Human Factors Influencing Visual Statistical Inference

Mahbubul Majumder, Heike Hofmann, Dianne Cook [*]

Visual statistical inference is a way to determine significance of patterns found while exploring data with graphics. It relies on human observers inspecting a lineup: one data plot randomly placed among a set of decoy plots. Each individual is different in their cognitive psychology and judiciousness, which can affect the visual inference. The usual way of estimating the effectiveness of a statistical test is its power. The estimate of power of a lineup is controlled by combining evaluations from multiple observers. Factors that may affect the power of visual inference are the observers' demographics, visual skills, and experience, the sample of null plots taken from the null distribution, the position of the data plot in the lineup, and the signal strength in the data. This paper examines these factors. Results from multiple visual inference studies using Amazon's Mechanical Turk are examined to provide an assessment of these. The experiments suggest that individual skills vary substantially, but demographics do not have a huge effect on performance. There is evidence that a learning effect exists but only in the sense that observers become faster with repeated evaluations, while accuracy stays the same. The placement of data plot in the lineup does not affect inferential results.

Keywords: statistical graphics, non-parametric test, cognitive psychology, data visualization, exploratory data analysis, data mining, visual analytics.

1. INTRODUCTION

The lineup protocol introduced in Buja et al. (2009) can be used to quantify the significance of graphical findings in the exploratory data analysis process. This methodology is part of what is called visual statistical inference, which has been developed further and validated by Majumder et al. (2013) in simulation studies of head to head comparisons with conventional inference. One of their major contributions is to define the power of visual tests and to provide a method for obtaining the power for a particular lineup. In some scenarios, the power of a visual test was seen to outperform that of a conventional test.

In visual inference, the test statistic is a plot of the observed data, called the *(actual) data plot*. To create a lineup the data plot is placed randomly in a grid of null plots. Null plots are rendered from data generated as specified by a null hypothesis. Often, this null hypothesis assumes independence or no structure in the data. An observer is then asked to evaluate the lineup. If the actual data plot is detected by the observer, this counts as evidence against the null hypothesis. Based on the number of evaluations by independent observers and the number of times that the data plot is identified, we are able to evaluate the significance of the graphical relationship in the same way and the same rigor as a conventional hypothesis test.

Figure 1 displays an example lineup, one of the plots is based on the observed data, while the remaining 19 plots show data generated from a null model. Which one of the 20 plots is the most different from the others? When asked this question, 12 out of 72 observers picked the data plot located in panel number $4^2 - 3$ in the lineup[1]. The corresponding p-value is 0.0077, indicating sufficient evidence to reject the null hypothesis.

How do we interpret this finding, though? For that, we need to know the context of the data and we need to have more information about how the null plots were generated. This particular example investigates the results from the 2012 US presidential election in comparison to poll results just prior to the election. Although this example is more simplistic than most of the tests conducted to date, it serves the purpose of illustrating the lineup protocol. The data consists of the difference in poll results between the two (major) presidential candidates, Obama and Romney, for all states. Each panel in Figure 1 shows an 'electoral building' (Mosley et al., 2010) where each state in the

---

[1]The little piece of calculus used for describing the location of the data plot in the lineup provides a small cognitive hurdle that is supposed to enable the interested reader to inspect the lineup without being biased by knowing the location of the data plot.
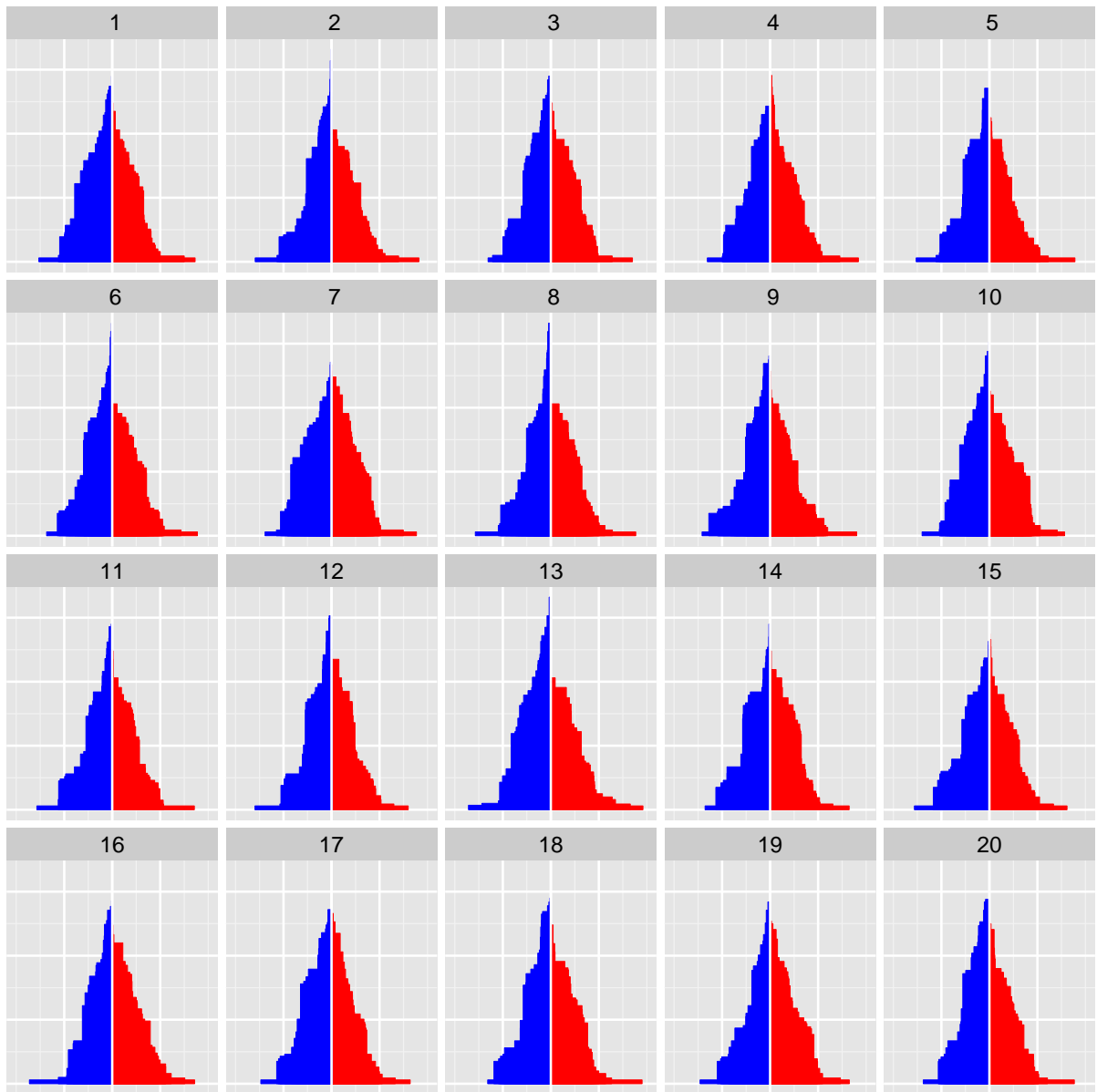
Figure 1. *Which one of the plots, labelled 1 through 20 from top left to bottom right, is the most different from the others?*

union is represented by a rectangle. The difference in poll results is plotted horizontally, while the height of each box corresponds to the number of that state's electoral votes. Color indicates party affiliation. The null hypothesis is that "the election results are consistent with the polls". Therefore the polling results provide the null model from which data is simulated. A normal model with mean and standard deviation based on a poll and its margin of error is used to simulate different scenarios that might have resulted on election day, if the polls were on target. Each null data set is generated as a set of draws from this model. These samples are plotted as electoral buildings, and the plot based on the actual election results is placed randomly among the null plots in a lineup of size 5×4. If the null hypothesis is true, the data plot should look like any of the other plots, and not be identifiable by an observer. Figure 2 shows the plot of the electoral building with the additional context information and labels.
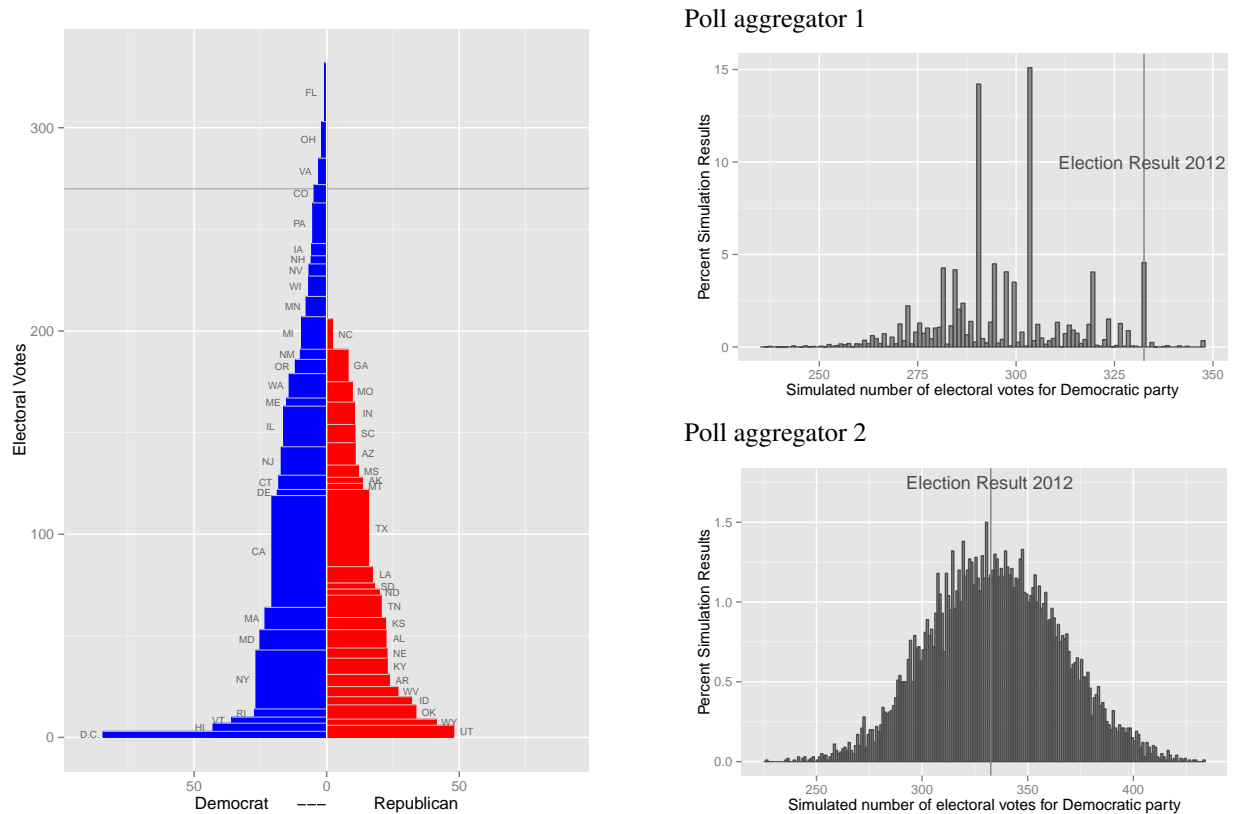

Figure 2.   Electoral building plot of the results of the 2012 U.S. Presidential Election (left). On the right two histograms showing simulated poll results for the Democratic presidential candidate using polling averages collected from two different sources, Freedom's Lighthouse Averages (top), RealClearPolitics (bottom). If the polls matched the electoral result the simulated results would be centered at the electoral result. This happens for the RealClearPolitics polls (bottom) but not for Freedom's Lighthouse. For the latter, only 5.33% of simulated results were as high as the actual election result, which converts to a p-value of 0.0533, indicating an almost statistically significant difference. It hints at the polls from this source under-estimating the vote for Obama. This source was used to generate the null plots in the lineup in Figure 1. The p-value from the lineup protocol, was 0.0077, which more strongly concludes that these polls mismatched the actual result.

The lineup of Figure 1, as well as the other lineups in the manuscript are created using the nullabor package by (Wickham et al., 2014).

A lineup can be evaluated by a single person or multiple observers. A distribution similar to the binomial distribution, but adjusted for dependencies introduced by the lineup scenario, is used to calculate the p-value based on the number of times observers identify the actual data plot, which provides the information needed to make a decision on rejecting or failing to reject the null hypothesis (Hofmann and Röttger, 2015). To avoid expectation errors (Meilgaard et al., 2006), particularly in emotionally charged areas such as election results, observers should not be aware of the data that constitutes a lineup, and should therefore not have seen the actual data plot before inspecting the lineup.

This is the reason that we revealed the context of the election problem only after the interested reader had a chance to evaluate the lineup.

When presenting a lineup to an observer the accompanying question should be phrased in as general a manner as possible, effectively asking the observer to pick the plot that is most different, and allowing observers to provide their own reasons for their choice. This ensures all possible deviations from the null hypothesis to be detected. However, if lineup tests are run in head-to-head comparisons with conventional statistical tests, such as the experiments in Majumder et al. (2013) or Yin et al. (2013), these questions have to be phrased much more specifically, in order to avoid type III errors (Mosteller, 1948) consisting of correctly rejecting the null hypothesis for a wrong reason. In those experiments, the structure in the data is strictly controlled in the simulation process, which allows for specific questions to be asked. In the election example, observers were asked "which plot is the most different?". The type of plot, showing two (modified) stacked bar charts in different colors should suggest to the observer that the interesting feature is the difference between the two heights. Most observers based their decision on this – besides 'asymmetry' between the towers (a reason suggested by us), most free-form responses for the reason of choice describe a comparison between the red and blue towers. In the responses for this lineup we see something that is very typical of responses in lineups: it is only a few panels from the lineup that observers pick, while most panels do not get picked or get picked only a few times. Out of a total of 72 responses, 23 observers picked panel #8, 12 observers picked the (data) panel #13, 10 observers picked panel #4, 6 observers picked panel #7, and all of the remaining panels were picked at most three times. From the panels picked, we see that the top three choices are all extreme with respect to heights of blue versus red towers: both panels #8 and #13 show a large difference between the height of the towers, while panel #4 is the only panel, in which the red tower is higher than the blue one.

The lineup protocols allows us to calculate all relevant properties that we are familiar with from conventional statistical tests. In particular, the power of a lineup is calculated as the *detection rate* at which observers identify the actual data plot. Visual power depends unlike power in conventional tests not only on the strength of the signal, but also on individuals' ability. The ability of individual observers varies, and the effects that might influence this ability are the focus of this paper.

We have conducted a series of experiments (see Table 1 for an overview) using Amazon's Mechanical Turk (Amazon, 2010) (MTurk) for a variety to purposes: validating the protocol against existing tests, comparing plot designs, and evaluating structure in data analysis problems. In all of the experiments we collected demographic details, such as gender, age bracket, and educational background. Mining this data will provide a basis for evaluating the effects of different covariates on an individual's ability to evaluate lineups. The design of some of the experimets allow a further investigation into aspects beyond participants' demographics: experiments 5, 6 and 7 can be used to investigate short term learning trends, while experiment 9 allows insight on whether the positioning of the data plot in the lineup has an effect on detection rate. Section 2 discusses human factors that we expect to be influential on the performance of the observer. Section 3 describes the methods used to assess the effects, and Section 4 summarizes the results of what we have learned about human factors affecting lineup evaluation.
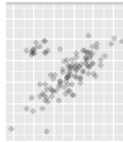
2. FACTORS POTENTIALLY AFFECTING VISUAL INFERENCE

Visual statistical inference relies not only on the strength of the signal in the data, but also on how this information is presented in the plot, on the lineup design and on the abilities of human observers. It is important to understand how these factors might affect results. Here, we provide a brief discussion of the factors that are expected to have some impact on the performance of visual statistical inference.

- **Choice of Visual Test Statistic:** Data can be plotted in a variety of ways and we can enhance plots by axes, legends or representations of models. The choice of plot primarily should be appropriate for the problem being investigated, as the data plot is the visual test statistic. When for example studying the association between two quantitative variables a scatterplot is usually considered best practice. Enhancements, such as a line overlaying a scatterplot to represent a fitted model, might be appropriate for studying models for the data. Some choices of visual test statistics provide better results than others when studying the same problem (Hofmann et al., 2012), because some plot types make it easier for certain structures to be seen.

  Table 1 shows the large variety of visual test statistics used in the Turk studies completed so far. These experiments were conducted to examine many different visual tasks. Side-by-side boxplots were used in experiment 1 to study the importance of a categorical variable in a regression model. For both experiments 2 and 3, a scatterplot is the visual test statistic, but a regression line fitted through the points is overlaid for experiment 2. The ex-

4

Table 1. Overview of 10 different Turk experiments, from which data was collected to study human factor effects. All of the experimental data was used to estimate the effect of demographic factors (DF) on visual inference while three were suitable for assessing learning trend (LT) and location effect (LE) was possible to assess using just one specially designed study.

	Experiment	Test Statistic	Lineup question	Used in study of		
1	Box plot		Which set of box plots shows biggest vertical difference between group A and B?	DF		
2	Scatter plot		Of the scatter plots below which one shows data that has steepest slope?	DF		
3	Contaminated plot		Of the scatter plots below which one shows data that has steepest slope?	DF		
4	Polar vs Cartesian		Which plot is different?	DF		
5	Hist vs density		In which plot is the blue group furthest to the right?	DF	LT	
6	Violin vs boxplot		In which plot does the blue group look the most different from the red group?	DF	LT	
7	Group separation		Which of these plots has the most separation between the coloured groups?	DF	LT	
8	Sine Illusion		In what picture is the size of the curve most consistent?	DF		
9	Gene expression		In which of these plots is the green line the steepest, and the spread of the green points relatively small?	DF		LE
10	Test normality		Which of these plots is most different from the others?	DF		

periments studied slightly different problems, for which the difference in plot type was important. Experiment 4 compared polar coordinates to euclidean coordinates on the reading of structure in a barchart. Experiment 5 examined the style of plot for a bivariate data problem, by comparing results when using a scatterplot, side-by-side dot plots, overlaid densities or side-by-side boxplots. Experiment 6 studied the difference between variations of boxplots. Experiment 7 studied the effects of high dimension low sample size on the separation of groups in low-dimensional projections. The sine illusion, where people read the strength of differences between curves as different when there is a strong sinusoidal structure, is examined in experiment 8. Experiment 9 was conducted to test if there was any structure in an RNA-seq data set measured on soybeans, and used interaction plots classically constructed to examine interaction effects in 2×2 factorial designs. Experiment 10 used different sorts of residual plots for hierarchical linear models as the visual test statistics.

In experiments 4, 5 and 6, different visual test statistics were used on the same data, providing the ability to examine the effect of these choices, the use of lineup tests for comparisons and examples are discussed in Hofmann et al. (2012); Loy et al. (2015)

- **Signal in the Data:** A visual test statistic is chosen for a particular task. If it is well-designed departures from the null hypothesis will be visible in the plot, especially if the signal is strong. An important factor that helps an observer to identify the actual data plot in a lineup is the strength of the signal in the data. Simulated data was generated for all of the experiments, except exps. 4 and 9, enabling the study of signal strength and actual data plot detection. In experiment 4 we took a different approach by varying the sampling size, which also allowed us to investigate the effect a design had on detection rate while varying signal/noise ratio. Primarily the results were as expected, that as the signal strength increased observers more frequently picked the actual data plot as different from the null plots.

- **Question Design:** A primary purpose of visual inference is to preclude the necessity to pre-specify discoverable features in data that is required by classical statistical inference. Visual inference supports discovery of structure, enabling the unexpected to surprise us. This is tightly connected to the task that the human observer is asked to perform. For most purposes the observer is asked a very general question, such as to pick the plot that is most different from the others, and explain the reasons why they see it as different. By being general, all possible discoverable structures are included in the significance calculations (see also Buja et al., 2009).

In the initial Turk experiments (1,2,3) conducted with simulated data to compare the results from visual inference with those obtained by classical inference, the questions posed to the observers were more targeted. This was important to allow a direct comparison of the results to those from conventional testing. For latter studies (4, 7, 10), questions posed to the observer were more general. Table 1 lists the questions used for each experiment.

- **Demographics:** During each of the experiments, data on age, gender, education level and geographic location was collected. Each observer self-reported gender, age in roughly five year intervals, education level as high school or less, some college courses, an undergraduate degree, some graduate courses or a graduate degree. The IP address of the computer afforded the geographical location of the subject. The purpose of collecting this information with each experiment is to examine the effect that they have on the results of visual inference – ideally very little.

- **Learning Trend:** One might expect that as an observer evaluates more lineups they become more skillful in their evaluations. Each participant in experiments 5, 6 and 7 was asked to evaluate a block of ten lineups of the same type of data plot. The ten lineups were randomly chosen from the lineups produced for each study. Before evaluating the first lineup, the observer needs to read instructions and become accustomed to the type of plot used in the lineup. In subsequent lineups the type of plot is the same. It is possible that the observer becomes more skillful at recognizing the most different plot, either by more often detecting the actual data plot or more quickly reporting their choice. These two ways of measuring learning trend are evaluated on data collected from experiments 5, 6 and 7.

- **Location of Actual Data Plot in the Lineup** For all of the lineups used in the experiments a 5×4 grid of 20 panels is used. A random number generator is used to determine the position where the actual data plot is placed in each lineup. Eye tracking experiments (Zhao et al., 2013) suggest that some observers traverse lineups in horizontal direction, while others have a vertical up-and-down approach to their search. Almost universally, observers start at the top left of a lineup. This leads to some concern that observers might be able to more easily identify the actual data plot, if it is placed at the top left of a lineup than when it the actual data is placed at the bottom right. Ideally, this does not happen. Experiment 9 is designed to allow us to investigate the effects of

location on detection. Five different locations in the lineup were used to assess how fast and accurately observers identified the actual data plot.

- **Sample of Null Plots:** In classical inference, the test statistic is compared to the full null distribution, to decide if it is extreme or not. In visual inference the actual plot is compared to plots of a finite number of samples from the null distribution. In the lineups used in the Turk experiments, the actual data plot is compared on 19 null plots. These null samples are random draws, and there is a chance that one or more null plots in the lineup may be similar or even more extreme than the actual data plot. The sample of null plots can affect the observer's decision. This is tested with the data from experiment 9, where the experiment was set up with lineups made from different null plots.

- **Individual Skill or Ability:** Each person may have different aptitude for reading statistical plots, and their visual skill sets might be differently developed. We can examine the effect of individuals' skills and abilities because multiple subjects evaluated the same lineups, which allows us to estimate, if some subjects consistently detect the actual data plot more often than others.

3. EXPERIMENTAL METHODS

Two of the factors, signal in the data and individual abilities, were studied in Majumder et al. (2013). The choice of visual test statistic was examined in Hofmann et al. (2012). In each of these analyses demographic factors were given a cursory glance, to ensure that they did not have large effects on the results. The design of experiments 5, 6 and 7 enables the examination of learning trend, which is studied in this paper. Experiment 9 was a real test case for visual statistical inference, and in order to understand the significance of the structure in the genomic data, multiple lineups were made in which location of the actual data plot, and the sample of nulls, were randomized. This enables the assessment of the effect of these factors on the results. This section describes the experimental methods used to examine the effects of demography, placement of the actual data plot, sample of null plots and the existence of a learning trend.

3.1 Demographic Factors

For all ten experiments shown in Table 1, the following demographic information was collected from subjects:

1. *Age group*, with categories set to be 18-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, above 60.
2. *Gender*, male or female.
3. *Education level*, with levels being high school or less, some undergraduate courses, undergraduate degree, some graduate courses, and graduate degree.
4. *Geographical location*, collected from the IP address of the participants' computer, as latitude, longitude, city and country.

Let Y_{ij} denote the response from observer i on a lineup j, with $Y_{ij} = 1$ if the actual data plot is chosen, otherwise $Y_{ij} = 0$. The factors are examined in association with the observer's response using a logistic regression with random effect terms:

$$g(\pi_{ij}) = \mu + \alpha_{k(i)} + \gamma_{l(i)} + \tau_{m(i)} + \kappa_{s(i)} + \ell_j, \tag{1}$$

where $\pi_{ij} = E(Y_{ij})$ is the probability that observer i picks the actual data plot from lineup j, μ is an overall population average, α, γ, τ and κ are the effects of age group $k(i)$, gender $l(i)$, education level $m(i)$ and country name $s(i)$, respectively, for observer i. The term ℓ_j is a random intercept predicting lineup difficulty level and we assume independence and normality of the errors, i.e. $\ell_j \sim N(0, \sigma_\ell^2)$. $g(.)$ denotes the *logit* link function $g(\pi) = \log(\pi) - \log(1 - \pi); 0 \leq \pi \leq 1$.

Similarly, we model the time an observer takes to identify a panel from a lineup. Let Z_{ij} denote the logarithm of time taken for observer i to evaluate lineup j. Let $\mu_{ij} = E(Z_{ij})$ be the average of the (log) time taken by observer i to pick a panel from lineup j. We model this in a mixed effects model of the same structure as model (1) given as:

$$Z_{ij} = \mu + \alpha_{k(i)} + \gamma_{l(i)} + \tau_{m(i)} + \kappa_{s(i)} + \ell_j + \epsilon_{ij}, \tag{2}$$

where μ represents overall average of log time taken by an observer to evaluate a lineup, α, γ, τ and κ are as described in model (1), ℓ_j is a lineup-specific random effect for the time needed to evaluate a lineup, with $\ell_j \sim N(0, \sigma_\ell^2)$ and the overall error $\epsilon_{ij} \sim N(0, \sigma^2)$.

3.2 Learning Trend

Learning trend of a subject can be observed in terms of performance over successive responses when multiple lineups are shown for evaluation. Experiments 5, 6 and 7 were used for this. Each subject was shown a total of 10 lineups randomly selected from a pool of lineups. The lineups are not necessarily of the same difficulty level, but the order of lineups was randomized. The responses of the lineups were recorded by attempt 1 through 10. Attempt 1 means that the response is for the first lineup the observer evaluates and attempt 10 refers to the response for the 10th lineup. The goal is to estimate whether performance of the observer improves, or changes, from attempt 1 to attempt 10.

It should be noted that we are examining the observer's performance, when we model response as detected or not, but this is not the goal of visual inference. Visual inference is constructed to measure the significance of structure discovered in data. It is expected that some observers will be more skilled at reading data plots, and hence, more readily detect the plot that is different. It is also expected that as observers gain experience in evaluating lineups that they become more proficient in reading data plots, particularly if feedback is given on whether the actual data plot was chosen or not. Choosing the actual data plot will be more difficult in some lineups than others, and indeed should happen purely by chance in some lineups. So in this context, detected, or not, is used as a response to examine individual differences.

Let Y_{ijk} denote the response from observer i on lineup j at their kth evaluation attempt, where $Y_{ijk} = 1$ if the observer detected the actual data plot otherwise $Y_{ijk} = 0$. Let $\pi_{ijk} = E(Y_{ijk})$ be the probability that observer i picks the actual data plot from lineup j in their kth attempt. Learning trend is assessed using a generalized mixed effects model of the form

$$g(\pi_{ijk}) = \mu + \alpha_k + u_i + a_i k + \ell_j, \tag{3}$$

where μ is an overall population average, α_k is the effect of the kth attempt on the probability, using the first attempt as reference, $\alpha_1 = 0$, and $k = 1, ..., K$, u_i and a_i are observer specific random effects, $i = 1, ..., I$. The term, u_i is a random intercept, describing a basic subject-specific ability, with $u_i \sim N(0, \sigma_u^2)$. The term a_i is a random slope capturing an individual's specific learning effect over the course of K attempts, where $a_i \sim N(0, \sigma_a^2)$. For ℓ_j a normal distribution, $N(0, \sigma_\ell^2)$, is assumed, and ℓ_j is a random intercept predicting lineup difficulty level. $g(.)$ denotes the *logit* link function $g(\pi) = \log(\pi) - \log(1 - \pi); 0 \le \pi \le 1$. The inverse link function, $g^{-1}(.)$, from equation 3 leads to the estimate of the subject and the lineup specific probability of successful evaluation in the kth attempt by a single observer as

$$\hat{p}_{ijk} = g^{-1}(\hat{\mu} + \hat{\alpha}_k + \hat{u}_i + \hat{a}_i k + \hat{\ell}_j). \tag{4}$$

When time taken to evaluate a lineup is used as the response, let Z_{ijk} denote the logarithm of time taken for an observer i to evaluate a lineup j in his/her kth attempt. Let $\mu_{ijk} = E(Z_{ijk})$ be the average of the (log) of time taken by observer i to choosing a panel from lineup j in his/her kth attempt. We evaluate this in a mixed effects model of the form

$$Z_{ijk} = \mu + \alpha_1 + \alpha k + u_i + a_i k + \ell_j + \epsilon_{ijk}, \tag{5}$$

where μ represents overall average of log time taken by an observer to evaluate a lineup, α is the average change in log time taken for each additional attempt, α_1 is an offset in log time taken for the first attempt. All other effects are random effects: u_i is a subject-specific intercept representing individual speed of an observer with $u_i \sim N(0, \sigma_u^2)$, a_i is a subject-specific slope representing the deviation of the speed-up (or -down) by attempt k, with $a_i \sim N(0, \sigma_a^2)$, ℓ_j is a lineup-specific random effect for the time needed to evaluate a lineup, $\ell_j \sim N(0, \sigma_\ell^2)$ and the overall error $\epsilon_{ijk} \sim N(0, \sigma^2)$. Equation 5 leads to the estimate of the subject and the lineup specific time taken for an evaluation in kth attempt by a single observer as

$$\hat{\mu}_{ijk} = \hat{\mu} + \hat{\alpha_1} + \hat{\alpha} k + \hat{u}_i + \hat{a}_i k + \hat{\ell}_j. \tag{6}$$

To fit all these mixed effect models the function `lmer()` is used from R package `lme4` by Bates et al. (2015a,b). We employ a normal approximation to obtain p-values corresponding to fixed effect parameters estimates.

3.3 Location Effect

Experiment 9 studied significant expression in an RNA-seq study, and was designed so that location effect of the actual data plot in a lineup could also be assessed. The data used, documented in Atwood et al. (2013), measures gene

expression of soybean by RNA-seq methods. Two factors were of primary interest a main effect for genotype and an interaction effect between genotype and treatment condition.

In large studies such as this there is a valid question whether the data exhibits any structure at all, or if the small p-values are simply occurring by chance, from the massive multiple testing. This overall significance is studied using visual inference in Yin et al. (2013).

In order to study the effect that location has on evaluating lineups, we used multiple lineups for each of the data plots. For each data plot, five sets of null plots were generated and the actual plot was randomly placed in one of five different locations in a lineup of size 20. For the genotype effect, the locations were 1, 8, 12, 17, 20 and locations 2, 9, 12, 16, 20 were used for the interaction effect. Overall this created a total of 25 lineups for studying the genotype effect, and another 25 lineups for studying the interaction effect. Each observer saw three lineups, one for genotype, one for interaction, and one easy lineup that was used to help clean the data.

To examine if the difference in detection rate among the locations is statistically significant a one-way multivariate analysis of variance (MANOVA) model is fit to the data. Let $\mathbf{Y} = (Y_1, Y_2, ..., Y_p)^\top$ be a vector of random variables with dimension $p = 5$, the total number of null sets, and let $\mathbf{Y}_{ij}$ represent the jth vector response for location i with $i = 1, 2, ..., I$ for $U = 5$. Because the same data plot is shown in each lineup, it is assumed that there could be some association between the responses for each null set, which suggests the MANOVA model rather than a univariate ANOVA. The MANOVA model

$$\mathbf{Y}_{ij} = \mu_i + \epsilon_{ij} \tag{7}$$

where $\mu_i = (\mu_{1i}, \mu_{2i}, ..., \mu_{pi})^\top$ is the mean vector for location i and $Var(\epsilon_{ij}) = \Sigma$, tests for significant difference between the means.

3.4 Data Collection Methods

Human subjects were recruited to evaluate the experimental lineups through MTurk (Amazon, 2010). It is an online work place where people from around the world can sign up for so-called 'HIT's, human intelligence tasks, generally short tasks that are humans are typically better at solving than computers. Usually tasks are very simple and no specialized training is required to do them. Tasks are designed for anyone to do but some tasks may require some skills depending on the recruiters' need. For completing a HIT workers are paid a small amount of money, on the order of minimum wage in the USA.

We designed and developed a web application which enables the display of lineups to observers as per experimental need. The MTurk workers were re-directed to this web application to complete their assigned tasks. Responses were collected, stored automatically into a local database server, along with demographic information, age group, gender and education level. The time taken for each evaluation is computed based on the time the plot was shown and the time the feedback was received. Location of an observer is determined based on the ip address of the observer.

4. RESULTS

4.1 Overview of the Data

A total of 2321 participants provided feedback data on the lineups in ten different experimental studies. Figure 3 displays the locations of participants around the world. Most of the participants were from the United States and India. There were 76 other different countries represented. This provides a diverse pool of participants. The diversity in not only geographic but also in gender, age group and education level as can be seen in Table 2. The large number of female participants from all countries was a pleasant surprise to us.

Besides the United States and India, countries such as Canada, Romania, the United Kingdom and Macedonia have more than 10 participants each. The remaining 70 countries have fewer than 10 participants each. The distribution of participants is similar in all ten experiments.

The largest number of participants falls within the age group of 18 to 25, with the majority being between 18 to 35. Many older participants also took part in the experiments. The United States shows a more uniform participation beyond age 30, participants from India are primarily younger.

In terms of academic background, the majority of participants had a graduate degree (a total of 902 participants, about 38.51%). Examining the participants' location in conjunction with the academic background reveals that there is an above average number of participants from India reporting to have a graduate degree. In retrospect this turned out to be a culturally based misunderstanding, as any university degree in India is considered to be a graduate degree,
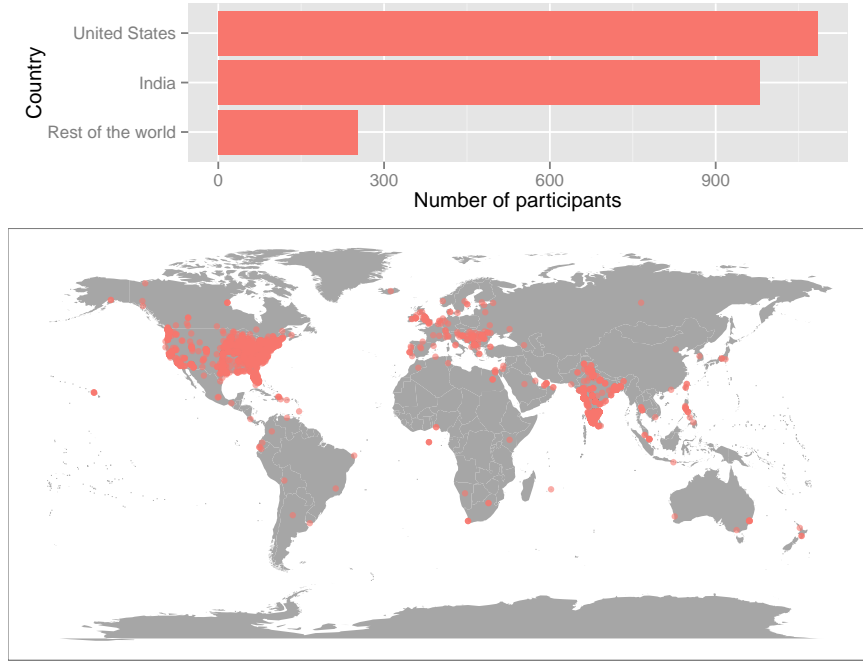
Figure 3. Location of Amazon Mechanical Turk workers participating in the experiments. Most subjects came from India and the United States, but subjects from countries around the world took part.

Table 2. Demographic information of the subjects participating in the MTurk experiments. Average time taken for evaluating a lineup is shown in seconds.

Factor	Levels	Participants Total	%	Average Time	Number of Responses
Gender	Male	1348	57.63	48.51	13493
	Female	991	42.37	43.75	10564
Education	High school or less	193	8.24	37.21	2241
	Some under graduate courses	418	17.85	42.84	4070
	Under graduate degree	584	24.93	44.29	5775
	Some graduate courses	245	10.46	43.43	2460
	Graduate degree	902	38.51	52.18	9511
Age	18-25	740	31.61	42.97	7311
	26-30	547	23.36	46.27	5585
	31-35	376	16.06	44.27	3923
	36-40	257	10.98	55.03	2714
	41-45	141	6.02	43.90	1519
	46-50	95	4.05	49.29	1003
	51-55	83	3.54	48.67	867
	56-60	64	2.73	59.73	678
	above 60	38	1.62	48.67	457
Country	United States	1087	46.83	39.64	10769
	India	980	42.22	52.63	10227
	Rest of the world	254	10.94	46.86	2819

while the North American definition of a graduate degree refers to a Masters level education or above. Most of the participants from the U.S. report to have an undergraduate degree or at least have some undergraduate courses.

Compared to data from the 2010 US Census Bureau (2010), participants were both relatively younger, and more highly educated than the general US population. Being turkers implicitly assures that the participants in these experiments have a relatively high level of technological fluency. In fact, the demographics of participants align somewhat better with demographics of internet users published by Center (2014).

The distribution of male and female participants are similar among all age groups except for age group 18-25 in India where the proportion of female participants is lower. The distribution of education levels also differs slightly across countries for this age group.

A total of 1911 lineups were evaluated in the ten experiments. Each person evaluated at least 10 lineups except for experiment 9 where each participant evaulated three lineups. As a measure to ensure a high quality in the evaluations, a test plot was shown to each observer, and the performance on this test plot was used for screening purposes. Test plots were chosen in a way that any serious attempt at answering the lineup would result in a correct answer. Data from participants who gave the wrong answer on the test plot led to the removal of all of their answers from the analysis. Results from test plots were also not considered in any of the analyses. In some cases participants chose to not provide their demographic information. Also, for some ip address, the actual geographical locations could not be retrieved. This resulted in some missing demographic information.

4.2 Demographic Factors

To explore the significance of the demographic factors model (1) is fit to the data, with age, country, education and gender as fixed effects. To estimate the significance of each of the factors, reduced models are fit with that factor removed from the model. Table 3 summarizes the results. Except for gender, all of the demographic factors show significant differences in detection rates. With respect to time taken, all of the demographic factors are significant.

Table 3. Analysis of variance (ANOVA) table comparing full model, all the demographic factors, with the reduced models, obtained by removing respective factor variable. Gender does not have a significant effect on detection rate, but does on time to respond. All factors significantly affect time to respond.

	Log Time (model 2)				Detection rate (model 1)			
	Deviance	χ^2	df	p-value	Deviance	χ^2	df	p-value
Full	51872.4	—	—	—	23792.0	—	—	—
- Age	51329.3	543.2	6	<0.001	23766.4	26.0	6	<0.001
- Country	51496.4	376.1	2	<0.001	23765.9	26.0	2	<0.001
- Degree	51756.2	116.2	4	<0.001	23768.2	23.8	4	<0.001
- Gender	51804.3	68.1	1	<0.001	23789.7	2.2	1	0.140

Parameter estimates from the model fits are shown in Table 4. The first level of each factor serves as the baseline for the model. The age group 36-40 has a significantly higher detection rate than the 18-25 age group, and to a lesser extent this is also true for age groups 31-35 and above 50. Participants from India have a similar detection rate to those from the USA, but participants from the rest of the world show a significantly higher detection rate. Subjects reporting to have a graduate degree had a significantly higher detection rate. For time to respond, all age groups were significantly slower than 18-25 year olds. Participants from India were slower, while those from the rest of the world were faster than US Americans. A higher degree corresponded to longer time, and males were on average slower than females.

Figure 4 supports the model results. The average time taken to respond (natural log) and detection rate are computed by lineup for the different demographic factor levels, and displayed using boxplots. Mean values are shown as dots inside the boxplot, as this statistic rather than the median is used in the models. The means and medians for response time (top row) are essentially equal. There are some differences between the two in the detection rate plots (bottom row). The differences in these values for different factors, that was tested by the models, can be seen in the plots. However, particularly for detection rate, the variation is huge. A large component of the variation is difficulty of the lineup. In some lineups the actual data plot was distinctively different from the null plots, and we would expect that the detection rate would be close to 100%. In other lineups there was no difference, making it very difficult to evaluate, with low detection rate and longer time to respond. In Table 4 notice that lineup specific error is estimated as 2.29 which is much higher than any of the other parameter estimates in the model. This indicates that the major

Table 4. Parameter estimates of models (2) and (1) fitted for average log time taken and detection rate, respectively. For time taken all the demographic factors are significant. For detection rate, age group 36-40, rest of the world and graduate degree are significantly different. For gender no difference in performance is observed. Lineup variability is estimated to be very large for model (1).

Demographic Factor	Level	Log Time (model 2)			Detection rate (model 1)		
		Est	(2.5%, 97.5%)		Est	(2.5%, 97.5%)	
Fixed Effects	Average μ	3.16	(3.12, 3.20)	***	-0.61	(-0.79, -0.43)	***
Age Category (α)	18-25	0.00			0.00		
	26-30	0.06	(0.03, 0.08)	***	0.06	(-0.03, 0.16)	
	31-35	0.07	(0.04, 0.10)	***	0.11	(0.01, 0.22)	*
	36-40	0.23	(0.20, 0.26)	***	0.31	(0.19, 0.43)	***
	41-45	0.18	(0.14, 0.22)	***	0.16	(0.00, 0.31)	*
	46-50	0.27	(0.22, 0.32)	***	0.14	(-0.05, 0.33)	
	above 50	0.35	(0.32, 0.39)	***	0.15	(0.01, 0.28)	*
Country (κ)	U.S.	0.00			0.00		
	India	0.23	(0.21, 0.25)	***	-0.18	(-0.27, -0.10)	***
	Rest of the world	0.14	(0.11, 0.17)	***	0.08	(-0.04, 0.20)	
Education (τ)	High School	0.00			0.00		
	Under grad courses	0.13	(0.09, 0.16)	***	-0.03	(-0.18, 0.11)	
	Under grad degree	0.13	(0.10, 0.17)	***	0.12	(-0.01, 0.26)	.
	Graduate courses	0.01	(-0.04, 0.05)		-0.17	(-0.33, -0.01)	*
	Graduate degree	0.13	(0.09, 0.16)	***	-0.06	(-0.19, 0.07)	
Gender (γ)	Female	0.00			0.00		
	Male	0.08	(0.06, 0.10)	***	0.06	(-0.02, 0.13)	
Random Effects	lineup (σ_ℓ)	0.29			2.29		
	Error (σ)	0.69			0.89		

Signif. codes: $*** < 0.001 \leq ** < 0.01 \leq * < 0.05 \leq . < 0.1 \leq$ ' ' < 1

and most important factor affecting the detection rate is lineup difficulty. So although the demographic factors emerge from the model to be statistically significant, the practical significance is minimal. We illustrate this with the following example of graduate degree.
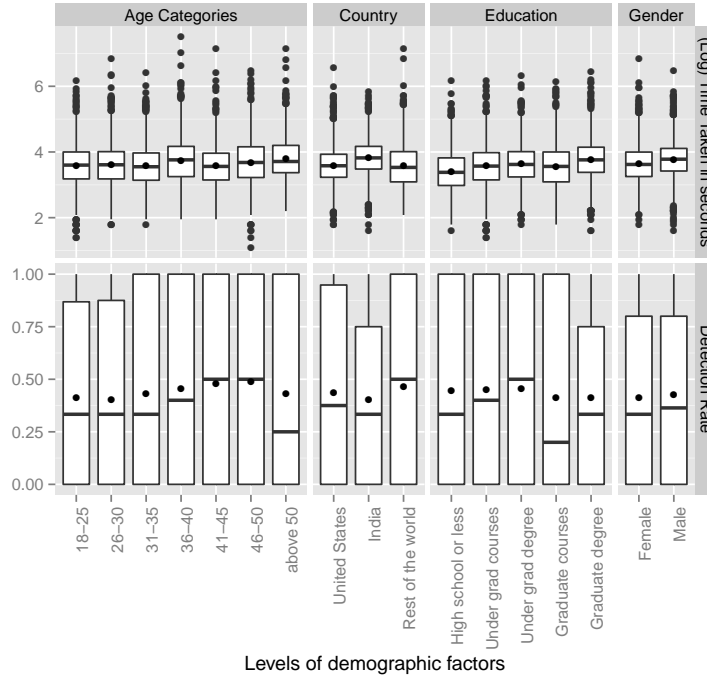

Figure 4. Boxplots of average log time taken and proportion correct responses (detection rate) of all the lineups plotted for each demographic factor levels. The dots inside the boxes represent means. Some differences in means of various demographic factors are observed. Variability in detection rate indicates large variability in lineup difficulties.

While some of the demographic factors are strongly statistically significant, the main source of variation in detection rate is the lineup difficulty. For example, let's examine the effect of graduate degree. To see just how large the effect is, we examine the change in detection rate for a (hypothetical) 18-25 year old female in the United States, with some graduate course work as compared to an undergraduate degree, for an average difficulty lineup (random effect of zero). Plugging in the relevant quantities to the fitted model gives a difference equal to:

$$\frac{\exp(-0.61-0.17)}{1+\exp(-0.61-0.17)} - \frac{\exp(-0.61+0.12)}{1+\exp(-0.61+0.12)} = 0.314 - 0.379 = -0.065.$$

The person with some graduate course work picks the data plot on average in 33.5% of lineups of average difficulty, as compared to 37.9% if they have an undergraduate degree. This difference is reduced to 2% for a lineup with one standard deviation order of magnitude difference in difficulty. For two standard deviations it further reduces to 0.3%. Thus, although there is a statistically significant difference in the proportion at which participants identify the data plot for some demographic factors, these are not significant in any practical sense. Figure 5 illustrates this example showing fitted models for a US 18-25 female with either a high school education or a graduate degree. Similar calculations show the same negligible impact of age level 36-40 (0.073 at most) and country (0.058 at most) on the probability of a correct response. Thus even though some of the demographic factors are statistically significant, practically, demographics do not substantially influence the results.

4.3 Learning Trend

Models (3) and (5) are fitted to the data from experiments 5, 6, and 7 separately. In this model, attempt is fitted as a factor variable to allow for all possible non-linear learning trends. It should be noted that we also examined an alternative to model (3) where attempt was linearly fitted as a continuous covariate, but this also was not significant.
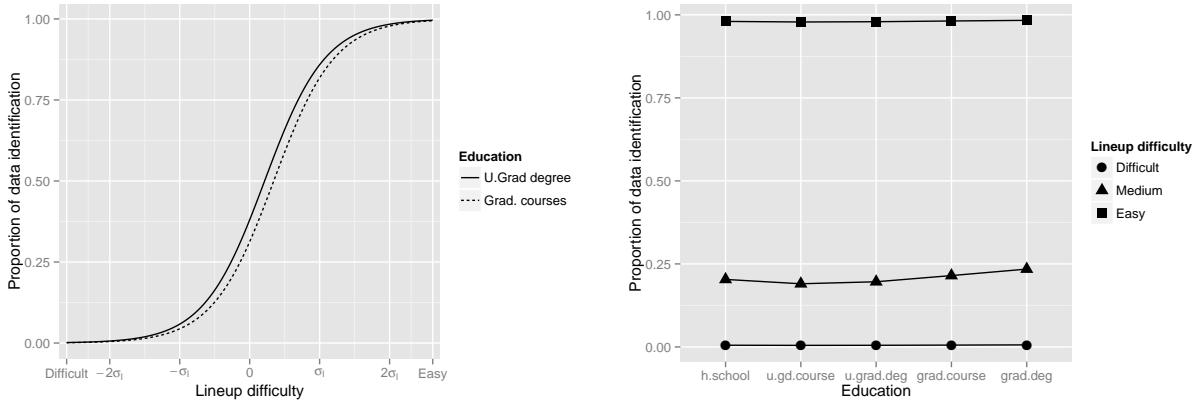
Figure 5. Proportion of estimated data identifications by an 18-25 year old female in the United States with a graduate degree compared to a high school degree. Even though having a graduate degree leads to statistically significant increase, the effect size in the dependent variable is 0.045 which is negligible from a practical point of view. The difference diminishes as we move away one or two standard deviations ($\sigma_\ell = 2.29$) of lineup variability.

Table 5 displays the parameter estimates and p-values of fixed effect estimates of model (3) examining detection rate. Only for experiment 6 there is some evidence of a learning curve. There are marginally small p-values for most of the attempt levels, and positive parameter estimates, suggesting that more attempts increases detection rate.

Table 5. Parameter estimates of models (3) fitted to data from three different experiments using detection rate as the response to assess learning trend. Attempt number is fitted as a factor to enable modeling any sort of non-linear learning trend. Only experiment 6 shows some evidence of a learning trend, with detection rate essentially increasing as attempts increase.

	Experiment 5			Experiment 6			Experiment 7		
Effect	Est	(2.5%, 97.5%)		Est	(2.5%, 97.5%)		Est	(2.5%, 97.5%)	
Fixed									
μ	-1.30	(-1.69, -0.92)	***	-0.22	(-0.51, 0.07)		-1.71	(-2.77, -0.65)	**
α_2	0.27	(-0.19, 0.73)		0.26	(-0.06, 0.58)		-0.49	(-1.31, 0.32)	
α_3	-0.18	(-0.65, 0.29)		0.34	(0.02, 0.66)	*	-0.19	(-1.01, 0.63)	
α_4	0.08	(-0.39, 0.55)		0.36	(0.04, 0.68)	*	-0.47	(-1.28, 0.35)	
α_5	0.30	(-0.17, 0.77)		0.37	(0.05, 0.70)	*	-0.16	(-0.96, 0.65)	
α_6	0.04	(-0.44, 0.53)		0.24	(-0.07, 0.56)		-0.03	(-0.87, 0.82)	
α_7	0.28	(-0.20, 0.77)		0.16	(-0.16, 0.48)		0.02	(-0.81, 0.84)	
α_8	-0.04	(-0.54, 0.45)		0.34	(0.02, 0.66)	*	-0.08	(-0.89, 0.73)	
α_9	-0.20	(-0.69, 0.30)		0.38	(0.05, 0.70)	*	0.15	(-0.71, 1.02)	
α_{10}	0.51	(0.03, 1.00)	*	0.19	(-0.14, 0.52)		-0.28	(-1.14, 0.58)	
Random									
σ_u^2	0.85			0.90			0.76		
σ_a^2	0.02			0.03			0.06		
σ_l^2	1.48			1.42			3.37		

Signif. codes: *** < 0.001 ≤ ** < 0.01 ≤ * < 0.05 ≤ . < 0.1 ≤ ' ' < 1

To visualize how detection rates change over successive attempts, we fitted model (3) excluding the covariates related to attempt from the model and computed the residuals. Least square regression lines were fitted through the subject specific residuals as shown in Figure 6. The averages of these residuals for each of the attempts are shown as dots. Two important features were observed; one is subject specific variability and the other is random slope with attempts which indicates subjects specific learning trend. Some subjects show improvement over time and some show the decrease in performance. Although, the model fit experiment 6 suggested a statistically significant learning curve the plots indicates that it is minimal.
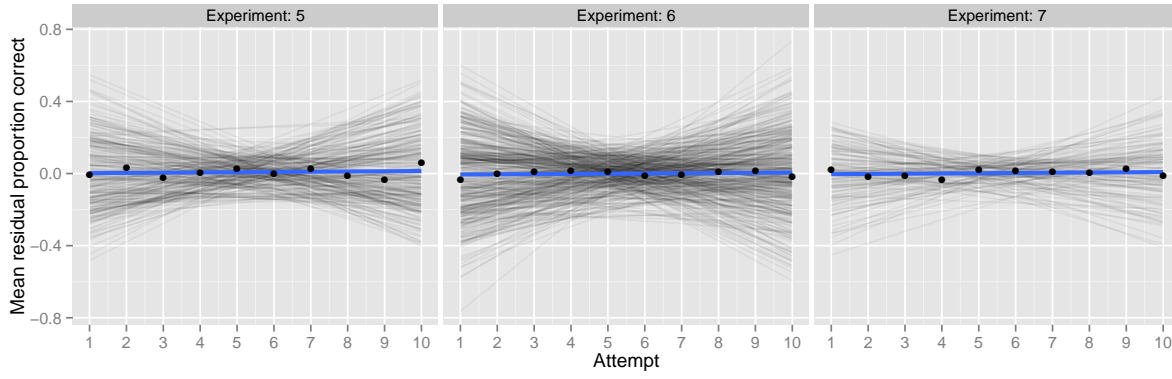
Figure 6. Least square lines fitted through the subject specific residual proportion correct obtained from model (3) fitted without attempt are plotted against attempt. Subject specific positive and negative slopes are observed. Mean residuals are shown as dots and least square regression lines fitted through the points show no overall learning trend in each of the three experiments.

Table 6 presents the results of model (5) where response time is examined with respect to attempt. Attempt is modeled as a linear covariate here, with a shift factor used to adjust for additional time on the first lineup evaluated. Interestingly, time to respond significantly decreases as number of attempts increases, for all three experiments. The parameter α for fixed effect covariate attempt is highly significant in all the experiments. The negative estimates suggest that on average later attempts took less time. Even though observers did not improve in their performance in successive attempts, they became more efficient in their response. The parameter α_1 for first attempt is also highly significant. The positive estimates of α_1 indicates that first attempt made by an observer required much more time than any of the other attempts. One explanation might be that participants take the chance to carefully read through instructions and familiarize themselves with the types of plots used in the experiment before evaluating their first lineup, which would inflate the time taken on the first lineup. Each of the experiments asks observers to give a reason for their choice (out of a list of pre-defined answers. For each lineup evaluation, participants are also asked to state how confident in their choice they are (on a scale of 1 to 5). Later pages of the web site were identical except for the lineup.

Table 6. Parameter estimates of model (5) fitted for log time taken to evaluate a lineup. Both fixed effects parameters of Attempt (α_1 and α) are highly significant for all three experiments 5, 6 and 7.

	Experiment 5			Experiment 6			Experiment 7		
Effect	Est	(2.5%, 97.5%)		Est	(2.5%, 97.5%)		Est	(2.5%, 97.5%)	
Fixed									
μ	3.82	(3.74, 3.89)	***	3.90	(3.84, 3.97)	***	3.73	(3.62, 3.84)	***
α_1	0.33	(0.26, 0.39)	***	0.34	(0.28, 0.39)	***	0.28	(0.18, 0.38)	***
α	-0.04	(-0.05, -0.03)	***	-0.04	(-0.05, -0.03)	***	-0.03	(-0.04, -0.02)	***
Random									
σ_u^2	0.51			0.49			0.37		
σ_a^2	0.03			0.04			0.05		
σ_l^2	0.09			0.20			0.23		
σ^2	0.46			0.50			0.45		

To visualize how the time taken reduces over the successive attempts, we fitted model (5) excluding the covariate attempt from the model and computed the residuals. Least square regression lines are fitted through the subject specific residuals. Subject specific slopes are much different in each of the three experiments as we see in Figure 7. Some subjects improved over attempts by taking less time in the later attempts while others got worse. The averages of these residuals for each attempt are plotted as dots. Least square regression lines are fitted to these points excluding the first attempt since for first attempt we fitted an indicator covariate. The downward trends are evident in the plots. All the slopes are highly significant. As expected we observed large positive residuals for each of the experiments for first
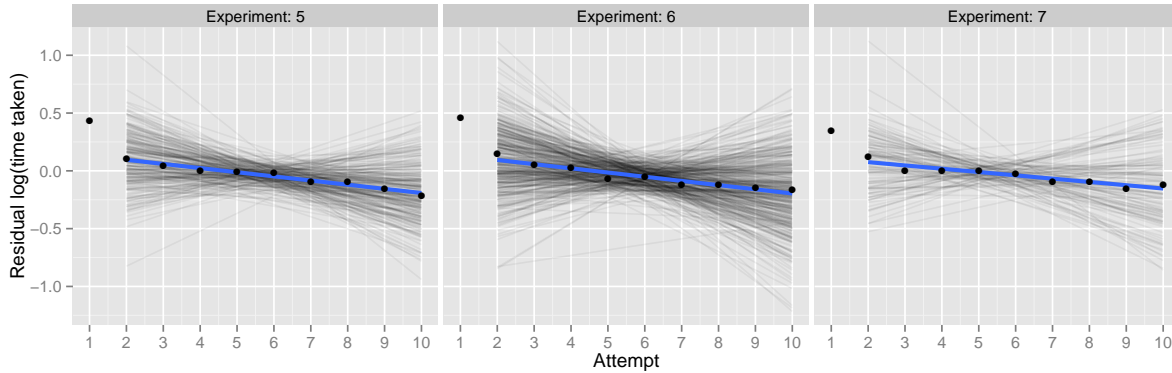
15

attempt.

Figure 7. Least square regression lines fitted through the subject specific residuals obtained by fitting model (5) without covariate attempt. Differences in subject specific slopes are observed. Some of the subjects did worse over successive attempts while others did better. Averages of these residuals are plotted as dots and least square regression lines are fitted to obtain overall trends. For all the three experiments the overall downward slopes are statistically significant which indicates that MTurk workers become more efficient as they progress through their attempts.

4.4   Location Effect

A total of 111 subjects was recruited to evaluate lineups designed to investigate the location effect of the actual data plot in the lineup as described in Section 3.3. Each subject evaluated two lineups; one for Interaction effect and the other for Genotype. In total there were 222 responses for 50 lineups. The data on the test lineup were excluded from the analysis.

We are investigating location effect in two ways: (1) location effect of each individual plot location, and (2) location effect of inside/outside panel positions.

Figure 8 shows an overview of detection rates for both of the effects studied in experiment 9. Each dot corresponds to one lineup. The size of each dot gives the number of evaluations. Lineups with the same set of null plots are connected by lines. The overall average detection rate for each location is shown using dashed lines. Size of the points represent the number of responses. For some locations we have as many as 10 responses. For location 1, we did not have any responses for null set 1 in one of the interaction lineups.

We observe some variability in the performance due to different null sets even though the same actual data plot was used for all of these null sets. The lineup protocol is using a finite, small sample (19 in these experiments) of all possible null sets, so it is possible to see some variability in performance depending on the null set. If null sets come from the "extremes" of the sampling distribution they may have structure that is more extreme than the actual data plot, making the lineup more difficult to evaluate. From the plot we can see that null set 3 appeared to be more difficult in the Interaction lineup, because the detection rate was lower regardless of the position of the actual data plot. This null set effect was tested, and a significant difference was found for null set 3 of the interaction lineup. The rest of the null sets do not show any differences at a 1% significance level.

Model (7) was fit to the data to test, if detection rates were significantly different for the different locations. For this we use the *manova( )* function provided in the *stats* package of the software R (R Core Team, 2015). The results are shown in Table 7. The p-values for both the Interaction and the Genotype effect are much bigger than the conventional threshold of 0.05 suggesting that location of the actual data plot in the lineup does not significantly affect performance.

For the second approach in investigating the location effect, we distinguish between inner panels 7, 8, 9, 12, 13, 14 and outer panels (any panel with an outside edge). From Figure 8 we see that location 9, 12 are inner panels for Interaction effect and locations 8, 12 are inner panels for Genotype. Model (7) fitted with two locations effects, inner and outer, did not result in any significant differences.
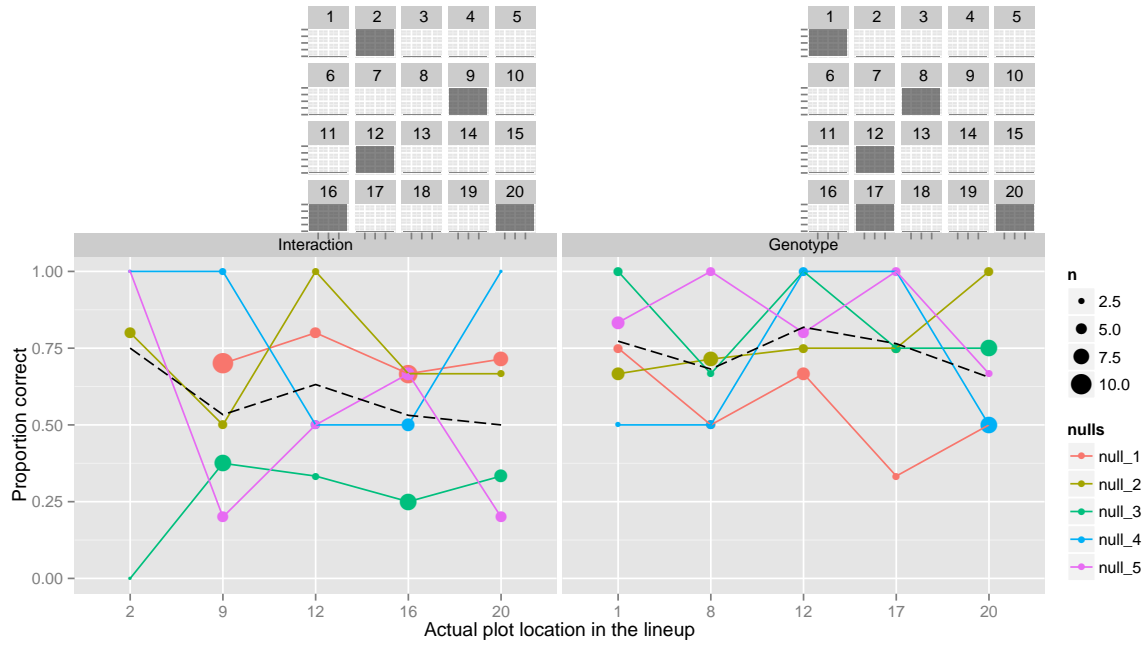
16

Figure 8. Location of data plot in the lineup and proportion correct for both Interaction and Genotype effect. Each colored line represents a null set and the size of the dots represents number of responses. The overall average proportions are shown by dashed lines. The actual data plot locations are shaded grey on the top panels to demonstrate their relative positions on a lineup.

Table 7. The summary of the results obtained by fitting MANOVA model (7) using Wilks test.

Location				Degrees of Freedom			F test
Effect	DF	Wilks	Approx. F	Numerator	Denominator	Residual	*p*-value
Interaction	3	0.07986	0.592	15	5.92	6	0.8082
Genotype	4	0.01689	1.722	20	14.22	8	0.1488

5. CONCLUSION

Human demographics have the potential to influence performance when the lineup protocol is used for statistical inference. In our study of the demographic effects on performance across a set of different experiments we found some statistically significant effects. Age group 36-40, countries other than India and United states, people who have a graduate degree have a significantly higher detection rate. Gender does not have any significant effect. However, the effects are minimal, on the order of a few percentage points different from the average. These results are very important for the power of visual tests as they demonstrate the robustness of the test against different human factors.

Individual learning trend is observed in time taken but not so much in observer performance. Some individuals improved on their performances while others showed a decrease on successive attempts. This could be interpreted as good, in the sense that it means that observers could realistically be recruited from sources like MTurk, and that substantial training is not necessary in order to obtain useful evaluations of lineups in practice. However, we had hoped that lineups may be a useful teaching tool, to improve students ability to read data plots, and the lack of a learning trend diminishes this idea, at least for short term learning.

The simulation experiment reveals that there is no significant effect of location on the actual data plot in the lineup. This is important as the visual statistical inference procedure prescribes that the data plot be placed at random in the lineup. This paper suggests that any random place in a lineup is as good as other places in the lineup. Even though there are variations on the performance depending on different null sets, their impact on probability to correctly evaluate a lineup is very negligible.

There are many more ways that the lineup protocol might be tested to learn what we might expect about its performance for tackling real data mining problems. Possible changes in the lineup protocol in the pipeline are allowing observers to select more than one plot, and to vary the size of the lineup from 20 to a smaller number, and have observers all see different lineups with different null sets. This paper assessed the effect of human factors on the experiments conducted to date. As changes to the lineup protocol are suggested by other experiments the human factor effects may need to be examined again.

References

Amazon (2010), "Mechanical Turk," https://www.mturk.com/mturk/welcome.

Atwood, S. E., O'Rourke, J. A., Peiffer, G. A., Yin, T., Majumder, M., Zhang, C., Cianzio, S., Hill, J. H., Cook, D., Whitham, S. A., Shoemaker, R. C., and Graham, M. A. (2013), "Replication protein A subunit 3 and the iron efficiency response in soybean." *Plant Cell and Environment*, 37, 213–234.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015a), *lme4: Linear mixed-effects models using Eigen and S4*, r package version 1.1-9.

Bates, D., Maechler, M., Bolker, B. M., and Walker, S. (2015b), "Fitting Linear Mixed-Effects Models using lme4," *Journal of Statistical Software*, in press.

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 A*, 367, 4361–4383.

Bureau, U. C. (2010), "United States Census 2010," .

Center, P. R. (2014), "Internet User Demographics," .

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.

Hofmann, H. and Röttger, C. (2015), *vinference: Inference under the lineup protocol*, R package version 0.1.1.

Loy, A., Follett, L., and Hofmann, H. (2015), "Variations of Q-Q Plots – the Power of our Eyes!" *The American Statistician*, 4, 1–36.

Majumder, M., Hofmann, H., and Cook, D. (2013), "Validation of Visual Statistical Inference, Applied to Linear Models," *Journal of the American Statistical Association*, 108, 942–956.

Meilgaard, M. C., Carr, B. T., and Civille, G. V. (2006), *Sensory Evaluation Techniques*, CRC Press, 4th ed.

Mosley, L., Cook, D., Hofmann, H., Kielion, C., and Schloerke, B. (2010), "Visually Monitoring the 2008 Election," *CHANCE*, 23, 4–4.

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

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Wickham, H., Roy Chowdhury, N., and Cook, D. (2014), *nullabor: Tools for Graphical Inference*, R package version 0.3.1.

Yin, T., Majumder, M., Roy Chowdhury, N., Cook, D., Shoemaker, R., and Graham, M. (2013), "Visual Mining Methods for RNA-Seq Data: Data Structure, Dispersion Estimation and Significance Testing," *Journal of Data Mining in Genomics & Proteomics*, 4.

Zhao, Y., Cook, D., Hofmann, H., Majumder, M., and Roy Chowdhury, N. (2013), "Mind Reading: Using an Eye-Tracker to See How People are Looking at Lineups," *International Journal of Intelligent Technologies & Applied Statistics*, 6, 393–413.