Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design

Jeffrey Heer and Michael Bostock
Computer Science Department
Stanford University
{jheer, mbostock}@cs.stanford.edu

ABSTRACT

Understanding perception is critical to effective visualization design. With its low cost and scalability, crowdsourcing presents an attractive option for evaluating the large design space of visualizations; however, it first requires validation. In this paper, we assess the viability of Amazon's Mechanical Turk as a platform for graphical perception experiments. We replicate previous studies of spatial encoding and luminance contrast and compare our results. We also conduct new experiments on rectangular area perception (as in treemaps or cartograms) and on chart size and gridline spacing. Our results demonstrate that crowdsourced perception experiments are viable and contribute new insights for visualization design. Lastly, we report cost and performance data from our experiments and distill recommendations for the design of crowdsourced studies.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces—Evaluation/Methodology

General Terms: Experimentation, Human Factors.

Keywords: Information visualization, graphical perception, user study, evaluation, Mechanical Turk, crowdsourcing.

INTRODUCTION

"Crowdsourcing" is a relatively new phenomenon in which web workers complete one or more small tasks, often for micro-payments on the order of $0.01 to $0.10 per task. Such services are increasingly attractive as a scalable, low-cost means of conducting user studies. Micro-task markets lower the cost of recruiting participants, offering researchers almost immediate access to hundreds (if not thousands) of users. Similarly, by reducing the burden of participation, the subject pool is greatly increased and diversified [13].

The reduced cost structure of crowdsourced evaluations is particularly attractive in visualization, where the design space of possible visual encodings is large and perceptually interconnected [2, 7, 10, 19, 27, 34]. Crowdsourcing may enable experimenters to canvas a wide range of subjects using their standard displays, effectively swapping experimental control for ecological validity. Crowdsourced experiments may also substantially reduce both the cost and time to result.

Unfortunately, crowdsourcing introduces new concerns to be addressed before it is credible. Some concerns, such as ecological validity, subject motivation and expertise, apply to any study and have been previously investigated [13, 14, 23]; others, such as display configuration and viewing environment, are specific to visual perception. Crowdsourced perception experiments lack control over many experimental conditions, including display type and size, lighting, and subjects' viewing distance and angle. This loss of control inevitably limits the scope of experiments that reliably can be run. However, there likely remains a substantial subclass of perception experiments for which crowdsourcing can provide reliable empirical data to inform visualization design.

In this work, we investigate if crowdsourced experiments insensitive to environmental context are an adequate tool for graphical perception research. We assess the feasibility of using Amazon's Mechanical Turk to evaluate visualizations and then use these methods to gain new insights into visualization design. We make three primary contributions:

- We replicate prior laboratory studies on spatial data encodings and luminance contrast using crowdsourcing techniques. Our new results match previous work, are consistent with theoretical predictions [21], and suggest that crowdsourcing is viable for testing graphical perception.
- We demonstrate the use of crowdsourcing to generate new perception results. We conduct experiments investigating area judgments, chart size and gridline spacing. The results provide novel insights for optimizing display parameters.
- We analyze the performance and cost of Mechanical Turk across our experiments and distill recommendations for experimenters. For example, we find that qualification tasks and verifiable questions help ensure high-quality responses and that experimenters can accelerate the time to results by increasing the compensation level. Although we focus on evaluating visualizations, we believe these latter results generalize to a variety of crowdsourced studies.

GRAPHICAL PERCEPTION

A great deal of prior research has investigated how visual variables such as position, length, area, shape, and color impact the effectiveness of data visualizations. Following Cleveland [7], we use the term *graphical perception* to denote the ability of viewers to interpret such visual encodings

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.

and thereby decode information in graphs. Assessing the impact of visual encodings on graphical perception enables designers to optimize their visualizations and is vital to the design of automatic presentation software [21, 22].

Inspired by Bertin's [2] systematic treatment of visual variables, researchers in cartography [19, 27], statistics [7], and computer science [21] have derived perceptually-motivated rankings of the effectiveness of variables such as position, length, area, and color for encoding quantitative data. Some have further tested their predictions via human subjects experiments. For example, subjects in Cleveland & McGill's [7] seminal study were shown charts and asked to compare the values of two marks by estimating what percentage the smaller value was of the larger. This accuracy measure was then used to test and refine the ranking of visual variables.

Many researchers have applied experimental methods to graphical perception tasks, for example to test differences across chart types [28, 29], shape discrimination in scatter plots [17, 31], and the effects of viewing angle and perspective distortion [36]. These studies measure how an individual visual encoding variable affects the accuracy and/or response time of estimating values of the underlying data.

Researchers have also investigated interactions between visual variables [10, 34]. Viewers decode *separable dimensions* such as position and shape largely independently, while perception of *integral dimensions* such as color hue and saturation are correlated [34]. For example, a redundant encoding using integral dimensions may incur performance improvements (e.g., *redundancy gain*) or deficits (e.g., *filtering interference*). The interaction of visual variables complicates our characterization of the design space, as extrapolating the results from studies of isolated visual variables is unreliable.

Graphical perception is also affected by other design parameters and data characteristics, including contrast effects (e.g., due to background luminance [30]), plotting density [11, 30], and changes to chart size [12], scale [6], or aspect ratio [1, 5]. Such contextual cues need not be purely visual; some studies suggest that environmental context (e.g., calm or busy [24]) or textual prompts priming specific visual metaphors [37] may also affect the decoding of visualized data.

The above considerations reinforce the need for empirical assessment of visualizations to validate theory, replicate prior results, and evaluate real-world applications. We aim to establish the viability of crowdsourcing as a low-cost adjunct to laboratory experiments. Moreover, as visualizations become increasingly prominent online [33, 35], web-based experimentation may improve ecological validity by reaching a diverse population of subjects and display configurations.

WEB-BASED EXPERIMENTS AND MECHANICAL TURK

The web is increasingly being used for experimentation and research. For example, by silently presenting different interfaces to randomized subsets of users, companies study the impact of changes on user behavior through log analysis. Kohavi et al. [15] provide a brief survey of experiments and recommendations for web experiment design. Web-based experimentation is increasingly popular and accepted in social psychology [16], including research on the development of cultural markets [25] and the manipulation of incentives for online peer-production [4, 18].

In this work, we investigate the viability of crowdsourcing graphical perception experiments. To do so, we conducted a series of experiments on Amazon's Mechanical Turk (MTurk), a popular micro-task market. On MTurk, *requesters* post jobs (called *Human Intelligence Tasks* or *HITs*) for consideration by a pool of *workers* colloquially referred to as *Turkers*. Each HIT has an associated *reward*—typically a micro-payment of $0.01 to $0.10—and a set number of *assignments*—the maximum number of Turkers who can perform the task. HITs may also require one or more *qualifications*, such as having 95% or better HIT acceptance or successfully completing a quiz. Workers discover HITs through a keyword search interface that supports task previews and from which workers can elect to complete any number of tasks. The requester pays the workers for completed tasks, but retains the ability to reject responses deemed invalid. At any time MTurk has thousands of active HITs; at the time of writing the number was 97,212.

MTurk provides a convenient labor pool and deployment mechanism for conducting formal experiments. For a factorial design, each cell of the experiment can be published as an individual HIT and the number of responses per cell can be controlled by throttling the number of assignments. Qualification tasks may optionally be used to enforce practice trials and careful reading of experimental procedures. The standard MTurk interface provides a markup language supporting the presentation of text, images, movies, and form-based responses; however, experimenters can include interactive stimuli by serving up their own web pages that are then presented on the MTurk site within an embedded frame.

Recent research has investigated the use of MTurk for crowdsourcing labor, including user studies. Kittur et al. [14] used MTurk for collecting quality judgments of Wikipedia articles. Turker ratings correlated with those of Wikipedia administrators when the tasks included verifiable questions and were designed such that completing them meaningfully is as easy as not. Mason & Watts [23] studied the effect of compensation level for image sorting and word puzzle tasks. They found that raising the reward for each HIT increased the quantity of individual responses but not the quality (e.g., accuracy) of the work performed. The implication is that paying more results in faster, though not better, results.

Mechanical Turk has also been applied to perception experiments. Cole et al. [8] studied shape perception of 3D line drawings by asking Turkers to orient gauge figures indicating surface normals. They collected 275,000 gauge measurements from 550 Turkers, which they used to evaluate rendering techniques. Compensation and collection time were not reported, and the study did not validate the use of MTurk via comparison to results collected in a laboratory.

RESEARCH GOALS

Our first research goal was to assess the viability of crowdsourced graphical perception experiments by replicating previous laboratory-based studies. To cover a suitably inter-

esting set of perceptual tasks, we replicated Cleveland & McGill's [7] classic study (Exp. 1A) of proportionality estimates across spatial encodings (*position*, *length*, *angle*), and Stone & Bartram's [30] alpha contrast experiment (Exp. 2), involving transparency (luminance) adjustment of chart grid lines. Our second goal was to conduct additional experiments that demonstrate the use of Mechanical Turk for generating new insights. We studied rectangular area judgments (Exp. 1B), following the methodology of Cleveland & McGill to enable comparison, and then investigated optimal chart heights and gridline spacing (Exp. 3). Our third goal was to analyze data from across our experiments to characterize the use of Mechanical Turk as an experimental platform.

In the following four sections, we describe our experiments and focus on details specific to visualization. Results of a more general nature are visited in our performance and cost analysis; for example, we delay discussion of response time results. Our experiments were initially launched with a limited number of assignments (typically 3) to serve as a pilot. Upon completion of the trial assignments and verification of the results, the number of assignments was increased.

EXPERIMENT 1A: PROPORTIONAL JUDGMENT

We first replicated Cleveland & McGill's seminal study [7] on Mechanical Turk. Their study was among the first to rank visual variables empirically by their effectiveness for conveying quantitative values. It also has influenced the design of automated presentation techniques [21, 22] and been successfully extended by others (e.g., [36]). As such, it is a natural experiment to replicate to assess crowdsourcing.

Method

Seven judgment types, each corresponding to a visual encoding (such as *position* or *angle*) were tested. The first five correspond to Cleveland & McGill's original position-length experiment; types 1 through 3 use position encoding along a common scale (Figure 1), while 4 and 5 use length encoding. Type 6 uses angle (as a pie chart) and type 7 uses circular area (as a bubble chart, see Figure 2).

Ten charts were constructed at a resolution of 380×380 pixels, for a total of 70 trials (HITs). We mimicked the number, values and aesthetics of the original charts as closely as possible. For each chart, N=50 subjects were instructed first to identify the smaller of two marked values, and then "make a quick visual judgment" to estimate what percentage the smaller was of the larger. The first question served broadly to verify responses; only 14 out of 3,481 were incorrect (0.4%). Subjects were paid $0.05 per judgment.

To participate in the experiment, subjects first had to complete a qualification test consisting of two labeled example charts and three test charts. The test questions had the same format as the experiment trials, but with multiple choice rather than free text responses; only one choice was correct, while the others were grossly wrong. The qualification thus did not filter inaccurate subjects—which would bias the responses—but ensured that subjects understood the instructions. A pilot run of the experiment omitted this qualification and over 10% of the responses were unusable. We discuss this observation in more detail later in the paper.

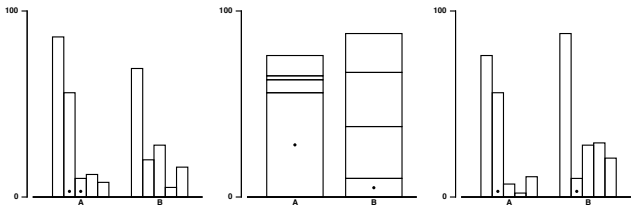

Figure 1: Stimuli for judgment tasks T1, T2 & T3. Subjects estimated percent differences between elements.

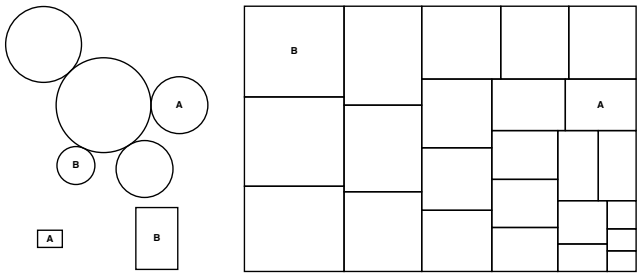

Figure 2: Area judgment stimuli. Top left: Bubble chart (T7), Bottom left: Center-aligned rectangles (T8), Right: Treemap (T9).

In the original experiment, Cleveland & McGill gave each subject a packet with all fifty charts on individual sheets. Lengthy tasks are ill-suited to Mechanical Turk; they are more susceptible to "gaming" since the reward is higher, and subjects cannot save drafts, raising the possibility of lost data due to session timeout or connectivity error. We instead assigned each chart as an individual task. Since the vast majority (95%) of subjects accepted all tasks in sequence, the experiment adhered to the original within-subjects format.

Results

To analyze responses, we replicated Cleveland & McGill's data exploration, using their log absolute error measure of accuracy: $\log_2(|\text{judged percent - true percent}| + \frac{1}{8})$. We first computed the midmeans of log absolute errors[1] for each chart (Figure 3). The new results are similar (though not identical) to the originals: the rough shape and ranking of judgment types by accuracy (T1-5) are preserved, supporting the validity of the crowdsourced study.

Next we computed the log absolute error means and 95% confidence intervals for each judgment type using bootstrapping (c.f., [7]). The ranking of types by accuracy is consistent between the two experiments (Figure 4). Types 1 and 2 are closer in the crowdsourced study; this may be a result of a smaller display mitigating the effect of distance. Types 4 and 5 are more accurate than in the original study, but position encoding still significantly outperformed length encoding.

We also introduced two new judgment types to evaluate angle and circular area encodings. Cleveland & McGill conducted a separate position-angle experiment; however, they used a different task format, making it difficult to compare

[1]The midmean–the mean of the middle two quartiles–is a robust measure less susceptible to outliers. A log scale is used to measure relative proportional error and the $\frac{1}{8}$ term is included to handle zero-valued differences.

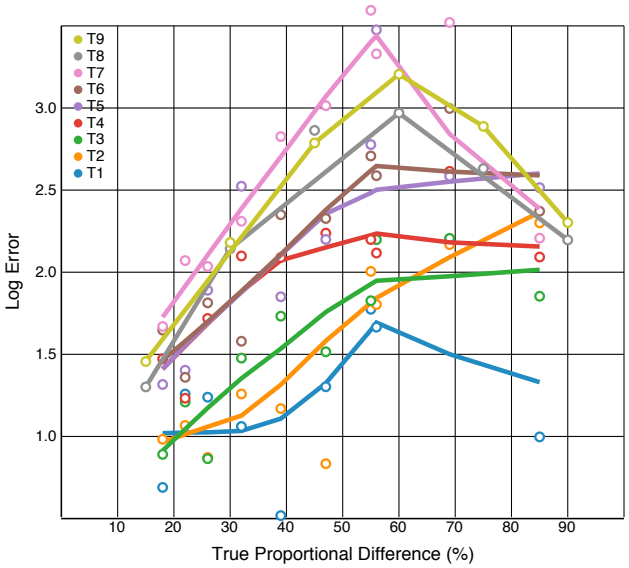

Figure 3: Midmeans of log absolute errors against true percentages for each proportional judgment type; superimposed are curves computed with *lowess*.

the results for the position-angle experiment to those for the position-length experiment. By designing judgment types 6 and 7 to adhere to the same format as the others, the results should be more apt for comparison. Indeed, the new results match expectations: psychophysical theory [7, 34] predicts area to perform worse than angle, and both to be significantly worse than position. Theory also suggests that angle should perform worse than length, but the results do not support this. Cleveland & McGill also did not find angle to perform worse than length, but as stated their position-angle results are not directly comparable to their position-length results.

EXPERIMENT 1B: RECTANGULAR AREA JUDGMENTS

After successfully replicating Cleveland & McGill's results, we further extended the experiment to more judgment types. We sought to compare our circular area judgment (T7) results with rectangular area judgments arising in visualizations such as cartograms [9] and treemaps [26]. We hypothesized that, on average, subjects would perform similarly to the circular case, but that performance would be impacted by varying the aspect ratios of the compared shapes. Based on prior results [19, 34], we were confident that extreme variations in aspect ratio would hamper area judgments. "Squarified" treemap algorithms [3, 35] address this issue by attempting to minimize deviance from a 1:1 aspect ratio, but it is unclear that this approach is perceptually optimal. We also wanted to assess if other differences, such as the presence of additional distracting elements, might bias estimation.

Method

We again used Cleveland & McGill's proportional judgment task: subjects were asked to identify which of two rectangles (marked **A** or **B**) was the smaller and then estimate the percentage the smaller was of the larger by making a "quick visual judgment." We used a 2 (display) × 9 (aspect ratios) factorial design with 6 replications for a total of 108 unique trials (HITs). In the first display condition (T8) we

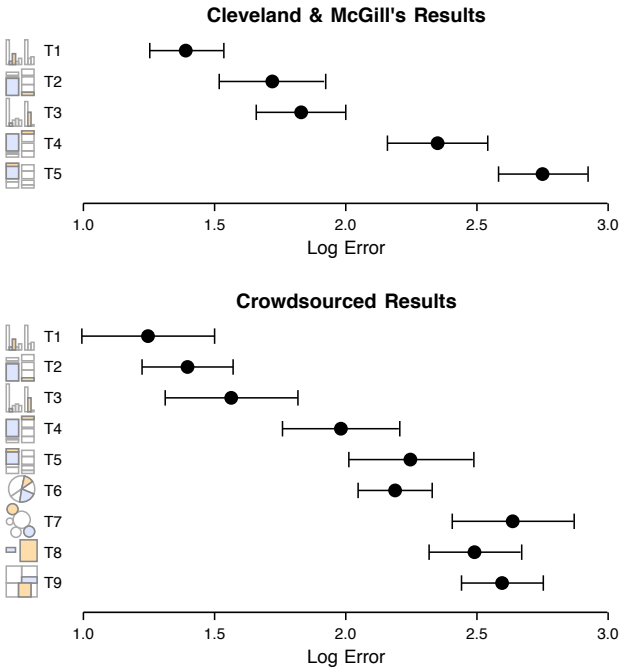

Figure 4: Proportional judgment results (Exp. 1A & B). Top: Cleveland & McGill's [7] lab study. Bottom: MTurk studies. Error bars indicate 95% confidence intervals.

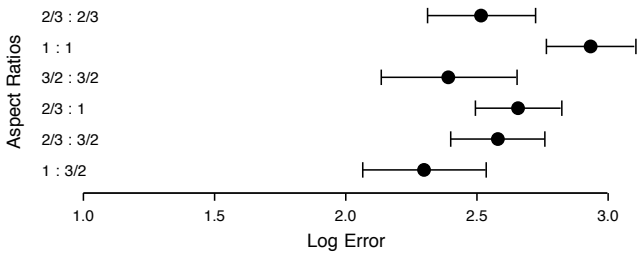

Figure 5: Rectangular area judgments by aspect ratios (1B). Error bars indicate 95% confidence intervals.

showed two rectangles with horizontally aligned centers; in the second display condition (T9) we used 600×400 pixel treemaps depicting 24 values. Aspect ratios were determined by the cross-product of the set $\{\frac{2}{3}, 1, \frac{3}{2}\}$ with itself, roughly matching the mean and spread of aspect ratios produced by a squarified treemap layout (we generated 1,000 treemaps of 24 uniformly-distributed random values using Bruls et al.'s layout [3]: the average aspect ratio was 1.04, the standard deviation was 0.28). We systematically varied area and proportional difference across replications. We modified the squarified treemap layout to ensure that the size and aspect ratio of marked rectangles matched exactly across display conditions; other rectangle areas were determined randomly.

As a qualification task, we used multiple-choice versions of two trial stimuli, one for each display condition. For each trial (HIT), we requested N=24 assignments. We also reduced the reward per HIT to $0.02. We chose this number in an attempt to match the U.S. national minimum wage (assuming a response time of 10 seconds per trial).

Results

To facilitate comparison across studies, we used Cleveland & McGill's log absolute error measure. We omitted 16 responses (0.62%), for which the subject's estimate differed from the true difference by more than 40%. Midmeans for each display type are included in Figure 3. We see a dependence on the true proportions: judgments become easier towards the extremes of the scale (0 or 100%). Confidence intervals are shown in Figure 4. The results confirm our hypothesis that, on average, the accuracy of rectangular area judgments matches that of circular area judgments.

We found a significant ($p < 0.05$) effect of aspect ratio on judgment accuracy, as shown in Figure 5. Somewhat surprisingly, comparisons of rectangles with aspect ratio 1 exhibited the worst performance, a result robust across both the rectangle and treemap display conditions. This finding suggests that viewers actually benefit from the inability of a squarified treemap algorithm to perfectly optimize the rectangles to 1:1 aspect ratios. The result is consistent with the hypothesis that viewers use 1D length comparisons to help estimate area: comparing the lengths of sides as a proxy for area leads to maximal error when comparing squares. Additional experimentation is needed to form an accurate perceptual model.

We found no significant difference between the rectangle (T8) and treemap (T9) conditions, suggesting that other elements in a treemap display do not interfere with judgment accuracy. That said, we might extend the study to comprehensively test for interference effects by including rectangles of varying color intensity. However, as we lack control over subjects' display configuration, we must first establish the reliability of crowdsourced studies involving luminance contrast. We take up this issue in our next experiment.

EXPERIMENT 2: GRIDLINE ALPHA CONTRAST

The previous experiments examined spatial encodings using black and white images. We now turn to a different set of perceptual tasks: separation and layering via luminance contrast. To do so, we replicated an alpha contrast experiment by Stone & Bartram [30] in which subjects configure the alpha (transparency) of scatter plot gridlines across variations of background darkness and plot density. The experiment seeks to bound the range of acceptable luminance contrast settings for visual reference elements such as gridlines. The results can inform smart defaults for the presentation of reference elements within display software.

As this experiment involves careful calibration of luminance contrast within visualization displays, a successful replication would help establish the utility of crowd-sourced experiments for a broader range of perception tasks. We expect monitor display settings and lighting conditions to affect the outcome of this task. While we lose control over such details when crowdsourcing, we might simultaneously gain a more representative sample of web users' displays: results may exhibit higher variance, but with means suitable for a larger user population. Accordingly, the goals of this replication were to (a) compare our crowdsourced results with those gained in the laboratory and (b) determine which display configuration details we can unobtrusively collect and assess to what degree they impact the results.

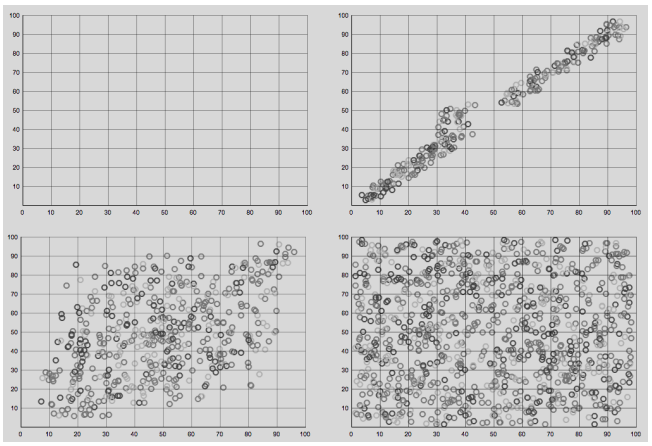

Figure 6: Density conditions for alpha contrast experiment (left-to-right): none, sparse, medium, dense.

Figure 7: Background intensities for alpha contrast experiment: #f3, #d8, #be, #a5, and #8e.

Method

We asked users to parameterize the display of chart gridlines drawn over a plotting area. In task **L**, we asked subjects, "Adjust the grid so that it is as light as possible while still being usably perceptible." In task **D**, we instructed them, "Adjust the grid strength to meet your best judgment of how obvious it can be before it becomes too intrusive and sits in front of the image; some users have called this a 'fence'."

As the experiment requires interactivity, we could not use the standard MTurk markup to create our HITs. Instead, we hosted a Flash application, presented to subjects in an embedded frame. The interface consisted of a chart display and alpha adjustment controls. "Lighter" and "Darker" buttons adjusted the alpha contrast by a value of 2 units on a 0-255 scale; holding a button resulted in an accelerated adjustment. By hosting the task ourselves, we were also able to use custom JavaScript to collect display configuration data, an option unavailable in the standard MTurk interface.

As a qualification task, subjects were asked to adjust a sample display so that the grid was fully transparent (alpha=0) or fully opaque (alpha=1), thereby ensuring that the subject could successfully run our Flash applet and adjust the grid contrast. We also considered eliciting additional display configuration information (such as monitor gamma), either by asking explicitly or with a calibration task. While a number of devices for facilitating user-provided perceptual estimates of monitor gamma exist, they are unreliable. For example, many LCD monitors are direction sensitive, with changes of viewing angle of just a few degrees causing a significant shift in perceived contrast. However, a rough estimate of gamma can be made using the web browser's "User-Agent" field to infer the operating system: most PC systems use a gamma of 2.2 while Mac OS X (prior to 10.6) uses 1.8.

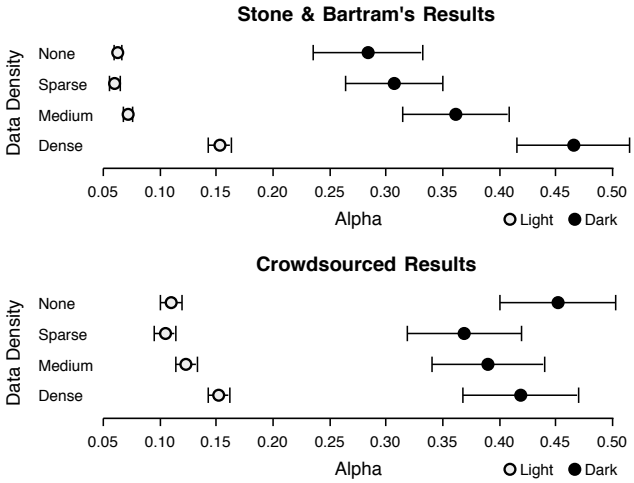

Figure 8: Alpha contrast results (2L & 2D). Top: Stone & Bartram's [30] lab study. Bottom: Our MTurk study. Error bars indicate 95% confidence intervals.

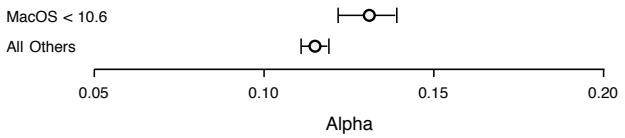

Figure 9: MTurk results for "Light" task (2L), grouped by operating system to estimate effect of monitor gamma.

We used a 5 (background) $\times$ 4 (density) factorial design with 3 replications, resulting in 60 trials (HITs) per task. Figures 6 and 7 illustrate these conditions. Each plot was sized at 450$\times$300 pixels, and displayed within a frame 700 pixels tall. The background of the frame was varied with the trial and sized to fill the majority of a standard laptop display.

For each trial, we recorded the alpha value, time to completion, and the subject's screen resolution, color depth, and browser type ("User-Agent"), as reported by JavaScript. We posted 60 HITs each for tasks L and D with N=24 assignments. Subjects were paid $0.02 per HIT.

Results

We analyzed 60$\times$24 = 1,440 responses to task L, and 1,126 responses to task D. The missing responses ($\sim$22%) to task D were due to the expiration of our HITs on MTurk; we describe the reason why later. For task L, we omitted 9 results (0.62%) for which alpha=0 or alpha$>$0.4. For task D, we omitted 4 results (0.4%) for which alpha=0 or alpha=1.

Our results are shown in Figure 8, juxtaposed with the results of Stone & Bartram. Applying analysis of variance, we found a significant effect of plot density (F(3,2413) = 3.49, p = 0.015) but not of background intensity (F(4,2413) = 0.44, p = 0.779), consistent with Stone & Bartram's findings. Alpha values in task L are higher in our experiment. Stone & Bartram note surprise at how low their values are; we surmise that crowdsourced results may be more representative of web users than a single laboratory display. Alpha values for task D have a much higher variance than those of task L, again consistent with past results. Our results corroborate Stone & Bartram's recommendation of alpha = 0.2 as a "safe" default.

We also examined the effect of display configuration on alpha values in task L. (We limited our attention to task L because it was more clearly defined and resulted in notably less variance than task D.) We found a weak positive correlation (r(1431) = 0.07, $p < 0.01$) between alpha values and screen resolution (measured in total pixels; resolutions varied from 1024$\times$768 to 1920$\times$1200). Thus as the resolution increased, users tended to make the (likely thinner) gridlines slightly darker. Unsurprisingly, we also found a negative correlation (r(1431) = -0.176, $p < 0.01$) between alpha values and monitor color depth (one of 16, 24, or 32 bits): subjects tended to select lighter alphas on displays with greater color resolution, presumably due to better contrast.

We found a significant effect of operating system (F(1,1391) = 10.24, $p < 0.001$), as determined via the browser-reported User-Agent field (Figure 9). The darker alpha values for Mac OS X prior to 10.6 (220 responses) versus other operating systems (1211 responses) are consistent with a more "washed-out" monitor gamma of 1.8, indicating that the User-Agent field provides some predictive power.

EXPERIMENT 3: CHART SIZE AND GRIDLINE SPACING

Our next experiment focuses on a design variable that is difficult to control in a crowdsourced study: visualization size. While pixel size can easily be varied, the subjects' physical display size, resolution, and viewing distance can not be measured reliably. Still, by canvassing a diversity of web users, we might determine pixel-based settings to optimize presentation. Our goal was to assess the use of crowdsourcing for experiments involving variations in chart sizing.

We investigated the effects of chart size and gridline spacing on the accuracy of value comparisons in a chart. The experiment design was inspired by Heer et al.'s [12] study of time-series visualizations, which found that as chart heights were decreased (from a starting height of 48 pixels, or 135 mm on Heer et al.'s displays), subjects initially responded more quickly without diminished accuracy, implying that there are optimal sizes that maximize the speed and accuracy of graphical perception. However, they did not investigate the effect of further increasing chart height or introducing gridlines. In this experiment, we sought to determine optimized sizing and spacing parameters for web-based display.

Method

Subjects were shown a chart and asked to first indicate which marked element (the left or the right) was smaller and then

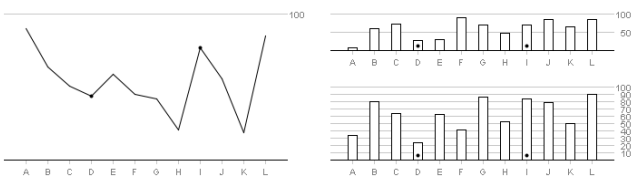

Figure 10: Experiment 3 stimuli varying chart type, chart height, and gridline spacing.

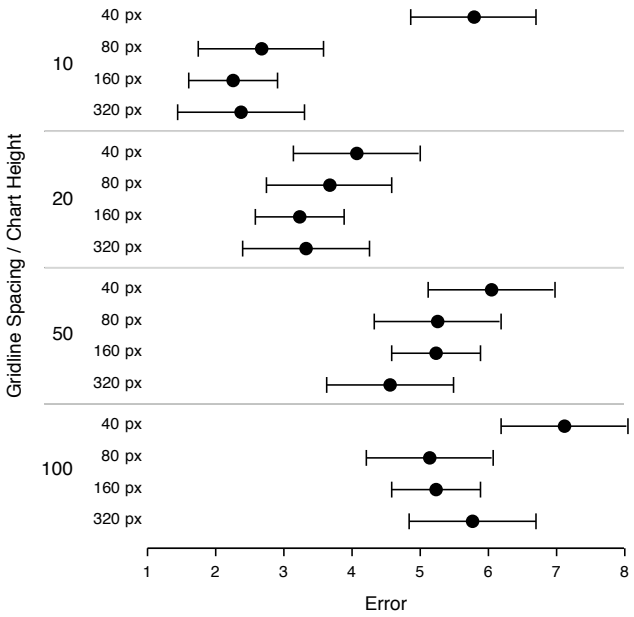

Figure 11: Chart height and gridline spacing results (3A & B). Error bars indicate 95% confidence intervals.

estimate the difference between the two (i.e., the value of the larger minus the smaller). Subjects were instructed to be as accurate as possible while making a "quick visual judgment."

We used a 2 (chart) $\times$ 3 (height) $\times$ 4 (gridline spacing) factorial design with 3 replications, for a total of 72 trials (HITs). Either a bar chart or a line chart was shown with a height of 40, 80, or 160 pixels; gridlines were drawn at intervals of 10, 20, 50, or 100 units (see Figure 10 for examples). Each chart consisted of 12 values distributed on a range of 0-100 units. Each value was labeled A-L and values D and I were always the compared values (c.f., [12]). As a qualification task, we used multiple-choice variants of two trial stimuli: one bar chart and one line chart, each 80 pixels tall.

For each experimental trial, we recorded estimation error as |judged difference - true difference|. We chose this error measure to facilitate comparison of our results with those of Heer et al. [12]; however, the unreliability of response times (discussed later) curtailed a deeper analysis of speed-accuracy trade-offs along these lines.

We requested N=24 assignments and paid $0.02 per HIT. We subsequently conducted a second experimental run, denoted as 3B. The extended experiment used chart heights of 160 and 320 for a total of 48 HITs. We again requested N=24 assignments, but raised the reward to $0.04 per HIT.

Results

We analyzed a total of 2,880 responses from the two experimental runs. We omitted 46 responses (1.60%) with error $>$ 40. We then ran an ANOVA on the error results. We found statistically significant effects for chart height ($F(3,2802) = 14.16$, $p < 0.001$), gridline spacing ($F(3,2802) = 31.98$, $p < 0.001$), and an interaction of height and spacing ($F(9,2802) = 2.11$, $p < 0.026$). Figure 11 plots these results.

Using Bonferroni-corrected post-hoc tests, we found that charts 40 pixels tall resulted in significantly more error ($p < 0.001$ in all cases), but found no significant difference between the other heights. The results confirm our hypothesis that accuracy plateaus as chart heights increase, and suggest little benefit for increasing chart height beyond 80 pixels when using a 0-100 scale. This size roughly coincides with the point at which the pixel and data resolutions match.

Adding gridlines improved accuracy, though post-hoc tests found no significant difference between 10 and 20 gridlines ($p = 0.887$) or between 50 and 100 ($p = 0.905$). Error increased steeply in charts with a height of 40 pixels and gridline spacing of 10 units. Presumably the dense packing of gridlines impedes accurate tracing to their labels. The results suggest that gridlines be separated by at least 8 pixels.

MECHANICAL TURK: PERFORMANCE AND COST

In this section, we analyze subject performance and experimental costs across our experiments, investigating subject overlap, task completion rates, quality of results, and the money and time costs of running studies on Mechanical Turk.

Turkers Overlap Across Studies

A total of 186 different Turkers participated in our experiments. Experiment 1A was launched in June 2009 as four simultaneously deployed collections of HITs grouped by judgment type. Participation across HIT groups was highly overlapping: of the 82 Turkers participating, 93% (76) contributed to multiple HIT groups and over half (45) contributed to all four. Experiment 1A consisted of a total of 70 HITs, so completing all HITs in a single session was easily achieved. The remainder of our experiments launched in September 2009 as five HIT groups, one each for experiments 1B, 2L, 2D, 3A, and 3B. HIT totals per group ranged from 48 to 108. These experiments netted 117 subjects. In our analyses we treat all experiment 1A runs as one group, as they match single HIT groups in the remaining experiments.

Figure 12 shows the cumulative distribution of Turkers by the number of experiments to which they contributed. Across experiments, 31% of Turkers (58/186) contributed to two or more experiments, and 15% (28) contributed to three or more. Only 1 Turker participated in all experiments and only 7% of Turkers (13) who participated in experiment 1A later participated in any of the other studies. In summary, there was substantial variability in the subject pool across experiments and very little overlap in studies separated by 3 months. For any given study, an average $\sim\frac{1}{3}$ of subjects also participated in another experiment.

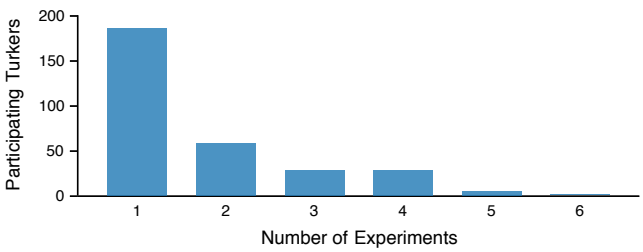

Figure 12: Cumulative number of subjects participating in our crowdsourced experiments.

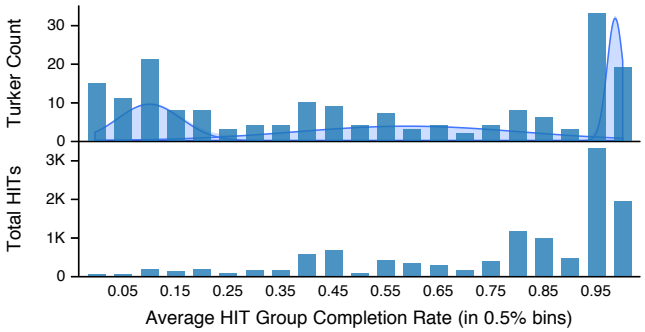

Figure 13: HIT completion results. Top: Turker count by HIT completion rate: histogram and best-fit Gaussian mixture model. Bottom: Total HITs submitted.

HIT Completion Rates: "Samplers" and "Streakers"

Do Turkers randomly sample tasks across HIT groups, or do they plow through every task in a group? Given the overhead of learning a new task, it would make economic sense to complete related tasks in batch, and the Mechanical Turk interface facilitates this process. However, we found that the number of trials completed by a subject varied substantially.

An "average" Turker completed 62 HITs (σ = 71.4) across all experiments—roughly one full study. However, as Figure 13 illustrates, the distribution of study completion rates is bi-modal. The histogram groups Turkers by their average HIT completion rate, which we calculate as the weighted percentage of HITs completed within participating HIT groups. Thus, if a Turker never participated in experiment 1B, the lack of HITs for that group is not factored into the average.

To analyze participation, we fit the data using Gaussian mixture models. A three cluster model provides the best fit according to AIC and BIC selection measures. The model confirms that Turkers cluster around low and high rates of completion. One cluster centers at a 10% completion rate, representing Turkers who sample only a few HITs in a group. The other localized cluster centers above 95% and represents Turkers who complete nearly all HITs in a consecutive streak. It is these "streakers" who do the lion's share of the work: almost half of all trials (45.7%) were completed by the 52 Turkers with an average completion rate of 95% or higher.

It is difficult to state definitively the implications of these results for study design. Certainly, these patterns do not result in strict between-subjects or within-subjects designs. However, in terms of user attention, these results suggest an interesting cross-slice of task behaviors. Real-world interfaces often have both dedicated and sporadic users, and it is possible that Turker completion patterns reflect similar distinctions. Further study is needed to evaluate these distinctions and also to assess how participation varies by task.

With Qualification, Turkers Provide High-Quality Results

Given the variety of completion rates, does the quality of Turker results vary? Overall, we found the quality of Turkers' responses to be high: rejected outliers constituted only 0.75% of responses. Though crowdsourced responses exhibited higher variance, our replicated studies (1A & 2) match prior results and imply identical design recommendations.

We found that the combined use of (a) qualification tasks to ensure subject understanding, and (b) clearly worded tasks with verifiable answers, encourages accurate crowdsourced results. Trial runs of Experiment 1 omitted the qualification task, and over 10% of the responses were unusable. We attribute this degradation in quality to confusion rather than "gaming" of the system. The use of verifiable answers (also advocated elsewhere [14]) serves to dissuade gaming, as wildly incorrect answers can be rejected outright, stripping Turkers of their pay. There is little incentive for crafting subtly incorrect answers; one might as well perform the task.

Standard HITs Frustrate Fine-Grained Timing

Although we found crowdsourcing to provide high-quality responses, the standard MTurk interface makes it difficult to collect fine-grained timing data. In a laboratory setting, we estimate that the trials in our experiments take a few seconds on average. In our crowdsourced studies, however, the average timing data was significantly higher. Rather than a few seconds per trial, the median response time was 42s (μ=54s, σ=41s). We observed a minimum time of 5 seconds, yet many responses took multiple minutes. There is simply not enough control: it is unclear how much time is due to page loading, scrolling, user inattention, and response submission.

Despite these limitations, significant effects due to time may still be found in the data. In experiment 2L, subjects spent an average of 5 extra seconds adjusting alpha contrast on dense plots ($F(3,1391) = 3.25$, $p = 0.021$). However, due to the inordinately high means and large variation, we forego making any predictions or recommendations based on such results.

If fine-grained timing is needed, experimenters should implement their own task interface and present it in MTurk as an embedded frame. One option is to maintain the typical micro-task format, but include "ready-set-go" phases at the beginning of each task and record response times using JavaScript. Another option is to use a "macro-task" format by batching a number of trials into a single HIT with higher compensation. While such a format might enforce within-subjects participation, pacing, and timing accuracy more similar to a lab study, it violates standard usage. Further study is needed to assess how such "macro-tasks" impact the performance and scalability of crowdsourced experiments.

Reward Level Affects Study Completion Time

How long does it take to run an MTurk study? Are completion time or result quality affected by the reward? For each experimental run, Figure 14 plots HITs completed vs. time elapsed since launch. Runs priced $\geq$ \$0.04/HIT are

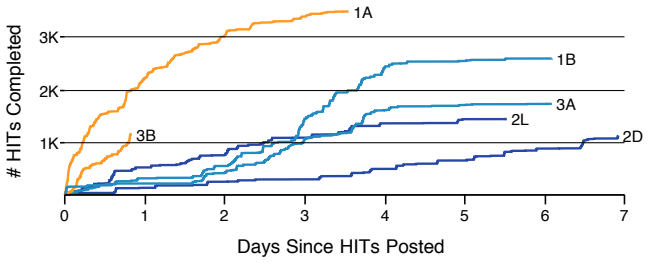

Figure 14: HIT completion rate per experiment.

shown in orange, those priced at $0.02 in shades of blue. Note the steeper slope, and thus faster completion rate, for tasks with higher rewards. One complicating factor is that the low-reward studies were launched on a holiday; however, the pattern holds even if timing values are shifted by one day.

We note that submissions in Experiment 2D lag those of 2L; this resulted in HITs for 2D expiring prior to completion. We attribute the lag to a naming error: the HIT titles for tasks 2L and 2D included the words "Part 1" and "Part 2", respectively. Turkers took us at our word and allocated more effort on "Part 1". Experimenters should take care to avoid such mistakes when studies need not be performed in sequence.

We analyzed the elapsed time from experiment launch to HIT completion across all studies (correcting to account for the holiday). An ANOVA found a significant effect ($F(1,11521) = 2817.28$, $p < 0.001$) of reward, as elapsed time averaged 0.8 days in the high-reward case and 1.9 days in the low-reward case. Our separate runs of Experiment 3A and 3B—priced at $0.02 and $0.04 respectively—also allowed us to inspect the affect of reward on result accuracy. We analyzed HITs with a chart height of 160 pixels, which we intentionally overlapped across runs. We again found a significant effect of reward on elapsed time ($F(1,1136) = 1035.56$, $p < 0.001$): Turkers completed more tasks when paid more. On the other hand, reward did not affect the time spent completing an individual HIT ($F(1,1136) = 0.08$, $p = 0.778$), only the total rate of HIT completion. We also found a small but significant effect of reward on accuracy ($F(1,1136) = 7.79$, $p < 0.005$): Turkers were less accurate ($\Delta\mu = 1.4$ units) when paid more. The difference does not alter the design implications of experiment 3. Our results corroborate those of Mason & Watts [23]: paying more does not substantially affect the quality of results, but does increase the rate of HIT completion. By raising the reward, experimenters can decrease the time to results.

Crowdsourcing Reduces Money and Time Costs

The total expenditure for our crowdsourced experiments was $367.77. Had we instead run five laboratory studies (one each for experiments 1A, 1B, 2, 3A, and 3B), using the same number of subjects (assignments) and paying a typical compensation of $15, the cost would have been $2,190. Thus our crowdsourced studies realized a cost savings factor of 6. Had we run all crowdsourced experiments with a $0.02 reward, this increases to a factor of 9 and thus order of magnitude savings are possible. However, experimenters should also consider the equitable treatment of Turkers. Our own misestimation of the average response time led us to compensate Turkers at decidedly less than minimum wage.

Crowdsourcing also provides opportunities beyond simple cost-cutting. Mechanical Turk largely eliminates recruiting effort, makes it easy to extend or modify a study, and automates administration. These result in substantial savings of time and effort: in just a few days (for Exp. 3B, a single day) we were able to run studies that normally would have taken two weeks due to recruiting and scheduling. Moreover, crowdsourcing can scale to large samples that would otherwise be prohibitively large (e.g., 550 Turkers in [8]), greatly expanding the space of feasible study designs.

FINDINGS AND FUTURE WORK

The results from Mechanical Turk demonstrate that crowdsourced graphical perception studies can be viable. We successfully replicated prior experiments on proportional judgments of spatial encodings [7] and alpha contrast adjustment of chart gridlines [30], with our crowdsourced results providing a good match and identical design guidelines to prior work. The increased variation of our results compared to previous results may be compensated by the platform's scalability: for the same cost, many more subjects can participate. We also found that operating system and monitor details reported by JavaScript, though supporting only incomplete and approximate inference of subjects' display configuration, can be predictive of results and so should be recorded if possible.

The results also demonstrate the use of Mechanical Turk to gain new insights into visualization design. Our rectangular area judgment experiment (1B) revealed that comparison of rectangles with aspect ratios of 1 led to higher estimation error than other aspect ratio combinations. This result suggests that the "squarified" optimization objective of leading treemap algorithms [3, 35] may rest on tenuous perceptual footing, and that viewers benefit from the inability of the algorithm to achieve its objective. Future work may lead to improved layout algorithms. Our chart height and gridline spacing experiment (3) suggests optimized parameters for displaying charts on the web: gridlines should be spaced at least 8 pixels apart and increasing chart heights beyond 80 pixels provides little accuracy benefit on a 0-100 scale.

Our results help characterize the use of Mechanical Turk for conducting web-based experiments. Experimenters can expect significant subject overlap when running simultaneous studies, and unreliable response times when using the standard HIT interface. By using qualification tasks and verifiable questions, one can increase the likelihood of high-quality responses. As higher rewards led to faster completion rates with little substantive difference in response quality, experimenters can use payment level to influence study completion time. To facilitate replication, we recommend that experimenters describe qualification tasks and compensation rate when publishing the results of crowdsourced studies.

Finally, we identified benefits for crowdsourcing over laboratory experiments. We found that crowdsourcing can provide *up to an order of magnitude cost reduction*. Such savings could be reinvested in more subjects or more conditions. For constant dollars, we might run better experiments. We realized a *faster time to completion*. This is separate from cost and can also be used to enrich experimental design, especially when experiments are run in stages. We can also gain *access to wider populations* [13]. Many experiments are done on college undergraduates due to the difficulty of recruiting wider populations. Crowdsourcing reduces this cost.

We believe crowdsourcing will be particularly useful in combination with other methods. There is something wrong with every methodological technique, which can often be compensated by combining techniques. Small-scale traditional laboratory experiments can be paired with Mechanical Turk experiments with overlapped conditions. In this way the results of laboratory experiments and crowdsourced exper-

iments can cross check each other, using the two in tandem to leverage their respective strengths.

Future research is needed to develop better tools for crowdsourced experimentation. The facilities for conducting user studies on Mechanical Turk are still rudimentary. Dynamic task generation and easier access control would help researchers conduct adaptive studies, enforce between-subjects designs, and prevent subject overlap across experiments. Already, tools such as Turkit [32] are being developed to close this gap. We believe these tools have an important role to play beyond simplifying study administration. By collecting and aggregating statistics of Turker performance, these tools might provide a means of tracking a dynamic market place, helping researchers make more informed estimates of participation, time to completion, and appropriate compensation.

By integrating crowdsourcing tools with web-based experiment design tools [20], an entire class of user studies may be subject to cheap, scalable web-based design and deployment. Moreover, by archiving and disseminating HIT definitions, such tools might also greatly facilitate study replication, comparison, or modification. In this spirit, all materials used for the studies in this paper can be downloaded from `http://hci.stanford.edu/gp/chi10.zip`.

Of course, crowdsourcing is far from a panacea. Some studies, particularly those dependent on physical or environmental context (e.g., [24, 36]) are simply ill-suited to the web. Crowdsourcing results might also be insensitive to factors such as color blindness or limited visual acuity. Despite these limitations, it is clear that crowdsourcing offers a cheap and scalable way to conduct a valuable range of graphical perception experiments. The time is ripe for investigating more subtle aspects of visualization design.

REFERENCES

1. V. Beattie and M. J. Jones. The impact of graph slope on rate of change judgements in corporate reports. *ABACUS*, 38:177–199, 2002.
2. J. Bertin. *Sémiologie Graphique*. Gauthier-Villars, 1967.
3. D. M. Bruls, C. Huizing, and J. J. van Wijk. Squarified treemaps. In *Data Visualization 2000, Eurographics/IEEE TVCG Visualization Symp.*, pp. 33–42, 2000.
4. C. Cheshire. Selective incentives and generalized information exchange. *Social Psychology Quarterly*, 70, 2007.
5. W. S. Cleveland. *Visualizing Data*. Hobart Press, 1993.
6. W. S. Cleveland, P. Diaconis, and R. McGill. Variables on scatterplots look more highly correlated when the scales are increased. *Science*, 216:1138–1141, Jun 1982.
7. W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. Am. Statistical Assoc.*, 79:531–554, 1984.
8. F. Cole, K. Sanik, D. DeCarlo, A. Finkelstein, T. Funkhouser, S. Rusinkiewicz, and M. Singh. How well do line drawings depict shape? In *ACM SIGGRAPH*, pp. 1–9, 2009.
9. B. D. Dent. *Cartography: Thematic Map Design*. William C. Brown Publishing, 1998.
10. W. R. Garner. *The processing of information and structure*. Erlbaum, 1974.
11. P. P. Gilmartin. Influences of map context on circle perception. *Annals Assoc. of Am. Geographers*, 71:253–258, 1981.
12. J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *ACM CHI*, pp. 1303–1312, 2009.
13. P. Ipeirotis. Mechanical Turk: The Demographics, 2008. behind-the-enemy-lines.blogspot.com/2008/03/mechanical-turk-demographics.html.
14. A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *ACM CHI*, pp. 453–456, 2008.
15. R. Kohavi, R. M. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *ACM KDD*, pp. 959–967, 2007.
16. R. Kraut, J. Olson, M. Banaji, A. Bruckman, J. Cohen, and M. Couper. Psychological research online: Report of board of scientific affairs' advisory group on the conduct of research on the internet. *American Psychologist*, 59:105–117, 2004.
17. S. Lewandowsky and I. Spence. Discriminating strata in scatterplots. *J. Am. Statistical Assoc.*, 84:682–688, Sep 1989.
18. K. Ling, G. Beenen, P. Ludford, X. Wang, K. Chang, X. Li, D. Cosley, D. Frankowski, L. Terveen, A. M. Rashid, P. Resnick, and R. Kraut. Using social psychology to motivate contributions to online communities. *J. CMC*, 10, 2005.
19. A. M. MacEachren. *How Maps Work: Representation, Visualization, and Design*. Guilford, 1995.
20. W. E. Mackay, C. Appert, M. Beaudouin-Lafon, O. Chapuis, Y. Du, J.-D. Fekete, and Y. Guiard. Touchstone: exploratory design of experiments. In *ACM CHI*, pp. 1425–1434, 2007.
21. J. D. Mackinlay. Automating the design of graphical presentations of relational information. *ACM TOG*, 5:110–141, 1986.
22. J. D. Mackinlay, P. Hanrahan, and C. Stolte. Show Me: Automatic presentation for visual analysis. *IEEE TVCG*, 13:1137–1144, Nov/Dec 2007.
23. W. A. Mason and D. J. Watts. Financial incentives and the "performance of crowds". In *KDD-HCOMP*, 2009.
24. D. F. Reilly and K. M. Inkpen. White rooms and morphing don't mix: setting and the evaluation of visualization techniques. In *ACM CHI*, pp. 111–120, 2007.
25. M. J. Salganik and D. J. Watts. Web-based experiments for the study of collective social dynamics in cultural markets. *Topics in Cognitive Science*, 1:439–468, 2009.
26. B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM TOG*, 11:92–99, 1992.
27. B. G. Shortridge. Stimulus processing models from psychology: Can we use them in cartography? *The American Cartographer*, 9:155–167, 1982.
28. D. Simkin and R. Hastie. An information-processing analysis of graph perception. *J. Am. Stat. Assoc.*, 82:454–465, Jun 1987.
29. I. Spence and S. Lewandowsky. Displaying proportions and percentages. *Applied Cog. Psych.*, 5:61–77, 1991.
30. M. Stone and L. Bartram. Alpha, contrast and the perception of visual metadata. In *Color Imaging Conf.*, 2009.
31. L. Tremmel. The visual separability of plotting symbols in scatterplots. *J. Comp. and Graph. Stat.*, 4:101–112, Jun 1995.
32. Turkit. http://groups.csail.mit.edu/uid/turkit/.
33. F. B. Viégas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. ManyEyes: A site for visualization at internet scale. *IEEE TVCG*, 13(6):1121–1128, 2007.
34. C. Ware. *Information Visualization: Perception for Design*. Morgan-Kaufmann, 2nd edition, 2004.
35. M. Wattenberg. Visualizing the stock market. In *ACM CHI Extended Abstracts*, pp. 188–189, 1999.
36. D. Wigdor, C. Shen, C. Forlines, and R. Balakrishnan. Perception of elementary graphical elements in tabletop and multi-surface environments. In *ACM CHI*, pp. 473–482, Apr 2007.
37. C. Ziemkiewicz and R. Kosara. The shaping of information by visual metaphors. *IEEE TVCG*, 14:1269–1276, Nov/Dec 2008.