Validation of Visual Statistical Inference, Applied to Linear Models

Mahbubul Majumder [a] , Heike Hofmann [a] & Dianne Cook [a]

[a] Department of Statistics and Statistical Laboratory , Iowa State University , Ames , IA , 50011-1210

PLEASE SCROLL DOWN FOR ARTICLE

Validation of Visual Statistical Inference, Applied to Linear Models

Mahbubul MAJUMDER, Heike HOFMANN, and Dianne COOK

Statistical graphics play a crucial role in exploratory data analysis, model checking, and diagnosis. The lineup protocol enables statistical significance testing of visual findings, bridging the gulf between exploratory and inferential statistics. In this article, inferential methods for statistical graphics are developed further by refining the terminology of visual inference and framing the lineup protocol in a context that allows direct comparison with conventional tests in scenarios when a conventional test exists. This framework is used to compare the performance of the lineup protocol against conventional statistical testing in the scenario of fitting linear models. A human subjects experiment is conducted using simulated data to provide controlled conditions. Results suggest that the lineup protocol performs comparably with the conventional tests, and expectedly outperforms them when data are contaminated, a scenario where assumptions required for performing a conventional test are violated. Surprisingly, visual tests have higher power than the conventional tests when the effect size is large. And, interestingly, there may be some super-visual individuals who yield better performance and power than the conventional test even in the most difficult tasks. Supplementary materials for this article are available online.

KEY WORDS: Data mining; Effect size; Exploratory data analysis; Lineup; Nonparametric test; Practical significance; Statistical graphics; Visualization.

1. INTRODUCTION

Statistical graphics nourish the discovery process in data analysis by revealing unexpected things, finding structure that was not previously anticipated, or orthogonally by contrasting prevailing hypotheses. The area of graphics is often associated with exploratory data analysis, which was pioneered by Tukey (1977) and is particularly pertinent in today's data-rich world where discovery during data mining has become an important activity. Graphics are also used in many places where numerical summaries simply do not suffice: model checking, diagnosis, and in the communication of findings.

Several new developments in graphics research have been achieved in recent years. Early studies on evaluating how well statistical plots are perceived and read by the human eye (Cleveland and McGill 1984), have been repeated and expanded (Simkin and Hastie 1987; Spence and Lewandowsky 1991; Heer and Bostock 2010) with findings supporting the original results. The research by Heer and Bostock (2010) used subjects recruited from Amazon's Mechanical Turk (Amazon 2010) for their studies. This body of work provides a contemporary framework for evaluating new statistical graphics. In a complementary direction, new research on formalizing statistical graphics with language characteristics makes it easier to abstractly define, compare, and contrast data plots. Wilkinson (1999) developed a grammar of graphics that is enhanced by Wickham (2009). These methods provide a mechanism to abstract the way data are mapped to graphical form. Finally, technology advances make it simple and easy for everyone to draw plots of data, and particularly the existence of software systems, such as R (R Development Core Team 2012), enable making beautiful data graphics that can be tightly coupled with statistical modeling.

However, measuring the strength of patterns seen in plots, and differences in individual perceptual ability, is something that is difficult and perhaps handicaps graphics use among statisticians, where measuring probabilities is of primary importance. This has also been addressed in recent research. Buja et al. (2009) proposed a protocol that allows the testing of discoveries made from statistical graphics. This work represents a major advance for graphics, because it bridges the gulf between conventional statistical inference procedures and exploratory data analysis. One of the protocols, the lineup, places the actual data plot among a page of plots of null data, and asks a human judge to pick the plot that is different. Figure 1 shows an example lineup. Which plot do you think is the most different from the others? (The position of the actual data plot is provided in Section 5.1.) Wrapped in a process that mirrors conventional inference, where there is an explicit, a priori, null hypothesis, picking the plot of the data from the null plots represents a rejection of that null hypothesis. The null hypothesis typically derives from the task at hand, or the type of plot being made. The alternative encompasses all possible antitheses, all types of patterns that might be detected in the actual data plot, accounting for all possible deviations from the null without the requirement to specify these ahead of time. The probability of rejection can be quantified, along with Type I and Type II errors, and p-value and power can be defined and estimated.

The protocol has only been informally tested until now. In the work described in this article, the lineup protocol is compared head to head with the equivalent conventional test. Specifically, the lineup is examined in the context of a linear model setting, where we are determining the importance of including a variable in the model. This is not the envisioned environment for the use of the lineup—actually it is likely the worst case scenario for visual inference. The intended use of lineups is where there is no existing test, and unlikely ever to be any numerical test. The thought is though, that the conventional setting provides a benchmark for how well the lineup protocol works under controlled conditions, and will provide some assurance that they

Mahbubul Majumder is PhD student (E-mail: mahbub72@gmail.com), Heike Hofmann (E-mail: hofmann@iastate.edu) is Associate Professor, and Dianne Cook (E-mail: dicook@iastate.edu) is Professor in the Department of Statistics and Statistical Laboratory, Iowa State University, Ames, IA 50011-1210. This research is supported in part by the National Science Foundation Grant # DMS 1007697.
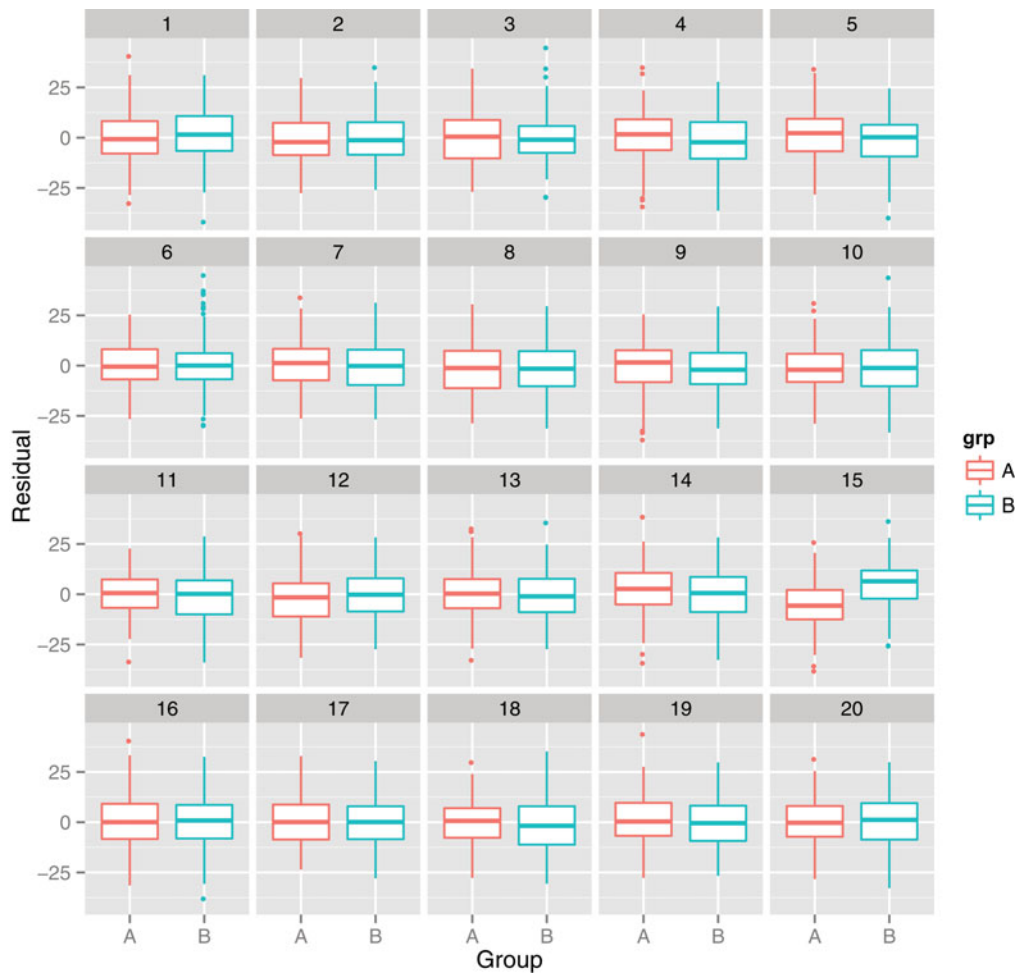
Figure 1. Lineup plot ($m = 20$) using side-by-side boxplots for testing $H_0 : \beta_k = 0$. One of these plots is the plot of the actual data, and the remaining are null plots, produced by simulating data from a null model that assumes H_0 is true. Which plot is the most different from the others, in the sense that there is the largest shift or location difference between the boxplots? (The position of the actual data plot is provided in Section 5.1.) The online version of this figure is in color.

will work in scenarios where there is no benchmark. Testing is done based on a human-subjects experiment using Amazon's Mechanical Turk (Amazon 2010), using simulation to provide controlled conditions for assessing lineups. The results are compared with those of the conventional test.

The article is organized as follows. Section 2 defines terms as used in visual inference, and describes how to estimate the important quantities from experimental data. The effect of the lineup size and number of observers on the power of the test is discussed in Section 3. Section 4 focuses on the application of visual inference to linear models. Section 5 describes three user studies based on simulation experiments conducted to compare the power of the lineup protocol with the equivalent conventional test, and Section 6 presents an analysis of the resulting data.

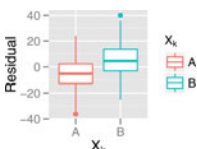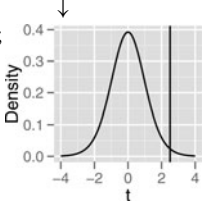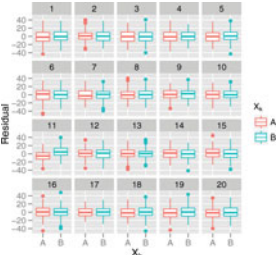## 2. DEFINITIONS AND EXPLANATIONS FOR VISUAL STATISTICAL INFERENCE

An illustration of the lineup protocol in relation to conventional hypothesis testing is presented in Table 1. Both methods start from the same place, the same set of hypotheses. The conventional test statistic is the t-statistic, where the parameter estimate is divided by its standard error. In the lineup protocol,

the test statistic is a plot of the data. Here, side-by-side boxplots are used, because the variable of interest is categorical and takes just two values. In conventional hypothesis testing, the value of the test statistic is compared with all possible values of the sampling distribution, the distribution of the statistic if the null hypothesis is true. If it is extreme on this scale then the null hypothesis is rejected. In contrast in visual inference, the plot of the data is compared with a set of plots of samples drawn from the null distribution. If the actual data plot is selected as the most different, then this results in rejection of the null hypothesis.

In general, we define θ to be a population parameter of interest, with $\theta \in \Theta$, the parameter space. Any null hypothesis H_0 then partitions the parameter space into Θ_0 and Θ_0^c, with $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. A test statistic, $T(y)$, is a function that maps the sample into a numerical summary, that can be used to test the null hypothesis. The hypothesis test maps the test statistic into $\{0, 1\}$, based on whether $T(y)$ falls into the acceptance region, or the rejection region, respectively. $T(y)$ is assessed relative to null values of this statistic $T(y_0)$, the possible values of T if $\theta \in \Theta$.

For visual inference, unlike in the conventional hypothesis test, the statistic is not a single value, but a graphical representation of the data chosen to display the strength of the parameter of

Table 1. Comparison of visual inference with conventional inference

	Conventional inference	Lineup protocol
Hypothesis	$H_0 : \beta = 0$ vs $H_1 : \beta > 0$	$H_0 : \beta = 0$ vs $H_1 : \beta > 0$
	$\downarrow$	$\downarrow$
Test statistic	$T(y) = \frac{\hat{\beta}}{se(\hat{\beta})}$	$T(y) =$
	$\downarrow$	$\downarrow$
Sampling Distribution	$f_{T(y)}(t)$;	$f_{T(y)}(t)$;
	$\downarrow$	$\downarrow$
Reject H_0 if	Actual T is extreme	Actual plot is identifiable

interest, θ. When the alternative hypothesis is true, it is expected that the plot of the actual data, the test statistic, will have visible feature(s) consistent with $\theta \in \Theta_0^c$, and that visual artifacts will not distinguish the test statistic as different when H_1 is not true. We will call a plot with this property a *visual statistic* for θ. More formally:

Definition 2.1. A *visual test statistic*, $T(.)$, is a function of a sample that produces a plot. $T(y)$ maps the actual data to the plot, and we call this the *(actual) data plot*, and $T(y_0)$ maps a sample drawn from the null distribution into the same plot form. These type of plots are called *null plots*.

Ideally, the visual test statistic is defined and constructed using the grammar of graphics (Wilkinson 1999; Wickham 2009), consisting of type and specification of aesthetics, necessary for complete reproducibility. The visual test statistic is compared with values $T(y_0)$ using a lineup, which is defined as:

Definition 2.2. A *lineup* is a layout of m randomly placed visual statistics, consisting of

- $m - 1$ statistics, $T(y_0)$, simulated from the model specified by H_0 (null plots) and
- the test statistic, $T(y)$, produced by plotting the actual data, possibly arising from H_1.

The $(m - 1)$ null plots are members of the sampling distribution of the test statistic assuming that the null hypothesis is true. If H_1 is true, we expect this to be reflected as a feature in the test statistic, that is, the plot of the data, that makes it visually distinguishable from the null plots. A careful visual inspection of the lineup by independent observers follows; observers are asked to point out the plot most different from the lineup. If the test statistic is identified in the lineup, this is considered as evidence against the null hypothesis. This leads us to a definition for the *p*-value of a lineup: under the null hypothesis, each observer has a $1/m$ chance of picking the test statistic from the

lineup. For K independent observers, let X be the number of observers picking the test statistic from the lineup. Under the null hypothesis $X \sim \text{Binom}_{K,1/m}$, therefore:

Definition 2.3. The *p*-value of a lineup of size m evaluated by K observers is given as

$$P(X \geq x) = 1 - \text{Binom}_{K,1/m}(x - 1)$$
$$= \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right)^i \left(\frac{m-1}{m}\right)^{K-i}$$

with X defined as above, and x is the number of observers selecting the actual data plot.

Note that for $x = 0$ the *p*-value becomes, mathematically, equal to 1. It might make more sense from a practical point of view to think of the *p*-value as being larger than $P(X \geq 1)$ in this situation. By increasing either m or K, the value at a higher precision can be determined. Table 2 shows *p*-values for different numbers of observers for lineups of size $m = 20$.

Definition 2.4. The *visual test*, V_θ of size m and significance level α, is defined as

- Reject H_0 if out of K observers at least x_α correctly identify the actual data plot, and
- Fail to reject H_0 otherwise.

where x_α is such that $P(X \geq x_\alpha | H_0) \leq \alpha$.

As with any test, there is the risk of Type I or II errors, which for visual inference are defined as follows:

Definition 2.5. The *Type I error* associated with visual test V_θ is the probability of rejecting H_0 when it is true; the probability for that is $P(X \geq x_\alpha)$, which is controlled by α. The *Type II error* is the probability of failing to identify the actual data plot, when H_0 is not true, $P(X < x_\alpha)$.

Table 2. Possible *p*-values for different numbers of observers, *K*, for fixed size *m* = 20 lineups

K	*x*	*p*-value	*K*	*x*	*p*-value	*K*	*x*	*p*-value	*K*	*x*	*p*-value	*K*	*x*	*p*-value
1	1	0.0500	2	1	0.0975	3	1	0.1426	4	1	0.1855	5	1	0.2262
			2	2	0.0025	3	2	0.0073	4	2	0.0140	5	2	0.0226
						3	3	0.0001	4	3	0.0005	5	3	0.0012
									4	4	<0.0001	5	4	<0.0001

Because X takes only discrete values we cannot always control exactly for α. For example, when there is only one observer, $1/m$ is the minimal value at which we can set α. It can be set to be smaller, even arbitrarily small, by increasing K, the number of observers. Type II error is harder to calculate, as is usually the case. In visual inference, individual abilities need to be incorporated to calculate Type II error. Here, we need to estimate the probability that an observer sees the actual data plot as different, when it really is different. This involves understanding the individual's visual skills. Thus, let X_i be a binary random variable with $X_i = 1$, if individual i $(= 1, \ldots, K)$ identifies the actual data plot from the lineup, and $X_i = 0$ otherwise. Let p_i be the probability that individual i picks out the actual data plot. If all individuals have the same ability, with the probability, p, for picking out the actual data plot, then $X = \sum_i X_i$ has distribution $\text{Binom}_{K,p}$, and we can estimate p by $\hat{p} = x/K$, where x is the number of observers (out of K), who pick out the actual data plot.

If there is evidence for individual skills influencing the probability p_i, then $X_i \sim \text{Binom}_{1,p_i}$ and X is a sum of independent Bernoulli random variables with different success rates p_i. This makes the distribution of X a Poisson-binomial by definition [see Butler and Stephens (1993) for details]. Ways to estimate p_i will be discussed in the following sections.

Definition 2.6. The *power* of a visual test, V_θ, is defined as the probability to reject the null hypothesis for a given parameter value θ:

$$\text{Power}_V(\theta) = \Pr(\text{Reject } H_0 \mid \theta).$$

An important difference between conventional and visual testing is that lineups will depend on observers' evaluation. Thus X, the number of observers who identify the actual data plot from the lineup, affects the estimation of power and the power is estimated by

$$\widehat{\text{Power}}_V(\theta) = \text{Power}_{V,K}(\theta) = 1 - F_{X,\theta}(x_\alpha - 1).$$

Here $F_{X,\theta}$ is the distribution of X and x_α is such that $P(X \geq x_\alpha) \leq \alpha$. Note that the distribution F_X depends on which hypothesis is true: under the null hypothesis, $X \sim \text{Binom}_{K,1/m}$, leading to:

$$\text{Power}_V(\theta, K) = 1 - \text{Binom}_{K,1/m}(x_\alpha - 1).$$

If the alternative hypothesis is true, with a fixed parameter value θ, we can assume that an individual's probability to identify the data plot depends on the parameter value, and $X_i \sim \text{Binom}_{1,p_i(\theta)}$. Assessing an individual's skill to identify the actual data plot will require that an individual evaluates multiple lineups.

Power is an important consideration in deciding which test to use for solving a problem. Here we use it to compare the performance of the visual test with the conventional test, but in practice for visual inference it will mostly be important in choosing plots to use. Analysts typically have a choice of plots to make, and a myriad of possible options such as reference grids, for any particular purpose. This is akin to different choices of statistics in conventional hypothesis testing, for example, mean, median, or trimmed mean. One is typically better than another. For two different visual test statistics of the same actual data, one is considered to be better, if $T(y)$ is more easily distinguishable to the observer. Power is typically used to measure this characteristic of a test.

3. EFFECT OF OBSERVER SKILLS AND LINEUP SIZE

3.1 Subject-Specific Abilities

Suppose, each of K-independent observers gives evaluations on multiple lineups, and responses are considered to be binary random variable, $X_{\ell i} \sim \text{Binom}_{1,p_{\ell i}}$, where $X_{\ell i} = 1$, if subject i correctly identifies the actual data plot on lineup ℓ, $1 \leq \ell \leq L$, and 0 otherwise. A mixed effects logistic regression model is used for $P(X_{\ell i} = 1) = p_{\ell i} = E(X_{\ell i})$, accommodating both for different abilities of observers as well as differences in the difficulty of lineups.

The model can be fit as:

$$g(p_{\ell i}) = W_{\ell i}\delta + Z_{\ell i}\tau_{\ell i}, \tag{1}$$

where $g(.)$ denotes the logit link function $g(\pi) = \log(\pi) - \log(1 - \pi); 0 \leq \pi \leq 1$. W is a design matrix of covariates corresponding to specifics of lineup ℓ and subject i, and δ is the vector of corresponding parameters. Covariates could include demographic information of individuals, such as age, gender, education level, etc., as well lineup-specific elements, for example, effect size or difficulty level. $Z_{\ell i}$, $1 \leq i \leq K$, $1 \leq \ell \leq L$, is a design matrix corresponding to random effects specific to individual i and lineup ℓ; and τ is a vector of independent normally distributed random variables $\tau_{\ell i}$ with variance matrix $\sigma_\tau I_{KL \times KL}$. τ will usually include a component incorporating an individual's ability or skill to evaluate lineups. Note that $\tau_{\ell i}$ usually only includes a partial interaction; for a full interaction of subjects' skills and lineup-specific difficulty we would need replicates of the same subject evaluating the same lineup, which in practice is not feasible without losing independence.

The inverse logit link function, $g^{-1}(.)$, from Equation (1) leads to the estimate of the subject and the lineup-specific probability of successful evaluation by a single observer as

$$\hat{p}_{\ell i} = g^{-1}(W_{\ell i}\hat{\delta} + Z_{\ell i}\hat{\tau}_{\ell i}). \tag{2}$$

3.2 Lineup Size, *m*

The finite number $m - 1$ of representatives of the null distribution, used as comparison against the test statistic, is a major difference between visual inference and the conventional testing. The choice of m has an obvious impact on the test.

The following properties can only be derived for the situation of a fully parameterized simulation study, as conducted in this article. They allow for a direct comparison of lineup tests against the conventional counterparts, and also to identify properties relevant for a quality assessment of lineups when they are used in practical settings. Two assumptions are critical:

1. the plot setup is structured in a way that makes it possible for an observer to identify a deviation from the null hypothesis,
2. an observer is able to identify the plot with the strongest "signal" (or deviation from H_0) from a lineup.

Evidence in support of the second assumption will be seen in the data from the study discussed in Section 5, the degree to which the first assumption is fulfilled is reflected by the power of a lineup. The better suited a design is for a particular task, the higher its power will be.

To compare the power of conventional and visual tests side-by-side, it is necessary to assume that we are in the controlled environment of a simulation with tests corresponding to a known parameter value $\theta \in R$ and associated distribution function F_t of the test statistic.

Lemma 3.1. Suppose $F_{|t|}(.)$ is the distribution function of an absolute value of t, the conventional test statistic. Suppose, the associated test statistic is observed as t_{obs} with p-value p_D.

The probability of picking the data plot from a lineup depends on the size m of the lineup and the strength of the signal in the data plot. Under the above assumptions, the probability is expressed as

$$P(p_D < p_0) = E[(1 - p_D)^{m-1}],$$

where p_D is the p-value associated with the data in the test statistic, and p_0 is the minimum of all p-values in the data going into null plots.

Proof. The proof and the details of the lemma are attached in the supplementary documents. □

The above lemma allows two immediate conclusions for the use of lineups. The probability that the observer correctly identifies the data plot is closely connected to the size of the lineup m, since the right-hand side of the above equation decreases for larger m, the probability of correctly identifying the actual data plot decreases with m. Further we see, that the rate of this decrease depends strongly on the distribution of p_D—if the density of p_D is very right skewed, the expectation term on right-hand side will be large and less affected by an increase in m. This can also be seen in Figure 2, which illustrates lemma 3.1. Figure 2 shows the probability of picking the actual data plot for lineups of different size: as m increases we have an increased probability to observe a more highly structured null plot by chance. It can also be seen that for a p-value, p_D, of about 0.15 for
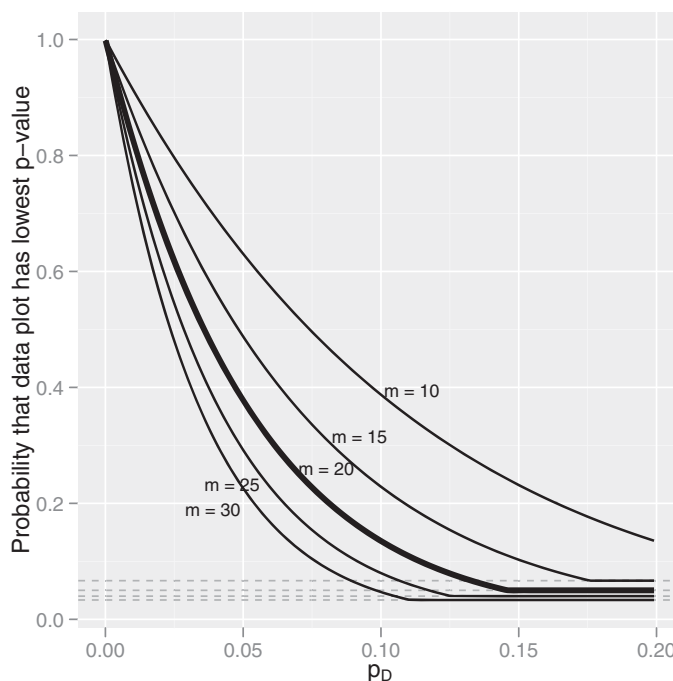

Figure 2. Probability that the data plot has the smallest p-value in a lineup of size m. With increasing p-value the probability drops—when it reaches $1/m$ a horizontal line is drawn to emphasize insufficient sensitivity of the test due to the lineup size.

the data plot, the signal in the plot is so weak that it cannot be distinguished from null plots in a lineup of size $m = 20$.

4. APPLICATION TO LINEAR MODELS

To make these concepts more concrete, consider how this would operate in the linear models setting. Consider a linear regression model
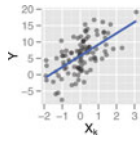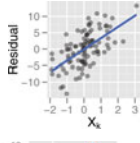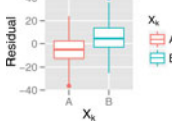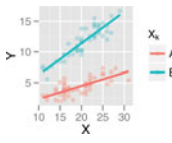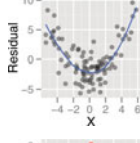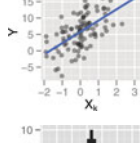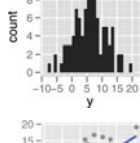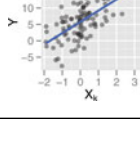
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \beta_3 X_{i1} X_{i2} + \cdots + \epsilon_i, \quad (3)$$

where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, $i = 1, 2, \ldots, n$. The covariates (X_j, $j = 1, \ldots, p$) can be continuous or discrete.

In this setting, there are many established graphics that are used to evaluate and diagnose the fit of a regression model (e.g., Cook and Weisberg 1999). Table 3 lists several common hypotheses related to the regression setting, and commonly used statistical plots that might be used as corresponding visual test statistics. For example, to examine the effect of variable X_j on Y, we would plot residuals obtained from fitting the model without X_j against X_j or for a single covariate we may plot Y against X_j (cases 1–4 in Table 3). To assess whether the assumption of linearity is appropriate we would draw a plot of residuals against fitted values (case 5 in Table 3). For the purpose of comparing visual against conventional inference, we focus on cases 2 and 3, with a continuous and categorical explanatory variable, respectively.

Suppose, X_k is a categorical variable with two levels, and we test the hypothesis $H_0 : \beta_k = 0$ vs $H_1 : \beta_k \neq 0$. If the responses for the two levels of the categorical variable X_k in the model are different, the residuals from fitting the null model should show a significant difference between the two groups. For a

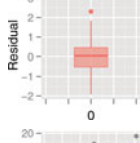Table 3. Visual test statistics for testing hypotheses related to the model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \beta_3 X_{i1} X_{i2} + \cdots + \epsilon_i$

Case null hypothesis	Statistic	Test statistic	Description
1 $H_0 : \beta_0 = 0$	Scatterplot		Scatterplot with least square line overlaid. For null plots we simulate data from fitted null model.
2 $H_0 : \beta_k = 0$	Residual plot		Residual vs X_k plots. For null plots we simulate data from normal with mean 0 variance $\hat{\sigma}^2$.
3 $H_0 : \beta_k = 0$ (for binary X_k)	Boxplot		Boxplot of residuals grouped by category of X_k. For null plots we simulate data from normal with mean 0 variance $\hat{\sigma}^2$.
4 $H_0 : \beta_k = 0$ (interaction of continuous and binary X_k)	Scatterplot		Scatterplot with least square lines of each category overlaid. For null plots we simulate data from fitted null model.
5 $H_0 : X$ Linear	Residual Plot		Residual vs predictor plots with loess smoother overlaid. For null plots we simulate residual data from normal with mean 0 variance $\hat{\sigma}^2$.
6 $H_0 : \sigma^2 = \sigma_0^2$	Boxplot		Boxplot of standardized residual divided by σ_0^2. For null plots we simulate data from standard normal.
7 $H_0 : \rho_{X,Y|Z} = \rho$	Scatterplot		Scatterplot of residuals obtained by fitting partial regression. For null plots we simulate data (mean 0 and variance 1) with specific correlation ρ.
8 $H_0 :$ Model Fits	Histogram		Histogram of the response data. For null plots we simulate data from fitted model.
9 For $p = 1$ only: $H_0 : \rho_{X,Y} = \rho$	Scatterplot		Scatterplot with least square line overlaid. For null plots we simulate data with correlation ρ.

visual test, we draw boxplots of the residuals conditioned on the two levels of X_k. If $\beta_k \neq 0$ the boxplots should show a vertical displacement.

The conventional test in this scenario uses $T = \hat{\beta}_k / \text{se}(\hat{\beta}_k)$ and rejects the null hypothesis, if T is extreme on the scale of a t distribution with $n - p$ degrees of freedom. It forms the benchmark upon which we evaluate the visual test. To calculate what we might expect for the power of the visual test, under perfect conditions, first assume that the observer is able to pick the plot with the smallest p-value from a lineup plot. This leads to the decision to reject H_0 when $p_D < p_0$, where p_D is the conventional p-value as details given in Lemma 3.1. Thus the expected probability to reject by a single observer ($K = 1$) in this scenario is

$$p(\beta) = Pr(p_D < p_0) \quad \text{for} \quad \beta \neq 0. \tag{4}$$

Figure 3 shows the power of the conventional test in comparison to the expected power of the visual test for different K (number of observers), obtained using $p(\beta)$ from Equation (4). Notice that the expected power of the visual test exceeds the power of the conventional test as K increases, and when β gets larger. Conversely, visual power is below conventional power for parameter values close to the null hypothesis. This is even more pronounced for large number of observers. At the same time, the point of intersection between visual and conventional power approaches the value of the null hypothesis as the number of
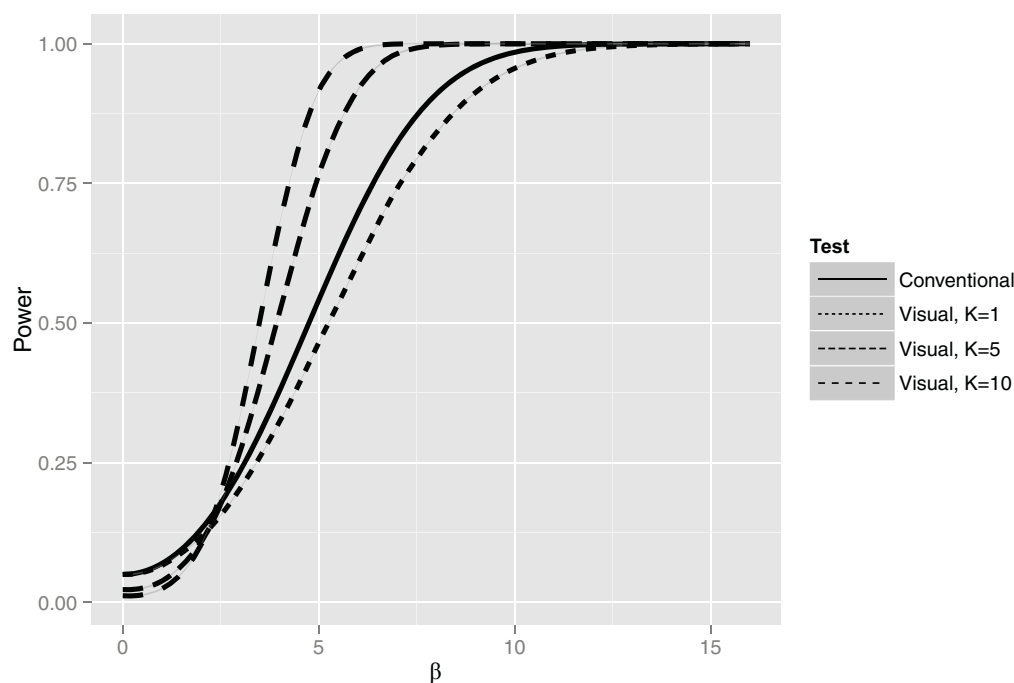
Figure 3. Comparison of the expected power of a visual test of size $m = 20$ for different K (number of observers) with the power of the conventional test, for $n = 100$ and $\sigma = 12$.

observers approaches infinity, leading to an asymptotically perfect power curve of zero in the null hypothesis and one for any alternative value. We observe this dichotomy of visual power in power estimates based on the data collected from user experiments, too. It features prominently in Figure 6.

5. HUMAN SUBJECTS EXPERIMENTS WITH SIMULATED DATA

Three experiments were conducted to evaluate the effectiveness of the lineup protocol relative to the equivalent test statistic used in the regression setting. The first two experiments have ideal scenarios for conventional testing, where we would not expect the lineup protocol to do better than the conventional test. The third experiment is a scenario where assumptions required for the conventional test are violated, and we would expect the lineup protocol to outperform the conventional test. (Data and lineups used in the experiments are available in the supplementary material.)

After many small pilot studies with local personnel, it was clear that some care was needed to set up the human subjects experiments. It was best for an observer or a subject to see a block of 10 lineups with varying difficulty, with a reasonable number of "easy" lineups. The explanations about each experiment (below) includes an explanation of how the lineups were sampled and provided to the subjects.

Participants for all the experiments were recruited through Amazon's online web service, Mechanical Turk (Amazon 2010). A summary of the data obtained for all three experiments are shown in Table 5. Participants were asked to select the plot they think best matched the question given, provide a reason for their choice, and say how confident they are in their choice. Gender, age, education, and geographic location of each participant are also collected. For each of the experiments, one

of the lineups was used as a test plot (easy plot) which everyone should get correct, so that a measure of the quality of the subjects effort could be made. Note that no participant was shown the same lineup twice.

5.1 Discrete Covariate

The experiment is designed to study the ability of human subjects to detect the effect of a single categorical variable X_2 (corresponding to parameter β_2) in a two variable ($p = 2$) regression model [Equation (3)]. Data are simulated using a range of values of β_2 or slopes as shown in Table 4, two different sample sizes ($n = 100, 300$) and two standard deviations of the error ($\sigma = 5, 12$). The range of β_2 values was chosen so that estimates of the power would produce reasonably continuous power curves, comparable to that calculated for the theoretical conventional test. Values were fixed for other regression parameters, $\beta_0 = 5$, $\beta_1 = 15$, and the values for X_1 were randomly generated from a Poisson ($\lambda = 30$) distribution, which is almost Gaussian. Three datasets were generated for each of the parameter values shown in Table 4 resulting in 60 different "actual datasets," and thus, 60 different lineups. For each lineup, the null model was fit to the actual dataset to obtain residuals and parameter estimates. The actual data plot was drawn as side-by-side boxplots of the residuals (Table 3, case 3). The 19 null datasets were generated by simulating from $N(0, \hat{\sigma}^2)$, and plotted in the same way. The actual data plot was randomly placed among these null data plots to produce the lineup. Figure 1 is an example of one of these lineups. It was generated for $n = 300$, $\beta_2 = 10$, and $\sigma = 12$. The actual data plot location is ($4^2 - 1$). For this lineup, 15 out of 16 observers picked the actual data plot.

The number of evaluations required for each lineup to provide reasonable estimates of the proportion correct ($\hat{p}$) is determined by the variance of the number of correct evaluations. Suppose,

Table 4. Combination of parameter values, β_2, n, and σ, used for the simulation experiments

Sample size (n)	Error SD (σ)	Experiment 1 Discrete covariate	Experiment 2 Continuous covariate	Experiment 3 Contaminated data
			Slope (β)	
100	5	0, 1, 3, 5, 8	0.25, 0.75, 1.25, 1.75, 2.75	0.1, 0.4, 0.75, 1.25, 1.5, 2.25
	12	1, 3, 8, 10, 16	0.5, 1.5, 3.5, 4.5, 6	
300	5	0, 1, 2, 3, 5	0.1, 0.4, 0.7, 1, 1.5	
	12	1, 3, 5, 7, 10	0, 0.8, 1.75, 2.3, 3.5	

γ denotes the conventional test power for each parameter combination shown in Table 4. Since the expected power of visual inference is very close to the power of conventional test (Figure 3 with $K = 1$) we consider $\gamma = p$. For a given proportion γ, it is desired to have a margin of error (ME) less than or equal to 0.05. Thus we have ME $= 1.96\sqrt{\gamma(1-\gamma)/n_\gamma} \leq 0.05$ which gives us the estimation of minimum number of evaluations

$$n_\gamma \geq \frac{\gamma(1-\gamma)}{(0.05/1.96)^2}.$$

Each subject viewed at least 10 lineups with the option to evaluate more. Depending on the parameter combinations, we group the lineups in different difficulty levels as easy, medium, hard, and mixed (actual numbers are given in the supplementary material). For each difficulty level a specific number of lineups was randomly picked for evaluation. This number is chosen so that total number of evaluations for each lineup for that group exceed the threshold n_γ. To satisfy this plan we needed to recruit at least 300 subjects.

5.2 Continuous Covariate

This experiment is very similar to the previous one, except that there is a single continuous covariate and no second covariate [Equation (3) with $p = 1$], following the test in Table 3, case 2. Data are simulated with two sample sizes ($n = 100, 300$), two standard deviations of the error ($\sigma = 5, 12$), and a variety of slopes (β), as given in Table 4. We arbitrarily set $\beta_0 = 6$, and values for X_1 are simulated from $N(0, 1)$. For each combination of parameters, at least three different actual datasets are produced, yielding a total of 70 lineups.

The actual data plot is generated by making a scatterplot of Y versus X_1 with the least squares regression line overlaid. To produce the null plots in the lineup null data were simulated from $N(X\hat{\beta}, \hat{\sigma}^2)$ and plotted using the same scatterplot method as the actual data. To select 10 lineups for a subject, each combination of sample size (n) and error SD (σ) is given a difficulty value based on the slope (β) parameters. For the smallest slopes, the difficulty is 4 (hardest) and for the largest slopes the difficulty is 0 (easiest). Figure 4 shows an example lineup for this experiment from difficulty level 4. This lineup is generated using a sample size (n) of 100, slope (β) of 1.25 and and error SD (σ) of 5. The actual data plot location is $(2^2 + 1)$. None of the 65 observers picked the actual plot while 46 observers picked plot 18 which has the lowest p-value among all the plots in this lineup.

For each combination of sample size and standard deviation, each participant is given five randomly selected lineups, one of each difficulty level. Another set of four lineups is chosen from a second tier of selected combinations of sample size and standard deviation, with difficulty levels 0 to 3. A last lineup was randomly selected from a set of lineups with difficulty level 0. The order in which the lineups are shown to participants is randomized.

5.3 Contaminated Data

The first two simulation experiments use data generated under a normal error model, satisfying the conditions for conventional test procedures. In these situations there exists a test, and there would, in general, be no need to use visual inference. The simulation is conducted in the hope that the visual test procedure, will at least compare favorably with the conventional test—without any ambition of performing equally well. This third simulation is closer to the mark for the purpose of visual inference. The assumptions for the conventional test are violated by contaminating the data. The contamination makes the estimated slopes effectively 0, yet the true value of slope parameter is not. The data are generated from the following model:

$$Y_i = \begin{cases} \alpha + \beta X_i + \epsilon_i & X_i \sim N(0, 1) & i = 1, \ldots, n \\ \lambda + \eta_i & X_i \sim N(\mu, 1/3) & i = 1, \ldots, n_c \end{cases},$$

where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma)$, $\eta_i \overset{\text{iid}}{\sim} N(0, \sigma/3)$ and $\mu = -1.75$. n_c is the size of the contaminated data. For the experiment, we consider $n = 100$ and $n_c = 15$ producing actual data with 115 points. Further, $\alpha = 0$, $\lambda = 10$, and σ is chosen to be approximately 3.5, so that error standard deviation across both groups of the data is 5. A linear model (Equation (3) with $p = 1$ and intercept $\beta_0 = 0$) is fit to the contaminated data. This experiment follows the test in Table 3, case 2. The actual data plot shows a scatterplot of the residuals versus X_1, and the null plots are scatterplots of null data generated by plotting simulated residuals from $N(0, \hat{\sigma}^2)$ against X_1.

Experiment 3 consists of a total of 30 lineups, made up of five replicates for each of the six slopes, as shown in Table 4. We use the slope directly as a measure of difficulty, with difficulty $= 0$ for the largest slope and difficulty $= 5$ for the smallest slope. Subjects were exposed to a total of ten lineups, with two lineups from each of the difficulty levels 0 through 3, and one lineup each from levels 4 and 5.

An example lineup for slope $\beta = 0.4$ is shown in Figure 5. Which plot is the most different from the others, in the sense that there is the steepest slope? The actual data plot location is $(3^2 - 2^3)$ and 13 out of 31 observers picked the actual plot.
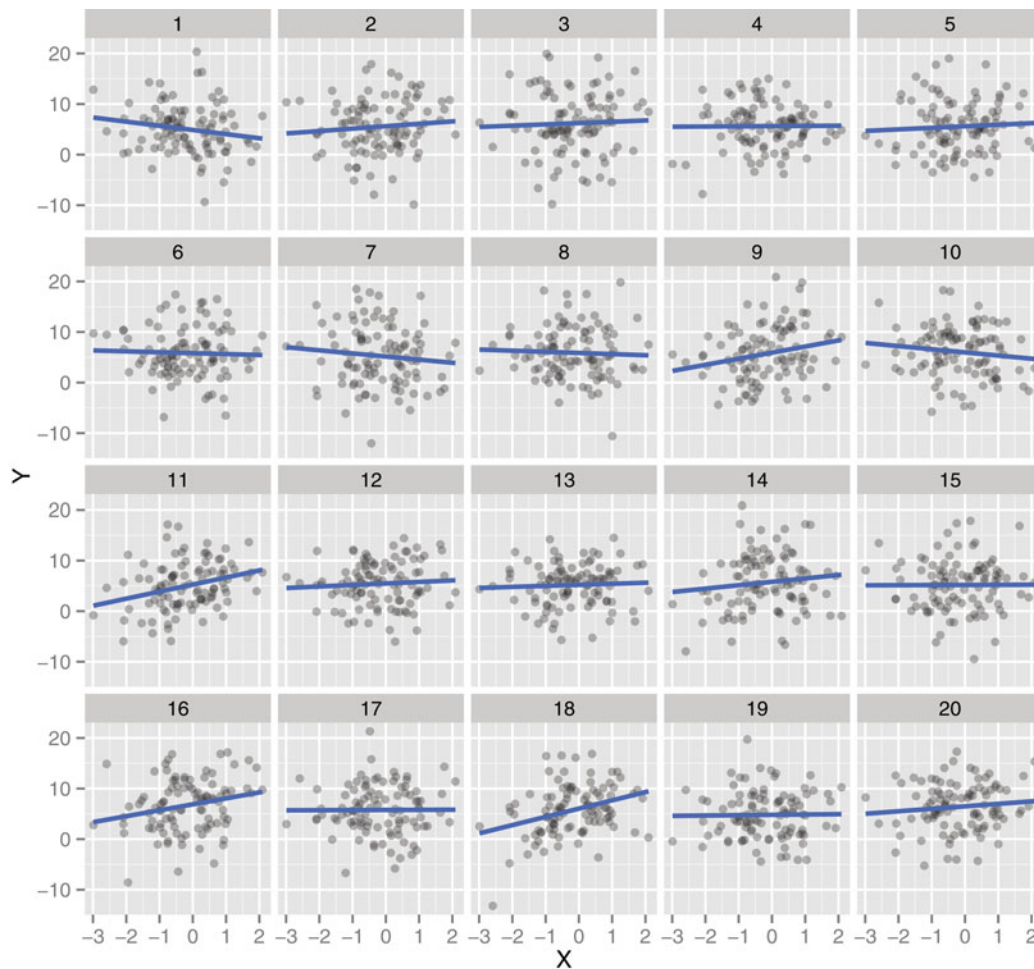
Figure 4. Lineup plot ($m = 20$) using scatterplots for testing $H_0 : \beta_k = 0$, where covariate X_k is continuous. One of these plots is the plot of the actual data, and the remaining are null plots, produced by simulating data from a null model that assumes H_0 is true. Which plot is the most different from the others, in the sense that there is the steepest slope? (The position of the actual data plot is provided in Section 5.2.) The online version of this figure is in color.

6. RESULTS

6.1 Data Cleaning

Amazon Mechanical Turk workers are paid for their efforts, not substantially, but on the scale of the minimum wage in the USA. Some workers will try to maximize their earnings for minimum effort, which can affect the results from the data. For example, some workers may simply randomly pick a plot, without actively examining the plots in the lineup. For the purpose of identifying these participants and cleaning the data, we use one of the very easy lineups that everybody was exposed to as a *reference lineup* and take action based on a subject's answer to this reference: if the subject failed to identify the actual data plot on the reference lineup, we remove all of this subject's data from the analysis. If the answer on the reference lineup is correct, we remove the answer for this lineup from the analysis, but keep all of the remaining answers. Table 5 tabulates the number of subjects, genders, and lineups evaluated after applying the data screening procedure.

6.2 Model Fitting

For each parameter combination, effect E is derived as $E = \sqrt{n} \cdot \beta / \sigma$.

The model in Equation (1) is fit using E as the only fixed effect covariate without intercept, that is, $W_{\ell i} = E_{\ell i}$. Instead of fitting an intercept, we make use of a fixed offset of $\log(0.05/0.95)$ so that the estimated power has a fixed lower limit at 0.05 (Type I error) when $E = 0$. Different skill levels of subjects are accounted for by allowing subject-specific random slopes for effect (E).

For Experiment 3, we do fit intercepts: both a fixed and subject-specific random effects, since forcing power to be fixed at 0.05 for $E = 0$ is not required by the experimental design.

Table 5. Number of subjects, gender, total lineups seen, and distinct lineups for all three experimental datasets. Note that in some of the lineups the number of male and female participants does not add up to the total number of participants due to missing demographic information

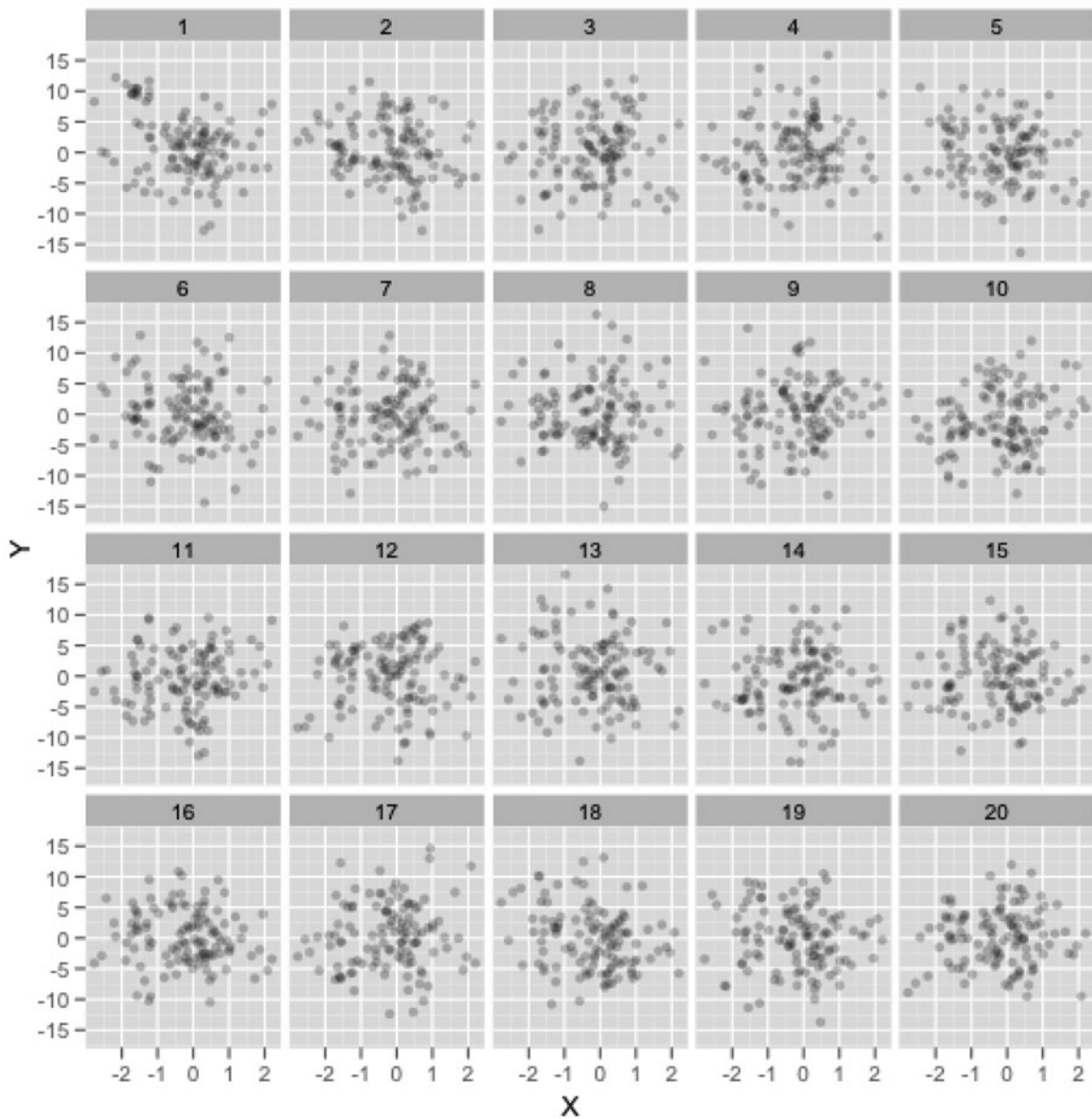| Experiment	Subject	Male	Female	Responses	Lineup
1	239	121	107	2249	60
2	351	185	164	3636	70
3	155	103	52	1511	29

Figure 5. Lineup plot ($m = 20$) using scatterplots for testing $H_0 : \beta_k = 0$, where covariate X_k is continuous but the inclusion of some contamination with the data spoils the normality assumption of error structure. One of these plots is the plot of the actual data, and the remaining are null plots, produced by simulating data from a null model that assumes H_0 is true. Which plot is the most different from the others, in the sense that there is the steepest slope? (The position of the actual data plot is provided in Section 5.3.)

For computation we use package `lme4` (Bates, Maechler, and Bolker 2011) and software R 2.15.0 (R Development Core Team 2012). p-value calculations are based on asymptotic normality.

Table 6 shows the parameter estimates of the mixed effects model of the subject-specific variation. The fixed effects estimates indicate that for all experiments the proportion of correct responses increases as the effect increases. This effect is less pronounced for Experiment 3. The subject-specific variability is smaller for Experiment 1, and relatively large for Experiment 3.

6.3 Power Comparison

Figure 6 shows an overview of estimated power against effect for the three experiments. Responses from each experiment are summarized by effect size and represented as dots, with size

indicating the number of responses. A loess fit to the data gives an estimate of the observed proportion correct $\hat{p}(E)$ for different effect sizes, with gray bands indicating simultaneous bootstrap confidence bands (Buja and Rolke 2011). $\hat{p}(E)$ is considered to

Table 6. Parameter estimates of model in Equation (1). Estimates are highly significant with p-value <0.0001 for all three experiment data

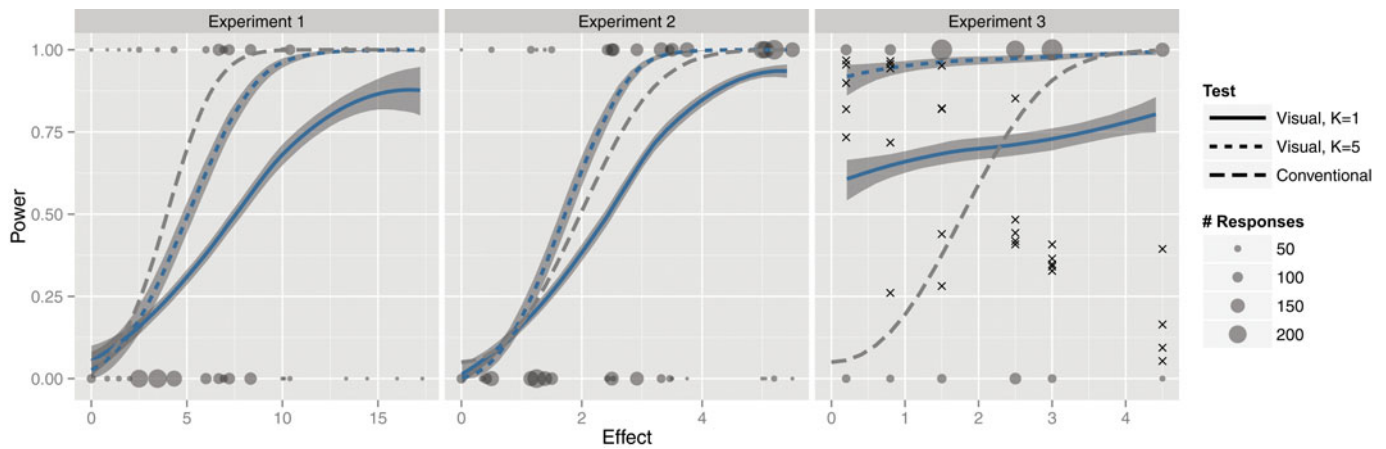| Experiment | Fixed effect | | Random effect |
	Estimate	Std. error	Variance
1	0.39	0.0094	0.0080
2	1.21	0.0197	0.0443
3	0.59 (Intercept)	0.1668	1.9917
	0.21 (Slope)	0.0511	0.0245
	−0.78 (correlation)		

Figure 6. Power in comparison to effect for the three experiments. Points indicate subject responses, with size indicating count. Responses are 1 and 0 depending on the success or failure, respectively, to identify the actual plot in the lineup. The loess curve (continuous line) estimates the observed proportion correct (power for $K = 1$), and surrounding bands show simultaneous bootstrap confidence band. Observed proportion is used to obtain power for $K = 5$. Conventional test power is drawn as a dashed line. For Experiment 3, conventional power is based on the slopes of the noncontaminated part of the data. Power of the conventional test for contaminated data is shown by cross marks. The online version of this figure is in color.

be the power for $K = 1$ and it is used to obtain power for $K = 5$. For comparison, the dashed lines show the corresponding power curves of the conventional tests. It is encouraging to see that visual inference mirrors the power versus effect relationship of conventional testing, in Experiments 1 and 2. In Experiment 3, the power of the visual test exceeds that for the conventional test, as expected. For larger values of K estimated power exceeds the power of conventional test. Note that for effect $E = 0$, the power is close to 0.05 (Type I error) for both Experiments 1 and 2, making the fixed offset a reasonable assumption.

Results for Experiment 3 are quite different. This is the situation where we expect to see the potential of visual inference, and indeed we do: the power of visual inference is always high, and much higher than the conventional test at small effect sizes. There is no actual conventional power in this situation, because assumptions are violated. The dashed line shows conventional power based on uncontaminated data, whereas the cross marks

show effective power based on the coefficient estimated from the contaminated data.

Results of Experiment 3 are curious insofar, as power of the visual test is largely independent of effect size. However, these results are based on correct identification of the actual data plot, regardless of reason. Although subjects were asked to select the plot that exhibited the highest association between the two variables, they might have cued in on the cluster of contaminated data. This will be explored further in Section 6.8.

6.4 Subject-Specific Variation

Subject-specific proportion correct $\hat{p}_i(E)$ is obtained using Equation (2) and it is used to obtain power for $K = 5$. Figure 7 shows power curves for both the overall experiment and subject-specific variations. The thick continuous line shows overall estimated power, the thinner lines correspond to subject-specific
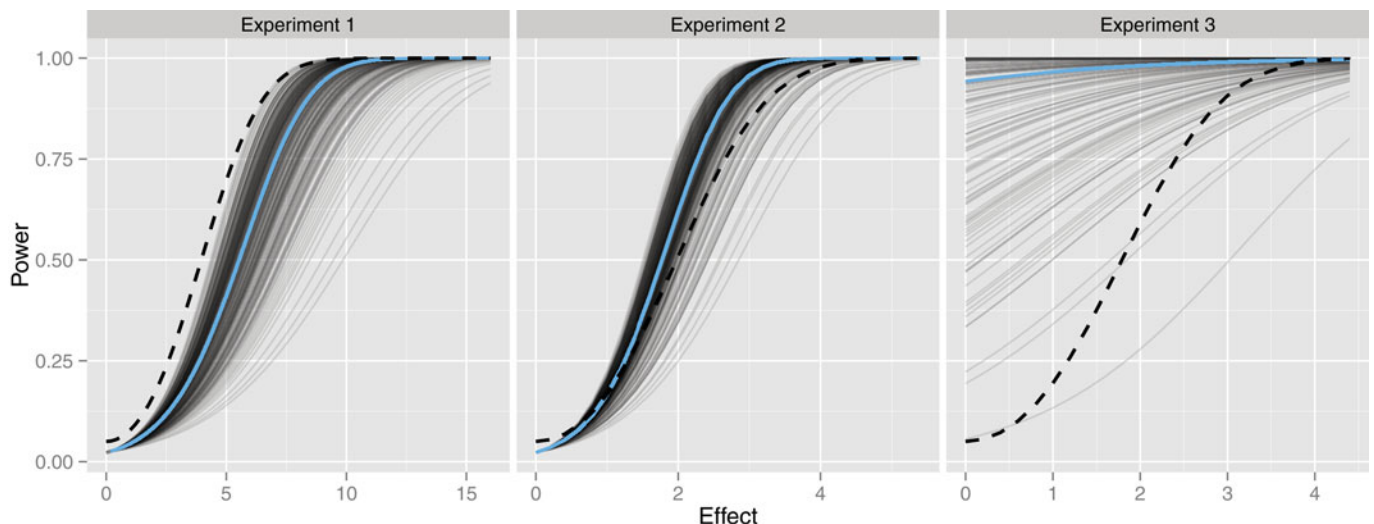

Figure 7. Subject-specific power for $K = 5$ obtained using the subject-specific proportion correct estimated from model 1. The corresponding power curve for conventional test (dashed line) is shown for comparison. The overall estimated average power curve is shown (light blue). The online version of this figure is in color.
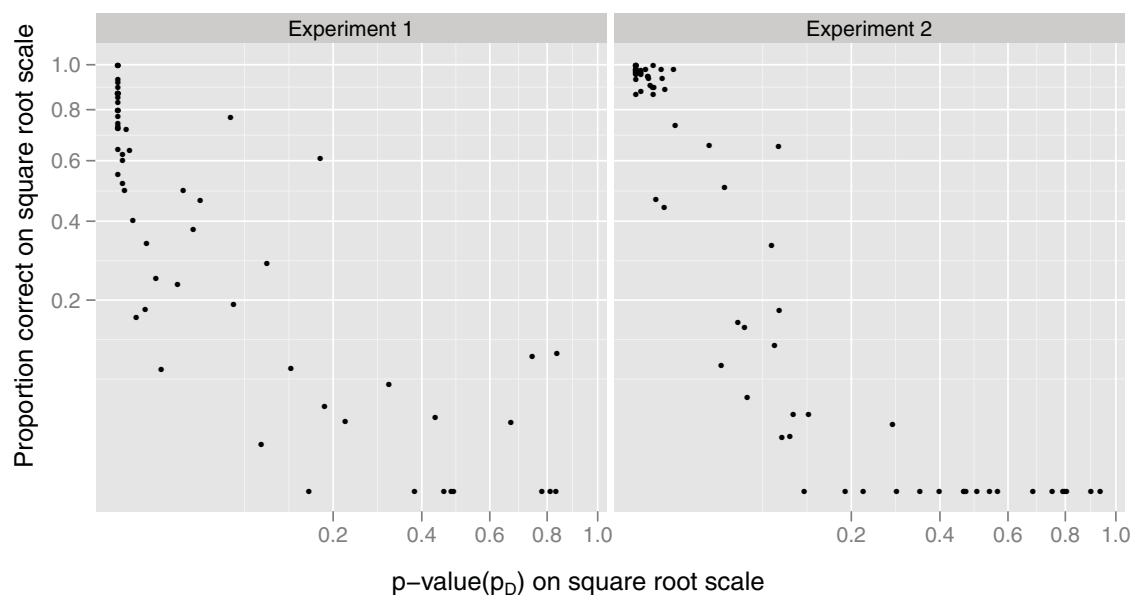
Figure 8. Proportion of correct responses decreases rapidly with increasing *p*-values. For *p*-values above 0.15, it becomes very unlikely that observers identify the actual plot. The theoretical justification of this is shown in Figure 2.

power curves. For comparison, the dashed lines show power curves of the conventional test. Subject-specific power is quite different between the three experiments. In Experiment 2, subjects performed similarly, and substantially better than the conventional test. In Experiment 1, there is more variability between subjects, with some doing better than the conventional test on large effects. In Experiment 3, there is the most subject-specific variation. Some subjects performed substantially better than the conventional test, and on average the visual test was better.

6.5 Estimating the *p*-value in the Real World

In the real setting, where visual inference is to be useful, there will be no conventional test *p*-values. Assessing the strength of perceived structure is a critical component of visual inference. In Experiments 1 and 2, there is a *p*-value associated with the actual data plot in each lineup. As the *p*-value increases the proportion of correct responses falls (Figure 8), which is evidence

of direct association between proportion of correct responses and conventional test *p*-values. For *p*-values larger than 0.15, it is very uncommon for subjects to correctly identify the actual data plot in the lineup.

From the experimental data, the visual *p*-values are estimated based on Definition 2.3. Figure 9 displays resulting estimates for each lineup against the conventional *p*-value. The pattern of visual *p*-values is interesting: for small *p*-values, the visual estimates tend to be very small, while lineups with larger *p*-values result in very large visual estimate, giving a clear indication to reject H_0 or not. This is why we do not see lot of visual *p*-values between 0.05 and 0.8 especially for Experiment 2. This guides the researcher to make decision confidently while conventional tests with marginal *p*-values make the decision whether to reject or not harder. For visual tests this is not common.

For Experiment 3, we see that the visual *p* values are very small no matter what the conventional *p*-values are. This is expected as the conventional test loses its power to reject H_0 even
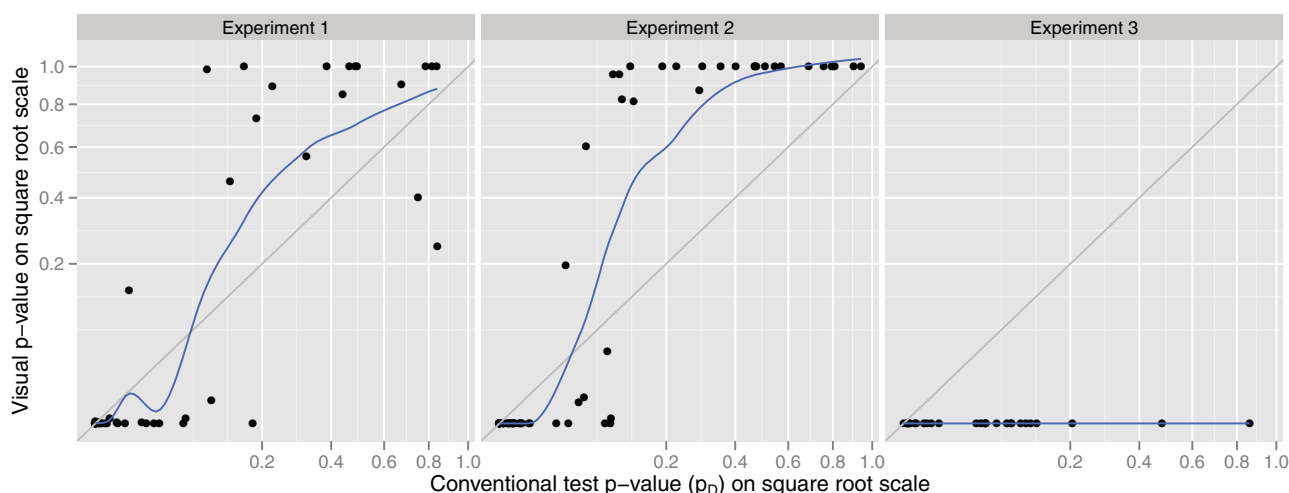

Figure 9. Conventional test *p*-value (p_D) vs visual *p*-value obtained from the definition. Values are shown on square root scale. The online version of this figure is in color.
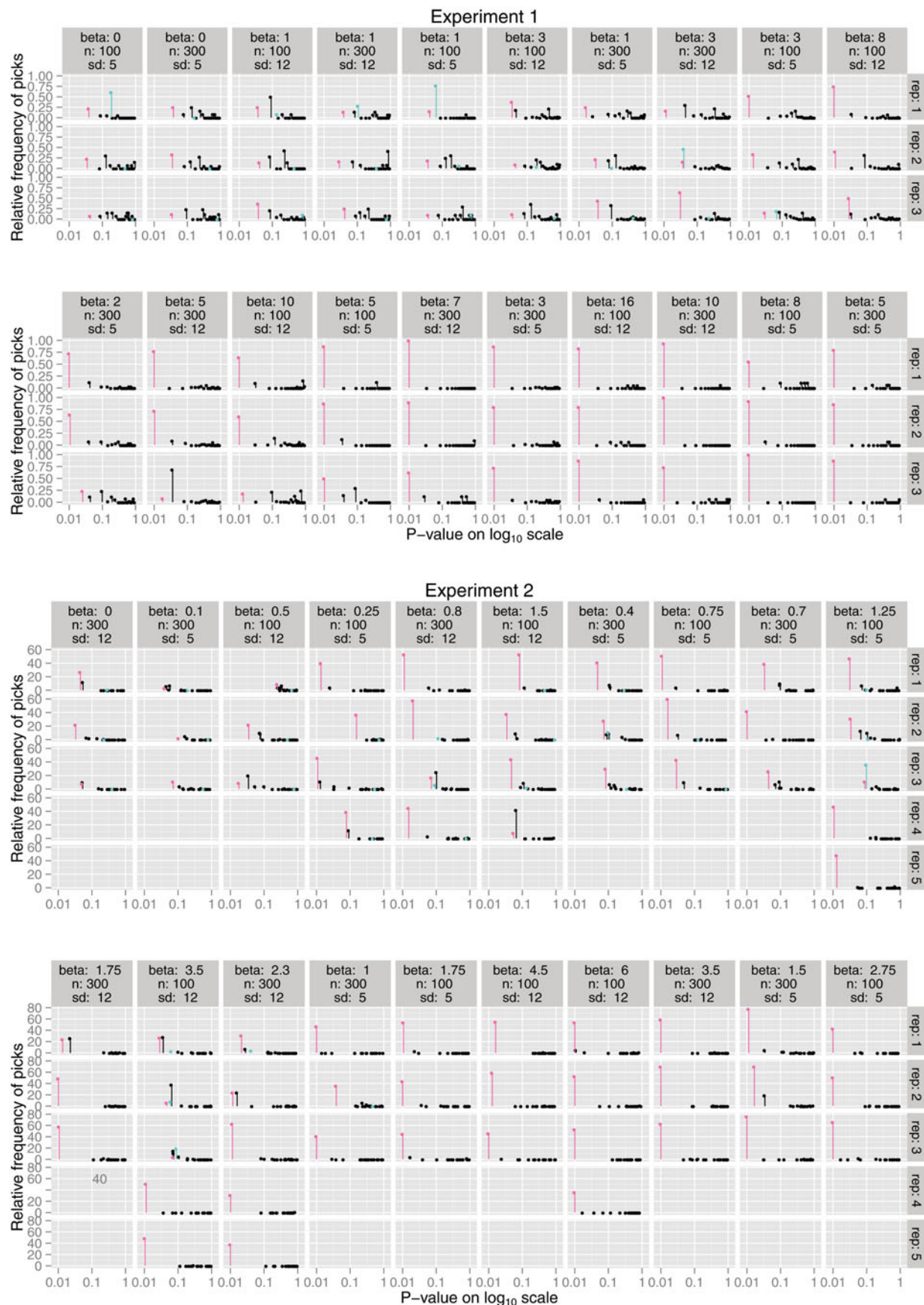
Figure 10. Relative frequency of plot picks compared to other plots in the lineup plotted against the *p*-value (on log$_{10}$ scale) of each plot for all individual lineups of both Experiments 1 and 2. Red indicates the plot with the lowest *p*-value, and blue indicates the actual data plot, when it is different from that with the lowest *p*-value. Columns are ordered according to effect size, with rows showing replicates of the same parameter combination on top of each other. Empty cells indicate combination of parameters that were not tested. Highest counts tend to be the plot in the lineup having the lowest *p*-value, more so for Experiment 2 than 1.

when the alternative is true, whereas the visual test performs well.

6.6 Do People Tend to Pick the Lowest *p*-value?

One assumption made to evaluate the effect of lineup size in the calculations of visual *p*-value and signal strength was that subjects would tend to pick the plot in the lineup that had the strongest signal. In Experiments 1 and 2, this corresponds to the plot with the smallest *p*-value. We examine the data collected from the first two experiments, to see if this assumption is, indeed, reasonable.

Figure 10 gives an overview of all selections in all lineups of Experiments 1 and 2. Each panel of the figure corresponds to a single lineup. Each "pin"—a short line topped by a dot—corresponds to one plot in the lineup. The *x*-location of the pin shows the plot's *p*-value on a log scale, its height is given by the number of observer choosing this plot. Columns are ordered according to effect size as defined in Section 6.2; rows show replicates for the same combination of parameters.

Red indicates the plot with the lowest *p*-value in the lineup. Blue indicates the plot of the actual data when it is different from that with the lowest *p*-value. In both experiments, people tended to select the plot with the lowest *p*-value. The results are clearer for Experiment 2, that used a continuous covariate. But even when subjects did not pick the plot with the lowest *p*-value they tended to oscillate their choices between the several low *p*-value plots. So for most subjects, the assumption that they pick the plot with the smallest *p*-value would appear to be reasonable, and the actual power of the visual test should be close to the expected power.

There are some noticeable exceptions to this rule. In Experiment 1, when $\beta = 0$, $n = 100$, $\sigma = 5$, rep = 1 people overwhelmingly chose a plot with much larger *p*-value, similarly, for parameters $\beta = 5$, $n = 300$, $\sigma = 12$, rep = 3, people tended to pick the plot with the second smallest *p*-value. For several of these exceptions, along with several easy lineups, a follow up experiment was conducted using an eye-tracker to examine which patterns or features participants are cueing on in making their choices (Zhao et al. 2012).

6.7 How Much Do Null Plots Affect the Choice?

Visual inference falls into the same framework as randomization tests, where the statistics from the data are compared with those from null data. Unlike randomization tests, visual inference is constrained to make the comparison with just a few draws ($m - 1$) from the null distribution. How this small set of null plots influences the subjects' choice is important for understanding the reliability of visual inference. If the actual data plot is very different from all of the null plots, then the null plots should not have much influence on the choice. Measuring the difference, generally, between plots is almost impossible. However, in this controlled setting we can use *p*-values of the test statistic calculated on the data used in each plot as a proxy for similarity of structure between the plots. If there is a null plot with a small *p*-value, or one close to that of the actual data plot, we would expect that subjects have a harder time detecting the actual data plot.

6.8 Type III Error

A little known error among statisticians is what was coined as Type III error in Mosteller (1948). Type III errors are defined as the probability of correctly rejecting the null hypothesis but for the wrong reason. Experiment 3 is prone to this type of error. Participants were asked to identify the plot with the largest absolute slope. But the actual data plot featured a cluster of points, the contamination that made the conventional test fail to see any trend. For the human eye, this cluster of points is as visible as the association between the remaining points, enabling the observer to identify the actual data plot by looking for the cluster instead of the slope. This would be considered a Type III error because it leads to a correct rejection of the null hypothesis, but is not related to the value of the slope parameter.

For visual inference, making a Type III, is not actually a problem. It is only a possibility in this experiment because we are working with known structure. In the real setting, we are excited to see observers detecting the actual data plot, and curious about how they detect it, with all possible reasons encapsulated in the alternative hypothesis. However, this highlights the importance of getting qualitative reasoning from observers for their choices.

7. CONCLUSIONS

This article has demonstrated that statistical graphics can be used in statistical inference and validates the lineup protocol proposed by Buja et al. (2009). Specific terminology was defined, and methods for obtaining the *p*-value and estimating the power of visual tests were introduced. To calculate the theoretical power, it was assumed that observers will select the plot having the strongest signal in the lineup, and the experimental data suggests that for most observers, this assumption holds. Results from visual inference in the controlled setting of the simulation study are comparable to those obtained by conventional inference. Visual inference is intended to provide valid tests where no conventional test exists, and our experiments in a controlled scenario suggest that it will perform as expected in the intended applications. The power of a visual test increases with the number of observers, which interestingly, leads to a result that the theoretical power of visual test can be better than that of conventional tests.

The lineup protocol operates similarly to statistical tests that have broad alternative hypotheses. If the null hypothesis is rejected, generally we can say that "there is something there" but not specifically what it is in the data that triggers the rejection. Follow-up questions on the reasons provide qualitative insight. In conventional testing, multiple comparisons are often done to refine and understand the test results, and perhaps some similar approaches might be developed for visual inference.

The performance of subjects was quite varied, but consistent. No restrictions were placed on Turk workers, in terms of abilities. There were clearly some subjects who performed very badly, but it was very interesting to see that there were some super-observers, people who detected the actual data plot at a rate better than that of the power of the best conventional test. It would be interesting to see how well-trained subjects might perform. Prior to the Turk experiments, we conducted pilot

studies using local graphics experts and obtained good results, indicating that training in data visualization might be helpful for visual inference. Future work might explore this.

Visual inference has been successfully used in two practical applications: to evaluate the power of competing graphical designs (Hofmann et al. 2012), and to detect signal presence in large p, small n data (Roy Chowdhury et al. 2011). It is hoped that the lineup protocol will prove to be valuable in data mining applications, and exploratory analyses, where there are no existing gauges of statistical significance.

SUPPLEMENTARY MATERIALS

Proof of Lemma 3.1, details of data collection and cleaning, longer discussion of effect of null plots and Type III error.

[Received August 2012. Revised April 2013.]

REFERENCES

Amazon (2010), "Mechanical Turk," available at *https://www.mturk.com/mturk/welcome*. [942,948]

Bates, D., Maechler, M., and Bolker, B. (2011), *Lme4: Linear Mixed-Effects Models Using S4 Classes*, R package version 0.999375-42. [951]

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., and Wickham, H. (2009), "Statistical Inference for Exploratory Data Analysis and Model Diagnostics," *Royal Society Philosophical Transactions*, Series A, 367, 4361–4383. [942,955]

Buja, A., and Rolke, W. (2011), "Calibration for Simultaneity: (Re)Sampling Methods for Simultaneous Inference with Applications to Function Estimation and Functional Data," unpublished manuscript. [951]

Butler, K., and Stephens, M. (1993), "The Distribution of a Sum of Binomial Random Variables," *Technical Report 467 Stanford University Stanford Calif USA April 1993 Prepared for the Office of Naval Research*. [945]

Cleveland, W. S., and McGill, R. (1984), "Graphical Perception: Theory, Experimentation and Application to the Development of Graphical Methods," *Journal of the American Statistical Association*, 79, 531–554. [942]

Cook, R. D., and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics*, New York, NY: Wiley Series in Probability and Statistics. [946]

Heer, J., and Bostock, M. (2010), "Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design," in *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, New York: ACM, CHI 10, pp. 203–212. [942]

Hofmann, H., Follett, L., Majumder, M., and Cook, D. (2012), "Graphical Tests for Power Comparison of Competing Designs," *IEEE Transactions on Visualization and Computer Graphics*, 18, 2441–2448. [956]

Mosteller, F. (1948), "A k-Sample Slippage Test for an Extreme Population," *The Annals of Mathematical Statistics*, 19, 58–65. [955]

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0. [942,951]

Roy Chowdhury, N., Cook, D., Hofmann, H., and Majumder, M. (2011), "Visual Statistical Inference for Large p, Small n Data," in *Proceedings of JSM 2011*, pp. 4436–4446. [956]

Simkin, D., and Hastie, R. (1987), "An Information Processing Analysis of Graph Perception," *Journal of the American Statistical Association*, 82, 454–465. [942]

Spence, I., and Lewandowsky, S. (1991), "Displaying Proportions and Percentages," *Applied Cognitive Psychology*, 6, 61–77. [942]

Tukey, J. W. (1977), *Exploratory Data Analysis*, Boston, MA: Addison-Wesley. [942]

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis, useR*, New York, NY: Springer. [942,944]

Wilkinson, L. (1999), *The Grammar of Graphics*, New York: Springer. [942,944]

Zhao, Y., Cook, D., Hofmann, H., Majumder, M., and Roy Chowdhury, N. (2012), "Mind Reading Using an Eyetracker to See How People Are Looking at Lineups," Technical Report 10, Iowa State University, Department of Statistics. [955]