. Stewart and K. M. O'connell. The role of categorization in visual search for orientation. *Journal of Experimental Psychology: Human Perception and Performance*, 18(1), 34, 1992.

[26] J. M. Wolfe and T. S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature reviews neuroscience*, 5(6), 495-501, 2004.