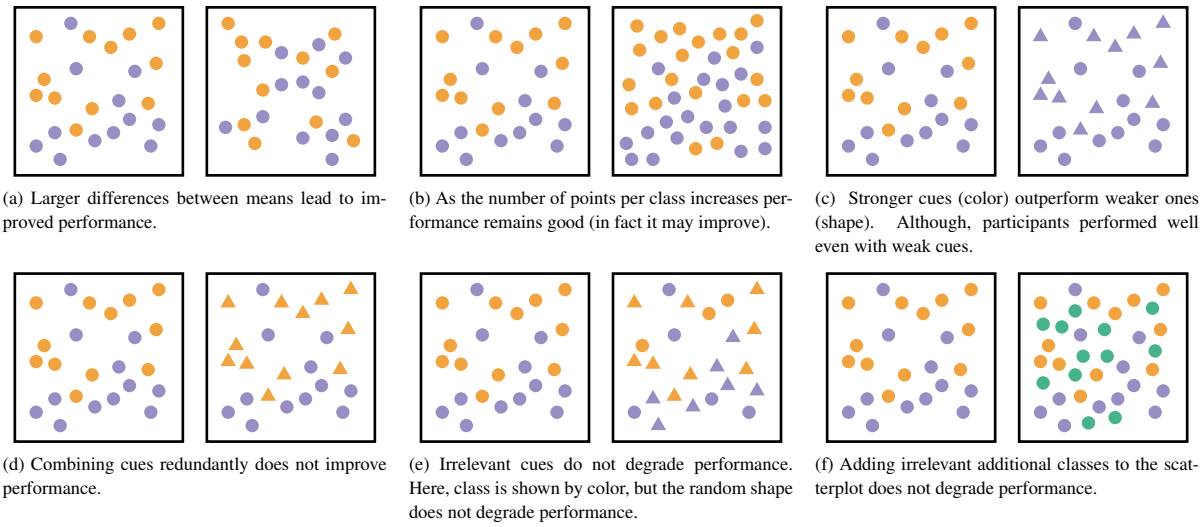
Perception of Average Value in Multiclass Scatterplots

Michael Gleicher, *Member, IEEE*, Michael Correll, *Student Member, IEEE*, Christine Nothelfer, and Steven Franconeri



(a) Larger differences between means lead to improved performance.

(b) As the number of points per class increases performance remains good (in fact it may improve).

(c) Stronger cues (color) outperform weaker ones (shape). Although, participants performed well even with weak cues.

(d) Combining cues redundantly does not improve performance.

(e) Irrelevant cues do not degrade performance. Here, class is shown by color, but the random shape does not degrade performance.

(f) Adding irrelevant additional classes to the scatterplot does not degrade performance.

Fig. 1. Summary of results: viewers can efficiently make comparative mean judgements, choosing the class with the highest average position in multiclass scatterplots across a wide variety of conditions and encodings.

Abstract—The visual system can make highly efficient aggregate judgements about a set of objects, with speed roughly independent of the number of objects considered. While there is a rich literature on these mechanisms and their ramifications for visual summarization tasks, this prior work rarely considers more complex tasks requiring multiple judgements over long periods of time, and has not considered certain critical aggregation types, such as the localization of the mean value of a set of points. In this paper, we explore these questions using a common visualization task as a case study: relative mean value judgements within multi-class scatterplots. We describe how the perception literature provides a set of expected constraints on the task, and evaluate these predictions with a large-scale perceptual study with crowd-sourced participants. Judgements are no harder when each set contains more points, redundant and conflicting encodings, as well as additional sets, do not strongly affect performance, and judgements are harder when using less salient encodings. These results have concrete ramifications for the design of scatterplots.

Index Terms—Psychophysics, Information Visualization, Perceptual Study

1 Introduction

Many visualization tasks require the viewer to create abstractions, or statistical summaries, over groups of marks. We use the term *visual aggregation* for such situations where the viewer "computes" the aggregate properties when presented with a collection of objects. In many cases, these abstractions can be constructed rapidly even for large numbers of objects (i.e.,"preattentively"). This ability has been studied extensively in the perception literature, leading to models of the mechanisms behind them as well as implications for visualization. However, models of aggregation from the perception literature are typically based on performance patterns for brief display exposures, leaving it unclear whether their implications apply to situations where viewers contemplate more complex displays across longer periods of time. Prior studies isolate individual mechanisms, but provide little insight on how these mechanisms may be combined.

In this paper, we explore aggregate judgement in visualizations using a realistic task: assessing the difference in class means in a scatterplot. The task involves accurate localization, and we permit viewers to take time to make accurate judgements. This differs from prior studies that use unrealistically short exposures in order to build models of efficient aggregation in the visual system.

Scatterplots are a common visual presentation. Viewer ability to rapidly and accurately assess trends has been studied (e.g. Doherty et al. [17] and Rensink & Baldridge [45]). Scatterplots often present multiple data classes simultaneously to aid comparison. Such displays are advantageous because they allow the viewer to see specifics and trends within each class, as well as to make relative judgements between classes. Li et al. [38, 39] demonstrate viewers' ability to make rapid judgements about multi-class scatterplots for several tasks. While there are many different ways to measure the difference between classes [50], comparison of the means of groups is common as it corresponds to many decision criteria (e.g. is one class better than another). The importance of mean separation has lead to view selection methods, such as [52] and [16], that maximize it.

It is often possible to present the descriptive statistics to the viewer

- *Michael Gleicher is with the Department of Computer Sciences, University of Wisconsin - Madison. Email: gleicher@cs.wisc.edu.*
- *Michael Correll is with the Department of Computer Sciences, University of Wisconsin - Madison. E-mail: mcorrell@cs.wisc.edu.*
- *Christine Nothelfer is with the Department of Psychology, Northwestern University. E-Mail: cnothelfer@u.northwestern.edu*
- *Steven Franconeri is with the Department of Psychology, Northwestern University. E-mail: franconeri@northwestern.edu.*

Author's preprint version.
To appear in *IEEE Trans. on Visualization and Comp. Graphics,* 19 (12), 2013.

(e.g. explicitly marking the means). However, allowing the viewer to make the judgement by aggregating the data *visually* can offer a number of advantages, such as not needing to know the viewer's needs, not needing to clutter the displays with another form of information, and providing a natural combination of the statistics with the details and trends. However, these potential benefits of visual aggregation can only exist if viewers are able to make reliable judgements.

The theory and evidence in the perception literature illuminates mechanisms that viewers can use for aggregation tasks. However, this prior work has typically focused on performance within relatively simple displays that are briefly flashed, in contrast to more complex visualizations that can be inspected over the course of several seconds. Even though viewers can make rapid judgements about multi-class scatterplots when forced (e.g. [38, 39]), they generally choose to take more time. The perception literature describes constraints on the visual system for rapid, simple tasks. If the same mechanisms are part of more complex judgements, these constraints make predictions about our tasks of interest, in situations where viewers take more time.

Because we are interested in viewer performance when they are not time constrained, our questions cannot be studied using the standard experimental paradigm that measures response time, either by asking participants to respond as quickly as possible or varying exposure time. Instead, we use an experimental design where participants are not time constrained (within limits), and we vary the challenge level of the trials by altering proprties of the stimulus. Fortunately, the multi-class scatterplot mean comparison problem affords many types of control over task hardness. Also, the fact that we do not need precise timing allows us to implement the experiment in a standard web browser, affording the use of crowd sourcing which gives us access to a large and diverse participant pool.

Based on the prior evidence of the limitations of visual mechanisms, we generated a set of predictions, described in §1.1 below, and tested these predictions by asking over 750 crowd-sourced participants to compare group means within scatterplots. These experiments share many common features, discussed in Section 3. The specific results of the experiments are discussed in Section 4. The first experiment, discussed in Section 4.1, uses a between-subjects design to establish the main points of our theory. However, since many of the important results are null results (i.e. we predict no performance differences between different sorts of scatterplots), this design does not provide the statistical power needed to make some conclusions with confidence. Therefore, we conducted a series of within-subjects experiments, described in Section 4.2, that reinforce these results with higher confidence. After presenting these results, we discuss their ramifications both in terms of providing an understanding of the perceptual mechanisms in visualization tasks, but also to the design of displays that support aggregate judgement.

We find that viewers can make efficient judgements about the means in scatterplots, and that constraints on this ability follow predictions based on the perception literature. Our key findings are summarized in Figure 1.

Contributions: We test a set of predicted constraints on visual aggregation tasks that involve relatively complex displays viewed over the course of several seconds. We present a large scale, crowd-sourced study that supports these predictions. Our task requires judgements within multi-class scatterplots, a very common visualization, and our work provides an empirical assessment of how viewers perform on aggregation tasks, as well as guidance on how to create visualizations that support these tasks.

1.1 Predictions of Performance

Both the perception and visualization communities have demonstrated that the visual system can accomplish a wide range of tasks efficiently, such as estimation of numerosity or mean value.

One common mechanism underlying performance across many of these tasks is attentional selection: the ability to amplify visual information that meet some criteria, while suppressing the rest. One criterion is location, set both by the position of the eyes, as well as the "spotlight" of attention (though these two are typically highly correlated; see [23] for review). Other criteria for selection are featural, constrained by the presence of existing tuning mechanisms that alter the weighting of particular features (e.g., certain colors or shapes). The visual system uses these features to create an abstract map of the information present. After a subset of visual information is selected from this map, the visual system can form abstractions over that subset [33, 34, 57]. Tasks can be done efficiently (with only a "flash" of exposure time) if the viewer is able to weight a map to select the appropriate features and make simple assessments.

Unfortunately, these prior models do not explain what a viewer does with more time, or how they might be able to use more time to achieve better performance. These prior studies outline limitations on what the human visual system is capable of selecting, ignoring, and aggregating, and here we test how observers perform when they are allowed sufficient time to use this architecture effectively, and in situations that are relevant to visualization.

For example, studies from the perception literature show that observers can efficiently average position across a handful of points [1], but what about the dozens of points in a typical scatterplot? Perception studies show that global selection of a single feature value (e.g. red) is possible (e.g., [49]), but can average position be extracted from these subsets? Perceptual studies show that some features are easier to select or localize than others (e.g., color vs. shape) [12, 56] - does this generalization hold when more time is provided to inspect a visualization? Perceptual theory suggests that selecting a value within a single dimension (e.g., red among colors) can be difficult (e.g., [58]), so will observers really select a value for a second dimension (e.g. circles among shapes) to take advantage of redundant encoding? Perceptual studies of selective attention suggest that irrelevant 'distractor' features are extremely difficult to ignore in briefly presented displays (e.g., [58]), but will this still be true when observers have sufficient time to tune their feature-selective filters?

These questions lead to more concrete predictions based on limitations in the basic mechanisms.

- As the means become closer, the task will become more difficult, and performance will degrade. This prediction is included as a check of our experimental design.
- Because feature selection acts globally over large collections, and some work has shown that center-judgements are possible over large collections of points, performance should not be impaired by larger collections. (Fig 1b)
- Because the effectiveness of selection is influenced by feature contrast, features that are easy to select (e.g. salient colors) will lead to better performance than features that are harder to select (e.g. shape, or less salient colors). (Fig 1c)
- Because features must be selected individually, redundant encodings that provide multiple features will not improve performance, beyond giving the viewer a choice of a feature to select. (Fig 1d)
- Because selecting one feature suppresses the others, conflicting encodings, such as adding variability along a different feature dimension, will not impair performance unless the conflicting feature is so salient that it interferes with the selection of the primary feature. (Fig 1e)
- Because the viewer selects specific values of the feature, adding other values does not cause significant distraction. For example, if the viewer selects purple, then orange, the existence of green dots should not interfere. (Fig 1f)
- Because selection requires choosing what to select and what to suppress, a sufficient diversity of distractions may impair performance.

Our premise is that some of the same underlying mechanisms used in rapid response tasks are used in longer time tasks. This does not imply that performance is the same: viewers may have different ways of

using the basic mechanisms. However, it does suggest that the fundamental limitations of the mechanisms (what can be selected and what can be extracted from the resulting subset) still apply at longer durations. Our experiments seek to confirm these performance predictions.

2 Background and Related Work

The literature on perceptual psychology provides inspiration for both the types of abstractions that can be constructed over sets of objects, and the types of cues that allow efficient segmentation of these sets. The human visual system can quickly construct many types of abstractions from sets, including numerosity (see [22], for review), and averages over dimensions like size [3, 10], orientation [11], motion direction [37], spatial frequency [2], and perhaps even more complex properties like facial emotion and gender [26] (but see [40], for caveats). Efficient segmentation of sets has been studied using tasks such as visual search [54], texture boundary identification [8,9], and number discrimination [27]. Most relevant to the present studies, observers can average spatial position over a set of objects [1], but this study only demonstrated this ability for a handful of objects. Other work shows that people can make saccades to the centerpoint of objects made up of large sets of dots that form a rough object contour [41], but it is not clear that this ability will generalize, or that this centerpoint estimate is consciously available.

These tasks have revealed many features that serve to segment sets of objects broadly and rapidly, including relative differences in hue, orientation, shape, and size [18, 56, 57], as well as some more complex visual properties such as lighting direction [20]. Some features are processed more efficiently than others [56] - e.g. tasks involving selecting lines with atypical colors in a display are faster and less error prone than tasks involving selecting lines with atypical concavity. Haroz and Whitney [29] also explore how the mechanisms of attention limit performance in various visual tasks.

The visual system can create abstractions (e.g. numerosity estimation, mean position, spatial envelope extraction) across the set of visual field locations that are currently selected by attention. Concretely, attentional selection is a relative amplification of visual information that meets certain criteria, such as being in a specific location (e.g., in the upper left of a display), or containing specific feature values (e.g., red, left tilted, curved, or two-inches-tall). The possible criteria are constrained by the presence of existing feature maps [24, 51] that index the presence or absence of that feature across the visual field, such that novel or arbitrary criteria are not available. Increasing the weight on one or more of these maps would lead to amplification of the visual information that is spatially correlated with the locations highlighted by that map. The perception literature explores many questions related to how this type of model operates, such as whether we can amplify multiple maps corresponding to values on the same dimension [33,37], or whether new maps can be constructed with practice [7].

For present purposes, the most pressing questions revolve around how well people can use these maps and the underlying mechanisms for performing more complex tasks. Most of this work relies on briefly presented visual search displays, and shows that ignoring particularly salient objects can be difficult in some types of displays, suggesting a default mode where people automatically weight maps with unique spots of activation (e.g., [5]). Other work shows that instruction or recent experience can alter the weights on these maps, leading to increased attentional control over what spatial locations or feature values contribute most to attentional selection (for review see [19]) .

While such results from the perception literature are informative, their conclusions about attentional selection do not necessarily generalize to the types of set segmentation needed within data visualizations. First, the tasks used typically require responses within less than a second, which may lead to an underestimate of the types of attentional selection control that may be possible given less rushed visualization tasks that extend over the course of several seconds. We believe the same constraints apply within more complex and extended visual operations that unfold beyond the first "preattentive" snapshot. Second, the tasks typically require observers to either find a single unique object (visual search) or compare relative size or numerosity across multiple sets of objects (e.g. [10, 27]), leaving open the question of whether other types of judgements (e.g. mean spatial position) can rely on the same mechanisms.

Researchers in visualization and graphics have investigated how known perceptual processes and features interact in more realistic displays. Healey, Booth & Enns varied features for encoding salmon migration data [31], finding that participants could successfully perform numerical estimation of items of a particular hue (with task-irrelevant orientation) and of a particular orientation (with task-irrelevant hue) quickly ($<$ 200 ms) and accurately. They also found no effect of interference from the task-irrelevant features, unlike previous studies ([8, 9]). Their displays were regular grids, and the values were contiguous regions, and therefore are quite different than scatterplots. More recently, others have investigated the best symbols for data encoding. Li et al. varied lightness and size of symbols in scatterplot displays from which participants performed several visual analytic tasks [38, 39]. Participant performance was used to model an optimal discriminability scale with equal perceptual separation between scatterplot symbol lightness and sizes.

This work is one of several in a recent trend towards using empirical methodology to analyze how aggregate statistics are perceived and compared in common visual displays. The proposed experimental task (comparison of mean values) and its connection with a relatively high-level aggregate statistical comparison of mean has been used by Foureizos et al. to analyze statistical decision-making in bar charts [21], where as Doherty et al. [17] and Rensink & Baldridge [45] have both looked at correlation coefficients in scatterplots. Two recent papers by Correll et al. have examined the visual perception, aggregation, and comparison of mean values, in both time series data and in paragraphs of tagged text [14, 15].

3 Experimental Design

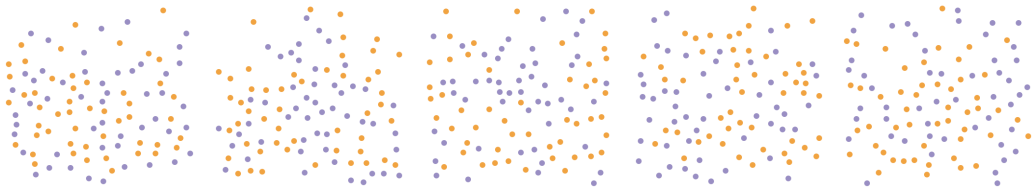
Our work involved a series of pilot studies and two main experiment sets[1]. This section describes the elements common among them.

The model task chosen for our study was to judge the average height of the classes in a multi-class scatterplot. We presented this to participants as a two-alternative forced choice – "which type of point is *on average* higher?"

Our performance measure was participant accuracy. We explicitly do not consider time as a performance criteria: we want to understand viewer performance when they can take the amount of time they feel is necessary to perform accurately. Participants were instructed to answer as accurately as possible, rather than as quickly as possible. We did bound the exposure time of the displays (to ten seconds), to ensure that the participants made sufficient progress and to thwart certain kinds of cheating. The experiment was instrumented to enforce the time limit and record time measurements. Very rarely did participants run into this time limit.

Multi-class scatterplots have a number of attributes that may affect performance on the mean estimation task. First is the number of classes. In our experiments, we only consider two-way comparisons, although for some conditions we add a third class as a distractor. Second is the number of points per class. For this study, we consider only cases where each class has the same number of points, as we did not wish to confound numerosity (which has been previously studied) with mean position. In pilot studies, we confirmed that performance was consistent over a range of numbers of points per class (15-75). We were limited in the range we could explore: if there are too few points, the viewer may be tempted to use a serial strategy; if there are too many points the display may become too dense for the points to be presented distinctly. In most experiments, we chose to use 50 points per class in each display, with exception of one condition where we have 75 points per class to assess the impact of number of points. A third issue is the degree that the classes are inter-mixed. If the different classes are highly disjoint then the comparative averaging task is triv-

[1]Further details of the experiments, including stimuli, instructions, and result tables are available at the project website `http://graphics.cs.wisc.edu/Vis/ScatterVis13/`.



(a) $\Delta = 4$ pixels. (b) $\Delta = 12$ pixels. (c) $\Delta = 20$ pixels. (d) $\Delta = 28$ pixels. (e) $\Delta = 36$ pixels.

Fig. 2. Example stimuli from our various levels of task difficulty parameter Δ, the difference (in pixels) between classes in our scatterplots. When $\Delta = 0$, both classes of points have the same average. In 2a,2c,2e purple points have the highest average. For the others, orange points have the highest average. For this set of stimuli orange points always have the highest absolute value, to disambiguate averaging and peak-finding tasks. Even for the lowest levels of Δ used in our experiments, aggregate performance was significantly better than chance.

ial. If the points are clustered, various visual mechanisms can simplify the task [29].

The primary "hardness" attribute in our design is the vertical distance between the means of each class of points. The closer these are together, the more accurately the means must be localized such that a correct comparison can be made. We call this parameter Δ, the distance between centers, and measure it in pixels. Figure 2 shows various stimuli at different levels of Δ. In the experiments below, we confirm that this parameter is correlated with performance. In pilot studies we observed that when the task was "sufficiently hard," people took a few seconds to make a judgement. There is an expected ceiling: when the task becomes sufficiently easy, most people can get the right answer most of the time. In pilot studies we observed that if we only showed participants hard examples, their performance on those hard examples was significantly worse than if we also showed them some easier examples. Through additional piloting we determined appropriate hardness levels that were used in subsequent experiments.

For each condition, each participant was shown a number of different hardness levels. For each hardness level, six different trials were shown, three of each class as the correct answer. Within each condition, the order was fully randomized.

After giving consent, participants were given a color vision deficiency test using Ishihara plates [28]. Participants failing this test were barred from participating in the main study. Those qualifying were shown a brief tutorial explaining the experimental task that emphasized that they were to identify the class with the highest average value, not the highest specific value. Participants were then shown a number of practice stimuli. The practice consisted of a set of "very easy" stimuli. If the participants correctly answered two of these set in a row they proceeded to a set of slightly more difficult stimuli. After correctly guessing two in a row of these stimuli, they were shown an example of a difficult stimulus. After an incorrect guess, participants were explicitly shown the right answer and then allowed to proceed (see Figure 3).

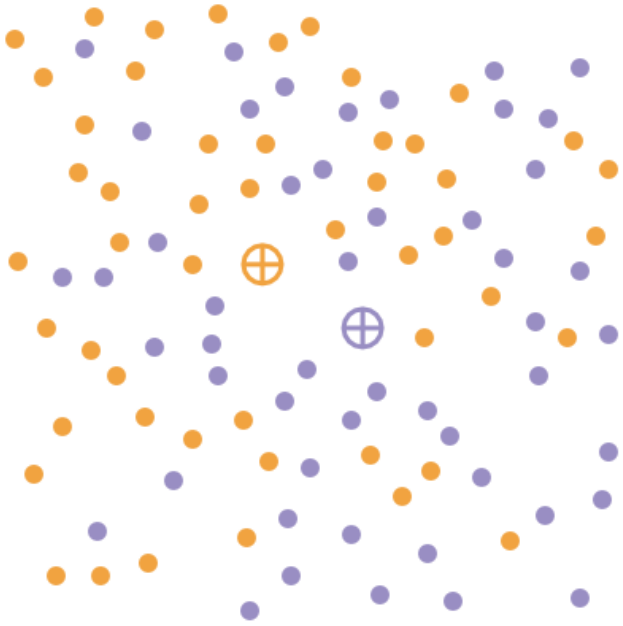


Fig. 3. An example stimulus from our experiments. Participants were asked "Which type of points are *on average* higher?" In this case the types of points being purple or orange circles. In this example the actual mean values are marked to make the difference between classes obvious.

3.1 Stimulus and Generation

Our stimuli are randomly generated multi-class scatterplots. The points were placed according to a uniform random distribution subject to constraints that the difference in the means between groups had the specified value of Δ, and that the points were spaced sufficiently such that no glyphs would overlap. To generate a set of points for a trial, the vertical center of the two classes was randomly selected to be somewhere in the middle third of the display. Randomly shifting the center discouraged strategies that involved considering the mid-point of the display. Given the center of the entire point set, the mean for each class is computed by displacing them above and below. To generate the random points, a dart-throwing approach [36] was used for the Poisson sampling (to prevent overlaps), and best-candidate sampling [42] was used to bias the random distributions to have the appropriate means. Darts were thrown alternating between the two classes to allow for random mixing. The points were adjusted by displacing the points of each class a small amount such that the difference in the means had exactly the required value. When a third class of points was necessary as a distractor this was added after the first two classes of points were generated using the same best-candidate dart throwing sampler. In cases where there was a conflicting cue, the specific level of the conflicting cue for each point was determined randomly.

In pilot studies, we noted that the class with the higher average was more likely to have one of its points be the highest on the chart, leading to a dominant strategy where participants would simply pick the highest point in the chart achieve good accuracy. To discourage this strategy, we purposefully de-correlate which class has the topmost (and bottom-most) point from which class has the higher average by making one class always have the highest point (and the other have the lowest).

In the pilot studies and initial trials, an alternate set of stimuli was created by flipping the stimuli vertically. We counter-balanced flipped and non-flipped cases in a between subjects design and found no significant difference between them. This further added to our confidence that the random process did not create artifacts that skewed the results.

The specific point values of all classes across all levels of Δ were

computed in advance. We used the same sets of points for different conditions between participants, to the extent possible. That is, all of the conditions in Experiment Set One that have two classes of 50 points used the same positions for those points, only altering the way the points were drawn. Similarly, all within-subjects experiments used the same data. Different sets were generated for each block of the experiments, but all experiments (except for the one with 75 points) used the same generated point sets.

Stimuli were generated for a variety of visual encodings. Color encodings used orange and purple for the two classes, chosen from a ColorBrewer [30] qualitative set. In pilot studies, we confirmed that there was neither bias between these colors, or to which one was listed first on the question page. In contrast, our pilot studies found that there was a bias towards red, which is consistent with effects seen in other experiments [13] [53]. Glyphs were usually drawn as filled circles, except when a shape encoding was used, in which case glyphs included triangles, pluses, or squares. The glyphs were sized such that they had (approximately) equal area.

Stimuli were pre-rendered as 400 pixel square images, with white backgrounds and without axes. They were delivered to the participants web browsers in a lossless image format to avoid differences in browser rendering. Circular glyphs were 6 pixels in radius, other shapes were adjusted to have similar area. Each glyph had a 10 pixel radius "exclusion zone" in which no other glyph could be drawn. Since these zones do not overlap, the minimum distance between two circular glyphs is 8 pixels. Stimuli were rendered using sub-pixel accurate anti-aliasing.

Practice and tutorial stimuli were generated using a similar procedure to the actual experimental stimuli. However, the dart throwing process was used to ensure that a space was left for the mean mark. This additional space may have skewed the distributions, but we were not concerned about this for the tutorial and practice images. Figure 3 shows an example stimulus generated by this method, as well as showing an example of a mean mark of the type seen by participants in practice images.

3.2 Crowdsourcing

We chose to use Amazon's Mechanical Turk crowdsourcing platform to conduct our experiments – previous research has shown that Turk offers a participant pool that is more diverse than would be recruited from a college campus [6,47], and that the relatively quick turn around of Turk studies fits our model of performing a large number of iterative, somewhat contingency-based studies. With proper care in experimental design to avoid "click through" or other cheating behavior, Turk studies can be a reliable source of human subjects data [43], and data for the analysis of the efficacy of designs in information visualization specifically [32, 35].

For our experiments only Turkers from the United States were eligible to participate, and once a participant completed an experiment they were added to a blacklist such that were not eligible for any of the other experiments discussed in this paper. To prevent click-through behavior we randomized question order, included non-obtrusive validation questions, and split questions across multiple pages and response types (e.g. binary choices, text fields, constrained response). Participants were paid at a standard rate ($6 per hour) for the estimated time the study took (10 minutes ($1) for the between subjects experiments, 15 minutes ($1.50) for the within subjects experiments). Participants completing the study were paid for the full expected time, even though most completed more quickly.

3.2.1 Demographics

We recruited 778 participants in total, 453 (58%) men and 325 (42%) women. Ages ranged from 18-65 (μ_{age} = 32.6, σ_{age}=10.5). 117 participants were recruited for pilot studies and other test experiments that are not included in the final analysis. While the age data seem close to the expected values of U.S. Turkers as a whole, our gender ratios do not match the self-reported demographics of U.S. Turkers as measured in previous studies [47], indicating either a shift in demographics for Turk as a whole, or a recruited population with a different profile than in previous census tasks (as an example our task was higher paying than tasks designed just for self-reporting of demographic data, which may attract a different participant pool).

4 Experiments

In Section 1.1, we made predictions about the mean comparison task. These lead to a set of hypotheses:

1. Our parameter Δ would be a useful metric for task hardness - across all experiments, Δ would be positively correlated with overall performance at the mean estimation task.
2. Within reasonable bounds, increasing the number of points would not significantly hurt performance.
3. Color, as a very strong cue, would have higher performance than other choices for primary cue, such as shape or orientation.
4. Using multiple cues to redundantly encode class membership would not significantly help performance - since efficient selection is accomplished using a single feature, so making the selection easier would not translate to improved accuracy.
5. Having a second cue which is non-informative as to class membership would not hurt performance, for similar reasons.
6. Adding additional classes which were non-informative to the binary forced choice would also not hurt performance.
7. A sufficient diversity of distractions will impair performance.

We ran an initial set of between-subjects experiments to confirm or disconfirm these hypotheses. Since many of our hypotheses (2,4,5,6) were suppositions about negative results, we also performed a set of within-subjects experiments where negative results would be stronger statements about the actual difference in means.

4.1 Experiment Set One (Between-Subjects)

We performed a series of eleven disjoint experiments that we treat as one between-subjects experiment [2] in order to initially explore the parameter space for encodings and confirm our hypotheses about the pre-attentive aspects of the mean estimation task. In each experiment participants were exposed to a single type of multiclass scatterplot.

Participants were shown 39 total stimuli in random order - six stimuli from each of six different levels of our proposed hardness parameter Δ (the pixel difference between means of the two classes in the scatterplot): 8, 16, 24, 32, 40, and 80 pixels. In addition there were three questions with a Δ of 0 pixels – that is, both classes had identical means. The $\Delta = 80$ questions were used for validation purposes – if participants did not get more than 50% of these questions correct then they were excluded from analysis. The $\Delta = 0$ questions were used to determine if there was systematic bias towards one answer or another (the expected distribution of answers should be approximately even at this level, since the participants would be essentially guessing). Both the $\Delta = 80$ and $\Delta = 0$ questions were otherwise excluded from the main analysis. 32 participants were recruited for each experiment. Participants not meeting the inclusion criteria were removed entirely from the main analysis, but no additional participants were recruited. Ultimately 40 exclusions were made using this criterion, out of a total participant pool of 352 people.

We performed eleven experiments with this general model, each with a different choice of class encoding. Figure 4 summarizes these choices of encoding. We had three main groups of encoding: one in which hue was used to encode class membership (in this case one class

[2] We note that this was a sequential series of experiments, and not a proper single between-subjects experiment, as we ran different conditions on different days at different times. Since the participant pool of Mechanical Turkers may vary widely depending on time of day, this was a potential source of variance in our results which we could not adequately model, although experiments were run at similar times each day.

Label	Encoding	Glyphs	Notes	Accuracy
a)	Hue	vs.	–	79.4%
b)	Hue	vs.	–	80.6%
c)	Hue	vs.	75 points per class	84.5%
d)	Hue	vs.	distractor points	83.8%
e)	Hue	vs.	Shape as redundant cue	82.8%
f)	Hue	vs.	Shape as conflicting cue	79.8%
g)	Hue	vs.	Orientation as conflicting cue	84.2%
h)	Hue	vs.	distractor points	80.0%
i)	Luminance	vs.	–	79.2%
j)	Shape	vs.	–	72.9%
k)	Shape	vs.	Color as conflicting cue	71.4%

Fig. 4. User performances in the first block of between-subjects experiments. Unless otherwise noted, participants saw 50 points of each class of glyph and were asked to guess which class was on average higher. Using the results of a Tukey's HSD, light green rows were statistically indistinguishable from each other but significantly different from the orange rows, and vice versa.

was orange and the other purple), shape encodings of class membership (one class had circular glyphs and the other triangular glyphs), and a single experiment where luminance was used (one class with light gray glyphs and the other dark gray glyphs). We also investigated the effect of having additional encodings layered on top of the main class encoding, which either supported the main cue (they were redundant with the main encoding) or provided no information (they conflicted with the main encoding).

4.1.1 Results

Our results across all experiments in this set confirmed hypothesis 1: as Δ increased, participants performed monotonically better. Figure 5 presents these results broken down by Δ. A one-tailed t-test confirmed that even at our lowest sampled level of Δ, participants still performed significantly better than chance (t(1868)=9.89, $p(\mu_x < 0.5) < 0.0001$.

Hypothesis 3 was also confirmed by this set of experiments. We conducted a two-way Analysis of Variance (ANOVA) to determine the effect of our eleven encoding/secondary cue choices on performance. We found a significant effect of this factor on performance (F(10,9328) = 9.91, $p < 0.001$). Post hoc analysis using a Tukey test of Honest Significant Difference (HSD) revealed two clusters of performance: one cluster where color (hue or luminance) was used as the main cue, which significantly outperformed the cluster where shape was used as the main cue.

This clustering also provides evidence that causes us to fail to reject our other hypotheses (2,4,5, and 6). There was no significant difference between performance when the number of points per class was increased from 50 to 75 (experiment 1c), nor when additional classes of points were included (experiment 1d,1h), or when secondary cues where used (either redundantly as in experiment 1e or in conflict with the main cue as in experiments 1f and 1k). While these similar levels of performance provided some evidence of the validity of these hypotheses, we decided that a between-subjects experiment was an insufficiently powerful model to capture these negative results. Figure 4 presents the results of experiment block one in detail.

4.2 Experiment Set Two (Within-Subjects)

In order to reconfirm our results with a stronger model of participant variance, as well as explore the inter-relation between more specific classes of encodings, we performed nine within-subjects experiments in which one participant saw two different sets of scatterplots with different sets of encodings.

The task and parameters of the stimuli were similar to the first set of experiments. One difference is that participants were given two "blocks," each containing 36 different scatterplots, for a total of 72 questions. In each block there were an equal number of stimuli with the following pixel differences (our Δ parameter) between the per-class

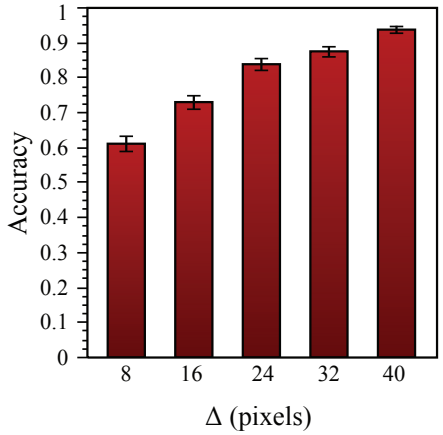


Fig. 5. The effect of our proposed hardness parameter Δ on hardness across all of our between-subjects experiments. Δ represents the difference, in pixels, between the mean values of each class in a particular scatterplot. As Δ increases the comparative judgement of which class of points is on average higher becomes easier.

means: 4, 12, 20, 28, 36, and 80. These Δ values were chosen to provide more information about the more difficult questions while still containing enough easier questions to not discourage participants. The $\Delta = 80$ questions were used for validation but otherwise excluded from analysis: if participants got 50% or fewer correct they were excluded. One block was a "baseline" block with a simpler choice of encoding, and the second had a more difficult encoding scheme. The presentation order was counter-balanced across all participants such that an equal number was exposed to each presentation order. If an exclusion was made additional participants were recruited dynamically, so that for the final analysis 16 participants were exposed to each presentation order for a total of 32 participants. 27 additional participants were recruited for this purpose, for a total subject pool of 319 people.

The choices of main encoding across these experiments, as well as the choices of difference between the first and second blocks, were aligned to provide additional evidence for hypotheses 2,4,5, and 6, for which we had some evidence of validity from our previous set of experiments, but no concrete measures of interaction between the baseline condition and individual factors (increasing the number of points, adding redundant cues, adding conflicting cues, and adding additional classes respectively). Figure 6 lists these choices of encoding in detail.

Label	Encoding	Block 1			Block 2			p-value
		Notes	Glyphs	Accuracy	Accuracy	Glyphs	Notes	
a)	Color	–	vs.	**79.5%**	75.5%	vs.	distractor points	$p = 0.03$
b)	Color	–	vs.	76.8%	**80.1%**	vs.	Shape as conflicting cue	$p = 0.07$
c)	Shape	–	vs.	**76.4%**	72.9%	vs.	Color as conflicting cue	$p = 0.07$
d)	Color	–	vs.	75.8%	**78.9%**	vs.	75 points per class	$p = 0.09$
e)	Color	distractor points	vs.	81.2%	79.7%	vs.	distractor points	$p = 0.39$
f)	Color	–	vs.	78.2%	76.9%	vs.	distractor points	$p = 0.46$
g)	Color	–	vs.	80.7%	79.7%	vs.	Color as redundant cue	$p = 0.55$
h)	Shape	–	vs.	78.5%	78.3%	vs.	Color as redundant cue	$p = 0.91$
i)	Shape	–	vs.	75.9%	75.7%	vs.	Color as conflicting cue.	$p = 0.91$

Fig. 6. An overview of results from our within-subjects experiments. Participants were presented with two different classes of stimuli in discrete blocks (presentation order was counterbalanced across participants). Unless otherwise noted, each stimulus had 50 points of each class. Statistical significance was determined via a two-way ANOVA. Green rows indicate significant difference ($p < .05$) in performance between the two blocks, light green indicates statistically marginal difference ($.1 > p > .05$), white indicates no significant difference between blocks.

4.2.1 Results

This block of experiments provided more evidence for hypothesis 2 (that additional points would not make the mean estimation task more difficult). We performed a repeated measures ANOVA (rANOVA) on experiment 2d to determine the effect of block (where one block had 75 points per class, and the other had 50 points per class) on performance. There was a marginal effect of adding additional points on performance ($F(1,2110)=2.83$, $p = 0.09$), but as the number of points increased performance was marginally *higher* (accuracy of 75.8% when there were only 50 points per class versus 78.9% when there were 75 points per class), not lower (as one would expect if adding additional points hurt performance).

Our experiments also provided more evidence for hypothesis 4: experiments 2g and 2h both dealt with redundancy, using color and shape simultaneously to encode class membership. Two rANOVAs were performed on experiments 2g and 2h to determine the effect on the inclusion of redundancy on performance. Performance was not significantly different comparing these redundant encodings to color encoding alone (2g, $F(1,1917)=0.36$, $p = 0.55$) or shape encoding alone (2h, $F(1,1917)=0.01$, $p = 0.91$).

Likewise experiments 2a, 2b, 2c, 2e, and 2i all featured conflicting cues, providing evidence concerning hypothesis 5. In experiments 2b, 2c, and 2i the addition of a cue which conflicted with the main class membership cue was the only difference between blocks. rANOVAs were performed on each of these experiments. For 2i (where the main cue was shape, and color of stroke was added as a conflicting cue) there was no significant effect of conflict on performance ($F(1,1917)=0.01$, $p = 0.91$). There were marginal effects of performance for 2b, where color was the main cue and shape was used as a distractor ($F(1,1917)=3.2$, $p = 0.07$), but it was the *presence* of conflict where performance was higher (80.1% accuracy where shape conflict was present, 76.8% where shape conflict was absent). Experiment 2c had shape as the primary cue with color as conflict: there was also a marginal effect of conflict on performance ($F(1,1917)=3.2$, $p = 0.07$), but in this case it was the absence of conflict where performance was marginally better (76.4% accuracy where there was no color conflict vs. 72.9% accuracy where conflict was present).

Experiments 2a, 2e, and 2f all featured a third distracting class of points that were meant to provide evidence for hypothesis 6. In experiment 2f the addition of a third class was the only difference between blocks. We performed an rANOVA on experiment 2f to test for the effect of the presence of an additional class of points on performance. We found no significant effect ($F(1,1917)=0.54$, $p = 0.46$).

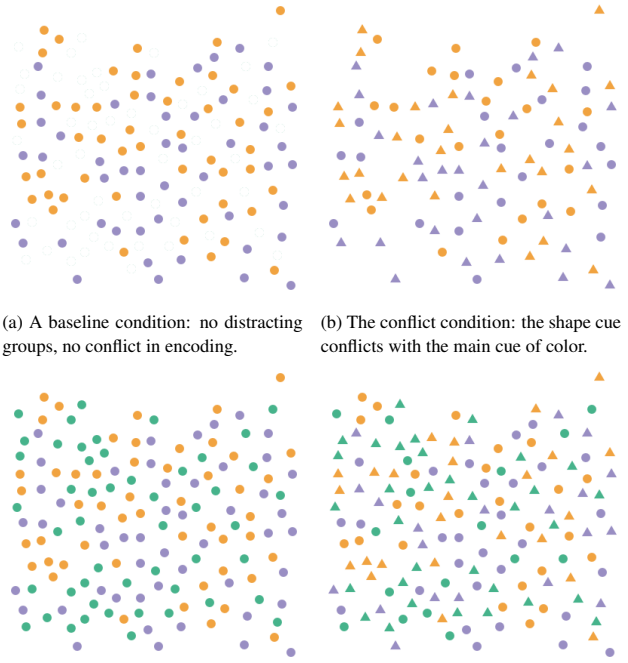
Experiments 2a and 2e dealt with the conjunction of hypotheses 5 and 6, and provide evidence for hypothesis 7. Experiments 2b and 2f discussed previously provided evidence that the addition of a shape as a conflicting cue, and the addition of another class of points, were by themselves not significant effects on performance. Likewise, an rANOVA performed on experiment 2e to check of the effect of shape conflict once a distractor class is already present on performance showed no significant effect ($F(1,1917)=0.74$, $p = 0.39$). Only when the jump was made from stimuli with neither additional classes nor shape conflict to stimuli where both were present (as in experiment 2a) was there a significant negative effect on performance (performance of 79.5% when no cue conflict or additional classes were present versus 75.5% when both cue conflict and additional were present). An rANOVA confirmed the statistical significance of this effect ($F(1,1917)=4.53$, $p < 0.03$). This is consistent with hypothesis 7.

5 Discussion

The mostly negative results of the first experiment set speak to the efficiency of this task: even with significant inter-participant variance from a diverse participant pool, and with a wide diversity of cues, aggregate performance across conditions differed only by as much as 13.1%. Within-cue variability was even lower: aggregate differences of 5.1% across eight experiments when hue was used as encoding, and 1.5% for both experiments where shape was used as the encoding.

Our experiments 2a, 2b, 2c, 2e, and 2i initially seem ambiguous as to the effect of conflicting encodings (additional cues beyond the main cue). While 2b, 2e, and 2i show no harm (or a marginal positive benefit) in the presence of conflicting encodings, 2a and 2c seem to point to a negative impact of cue conflict on performance. While marginal, the results of 2c might speak to our hypothesis that color is a much stronger cue for selection than other cues like shape (which is reflected in the literature, where color selection is more efficient in terms of accuracy and precision compared to many other potential choices of encoding [12, 56]). This would also explain why experiment 2i, which was very similar in design and choices of cue, did not see a similar degradation in performance – by using stroke color rather than fill color, the salience of color as a potentially distracting cue is reduced.

Experiment 2a represents an upper end to the complexity of the task: neither conflict (as in 2b) nor distractor classes (as in 2f) by themselves are sufficient to negatively affect performance. Figure 7 shows this visual complexity of these various levels – resulting in a stimulus which is very dense and very visually complex. Visual complexity (and specifically visual clutter) has been shown to lead to errors in judgement [4]. It is important to note that simply adding a few extra points does not add as much visual clutter as introducing visual heterogeneity to the existing points [46], which contributes to the negative result in experiment 2d. While it is likely that there is an upper limit (in terms of number of additional conflicting cues, or number of additional distracting classes) that would substantially reduce performance for the averaging task, in most real world settings the number



(a) A baseline condition: no distracting groups, no conflict in encoding.

(b) The conflict condition: the shape cue conflicts with the main cue of color.

(c) The distractor condition: there is an extra class of green points

(d) Both conflicting cues and a distractor class are present.

Fig. 7. An overview of the positive result in our within-subjects experiment. Performance was marginally better when conflicts in shape (7a vs. 7b) were included and statistically indistinguishable when distractor classes were included (7a vs. 7c), or when shape conflict was introduced to stimuli where a distractor class was already present (7c vs. 7d). Only when both potentially harmful conditions were introduced simultaneously (7a vs. 7d) was there a statistically significant impact on performance.

of classes to be distinguished are relatively small, and the number of dimensions which are simultaneously to be encoded is also small.

In experiments 2g and 2h, we found that redundant encodings (shape and fill color; shape and stroke color) did not improve performance. This contrasts with both common wisdom and past empirical findings that Ware [59] summarizes with the guideline "To make symbols in a set maximally distinctive, use redundant coding wherever possible" (guideline 5.11). Future work should re-evaluate this guideline.

These findings have ramifications for the design of multi-class scatterplots. First, viewers are capable of making judgements about the difference between classes, even when there are many points and the differences are small. This suggests that scatterplots can convey the inter-class differences without explicitly showing the means. The benefits of scatterplots in showing the data (e.g. distributions and trends) also afford communicating aggregate properties. Second, because conflicting cues do not hinder performance in the assessment of aggregates, layering information in multiple cues (e.g. using both color and shape to encode different properties) is likely to be an effective strategy. Third, because distractor classes have little effect on performance, multi-class scatterplots should not necessarily be avoided in favor of simpler ones. These guidelines are qualified by a number of limitations. For example they assume that the viewer has ample time to view the display and that the classes of points are sufficiently distinct. Other limitations are considered in the next section.

5.1 Limitations and Future Work

We have tried to generate data in ways that preclude dominant strategies where the experimental participants can easily figure out the correct answers, for example by crafting our data so that choosing the maximal points does not work. It is impossible to know that such a strategy does not exist. We did ask participants to self-report their strategies at the end of the experiment, although this is generally a poor assessment of how they actually do the task. These self-reports did not reveal any clear dominant strategy (except for the top-most strategy in the pilot study). However, if the participants were able to develop a strategy over the course of the experiment (without us providing feedback, and without them showing learning effects), they would probably develop similar strategies in realistic tasks as well. Indeed, some of the strategies that a participant might apply still work within our model, as they require pre-attentive selection and aggregation.

Our experiments only consider a single distractor task. In theory, our model suggests that more distractor classes should not hinder performance. In practice, however, this is difficult to test or exploit: as more classes are added, they are (necessarily) less distinctive. This loss of distinctness of classes would cause a degradation of performance, even if the increase of the number of classes does not.

A key limitation of our study is that we do not manipulate timing. While the lack of hard time constraints may be more realistic, our experiments cannot explain what a viewer does with this time, or even show that having time makes a difference. Participants chose to spend more time than they were given in prior studies of rapid response tasks, despite the fact that, as crowd-workers, they have financial incentive to finish their tasks quickly.

Our data show that the same attentional limits that apply in rapid response tasks also apply in the non-time-constrained mean comparison task. Given this, we might wonder how more time might be helpful. Viewers choose to take more time when asked to focus on accuracy, presumably because they believe it will help, and we have initial pilot data that suggests performance does vary with time. We do not believe this is a contradiction: while more time cannot improve the performance of the attentional mechanisms, it does give a viewer the opportunity to use these mechanisms differently.

The overall task of making mean judgements from a scatterplot is a higher-level task than considered in the perception literature. Breaking the task into smaller subtasks presumes a model of what the subtasks are and how they integrate. Sections 1.1 and 2 already describe some of the subtasks: selecting individual collections and constructing average position values from each. Our data are consistent with these mechanisms being involved: their limitations can be seen in performance on the compound task.

The literature suggests that the subtasks may be assembled as a serial process - one feature value (e.g. red, or circles), and thus one collection, is selected at a time. This property stems from theories and supporting evidence that keeping representations of different objects or subsets of objects separated requires that they be isolated over time. Selecting multiple things at once mixes their properties within the visual system's representation [44, 55]. Thus, processing the individual group centers of two collections required that each group be isolated serially [33, 37]. Without such serial selection, these theories predict that the visual system could only provide the center of the entire superset, which is a useful statistic for many purposes, but not for determining the higher collection. For the same reasons, determining the spatial relationship between the two collection centerpoints (is collection A higher than collection B) also likely requires serial selection over time [25, 48]

The serial selection process suggests theories of what a viewer may do with more time. For example, they may make several passes of selection across each feature dimension, selecting one group of points by color, abstracting its mean value, and then doing the same for the other group. For novices at the task, extended amounts of time may be beneficial as they develop routines of serial selection, and they become even more efficient at ignoring otherwise salient information from distractor collections. Even with practice, repeating this serial cycle may allow even expert observers to increase their accuracy for the collection difference, by combining information from several judgements, especially given that spatial memory for the means is likely to be noisy and time-limited.

A future goal is to better understand the specific subtasks as well as the mechanisms by which they are combined. We have begun pi-

lot explorations, both by considering the subtasks independently and by better instrumenting our observations of performance of the compound tasks. While our present studies focus on aggregation in scatterplots, we hope to consider other tasks to develop and validate a general model of non-time-constrained performance in visual aggregation tasks.

6 Conclusion

In this paper, we have studied empirically the human ability to meaningfully, efficiently, and accurately compare average value in multi-class scatterplots. Using stimuli with longer exposure times and more varied difficulty levels than in previous work we show that, within reasonable limits, this ability is robust across scatterplots with differing numbers of points per class, additional distracting classes, and with additional conflicting cues. Encoding data with redundant cues, which common wisdom would suggest would be helpful for tasks where users must select individual classes from a group, is likewise not a factor in performance. We consider limitations of the mechanisms of attentional selection that have been established in simpler tasks under time constraints, and show that they apply in this compound task without time pressure. We believe that our methodology, and our model of assembling more basic subtasks to achieve compound performance, is extensible to a wide range of common tasks in information visualization, where users must extract and possibly compare the aggregate statistics of different classes in a display.

Acknowledgments

This work was supported in part by NSF awards CMMI-0941013, BCS-1056730, SBE-1041707, DRL-0918409, DRL-1247262, and IIS-1162037 and NIH award R01 AU974787.

References

[1] G. A. Alvarez and A. Oliva. The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, 19(4):392–398, 2008.

[2] G. A. Alvarez and A. Oliva. Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences of the United States of America*, 106(18):7345–50, May 2009.

[3] D. Ariely. Seeing sets: Representation by statistical properties. *Psychological Science*, 12(2):157–162, 2001.

[4] S. Baldassi, N. Megna, and D. C. Burr. Visual clutter causes high-magnitude errors. *PLoS biology*, 4(3):e56, 2006.

[5] A. V. Belopolsky, L. Zwaan, J. Theeuwes, and A. F. Kramer. The size of an attentional window modulates attentional capture by color singletons. *Psychonomic bulletin & review*, 14(5):934–8, Oct. 2007.

[6] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[7] D. V. Buonomano and M. M. Merzenich. Cortical plasticity: from synapses to maps. *Annual review of neuroscience*, 21:149–86, Jan. 1998.

[8] T. C. Callaghan. Dimensional interaction of hue and brightness in preattentive field segregation. *Perception & psychophysics*, 36(1):25–34, July 1984.

[9] T. C. Callaghan. Interference and dominance in texture segregation: hue, geometric form, and line orientation. *Perception & psychophysics*, 46(4):299–311, Oct. 1989.

[10] S. C. Chong and A. Treisman. Statistical processing: computing the average size in perceptual groups. *Vision research*, 45(7):891–900, Mar. 2005.

[11] H. Choo, B. R. Levinthal, and S. L. Franconeri. Average orientation is more accessible through object boundaries than surface features. *Journal of Experimental Psychology: Human Perception and Performance*, 38(3):585, 2012.

[12] R. E. Christ. Review and Analysis of Color Coding Research for Visual Displays. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 17(6):542–570, Dec. 1975.

[13] W. S. Cleveland and W. S. Cleveland. A color-caused optical illusion on a statistical graph. *The American Statistician*, 37(2):101–105, 1983.

[14] M. Correll, D. Albers, S. Franconeri, and M. Gleicher. Comparing averages in time series data. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems (CHI 2012)*, pages 1095–1104. ACM, 2012.

[15] M. Correll, E. Alexander, and M. Gleicher. Quantity estimation in visualizations of tagged text. In *Proceedings of the 2013 ACM annual conference on Human Factors in Computing Systems (CHI 2013)*, pages 2697–2706. ACM, 2013.

[16] I. S. Dhillon, D. S. Modha, and W. Spangler. Class visualization of high-dimensional data with applications. *Computational Statistics & Data Analysis*, 41(1):59–90, Nov. 2002.

[17] M. E. Doherty, R. B. Anderson, A. M. Angott, and D. S. Klopfer. The perception of scatterplots. *Attention, Perception, & Psychophysics*, 69(7):1261–1272, 2007.

[18] M. D'Zmura. Color in visual search. *Vision research*, 31(6):951–966, 1991.

[19] H. E. Egeth, C. J. Leonard, and A. B. Leber. Why salience is not enough: reflections on top-down selection in vision. *Acta psychologica*, 135(2):130–2; discussion 133–9, Oct. 2010.

[20] J. T. Enns and R. A. Rensink. Influence of scene-based properties on visual search. *Science*, 247(4943):721–3, Feb. 1990.

[21] G. Fouriezos, S. Rubenfeld, and G. Capstick. Visual statistical decisions. *Attention, Perception, & Psychophysics*, 70(3):456–464, 2008.

[22] S. Franconeri, D. Bemis, and G. Alvarez. Number estimation relies on a set of segmented objects. *Cognition*, 113(1):1–13, 2009.

[23] S. L. Franconeri. The nature and status of visual resources. In D. Reisberg, editor, *Oxford Handbook of Cognitive Psychology*, chapter 10. Oxford University Press, 2013.

[24] S. L. Franconeri, G. A. Alvarez, and P. Cavanagh. Flexible cognitive resources: competitive content maps for attention and memory. *Trends in cognitive sciences*, 2013.

[25] 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):210–27, Feb. 2012.

[26] J. Haberman and D. Whitney. Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, 17(17):R751–R753, 2007.

[27] J. Halberda, S. F. Sires, and L. Feigenson. Multiple spatially overlapping sets can be enumerated in parallel. *Psychological science*, 17(7):572–6, July 2006.

[28] L. G. H. Hardy, G. Rand, and M. C. Rittler. Tests for detection and analysis of color blindness: I. an evaluation of the ishihara test. *Archives of Ophthalmology*, 34(4):295, 1945.

[29] S. Haroz and D. Whitney. How Capacity Limits of Attention Influence Information Visualization Effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2402–2410, 2012.

[30] M. Harrower and C. Brewer. Colorbrewer. org: an online tool for selecting colour schemes for maps. *Cartographic Journal, The*, 40(1):27–37, 2003.

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

[32] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 2010 ACM annual conference on Human Factors in Computing Systems (CHI 2010)*, pages 203–212. ACM, 2010.

[33] L. Huang and H. Pashler. A boolean map theory of visual attention. *Psychological review*, 114(3):599, 2007.

[34] L. Itti and C. Koch. Computational modeling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[35] R. Kosara and C. Ziemkiewicz. Do mechanical turks dream of square pie charts? In *Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization*, pages 63–70. ACM, 2010.

[36] A. Lagae and P. Dutré. A comparison of methods for generating poisson disk distributions. In *Computer Graphics Forum*, volume 27, pages 114–129. Wiley Online Library, 2008.

[37] B. R. Levinthal and S. L. Franconeri. Common-fate grouping as feature selection. *Psychological science*, 22(9):1132–1137, 2011.

[38] J. Li, J.-B. Martens, and J. J. van Wijk. A model of symbol size discrimination in scatterplots. In *Proceedings of the 2010 ACM annual conference on Human Factors in Computing Systems (CHI 2010)*, pages 2553–2562. ACM, 2010.

[39] J. Li, J. J. van Wijk, and J.-B. Martens. A model of symbol lightness

discrimination in sparse scatterplots. *2010 IEEE Pacific Visualization Symposium (PacificVis)*, pages 105–112, Mar. 2010.

[40] A. P. Marchant, D. J. Simons, and J. W. de Fockert. Ensemble representations: Effects of set size and item heterogeneity on average size perception. *Acta psychologica*, 142(2):245–50, Feb. 2013.

[41] D. Melcher and E. Kowler. Shapes, surfaces and saccades. *Vision Research*, 39(17):2929–2946, Aug. 1999.

[42] D. P. Mitchell. Spectrally optimal sampling for distribution ray tracing. *ACM SIGGRAPH Computer Graphics*, 25(4):157–164, July 1991.

[43] G. Paolacci, J. Chandler, and P. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419, 2010.

[44] R. A. Rensink. The Dynamic Representation of Scenes. *Visual Cognition*, 7(1):17–42, 2000.

[45] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. In *Computer Graphics Forum*, volume 29, pages 1203–1210. Wiley Online Library, 2010.

[46] R. Rosenholtz, Y. Li, and L. Nakano. Measuring visual clutter. *Journal of Vision*, 7(2), 2007.

[47] J. Ross, L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*, pages 2863–2872. ACM, 2010.

[48] J. C. Roth and S. L. Franconeri. Asymmetric coding of categorical spatial relations in both language and vision. *Frontiers in psychology*, 3:464, Jan. 2012.

[49] M. Sàenz, G. T. Buraĉas, and G. M. Boynton. Global feature-based attention for motion and color. *Vision Research*, 43(6):629–637, Mar. 2003.

[50] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A Taxonomy of Visual Cluster Separation Factors. *Computer Graphics Forum*, 31(3pt4):1335–1344, June 2012.

[51] J. T. Serences and G. M. Boynton. Feature-based attentional modulations in the absence of direct visual stimulation. *Neuron*, 55(2):301–312, 2007.

[52] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, June 2009.

[53] W. Tedford Jr, S. Bergquist, and W. Flynn. The size-color illusion. *The Journal of General Psychology*, 97(1):145–149, 1977.

[54] A. Treisman. Preattentive processing in vision. *Computer vision, graphics, and image processing*, 31:156–177, 1985.

[55] A. Treisman. The binding problem. *Current Opinion in Neurobiology*, 6(2):171–178, Apr. 1996.

[56] A. Treisman and S. Gormican. Feature analysis in early vision: evidence from search asymmetries. *Psychological review*, 95(1):15–48, Jan. 1988.

[57] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, Jan. 1980.

[58] S. Van der Stigchel, A. V. Belopolsky, J. C. Peters, J. G. Wijnen, M. Meeter, and J. Theeuwes. The limits of top-down control of visual attention. *Acta psychologica*, 132(3):201–12, Nov. 2009.

[59] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, 2013.