Graphical Tests for Power Comparison of Competing Designs

Heike Hofmann, Lendie Follett, Mahbubul Majumder, and Dianne Cook

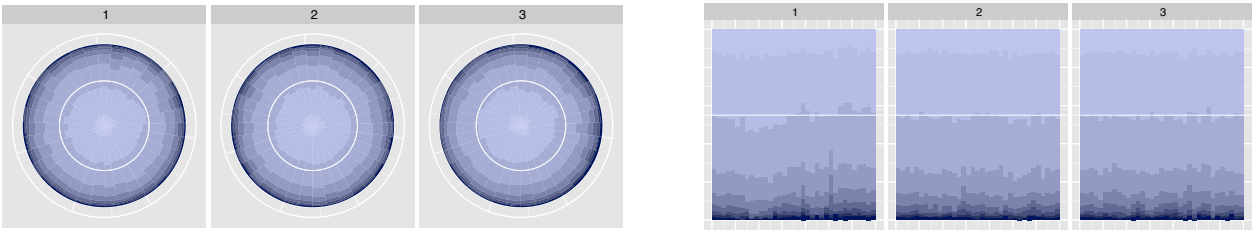

Fig. 1. Which of the three plots is different from the others? - Is this question easier to answer with the Polar Charts or the Barcharts?

Abstract—Lineups [4, 28] have been established as tools for visual testing similar to standard statistical inference tests, allowing us to evaluate the validity of graphical findings in an objective manner. In simulation studies [12] lineups have been shown as being efficient: the power of visual tests is comparable to classical tests while being much less stringent in terms of distributional assumptions made. This makes lineups versatile, yet powerful, tools in situations where conditions for regular statistical tests are not or cannot be met. In this paper we introduce lineups as a tool for evaluating the power of competing graphical designs. We highlight some of the theoretical properties and then show results from two studies evaluating competing designs: both studies are designed to go to the limits of our perceptual abilities to highlight differences between designs. We use both accuracy and speed of evaluation as measures of a successful design. The first study compares the choice of coordinate system: polar versus cartesian coordinates. The results show strong support in favor of cartesian coordinates in finding fast and accurate answers to spotting patterns. The second study is aimed at finding shift differences between distributions. Both studies are motivated by data problems that we have recently encountered, and explore using simulated data to evaluate the plot designs under controlled conditions. Amazon Mechanical Turk (MTurk) is used to conduct the studies. The lineups provide an effective mechanism for objectively evaluating plot designs.

Index Terms—Lineups, Visual inference, Power comparison, Efficiency of displays.

1 Introduction

So, you've just spent countless hours burning the late night candle wax and discovered several astonishing features in the data. Your exploratory graphics are quick to produce, interactive, flexible, and fabulous for discovery, but now you need to present the information to your boss, the co-workers, a class of students, or get it published in that top-rated journal. Those exploratory graphics are unlikely to communicate the discovered information to your audience efficiently, or elegantly. To decide on the best way to present the specific information there can be many types of plots to choose from. The decision process is assisted by past experience, personal aesthetic preferences, the characteristics of the audience, and a good working knowledge of general perceptual strengths and weaknesses of particular displays [6, 24, 25, 8]. Perceptual studies are thinly spread for data visualization and will probably only cover very broad perceptual principles, primarily because they tend to be based solely on simulated data. In order to study perceptual principles in a controlled experimental setting simulated data is unavoidable. This provides principles that may not apply closely enough to a specific analysis task. Faced with design decisions for a very specific task requires a closer level of detail. Focus groups [14] can find small problems with designs and move us closer to a final handful of possible designs. Usability studies and case studies (e.g. [20], [23]) typically focus on the use of particular software for solving data exploration tasks. These approaches lack broad audience validity and objectiveness. Several papers indicate continuing issues arising with, and suggestions for improving, evaluation of information visualizations [17, 7, 26, 15].

Here we suggest a new approach using recent research on visual inference [4, 28]. In visual inference, data plots are considered to be test statistics, and these are compared with plots of data generated from a null hypothesis using a lineup. The null hypothesis underlying a lineup is that there is no real structure visible in the data plot, that any patterns seen are consistent with randomness, or from a known model. If the null hypothesis is true then the plot of the data will not be distinguishable from the plots of null data. A lineup is similar to the police lineup, and this is where the name comes from. The suspect (the plot of the data) is placed randomly in a field of plots of null data, and an impartial viewer is asked to identify the plot that is the most different. If the viewer identifies the suspect it lends statistical significance to the conclusion that the data is not consistent with the null data. Figure 2 has an example of a lineup. Overlaid density plots are used to display the distribution of two groups (blue, red) of data. We are interested to know whether the data has a significant difference between the centers of the two groups. Plot number $5+2^3$ contains the data. Did you pick that one? In this plot the blue group has a flat density whereas the density of red group is more peaked and concentrated to the left of the blue.

The previous work [4, 28] describes the use of lineups for evaluating discoveries made in exploratory analysis. Here, we describe the use of lineups for evaluating plot design for communicating findings visually. Different lineups are created, using the same data, same null plots and positions within the lineup, but different plot designs. Comparing how long viewers take and how accurately they report the feature of interest will assess which design is better for the task. To investigate

- *Heike Hofmann is an associate professor in Statistics at Iowa State University, e-mail: hofmann@iastate.edu.*
- *Lendie Follett is an undergraduate student in Statistics at Iowa State University, e-mail: lfollett@iastate.edu.*
- *Mahbubul Majumder is a graduate student in Statistics at Iowa State University, e-mail: mahbub@iastate.edu.*
- *Dianne Cook is a full professor in Statistics at Iowa State University, e-mail: dicook@iastate.edu.*

Manuscript received 31 March 2012; accepted 1 August 2012; posted online 14 October 2012; mailed on 5 October 2012.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

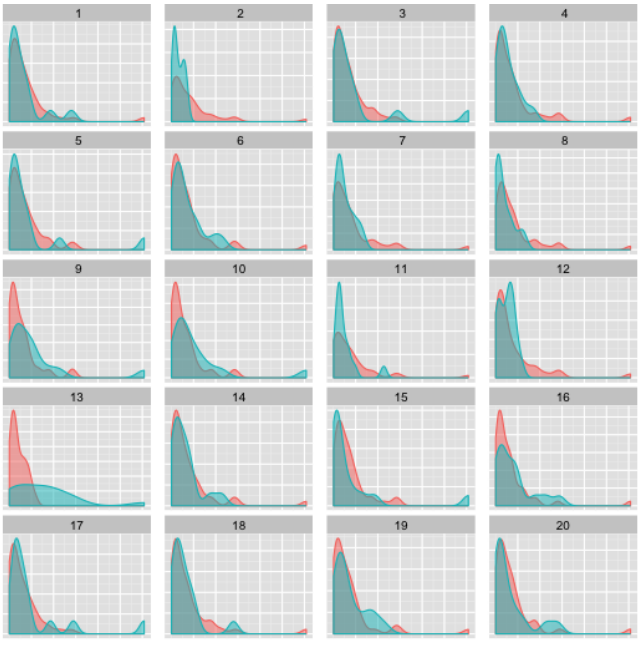

Fig. 2. Example lineup using density plots: one plot of data embedded with nineteen plots of null data. Which plot is most different from the others? (See the text for the answer.)

the benefit of lineups for plot design evaluation we have conducted an experiment that also contains simulated data, where the data is constructed to match and expand on the features of interest in the data. For example, in a simple situation, where the discovered feature of interest is a difference between the centers of two distributions, simulated data would be constructed roughly matching the distribution of the data that varies the distance between centers, and also spread and sample size. Lineups of these simulated data (embedded with simulated null data plots) are generated and evaluated along with the lineups of data. The simulation study provides a backdrop for the real data, which enables the broad applicability to be studied and a gauge for the strength of the pattern in the data.

Lineups provide a way to statistically quantify the significance of the finding [12]: think of the lineup as a set of m plots, one of which is the data (in Figure 2, $m = 20$). The probability that an observer picks the data plot just by chance, i.e. when it is really *not* different from the other null plots, is $1/m$. If an observer is able to identify the data plot from the lineup, we reject the null hypothesis. This sets our Type I error rate, α, at a level of $1/m$.

When there are multiple independent observers (n) we have more freedom in setting the significance level: assume that x out of those n observers picked the data plot. Let X be the corresponding random variable, i.e. X = # times out of n independent repetitions that the data plot is picked from the lineup. Under the null hypothesis, X has a Binomial distribution: $X \sim B_{n,1/m}$. We can then compute the p-value of a lineup as the probability to have x or more observers picking the data plot (under the assumption that the null hypothesis is true, i.e. the plot is not different):

$$p\text{-value} = P(X \geq x \mid H_0) = 1 - B_{n,1/m}(x-1).$$

When comparing different tests of the same quantity, we consider that test better if it has greater *power*. The power of a test is the probability to reject the null hypothesis, irrespective of whether it is true or false – in a lineup this is the probability that an observer identifies the plot of the real data. Analytically, power is usually difficult to calculate because it is requires specific use of an alternative hypothesis to calculate the probability. Work in [22] addresses this to some extent with measures on the quality of a lineup, how numerically different, as best it can be calculated, the data plot is from the null plots.

In our situation, though, it is fairly straightforward to *estimate* the power of a lineup:

> Let n be the number of independent observers and x_i the number of observers who picked plot i, $i \in \{1,...,m\}$, from the lineup. Then $(x_1, x_2, ..., x_m)$ follows a multinomial distribution $\text{Mult}_{\pi_1,\pi_2,...,\pi_m}(x_1, x_2, ..., x_m)$ with $\sum_i \pi_i = 1$, where π_i is the probability that plot i is picked by an observer, which we can estimate as $\widehat{\pi}_i = x_i/n$.

The power of a lineup can therefore be estimated as the ratio of correct identifications x out of n viewings. (More details on these derivations can be found in [13].)

The power of lineups is used in the work presented here to evaluate the effectiveness of different plot designs. This paper describes two examples where we had real data analysis problems and decisions to make in order to communicate the results. Section 2 explains the process of comparing designs. Section 3 describes two data analysis problems, the experiments conducted to evaluate the plot designs, and presents findings. We conclude in Section 4 and give suggestions for future use.

2 Comparison of Designs

We are going to make use of the signal strength gained from multiple viewings of a lineup in order to evaluate competing designs as follows:

1. **Create Lineup Data:** assuming that at least two variables, X and Y are involved in the design, we create data for a lineup of size m by creating $m-1$ permutations of Y or, in the case of a simulation study, drawing $m-1$ samples of size n (the number of rows in the data) from the null distribution. Add the original data to the lineup data randomly between 1 and m. The `R` package `nullabor` provides a framework for easy creation of lineup data.

2. **Create lineups from competing designs:** using the same data, render lineups of all competing designs.

3. **Evaluate Lineups:** by presenting the lineups to independent observers. Assess both signal strength and time needed by individuals to come to a decision. Note that each observer should only be exposed to each lineup data once.

4. **Evaluate Competing Designs:** differences in signal strength or time to decision are due to differences in the design. In the case that individuals were shown multiple lineups (as part of a bigger study), it is possible to correct outcome measurements for an individual's visual ability.

Comparing power of competing designs therefore involves comparing percentages of correct responses $\widehat{\pi}_1$ and $\widehat{\pi}_2$. An $\alpha{\cdot}100\%$ confidence interval for this comparison is given as

$$\widehat{\pi}_1 - \widehat{\pi}_2 \pm t_{1-\alpha/2,n-1}\sqrt{\widehat{\pi}_1(1-\widehat{\pi}_1)/n_1 + \widehat{\pi}_2(1-\widehat{\pi}_2)/n_2}, \qquad (1)$$

where n is the Welch-Satterthwaite [27] estimate of the degrees of freedom. Note that we use $\widehat{\pi}_i = (x_i+1)/(n_i+1)$ and n_i+2 for a better coverage of the confidence interval [1]. In the case of more than two competing designs, we have to additionally adjust the confidence intervals for multiple testing, e.g. using a Bonferroni adjustment which uses an adjusted confidence level $\tilde{\alpha} = \alpha/k$, where k are the number of confidence intervals involved (single adjustment) or the number of comparisons anticipated.

While this allows a direct comparison of the designs, we cannot adjust for the individuals' perceptual abilities. In the case that we have multiple responses from each person (i.e. data is collected on several *different* lineup tasks), we can estimate their perceptual ability and correct power differences between competing designs accordingly, e.g. by modeling power using a subject-specific random intercept in a generalized linear model.

3 EXPERIMENTS

Two recent data analyses motivated the setup of this experiment. The first one arises from an analysis of US flight traffic data, and a finding related to how wind direction affects the efficiency of an airport. The second one arose during the review of a paper claiming that it was impossible to statistically test for differences in center between two samples where sample size was small and from a non-normal population, with application to studying toxic waste sites. Using lineups containing dotplots we found it was possible to detect a difference, and we were curious to see if other display types (density, histogram, boxplot) might compare with dotplots.

3.1 Data Collection

Data for both studies was collected using Amazon's Mechanical Turk (MTurk) service. The website for the second study is available at `http://www.public.iastate.edu/~mahbub/feedback_turk5/homepage.html`. Figure 3 shows a screen shot of the website's layout for a dotplot lineup.

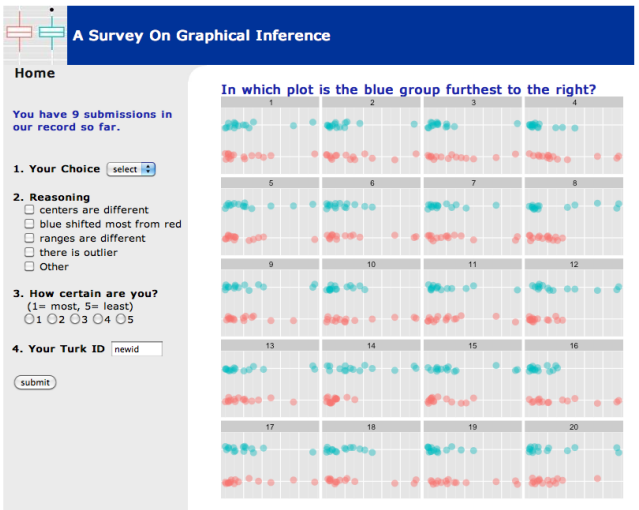

Fig. 3. Screenshot of the website for the second study.

Each participant is shown ten lineups and asked to identify the plot that is the most different from the other plots. In the second study the 'difference' is further specified as 'In which plot is the blue group the furthest to the right?". In both studies, though, the question is placed prominently above each lineup as a reminder of what participants are asked to look for. Additionally, participants are asked for a reason for their choice. They can select from a list of choices particular to the lineup design, why they chose this particular plot as well as how confident they are on a scale of one to five, where higher values indicate higher confidence. Personal information on age group, gender, and education is collected on a voluntary basis. The amount of time it takes the individual to answer is also recorded. This allows us to assess how design choices affect decoding [5] in terms of accuracy and speed.

In order to avoid people 'gaming the system' as found in earlier studies [8, 10], two of the ten lineups participants were shown were specially prepared as a baseline of performance – if participants fail to answer some of these baseline lineups, we require at least two correct answers for the remaining plots, which is indicative of a better-than-guessing performance, as inclusion criterion for the data evaluation.

3.2 Study I: Wind Direction and Airport Efficiency

3.2.1 Setup

The data contains all flights [21] in and out of Seattle/Tacoma International Airport (SEA) between July 2008 and June 2011. As a measure of airport efficiency we are using time between successive wheel-events (time at wheel take-off or touch-down), which we found to be independent of airline carrier and operating hours, as long as we restricted the data to 'regular' operating hours between 7 am and midnight and 'regular' weather conditions – i.e. we eliminated records associated with the top 5% wind speed measurements [16], leaving us with just under half a million flights.

In this scenario, all of the lineups are testing the null hypothesis

$$H_0 : \textit{wind direction has no effect on efficiency}$$

against the alternative H_a : *wind direction does have an effect.*

Statistical tests of mean efficiency by wind direction are not particularly helpful in this situation: the difference in mean efficiency between the wind direction at which the airport operates the most efficiently and its least efficient direction shows a statistically significant difference with a p-value $\leq 10^{-15}$. However, out of the 34 other wind directions, another 31 show significant differences in efficiency as well. This is much more a property of the large data size rather than practically usable differences. Mere significances also do not allow us an assessment of the underlying pattern.

In deciding on the design for displaying efficiency by wind direction we were using the fact that wind direction is circular, and displayed the data as (conditional) wind rose charts - i.e. for each of 36 wind directions we show the percentage of flights falling into one-minute intervals between successive flights, from zero minutes to eight minutes or more.

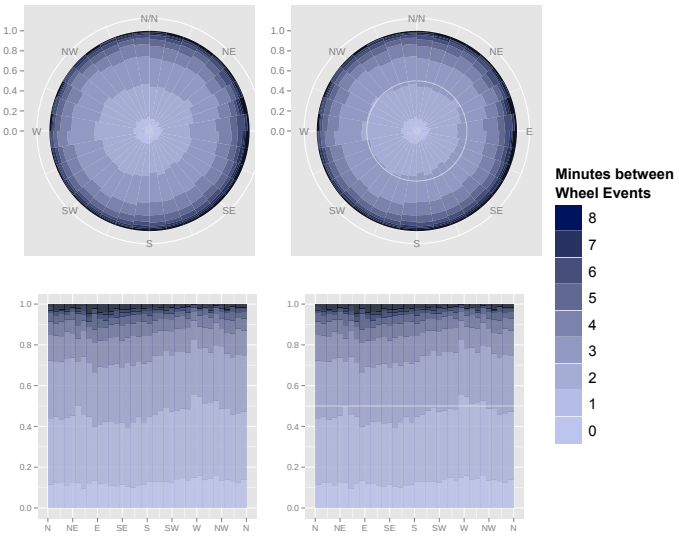

Fig. 4. All four competing designs: polar versus cartesian charts in top and bottom row, with (right) and without (left) reference lines at 50%.

In spite of the contextual circular property of wind direction, the pattern in efficiency did not seem to stand out well during the exploration process, which led us to using a corresponding cartesian design. Both designs were additionally equipped with a reference line at 50%. Figure 4 shows an overview of all four chart types in the study. During the exploration of the data, it became clear that Seattle airport functions most efficiently with winds coming from the east to southeast, while winds from the west seemed to be most troublesome, resulting in a wave-like pattern in the cartesian charts and a shift in center for the polar charts. Since runways can be used in both directions, changing the runway usage according to dominant wind direction for that day or time of the year might be a feasible solution in gaining efficiency.

An additional factor we were interested in this particular situation was to assess, how much of the data we actually needed to use in a design to have observers pick out the pattern. Clearly, a design is more efficient, if a smaller sample size is sufficient for showing the presence of a relationship. In order to investigate this, we took different size samples of the data and created lineup plots of all four designs for these subsets. The effect of sample size on the displays mostly stems from the additional variability introduced when using small samples, which might hide the pattern, if it is not displayed prominently.

Another perturbation to the data are 'shifts' in wind direction, i.e. we make use of the circular nature of the wind direction and adjust the x axis by using different offsets.

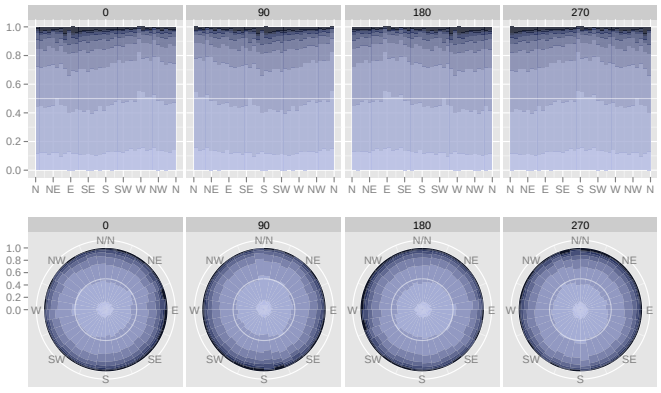

Fig. 5. (Top) Cartesian charts of the same sample, with shifts in wind direction by 90, 180, and 270 degrees. The shift by 90 and 270 leads to a 'valley' versus a 'mountain' pattern, whereas a shift by 180 degrees inverts the wave pattern from 'down-up' to 'up-down'. (Bottom) Shifts in wind direction in the polar chart lead to rotations of the display by 90 degrees.

This results in a shift of the wave in cartesian charts and a rotation in the polar charts as can be seen in Figure 5. Our initial thinking was that this would have no effect on the polar charts, but might have a deteriorating effect on the cartesian charts, in the case that the peak or the valley of the wave was at either end of the x axis, i.e. an offset of 90 or 270 degrees. While these simple perturbations and different sample sizes allow us to get insight into different aspects of the designs, they also allow us to get multiple responses from each observer to assess individual ability without the need to go outside the framework of the original data, i.e. we leave the joint relationship between x and y essentially unchanged. For all of these combinations we produced two replicates, resulting in a total of 192 different lineups (4 designs $\times$ 6 sample sizes $\times$ 4 offsets $\times$ 2 reps). Table 1 gives an overview of the number of times lineups in each combination of design and sample size were shown, and in how many of them the data plot was correctly identified.

Table 1. Breakdown of lineups: number correct/**number shown** (proportion correct). Each participant was shown eight lineups, and the same lineup was shown to multiple people (denominator in the table) reasonably well-spread among the treatment levels.

		sample size					
type of chart		2	4	6	8	10	24
cartesian	with	12/**24**	18/**27**	42/**44**	33/**41**	39/**43**	46/**47**
		(0.50)	(0.67)	(0.95)	(0.80)	(0.91)	(0.98)
	without	13/**41**	24/**32**	39/**44**	47/**51**	40/**45**	34/**37**
		(0.32)	(0.75)	(0.89)	(0.92)	(0.89)	(0.92)
polar	with	13/**49**	9/**43**	9/**39**	7/**35**	8/**40**	13/**38**
		(0.27)	(0.21)	(0.23)	(0.20)	(0.20)	(0.34)
	without	6/**51**	4/**34**	5/**34**	6/**39**	11/**43**	12/**37**
		(0.12)	(0.12)	(0.15)	(0.15)	(0.26)	(0.32)

3.2.2 Results for Study I

Figures 6 and 7 give an overview of the relationship between the three response variables: accuracy, speed, and confidence level.

Figure 6 shows histogram of the time participants needed to make a decision on each lineup. Correctness of answers is shown by color. Because of the skewness of the distributions, times were log-transformed. Polar charts take on average more time to answer, and are answered with much lower accuracy. The average amount of time spent on a cartesian lineup is $e^{3.53} = 34.1$ seconds compared to $e^{4.07} = 58.5$ seconds for a polar lineup. This is in stark contrast to accuracy: 76.92% of the cartesian lineups shown resulted in the a correct identification of the real data while only 20.31% of the polar charts were correct.

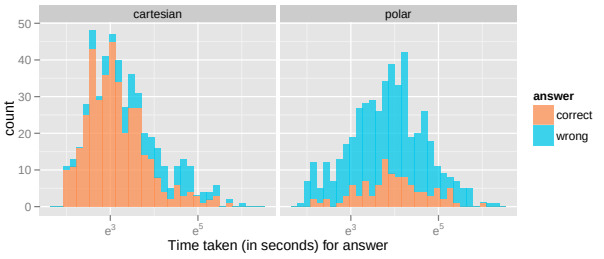

Fig. 6. Histograms of time taken for answering lineups. On average the cartesian design is answered faster and with higher precision.

Figure 7 shows two barcharts of confidence levels by task, again coloring is used for correctness of answers. Cartesian displays lead to a very bimodal distribution of confidence: participants are either very sure or not sure at all of their answer. Confidence levels in polar charts are distributed much more uniformly. Confidence levels seem to be independent from precision, though.

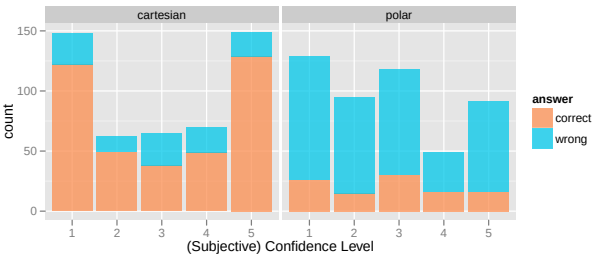

Fig. 7. Barcharts of self-reported confidence in answering correctly for each lineup. A strongly bimodal distribution is apparent, while confidence levels for polar charts are more uniform. Confidence levels are not indicative of accuracy: none of the differences between confidence levels show a significant difference in accuracy.

The perceptual involved in decoding the designs consist of comparisons along a common axis (in the barcharts) and a common origin (in the polar charts). The difference in designs is therefore based on how well we are able to judge deviations from a horizontal line compared to deviations from a circle. Based on this, the charts with added reference lines provide us with exactly the frame we compare with and should, therefore, be the 'better' designs - either in speed or accuracy.

Figure 8 shows a comparison of power for the four competing designs. 95% confidence intervals are Bonferroni adjusted for multiple testing. The bottom two confidence intervals in each panel show a 95% confidence interval of a direct comparison of the cartesian design versus the polar design with and without a reference line. With the exception of the 2% sample none of those confidence intervals include the zero, indicating a significantly higher power for the cartesian design than for the polar design.

However, all of these considerations are based on the - rather strong - assumption that all individuals have the same ability to detect the data plot from a lineup. In order to allow for individual differences in visual ability, we use a generalized linear mixed effects modeling approach [19] for each of the three response values, using the `R` package `lme4` [2]. For power predictions (cf. table 2) a logistic regression was fitted for the competing designs, including covariates *sample size* (2, 4,6, 8, 10, and 24 percent of the data) and *shift* in wind direction (offset of 0, 90, 180, and 270 degrees), both as main effects and in two-way interaction effects with design to assess their effect on the

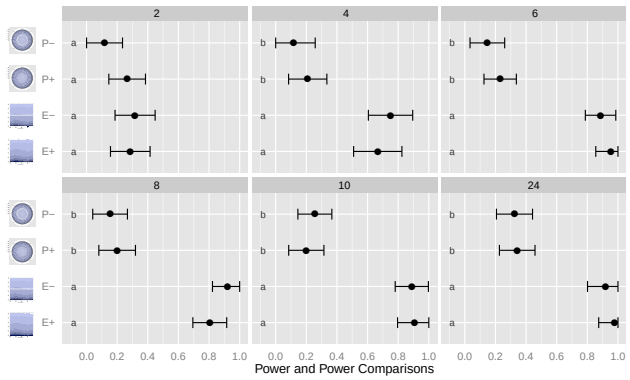

Fig. 8. Power results for four competing designs: polar versus cartesian, each with and without a reference line; panels are facetted by sample size (as percentage of original data). Dots show estimated power, surrounded by intervals of standard errors. The letters at the front of each panel allow comparisons across all designs [18]: all designs with different letters have significantly different power (at $\alpha = 0.05$). This is adjusted for all pairwise comparisons using a Benjamini-Hochberg adjustment [3, 9].

power of designs. To adjust for individuals' ability, a random intercept was included in the model.

Table 2. Output of a generalized linear mixed effects model for power of lineups (i.e. probability of identifying the data plot) for comparing designs. Included are two-way interactions with sample size and shifts in wind direction (offset). Cartesian designs without a reference line at offset 0 are used as baseline. Results are based on 976 lineup evaluations by 115 participants.

		Estimate	Error	z-value	p-value	
design	cartesian	-0.08	0.39	-0.21	0.84	
	polar	-1.98	0.32	-6.13	0.00	***
main effects						
	reference line	-0.14	0.26	-0.53	0.59	
	sample size	0.27	0.04	6.31	0.00	***
	offset: 90 degrees	-0.43	0.37	-1.18	0.24	
	180 degrees	-0.89	0.35	-2.51	0.01	**
	270 degrees	0.21	0.38	0.55	0.58	
interactions						
	polar:line	0.51	0.35	1.44	0.15	
	polar:sample size	-0.23	0.05	-5.02	0.00	***
	polar:offset 90	0.64	0.49	1.30	0.20	
	polar:offset 180	0.91	0.47	1.92	0.05	.
	polar:offset 270	-0.73	0.54	-1.35	0.18	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Overall, the results show huge differences in the power of designs between polar charts and cartesian: cartesian designs are significantly more powerful than polar charts, particularly so with small sample sizes. The reference line has surprisingly little influence, but it helps more for polar charts than for cartesian charts, An increase in sample size has a positive impact on power. Polar charts need a much bigger sample size to see an increase in power – only at about 24% of the original data do we see about the same power as for cartesian charts of a sample size of 2%. The changes in offset are significant – interestingly, borderline behavior (90 and 270 degrees) does not show a difference between polar and cartesian charts, whereas an inversion of the wave pattern (first up, then down), does show a difference. The power of cartesian lineups suffers significantly from this inversion whereas power of polar charts is unaffected. This is an unexpected finding, but is consistent throughout different lineups in the data. Figure 9 summarizes the results from the model: power predictions (y axis) are shown by sample size (x axis). The thick lines show average power by design for different shifts in wind direction. The thin lines represent power for individuals. What can be seen is the different impact of the offset by design: while an offset of 270 degrees (the 'mountain' pattern) has the highest power in cartesian charts, it comes out worst in polar charts.

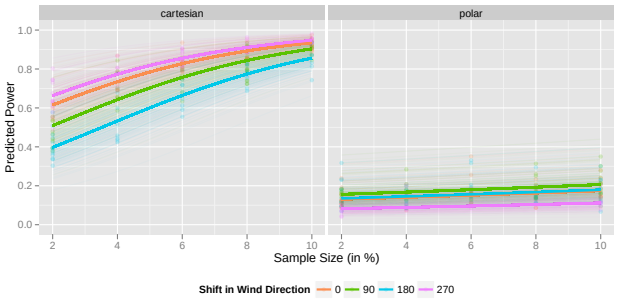

Fig. 9. Predicted Power of designs. The thin lines and the points on y axis show variability due to individuals' abilities. The saturated lines show average predicted power for each of the designs.

Time taken to answer was log transformed before the modeling process to de-emphasize the impact of very large values (up to 500 seconds). The findings are consistent with correctness - time taken shows big differences between polar and cartesian charts. The reference lines seem to increase the evaluation time, but not significantly. An increase in sample size decreases evaluation time, The inversion of the wave pattern at 180 degrees leads to a significant increase in time for cartesian charts, but not for polar charts (at least not significantly). Table 3 shows an overview of the model parameters and estimates.

Table 3. Model output for linear mixed effects model of (log) time taken, barchart design is baseline.

		Estimate	Error	t-value	approx p-value	
design	intercept	3.58	0.09	41.34	0.00	***
	polar	0.22	0.10	2.17	0.03	*
covariates						
	reference line	0.08	0.06	1.50	0.13	
	sample size	-0.03	0.00	-8.73	0.00	***
	offset: 90	-0.06	0.08	-0.72	0.47	
	180	0.16	0.08	2.07	0.04	*
	270	-0.04	0.07	-0.54	0.59	
interactions						
	polar:line	-0.03	0.08	-0.40	0.69	
	polar:sample size	0.03	0.01	6.25	0.00	***
	polar:offset 90	0.14	0.11	1.23	0.22	
	polar:offset 180	-0.17	0.11	-1.55	0.12	
	polar:offset 270	0.17	0.11	1.50	0.13	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Confidence levels are measured on a five point scale — they are subjective assessments by the participant 'how certain are you'. We model confidence using the same model structure as before, i.e. using sample size and offset as covariates and including up to two-way interactions, While there is a difference between the designs - participants reported higher confidence in dealing with the cartesian lineups – this is only a trend (i.e. not significant at 5% but below 10%). The only significant effect on reported confidence level is the use of reference lines in polar coordinates: participants report an increase in confidence (0.31 $\pm$ 0.13, p value = 0.01) when using the reference line. This, however, does not translate to a significant increase in accuracy of results, as we have seen before.

Cartesian coordinates resulted in a significantly more accurate identification of the real data set in significantly shorter time.

3.3 Study II: Comparing Distribution Centers

3.3.1 Study Setup

In a second experiment, we conducted a simulation study to investigate the power of four competing designs, as shown in Figure 11, in assessing a mean shift between distributions.

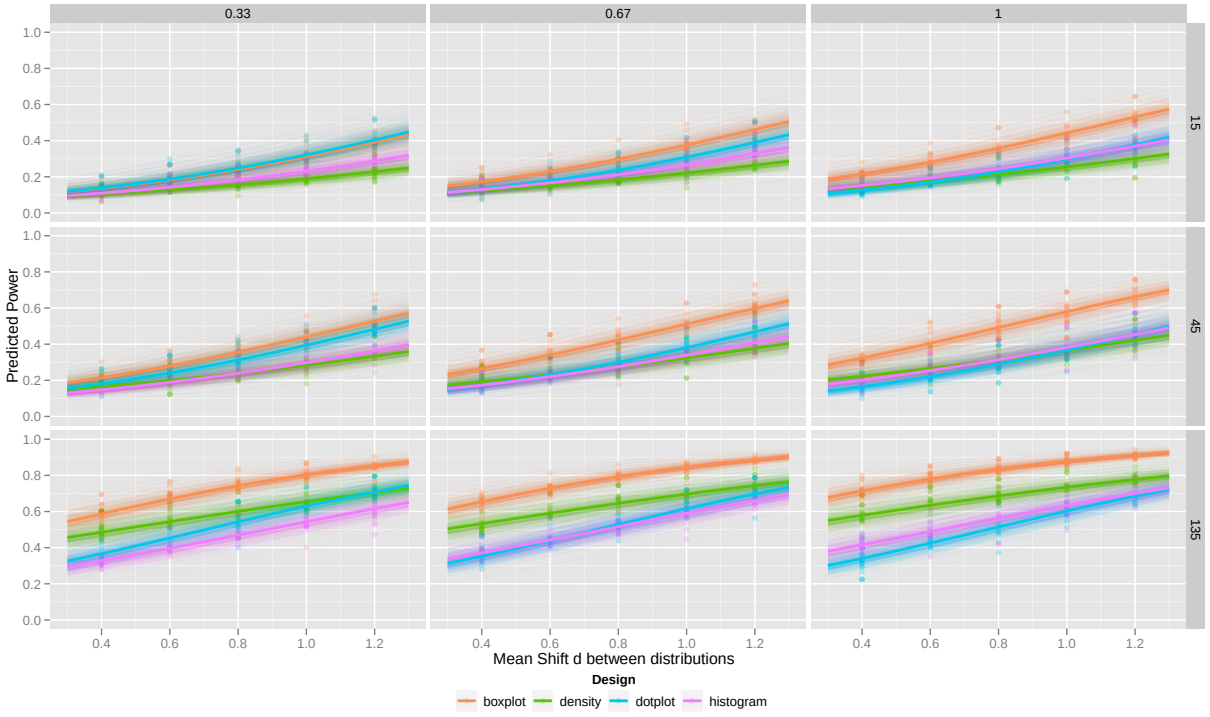

Fig. 10. Overview of power predictions for the four different designs. The fully saturated thick lines show average predicted power for each of the designs facetted by size of the red group (top to bottom) and relative size of the blue group to the red group (left to right). Thin lines represent variability due to subject-specific abilities.

The hypotheses for each lineup are

H_o : *centers of the two groups are the same*,
H_a : *center of the blue group is shifted to the right*.

As covariates for this experiment, we considered the *size of the shift*, d, between the means of the distributions, with $d = 0.4, 0.6, 0.8.1.0$, and 1.2, as well as the sample size. Since we are dealing with two groups of points, we varied both the *size of the larger group*, $n_1 = 15, 45$, and 135, as well as the *relative size* of the second group (*ratio*), with $n_2 = 1/3, 2/3$ and $3/3$ of n_1. For each of these combinations, we created three replicate data sets with values from an exponential distribution with rate parameter $\lambda = 1$ (for the first group) and rate parameter $\lambda_2 = 1/(d+1)$, which corresponds to a mean shift of d, for the second groups. Based on each data set we created four lineup charts, one for each of the competing designs. This resulted in 540 different lineup charts overall, which we sorted by p-value corresponding to the difference in means between the two groups in the simulated data. According to the p-value, we associated a 'difficulty' level with lineups from 1 to 9 (easy to hard). Ten lineup plots were shown to each participant - one out of each of the difficulty levels and one additional reference chart with a particularly easy lineup to allow for data quality checks.

All of the designs were chosen based on common usage. The main difference between the designs is their varying level of data summarization.

From most to least summarized we consider these designs:

boxplots: five-point summary statistics (and a number of outliers),
density plots: need – mathematically – all of the data points, but their smoothness and continuity makes them appear visually quite simple,
histograms: rendered by a frequency breakdown for a specified number of bins (10×2, in the study),
dotplots: all n data points necessary for rendering.

For the boxplot and dotplot, the vertical axis is free to be used for another variable, *group*, so the data can be split apart. For the histogram and density the vertical axis is used for count or frequency, and is not available for group information. Based on the assumption that charts are most effective [11], when they show enough information to solve the task at hand, but not too much to hamper the observer, we anticipated results to reflect the above order.

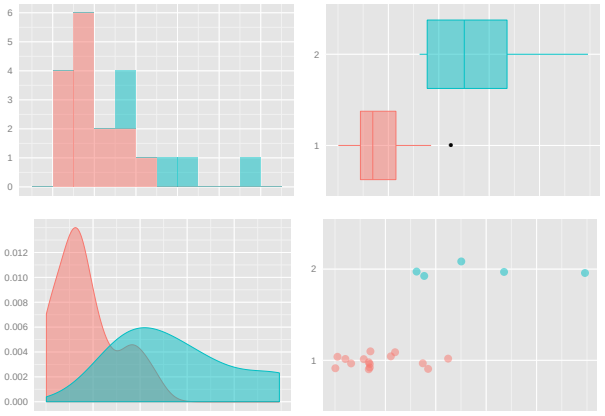

Fig. 11. Overview of all four competing designs for displaying differences in distributions.

3.3.2 Results

Since each participant provides results from nine lineups, we can use a generalized linear mixed effects model of power and account for individuals' abilities by including a subject-specific random intercept. All of the above described covariates are considered in the model, both as main effects and as two-way interaction with the design to assess their impact on the power of each of the designs. Figure 10 summarizes the modeling results as shown in table 4. Generally we can see that the power of all four designs increases with an increase in effect size d (on the x axis), as well as an increasing sample size n_1 (top to bottom

facetting). Generally, an increase of the second group also is beneficial to power. Dotplots are an exception to this – they actually lose power significantly as the second group increases. This might be due to visual complexity, as discussed above.

Table 4. Results from a Generalized Linear Mixed Effects Model of power of a lineup given design and all two-way interactions with effect size d, group size n_1, and ratio to second to first group size. Note that for all terms involving n_1 estimate and error are multiplied by a factor of 100. Reference group for all effects are boxplot designs. The model fit is based on 2513 lineup evaluations by 208 participants.

		Estimate	Error	t-value	approx	p-value
design	Intercept	-3.11	0.41	-7.61	0.00	***
	density	0.05	0.57	0.09	0.93	
	dotplot	0.46	0.57	0.81	0.42	
	histogram	0.06	0.58	0.11	0.91	
covariates						
	d	1.77	0.34	5.18	0.00	***
	$n1_{100}$	1.84	0.19	9.73	0.00	***
	ratio	0.84	0.33	2.56	0.01	**
interactions						
	density:d	-0.60	0.47	-1.27	0.20	
	dotplot:d	0.03	0.46	0.07	0.95	
	histogram:d	-0.28	0.48	-0.59	0.56	
	density:$n1_{100}$	-0.10	0.26	-0.36	0.72	
	dotplot:$n1_{100}$	-0.77	0.26	-3.00	0.00	***
	histogram:$n1_{100}$	-0.69	0.26	-2.64	0.01	**
	density:ratio	-0.28	0.47	-0.59	0.56	
	dotplot:ratio	-1.01	0.46	-2.19	0.03	*
	histogram:ratio	-0.28	0.47	-0.59	0.55	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Visually, density plots seem to The two highly summarized designs benefit the most from an increase in sample size: both density charts and boxplots profit significantly more than histograms and dotplots. This might, again, be a hint towards a trade-off between visual complexity and amount of information shown as the basis for making decisions.

Table 5 gives an overview of results from a linear model for (log) evaluation time of lineups. Subject-specific speed (due to spatial ability) are accounted for including a random intercept. Structurally, the model is the same as before, i.e. we regard effect size d, size n_1 of the first group, and the ratio n_2/n_1 between group sizes both as main effects and in two-way interactions with the four competing designs. To make estimate sizes comparable, we have again multiplied all estimates and standard errors involving the group size n_1 by 100. This means that we need to interpret those estimates as average change if the group size is increased by 100 elements.

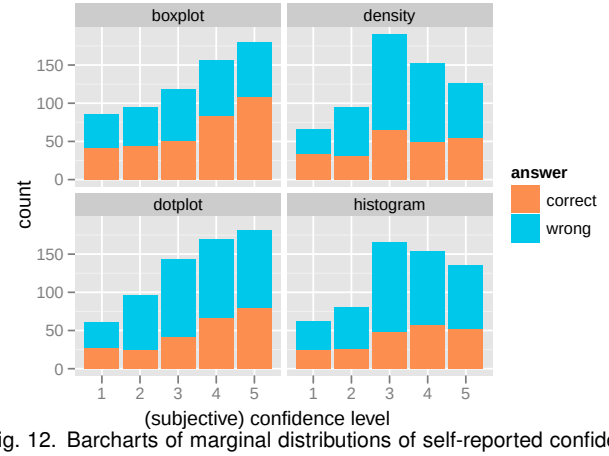

Fig. 12. Barcharts of marginal distributions of self-reported confidence by design. The distributions display prominent differences.

The results of this model are similar to the previous: evaluation

Table 5. Results from a Linear Mixed Effects Model of (log) time to answer a lineup given design and all two-way interactions with effect size d, group size n_1, and ratio to second to first group size. Estimate and error for all terms involving n_1 are multiplied by a factor of 100 for comparability. Reference group for all effects are boxplot designs.

		Estimate	Error	t-value	approx	p-value
design	Intercept	3.85	0.09	42.21	0.00	***
	density	0.01	0.12	0.07	0.94	
	dotplot	0.03	0.12	0.26	0.80	
	histogram	0.23	0.12	1.97	0.05	.
covariates						
	d	-0.23	0.08	-3.02	0.00	***
	$n1_{100}$	0.07	0.04	1.70	0.09	.
	ratio	-0.04	0.02	-1.48	0.14	
interactions						
	density:d	0.11	0.11	1.02	0.31	
	dotplot:d	0.20	0.11	1.86	0.06	.
	histogram:d	-0.01	0.11	-0.09	0.93	
	density:$n1_{100}$	-0.12	0.06	-2.11	0.03	*
	dotplot:$n1_{100}$	-0.16	0.06	-2.83	0.00	***
	histogram:$n1_{100}$	-0.05	0.06	-0.93	0.35	
	density:ratio	0.04	0.03	1.18	0.24	
	dotplot:ratio	0.02	0.03	0.65	0.52	
	histogram:ratio	0.00	0.04	0.04	0.97	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

speeds needed for the four different designs is very close, but histograms take on average a significantly longer time ($e^{0.23} = 1.26$ seconds) to evaluate than lineups of boxplots. Where an increase of sample size led to an increase of accuracy before, we now see that it increases time for evaluation of boxplot slightly, but not significantly so. What is interesting is that evaluation time decreases significantly with an increase in sample size, particularly for dotplots, but also for density charts. This is counter-intuitive to our previous argument on dotplots being affected by visual complexity the most. It seems, that observers of dotplot lineups come to a conclusion faster as sample size increases, albeit more often to the wrong conclusion. Rather than visual complexity, this might be indicative that over-plotting is problematic in dotplots. While we tried to accommodate for over-plotting by making use of alpha-blending and jittering, this might not have been sufficient for the large sample size.

Our third response variable is again the subjective assessment of confidence level. On average participants reported a confidence in their results of 3.41 (on a five point scale with 1=lowest and 5=highest), with a standard deviation of 0.149. Modeling confidence levels with a linear mixed effects model with the same structure as before resulted in only one significant effect beyond the intercept: participants report significantly higher confidence for increased number of points in dotplot lineups. On average an increase of 100 points leads to an increase in confidence of 0.22 (with a standard error of 0.082). This finding is significant with a p-value of 0.007. Together with the previous results, we can tell that something worthy of further investigation happens in the perception of dotplots as sample sizes increase.

Figure 12 shows barcharts of confidence assessments by design. It is obvious that the distribution of confidence is very different from one design to the next. Boxplots and dotplots show a similar pattern with an emphasis on high confidence levels, whereas histograms and density lineups have modes at a confidence level of three. One reason behind these differences in confidence distributions might be familiarity with the design - neither density charts nor histograms are particularly frequent displays for a general audience.

4 Conclusions

In this paper we are investigating lineups for evaluating competing plot designs based on their power modeled by a logistic regression with subject-specific random intercepts. Lineups provide a powerful tool for evaluating different designs in the framework of the data. We can show in two MTurk studies, that cartesian charts are significantly

more powerful than polar charts. The results also supported our choice of dotplots in the very small samples that triggered our investigation, but elucidated the visual complexity dilemma: showing much more information than necessary for a particular decision is detrimental to accuracy rather than supportive.

While we –theoretically– can't show the same data plot to a single individual, it does make in practice a quite convincing tool to do so, e.g. in the informal setting of a classroom or in interactions with collaborators, when we do not intend to collect data for further investigation of a problem. Using the MTurk service gives us fast feedback – both of the studies were completed within hours of their availability – but comes at the price that we have to deal with the occasional 'gamer'. For future studies we are planning on having a required entrance test to try to avoid those. Besides accuracy and speed, there is a variety of other aspects of the design, which could be collected in the same framework of MTurk studies, such as ease of use, familiarity with a particular type of design, or a quantitative visual pleasure rating of the lineups.

Different visual abilities of individuals make up a significant amount of variability in lineup evaluations. Further investigations on how these relate to known tests of cognitive skills, such as e.g. the paper folding tests.

ACKNOWLEDGMENTS

This work was supported in part by NSF DMS 1007697.

REFERENCES

[1] A. Agresti and B. A. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998.

[2] D. Bates, M. Maechler, and B. Bolker. *lme4: Linear mixed-effects models using S4 classes*, 2011. R package version 0.999375-42.

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57(1):289–300, 1995.

[4] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham. Statistical inference for exploratory data analysis and model diagnostics. *Royal Society Philosophical Transactions A*, 367(1906):4361–4383, 2009.

[5] W. S. Cleveland. *Elements of graphing data*. Hobart Press, 1994.

[6] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):pp. 531–554, 1984.

[7] G. Ellis and A. Dix. An explorative analysis of user evaluation studies in information visualisation. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, BELIV '06, pages 1–7, New York, NY, USA, 2006. ACM.

[8] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI Í0, pages 203–212, New York, NY, USA, 2010. ACM.

[9] T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.

[10] R. Kosara and C. Ziemkiewicz. Do mechanical turks dream of square pie charts? In *Proceedings BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*, pages 373–382. ACM Press, 2010.

[11] S. Kosslyn. *Graph design for the eye and mind*. Oxford University Press, 2006.

[12] M. Majumder, H. Hofmann, and D. Cook. Visual statistical inference for regression parameters. Technical Report 13, Iowa State University, Department of Statistics, 2011.

[13] M. Majumder, H. Hofmann, and D. Cook. Validation of visual statistical inference, with applications to linear models. Technical Report 4, Iowa State University, Department of Statistics, 2012.

[14] R. Mazza and A. Berre. Focus group methodology for evaluating information visualization techniques and tools. In *Information Visualization, 2007. IV '07. 11th International Conference*, pages 74–80, Los Alamitos, CA, USA, 2007. IEEE Computer Society.

[15] T. Munzner. A nested model for visualization design and validation. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):921 –928, nov.-dec. 2009.

[16] NOAA Climatic Data Center. Past weather. `http://weather.noaa.gov/weather/current/`, last accessed August 15, 2012, 2011.

[17] C. North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, pages 6–9, 2006.

[18] H.-P. Piepho. An algorithm for a letter-based representation of all-pairwise comparisons. *Journal of Computational and Graphical Statistics*, 13(2):456–466, 2004.

[19] J. Pinheiro and D. Bates. *Mixed-effects models in S and S-Plus*. Springer, 2000.

[20] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of Conference on Advanced Visual Interfaces*, pages 109–116, New York, NY, USA, 2004. ACM.

[21] RITA BTS Transtat. On-time performance. `http://www.transtats.bts.gov/Fields.asp?Table_ID=236`, last accessed August 15, 2012, 2011.

[22] N. Roy Chowdhury, D. Cook, H. Hofmann, M. Majumder, and Y. Zhao. Where's Waldo: Looking closely at a lineup. Technical Report 2, Iowa State University, Department of Statistics, 2012.

[23] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, BELIV '06, pages 1–7, New York, NY, USA, 2006. ACM.

[24] D. Simkin and R. Hastie. An information processing analysis of graph perception. *Journal of the American Statistical Association*, 82:454–465, 1987.

[25] I. Spence and S. Lewandowsky. Displaying proportions and percentages. *Applied Cognitive Psychology*, 6:61–77, 1991.

[26] F. B. Viegas and M. Wattenberg. Communication-minded visualization: A call to action [technical forum]. *IBM Systems Journal*, 45(4):801 –812, 2006.

[27] B. L. Welch. The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 34:28–35, 1947.

[28] H. Wickham, D. Cook, H. Hofmann, and A. Buja. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis '10)*, 16(6):973979, Nov.-Dec. 2010. 26% acceptance rate. Best paper award.