Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Statistical Significance Calculations for Scenarios in Visual Inference

Susan Vanderplas*[1] | Christian Röttger[2] | Dianne Cook[3] | Heike Hofmann[4]

[1]Department of Statistics, University of Nebraska-Lincoln, Nebraska, United States
[2]Department of Mathematics, Iowa State University, Iowa, United States
[3]Department of Econometrics and Business Statistics, Monash University, Victoria, Australia
[4]Department of Statistics, Iowa State University, Iowa, United States

Correspondence
*Susan Vanderplas, Email: susan.vanderplas@unl.edu

Summary

Statistical inference provides the protocols for conducting rigorous science, but data plots provide the opportunity to discover the unexpected. These disparate endeavors are bridged by visual inference, where a lineup protocol can be employed for statistical testing. Human observers are needed to assess the lineups, typically using a crowd-sourcing service. This paper describes a new approach for computing statistical significance associated with the results from applying a lineup protocol. It utilizes a Dirichlet distribution to accommodate different levels of visual interest in individual null panels. The suggested procedures facilitate statistical inference for a broader range of data problems.

KEYWORDS:
data visualization, statistical graphics, hypothesis testing, data science

1 | INTRODUCTION

Graphics provide the opportunity to understand statistical data at a (visual) sensory level. They are also important for data analysis because a plot can communicate more than summary statistics (Anscombe 1972; Matejka & Fitzmaurice 2017). Statistical graphics allow us to leverage the full bandwidth of the human visual system for implicit data analysis. Because this analysis is implicit, we often assume these visual analyses are not decisive in the same way that a hypothesis test is decisive. Graphics generally do not come with a significance threshold, and in many cases, we do not explicitly declare the hypothesis we might be testing before viewing the data plot. Buja et al. (2009) introduced a protocol for *visual inference* that allows for formal hypothesis testing using graphical displays and visual evaluation.

To test whether a chart shows a visually significant result, we can use the same machinery used by randomization tests: (1) construct a method to generate data consistent with the null hypothesis, such as using randomization or drawing samples from a model consistent with the null hypothesis, (2) generate many copies of the test statistic (in this case, the plot), and (3) see where the observed statistic falls in the distribution of artificially generated quantities. The conceptual framework was described in Buja et al. (2009), who introduced two protocols: a lineup, and a Rorschach (or null) lineup.

An assembly of several null plots with a target (or data) plot is called a *lineup*, named after the law-enforcement procedure to line up a suspect among a set of innocents to check if a victim can identify the suspect as the perpetrator of the crime. In its visual version, lineups are typically composed of $m - 1$ "null" plots (generated under the null hypothesis) and one data plot containing the observed data.

The Rorschach lineup is named after the ink-blot test (Exner & Erdberg 2003) historically used in psychoanalysis; as in the inkblot test, the Rorschach lineup provides ambiguous visual signals which are open to interpretation. Figure D1 shows several examples of each type of lineup. In a Rorschach lineup, all plots are null plots; the purpose is to assess the extent of visual variation that occurs in data generated by the null mechanism.

The fundamental premise of visual inference is that charts are visual statistics: summaries of data sets generated by mathematical functions. Underlying this premise is the concept of a grammar of graphics initially laid out by Wilkinson (1999), and further developed in Wickham (2009), which provides a functional mapping from variables (abstractly defined) into graphical elements of a plot. This abstract plot definition allows us to declare a hypothesis by specifying the relationship between the variables and the spatial elements of the plot. For example, the scatter plots

shown in the one target lineups in Figure D1b map variable 1 to the x-axis and variable 2 to the y-axis, with the intent to study the association between the two variables.

The natural next step is to consider using these visual statistics to conduct hypothesis tests. In the scatterplot example above, the null hypothesis is that there is no association between the two variables and the corresponding alternative hypothesis is that there is an association. Conducting a test using the observed data plot requires that we compare the data plot to a reference distribution. Any null generating mechanism, such as randomization or sampling from a known distribution, may be used to generate the data for the null panels in the lineup.

In a numerical statistical test, the summary statistics are naturally ordered, and the test statistic is compared to quantities consistent with the null hypothesis; often, a p-value is used to assess the probability that the observed statistic would arise by chance under the null hypothesis. When conducting visual inference, the statistics have no such natural ordering; instead, our statistics must be evaluated by human participants. Typically, graphical tests utilize a service like Amazon Mechanical Turk (Amazon 2005-2015) to acquire evaluations of lineups. An extreme visual test statistic is one that is easily distinguishable from other null plots in the lineup; that is, if the data panel is identified when the lineup is evaluated, we would reject the null hypothesis. Thus, lineups have the features of a *statistical test*. One notable difference between visual and numerical statistical tests is that visual tests are more comprehensive: individuals are asked to select one or more plots from the lineup which are “different", but typically, the specific type of difference is left unspecified. As a result, a visual test might evaluate several simultaneous characteristics of a plot, where the equivalent numerical assessment may involve multiple tests using different test statistics. Occasionally, because there is some inductive reasoning required of the participant, the feature used to select the “different" panel is not the feature under investigation by the experimenter.

Typically, lineups consist of $m = 20$ panels, $m_0 = 19$ of which are generated from the null distribution. Under this formulation, a single evaluation of a lineup in which the data panel is selected would have a visual p-value of $p \leq 0.05$. In most visual inference experiments, each lineup is evaluated by multiple individuals, and the aggregated results are used to generate a visual p-value using a Binomial model as described in Majumder, Hofmann, and Cook (2013). This approach is simple, but does not account for dependencies in the design of most visual inference experiments resulting from repeated evaluations of the same lineup.

In this paper, we propose a new approach for computing p-values using a Dirichlet-multinomial distribution to model the probabilities of selecting each panel and the observed participant selections. This model is more flexible than the binomial model used previously, and better represents the perceptual and statistical dependencies present when lineups are evaluated by multiple observers. Leveraging this model, we propose a method for estimation of the model parameters and present a diagnostic procedure for null plot generation models.

[**FIGURE** 1 about here.]

1.1 | Modeling Lineup Panel Selections

The lineup evaluation task boils down to a selection of one of m panels in a lineup. We model the probability of selecting panel i as $\theta_i, i = 1, ..., m$, where $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$. The m-dimensional Dirichlet distribution generates $\boldsymbol{\theta}$ distributed on the $m - 1$ simplex, that is, $\sum \theta_i = 1$. When $\alpha_1 = \alpha_2 = ... = \alpha_m$, the distribution is a symmetric Dirichlet distribution. As all null plots are generated using the same process, we model the picking probability $\boldsymbol{\theta}$ for a set of null plots using a symmetric Dirichlet distribution.

The density function of the symmetric Dirichlet distribution is given as

$$f(\boldsymbol{\theta} \mid \alpha) = \frac{(\Gamma(\alpha))^m}{\Gamma(m\alpha)} \prod_{i=1}^{m} \theta_i^{\alpha-1}, \tag{1}$$

where $\Gamma(.)$ is the Gamma function defined as

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx.$$

(Note that the Gamma function is equal to the factorial function $\Gamma(n) = (n-1)!$ for integer values of n.)

When $\alpha = 1$, the symmetric Dirichlet distribution is uniform on the $(m-1)$-dimensional simplex. When $\alpha < 1$, the mass of the distribution is along the edges of the simplex, where most values of θ_i will be close to 0. When $\alpha > 1$, the mass of the distribution is in the center of the simplex, with most of the θ_i having similar values. Figure D2 shows ternary plots (Hamilton & Ferry 2018) of values simulated from a 3-dimensional Dirichlet distribution which illustrate the effect of α on the sampled $\boldsymbol{\theta}$.

[**FIGURE** 2 about here.]

[**FIGURE** 3 about here.]

While graphical illustrations of the m-dimensional Dirichlet distribution are more difficult when $m > 3$, we can use simulation to assess the meaning of α as it relates to lineup panel selection probabilities θ_i. Figure D4(left) shows simulated selection probabilities θ_i, sorted such that the panel with the highest selection probability is first, for several values of α in a $m = 20$ panel lineup. It is evident that for $\alpha < 0.05$ only one panel of the lineup receives significant attention, while for $\alpha > .25$, participant attention is divided among several interesting panels of the lineup.
The right side of Figure D4 shows probabilities of null plot selections from three previous studies. Each line corresponds to one lineup. The vertical striping and the relatively large variability is due to the limited number of evaluations of each lineup (usually a lineup is not evaluation more often than 30 times). Clearly, each experiment is associated with a different distribution of θ that have characteristics corresponding to different α values. However, all three studies suggest α values of less than 1 when comparing with the simulated distributions.

[**FIGURE** 4 about here.]

From figures D2 and D4, it is evident that α provides some information about how many panels are likely to attract participant attention.

1.2 | Lineup Administration

In addition to different types of lineups, there are different ways that an analyst can run a lineup experiment. Three possibilities are detailed below and also illustrated in Figure D5.

Scenario 1 K different lineups are shown to K independent individuals. In this scenario, both the data and the null plots in each generated lineup are distinct from those in every other lineup. This scenario is only practical when using purely simulated data (for both the data and null plots) or data large enough to allow for sub-sampling to generate K different data plots. Under Scenario 1, we can consider the number of target plot selections t out of K total evaluations, where each lineup evaluation is a Bernoulli trial; the total number of data plot evaluations can then be modeled as a Binomial distribution with selection probability $1/m$ (for a single target lineup) (Majumder et al. 2013).

Scenario 2 K different sets of null plots are shown to K independent individuals; the same data plot is used in each lineup. Alternately, L sets of lineups are shown to $K > L$ individuals. In Scenario 2, there are dependencies introduced by the reuse of the data or the lineups, providing an intermediate case between the two extremes of Scenario 1 and Scenario 3.

Scenario 3 The same lineup is shown to K independent individuals. This scenario is the most common in the lineup experiments completed to date (Hofmann, Follett, Majumder, & Cook 2012; Loy & Hofmann 2015; Majumder et al. 2013; Roy Chowdhury, Cook, Hofmann, & Majumder 2012; VanderPlas & Hofmann 2017). In this scenario, lineup evaluations by independent viewers are not independent because the viewers are evaluating the same combination of null plots and data. Any peculiar features which arise in a null plot may cause participants to select that null plot over the data plot; it is likely that where one individual makes this choice, others might as well. Thus, in Scenario 3, which is the most common lineup experiment scenario, it is not reasonable to assume that all panels are equally likely to be selected under the null hypothesis.

[**FIGURE** 5 about here.]

Regardless of the scenario, we can formulate lineups as a hypothesis test in the following manner:

Null hypothesis the data plot is consistent with plots rendered from the null generating model. The probability for picking the data panel is (in distribution) the same as picking a null panel.

Alternative hypothesis the data plot is not consistent with plots rendered from the null generating model, i.e. based on the number of data plot selections c_t the data plot is found to be interesting by participants.

Let C_t be the random variable capturing the number of data (target) selections, $0 \leq C_t \leq K$. The visual p-value is defined as the probability that we see at least c_t evaluations in K independent lineup evaluations:

$$\text{visual } p - \text{value} = P(C_t \geq c_t) \tag{2}$$

The exact distribution of C_t depends on the scenario applied when administering the test. Scenario 3 is the primary motivation for the model which we will develop in the next section: while we cannot consider the panels in each lineup as equally likely to be selected, we can model the individual selection probabilities using a hierarchical model. A model for Scenario 2 would need to account for specific dependencies between the lineups; because there are many possible implementations of Scenario 2, we will leave explicit model development for those different scenarios to another paper.

2 | SIGNIFICANCE CALCULATION MODEL SPECIFICATION

In scenario 3 the distribution for selecting any plot from a lineup assuming the null hypothesis is true is given as $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$ for constant values of $\alpha_i = \alpha > 0$ for all panels $i = 1, ..., m$. Consider a lineup experiment under scenario 3, i.e. the same lineup containing m panels, independently evaluated K times. Define $c = (c_1, ..., c_m)$ to be the counts corresponding to the set of K evaluations, where c_i is the number of times a participant selected panel i, where $0 \leq c_i \leq K$ for all $1 \leq i \leq m$ and $\sum_i c_i = K$. The assumption is that each evaluator has to make a selection. In practice, evaluators get to make several selections – these choices are incorporated into counts as fractions adding to 1. In case an evaluator makes no choice, all panel counts are increased by 1/m. For notational convenience, we will assume that all c_i are integer counts. The results are immediately applicable to a non-integer number of selections, however.

The Multinomial (K, θ) distribution, with probability mass function given as:

$$f(c|K, \boldsymbol{\theta}) = \frac{K!}{c_1! \cdots c_m!} \prod_{i=1}^{m} \theta_i^{c_i} \tag{3}$$

is a natural model for the counts c_i, where K describes the number of evaluations and $\boldsymbol{\theta} = \theta_1, ..., \theta_m$ describes the probability that each panel $i = 1, ..., m$ is selected. As discussed in Section 1.1, the panel selection probabilities $\boldsymbol{\theta}$ can be modeled using a Dirichlet distribution with concentration parameter $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)$.

Under the null hypthesis, the data panel is not distinguishable from the null plots. We can therefore model the picking probabilities $\boldsymbol{\theta}$ for the panels of a lineup using a symmetric Dirichlet distribution, with $\alpha_i = \alpha, i = 1, ..., m$.

Thus, for the observed plot selections $(c_1, ..., c_m)$, and assuming parameter $\alpha > 0$, we specify the full model as:

$$\begin{aligned} \boldsymbol{\theta} \mid \alpha &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ c \mid \boldsymbol{\theta} &\sim \text{Multinomial}(\boldsymbol{\theta}, K) \end{aligned} \tag{4}$$

The joint distribution of $\mathbf{c}$, θ is then a Dirichlet-Multinomial mixture distribution. We will refer to this as **model DM**, for brevity in later explanations. If we were interested in the value of $\boldsymbol{\theta}$ and wanted to take a Bayesian approach, we would use the conjugate relationship between the Multinomial and Dirichlet distributions to get a posterior distribution of $\boldsymbol{\theta}$ given c and α as $\text{Dirichlet}(c + \boldsymbol{\alpha})$. However, the primary purpose is to obtain a visual p-value from this mixture model, which does not require that we conduct inference on $\boldsymbol{\theta}$.

Rather, if we assume that the Dirichlet distribution is symmetric, corresponding to the belief that while not all panels are equally likely to be selected, there is no a priori belief that the panels systematically differ, we can conduct a test to determine whether the observed counts are consistent with a single Dirichlet parameter α. Restated in hypothesis testing terms, our null hypothesis is that $\boldsymbol{\alpha} = \alpha * (1)_{m \times 1}$. If the lineup contains a target plot, that target plot should be overwhelmingly selected, producing counts that are not likely to occur under the assumption of symmetry in $\boldsymbol{\alpha}$.

This approach differs from the method in Majumder et al. (2013), where the number of target plot identifications was compared with the aggregate number of null plot identifications. However, that approach can be considered to be equivalent to a special case of the marginal distribution of c_i under DM. For a total of K evaluations, and $0 \leq c_i \leq K$ target plot evaluations, the marginal model simplifies to a Beta-Binomial (BB, in later discussions):

$$\begin{aligned} \theta_t \mid \alpha &\sim \text{Beta}(\alpha, (m-1)\alpha) \\ C_i \mid \theta_i &\sim \text{Binomial}(\theta_i, K), \end{aligned} \tag{5}$$

with the distribution of θ_i given c_i and α as $\text{Beta}(c_i + \alpha, K - c_i + (m-1)\alpha)$. Here, i is the index of the target plot, with null panel selections considered in aggregate.

Examining the parameters of either the full (Equation (4)) or marginal model (Equation (5)) specifications tells us that α provides the equivalent of pseudo-observations for each plot. That is, the effect of α is equivalent to adding α identifications to each panel in the lineup. When α is small, these pseudo-observations have relatively little influence, but when α is large, the pseudo-observations can quickly dwarf any information provided by the data. This is particularly true for the marginal BB model, where the equivalent of $(m-1)\alpha$ pseudo-observations are added. A lineup might be evaluated between 10 and 30 times, and with a $m = 20$ panel lineup, even $\alpha = 1$ can easily dominate any signal present in the participant selection data.

As discussed in Section 1.1, the value of α also determines how many panels will have a relatively high selection probability θ_i (and how many panels will have $\theta_i \approx 0$). This means that α also determines how many panels in a lineup are likely to attract participant interest. The next section explores the impact and interpretation(s) of α in the context of statistical lineups.

2.1 | Effect of α on Lineup Scenarios

Under Scenario 1, each participant sees an entirely different lineup: the data plot(s) and null plots are exchanged between each participant evaluation. Thus, while there may be differences in the visual interest of each panel in any one lineup, these differences do not carry over to the next lineup. While all panels in a single generated lineup may not be equally likely to be selected, we do not have any information or ability to estimate or quantify these differences. In fact, in Scenario 1, it does not make sense to track any information beyond whether or not the data panel was selected: each evaluation is in effect a separate, independent Bernoulli trial. Our selection of α under this scenario goes beyond what might be considered a non-informative $\alpha = 1$ (corresponding to uniformly distributed θ_i over the $m - 1$ simplex). Instead, because we are averaging over all panels which could be generated by the data and null models (as both the data plot and the null plots are exchanged every time), we can claim that every plot is strictly equally likely to be selected under H_0. As seen in Figure D6 the BB model converges for $\alpha \to \infty$ asymptotically to the Binomial model (with $\theta = 1/m$) proposed in Majumder et al. (2013).

[**FIGURE** 6 about here.]

Under Scenario 3, however, each participant evaluates the same lineup, with the same data and null plots. In this scenario, we have enough information that we can model the θ_i across different lineup evaluations. Because the perceptual mechanisms which determine visual interest are shared across participants, and the same plots are used, we must allow θ_i to vary. The DM model (Equation (4)) provides this flexibility through the introduction of the parameter α. The general formula for calculating a visual p-value under the BB model appropriate for use in Scenario 3 is:

$$\text{p-value} = P(C \geq c_i) = \sum_{x=c_i}^{K} \binom{K}{x} \frac{1}{B(\alpha, (m-1)\alpha)} \cdot B(x + \alpha, K - x + (m-1)\alpha) \tag{6}$$

where c_i is the number of times the data panel was picked in K evaluations of the lineup. $B(.,.)$ is the Beta function defined as:

$$B(a, b) = \int_0^1 t^{a-1} \cdot (1-t)^{b-1} dt \text{ where } a, b > 0.$$

The derivation is in Appendix A.

The visual p-value calculation using Equation (6) is dependent on the value of α. We know from past studies (Hofmann et al. 2012; Loy, Follett, & Hofmann 2015 2016; Loy & Hofmann 2013 2015; Loy, Hofmann, & Cook 2017; Majumder, Hofmann, & Cook 2014; Roy Chowdhury et al. 2012 2015; VanderPlas & Hofmann 2017; Yin et al. 2013; Zhao, Cook, Hofmann, Majumder, & Chowdhury 2013) that only a few lineup panels attract attention, even if all of the panels in a lineup are null plots. Combining this observation with Figure D4, we would expect that $\alpha \ll 1$.

We expect the value of α to depend on several factors: the null generating model and the type of plot, and other aesthetic choices, all of which affect the visual distinctiveness of the null and actual data. To calculate visual p-values for lineups evaluated under Scenario 3, we must match the underlying data generation method and visual evaluation processes with an appropriate value of α.

As it is difficult to design a null plot generating method which will result in a specific α value, in practice, we will need to select an appropriate α for a predetermined null plot generating model. The selected α modulates the calculated visual p-value, as shown in Figure D6: when α is low, there are likely a few visually quite distinctive null plots drawing the attention away from the data. This makes it difficult to attribute data panel selections to definitive visual differences between the null and data plots. When α is relatively high, however, there are likely to be more null plots that attract visual attention; in this situation, it is very easy to determine whether the data panel is visually distinct compared to the null panels. Clearly, the choice of α is critical.

3 | ESTIMATION OF α

While it is generally possible to use maximum likelihood to estimate α (Minka 2012; Robitzsch 2020) directly based on proportions observed from panel selections in a lineup, these approaches fail because of the large number of observed zeroes. When participants select only a subset of interesting null panels, there are naturally many panels that have no selections. The many zeroes are even more pronounced when α values are small and only one or two null panels attract participant interest. The consequence is that these methods for estimation of α are most likely to fail in precisely the region of the parameter space where lineup experiments typically operate.

Instead, we propose a method for *visual* estimation of α which accommodates zero-count values and a numerical method for estimating α based on the results of Rorschach lineup based testing. Both the visual and numeric methods for estimating α leverage the expected number of "interesting" panels in a lineup.

Definition 1. c-interesting

We define lineup panel i to be c-interesting if c or more participants selected the panel as the most different.

This definition gives us an objective way to let evaluators determine what is interesting. The threshold c does not have to be an integer value. Using this definition, we also define random variable Z_c, the number of panels which are c-interesting in K evaluations of an m-panel lineup. The expected number of panels selected at least c times, $E[Z_c]$, is calculated as:

$$E[Z_c(\alpha)] = \frac{m}{B(\alpha, (m-1)\alpha)} \cdot \sum_{x=\lceil c \rceil}^{K} \binom{K}{x} B(x+\alpha, K-x+(m-1)\alpha). \tag{7}$$

A derivation of $E[Z_c]$ from Equation (4) is provided in Appendix C. Note that c-interestingness and Z_c both depend on the experimental conditions – the number of times a lineup is evaluated, and the number of panels in the lineup. While it would be reasonable to normalize the definition of c-interesting by K or m, this would imply that we can compare across different scenarios. Unfortunately, because we are usually dealing with relatively small values of K and m, even when we normalize counts, we still deal with a set of discrete quantities.

In a lineup experiment, we typically compare the selections of the data panel relative to the *aggregate* selections of null panels, using the marginal BB model. While this makes the calculations much simpler, because there is no need to keep track of the locations and selections of individual null plots, it does disregard some information; namely, the distribution of participant interest in the null panels. The proposed estimation methods are predicated on utilizing that discarded information to estimate α.

We will first discuss the visual and numerical methods for estimation of α from lineup evaluations, and then discuss one consequence of these methods for improved lineup diagnostics.

3.1 | Estimation using null choices in a one target lineup

There have been several explorations of the use of visual statistics as a supplement or an alternative to statistical inference (Correll & Heer 2017; Lawrence & Makridakis 1989; Meyer & Shinar 1992; Mosteller, Siegel, Trapido, & Youtz 1981). Visual estimates of correlation and linear regression are known to differ systematically from the numerical estimates, but not because the visual system is inaccurate or misleading. Instead, estimates derived visually tend to discount the effect of outliers, producing a more robust estimate of the statistical quantity of interest. In this application, we expect that visual estimation of α based on the expected number of "interesting" panels will produce a more robust estimate of α than the numerically unstable maximum likelihood estimates.

Figure D4 illustrates that α is directly related to the number of panels that are selected as interesting by participants. The estimation of α is based on the relationship between α and the number of expected c-interesting panels as developed in the previous section. Figure D7 illustrates this relationship for the example of $K = 30$ evaluations of 19 null panels: the blue line corresponds to the expected number of 1-interesting null panels Z_1 as given in Equation (7). Each point in the figure is based on 10 simulations of a lineup of size (K, m).

This provides some variability around how many times each panel was selected, but the number of panels which attract participant attention is fairly consistent across multiple simulations. The light grey horizontal bands in Figure D7 mark different levels of c-interesting null panels. We consider situations with 2-3, 4-5, 6-7, and 8+ null panels that were selected at least once. The corresponding values for α are shown in the connected vertical bands. If only one null panel is selected more than c times, and there are a reasonable number of null panel selections overall, this may be a sign that the null plot generation method is unsuitable. A discussion of this case is deferred to Section 3.3.

We use Figure D7 for visually estimating α by executing a reverse lookup of the rate parameter based on the number of interesting null panels.

A natural threshold for sufficient participant interest is $c = K/m$; that is, setting the threshold c to be equal to the expected number of selections under the specified BB model. Using null panel selections hinges on a property of the Dirichlet distribution: when one category is removed from consideration, the remaining categories still maintain a reduced-dimension Dirichlet distribution with the same parameter α. A discussion of this property and its application to visual inference is provided in Appendix D.

[**FIGURE** 7 about here.]

The proposed visual estimation process is conducted using a simulation from the DM model in Equation (4). Furthermore, we will assume for this demonstration that the lineup experiment we are planning to conduct has a standard size lineup ($m = 20$ panel, $m_0 = 19$ null plot) with $K_0 = 30$ null panel selections out of $K = 40$ evaluations.

Identify the number of c-interesting null panels in the lineup of interest. Consider Figure D7. Generally, this figure needs to be adjusted to match the number of null panels, m_0, and the corresponding number of null panel selections, K_0. The `alpha_from_data_lineup()` function in the `vinference` package (available at https://github.com/heike/vinference) can be used to create this plot given values of c, m_0, and K_0.

Locate a y-axis range around the observednumber of c-interesting null panels in the data Using the simulated selected panel counts, on the vertical axis locate a band around the number of c-interesting null panels in the lineup. For instance, if there were three null panels with more than c selections, we would use a range of 2-3 panels.

Using a range instead of a single point estimate is intended to account for some of the variability resulting from the use of a single lineup with non-deterministic K_0 null panel evaluations. In a simulation, it is easy to increase the number for simulations N until results no longer change substantially; this is expensive and much more time consuming when using human evaluations and setting K instead of K_0 in the experimental design.

Select an $\hat{\alpha}$ value corresponding to the selected band and calculate the visual p-value. Using the chosen $\hat{\alpha}$ value, calculate a visual p-value using Equation (6). Assess the sensitivity of this p-value calculation to the choice of α values within the selected band. A plot such as the one shown in Figure D8 may be helpful when assessing the sensitivity of the visual p-value to different values of α in each band. Figure D8 is a segmented version of Figure D6; each panel corresponds to the bands of interesting panel counts selected in step 1. Locate the relevant panel of the plot, and, using the number of target panel selections, determine whether the visual p-value calculation is conclusive for every α in the panel.

For instance, if our selected lineup contained 3 interesting null panels, $\hat{\alpha} = 0.075$ approximately corresponds to $Z_c = 3$; we should compare to $\alpha = 0.01$ and $\alpha = 0.09$ (the approximate outer range of α for our band) to get a sense of the sensitivity of our p-value to α. Examining the second panel of Figure D8, which corresponds to 3 interesting null panels, we see that p-values will significant at the 0.05 level when the number of data panel selections is at least 20; if there are fewer than 15 data panel selections, then p-values will not be significant at the 0.05 level. In our example, we have 10 target selections ($K = 40, K_0 = 30$), corresponding to a visual p-value that will not be significant for any range of α in our selected band.

[**FIGURE** 8 about here.]

In Figure D8, we show the implications of the visual selection method for each band of α values in terms of the number of target plot selections necessary to achieve statistical significance at the $p = 0.05$ level. Due to the discretization of the expected number of panels with more than c selections, each range of α values is ambiguous for one or more potential data panel selection counts; these values are shown in red, with labeled thresholds for non-significance and significance. In the unfortunate situation where the estimated $\hat{\alpha}$ produces an inconclusive result, the experimenter has two options. The inexpensive, conservative approach is to declare any inconclusive results to be non-significant, in effect using the smallest α value corresponding to the approximate number of panels selected. Alternately, the experimenter could use the method of using additional Rorschach lineups as described in Section 3.2 to produce a more precise $\hat{\alpha}$ which would provide definitive results, at the cost of conducting a secondary study.

By estimating α, we produce visual p-values calibrated based on the specific null plot generation method. In most cases, we get the improved calibration for free, because we can obtain this information from the null panels in one or two target lineups. Occasionally, either because the signal in the data plot is too strong, or because the visual estimation method for α produces a range of p-values that are inconclusive given the number of target plot selections, we may need to invest additional effort to generate a precise $\hat{\alpha}$ estimate numerically.

In the case where a tested lineup has selections which overwhelmingly favor the data plot, we have an interesting dilemma: the lineup is likely significant (based on the overwhelming evidence that the data plot attracts the most visual interest), but we cannot estimate α because there are insufficient null panel selections. In this case, estimation of α will depend on the creation of a Rorschach lineup (that is, a lineup consisting entirely of plots generated under the null hypothesis). This process is described in more detail in Section 3.2.

As it is rare for the data plot to be the only panel to attract participant attention, we can usually recover information about α from the null panels in the same lineup. This is a more efficient use of participant time and experimenter resources, as we do not have to ask participants to evaluate two separate lineups to determine the visual p-value for a plot.

3.2 | Estimation using Rorschach lineups

Equation (7) can be used to generate a more precise estimate of α with a secondary study consisting of one or more Rorschach lineups evaluated K times each. This estimation method requires the following steps:

Experimentally evaluate some m-panel Rorschach lineups K times each. To ensure validity of the estimated α, the Rorschach lineups should use the same null data generating mechanism used in the standard lineup. It is also important that K be the same for each lineup which is evaluated, as Equation (7) depends on K.

Calculate the mean number of c-interesting panels in the Rorschach lineups. This empirical estimate of $E[Z_c]$ will be used in combination with Equation (7) to estimate α.

Numerically determine $\hat{\alpha}$ using $\overline{Z}_c$ and Equation (7). The reference distribution is one with $m = m_0$ null panels, because of the use of Rorschach lineups, in contrast to the reference distribution in Section 3.1.

The `alpha_from_null_lineup()` function in the `vinference` package (available at https://github.com/heike/vinference) can be used to estimate $\hat{\alpha}$ given values of $\overline{Z}_c$, c, m_0, and K_0.

Note that if on average only one null panel is selected, the estimated $\hat{\alpha}$ is not reliable: Equation (7) does not have the sensitivity to differentiate the results of a lineup with $\hat{\alpha} = 0.001$ from the results of a lineup with $\hat{\alpha} = 0.01$; this is also apparent on the left side of Figure D7. However, very small estimated values of $\hat{\alpha}$ signal that there is a problem with the null plot generation method; this case is described in more detail in Section 3.3.

This approach could also be used in situations where a precise $\hat{\alpha}$ is necessary, as it has greater numerical precision than the visual estimation method proposed in Section 3.1. An intermediate solution might be to include a few Rorschach lineups in the design of a lineup experiment, so that enough data is obtained from the Rorschach lineups to estimate α, without requiring the overhead of an additional study or a large number of additional participant evaluations. In this compromise approach, both the visual method and the numerical method could be used: the first to obtain an approximate interval, and the second to justify a specific value of α.

3.3 | Detecting inadequate null generation methods

When only one null plot in a Rorschach lineup is visually interesting, we encounter two problems: (1) $\hat{\alpha}$ cannot be estimated with precision, because many different α values below $\alpha \approx 0.025$ could give rise to a situation where only one panel is visually interesting; (2) any one-target lineup may be confounded if the data plot replaces the interesting null plot. If the interesting null is replaced by the data plot, we would expect that the count values would look similar to those produced when evaluating the Rorschach lineup, but even if it is not, the distribution of counts might not differ enough to show statistical significance when $\hat{\alpha}$ is small. The lineup hypothesis testing method is predicated on the ability to visually distinguish a Rorschach lineup from a lineup with a data target, so a method that sporadically generates overwhelmingly interesting null plots provides insufficient grounds for rejection of the null hypothesis.

Another interpretation of an extremely low $\hat{\alpha}$ value is to cast suspicion on the null generating process. Null generating methods should not inadvertently generate highly variable numbers of interesting nulls; rather one expects them to be uniformly dull. An example is described in VanderPlas and Hofmann (2017), where the generated null plots occasionally lacked important features critical to the study that made them inadvertently visually different. Estimation of α provides us with a way to screen for bad null plot generation methods, in addition to providing more accurate estimates of visual p-values.

We can see from Figure D6 that the proposed method will generate p-values that are more moderate than the simpler Binomial model that does not involve the extra step of estimating α. In the next section, we explore the practical implications of this method by examining actual lineups that were tested experimentally, and the visual p-values calculated using each method.

4 | EXAMPLE

This is a reassessment of the p-value computed for a lineup (Figure D9) used in Majumder et al. (2013). Using the DM model, computed visual p-values are more conservative than the previously computed using the Binomial model proposed in that work. The responses from evaluators were somewhat ambiguous: the data plot in panel 8 was selected most frequently by participants (14 times), but in total, null plots were selected more frequently than the target plot (22 times). In particular, the null plot in panel 4 attracted 10 selections. Using a Binomial model for the evaluation this plot produces a visual p-value on the order of 10^{-9}, which seems extremely small given that there were more null plot selections than data plot selections.

[**FIGURE** 9 about here.]

Under the DM model in Equation (4), the first step is to assess the expected number of null panels which attract attention: six of 19 null plots were selected from the lineup at least once. The corresponding simulation in Figure D10 suggests a value of α between 0.1 and 0.25.

[**FIGURE** 10 about here.]

Under the BB mixture model, the visual p-value for the data shown in the lineup in Figure D9 is 0.03 based on $\hat{\alpha} = 0.174$ corresponding to six interesting null plots. For a range of 5-7 interesting null plots, we have estimates for α between 0.1 and 0.25, corresponding to p-values between 0.02 and 0.04. All of these values are still significant at the 0.05 level, however, these values are larger than the p-value computed under the Binomial model. When we consider the actual results of the experimental evaluation of Figure D9, the p-value computed using the mixture model

is much more plausible: the data plot is the most favored of all of the panels in the lineup, but it is not overwhelmingly significant; at least one other panel is almost as popular, and overall, null panels were still selected more frequently than the data panel. This suggests that p-values generated using the BB mixture model are better calibrated than those generated by the Binomial model.

5 | DISCUSSION

In this paper, we have described three scenarios for visual inference experiments and developed a model for visual inference, that is effective even under Scenario 3, where dependencies complicate probability calculations. This model is a more general case of the model proposed in Majumder et al. (2013), but provides p-values that account for dependencies in successive evaluations of the same lineup by different participants. The proposed Dirichlet-Multinomial (DM) model must be calibrated using a parameter α which describes the number of null panels in a lineup which are visually interesting. Using expected quantities that are easily observable, we describe two methods for estimation of α: a visual method for use on lineups that contain data panels, and a numerical method for use on Rorschach lineups. The p-values derived from the proposed DM model are more conservative than those produced by Majumder's Binomial model. Even though the latter is appropriate for Scenario 1, with the new model the use of $\hat{\alpha}$, estimated from the null plot generation method, provides an additional benefit: we calibrate the p-value to the lineups by accounting for the *difficulty of the lineup evaluation under the null hypothesis*.

Visual inference leverages the strength of the perceptual system to test hypotheses, for a broad range of problems that cannot be addressed with traditional tests. This comes with a cost: we cannot easily describe the null distribution or determine its parameters from the experimental design. An exciting contribution of the visual and quantitative estimation methods proposed in this paper is that they allow us to determine the appropriate null distribution for a specific lineup experiment, with minimal cost to experimenters or participants. By controlling for the perceptual difficulty of a task using quantitative parameters, we can produce visual p-values that account for the demands of the task while still providing an assessment of the significance of the test statistic.

Visual estimation of α also allows us to screen out lineup generation methods that are problematic: if a null plot generation method produces one interesting plot out of a size $m = 20$ Rorschach lineup, it is functionally impossible to distinguish a data plot which has a large signal from a null plot which also has a large signal. Extremely small α values serve as a signal that the null plot generation method is not visually appropriate because it sporadically (not consistently) generates visual features that are likely to catch the attention of participants (possibly for the wrong reasons). We can guard against these anomalies by assessing α using a Rorschach pilot study, identifying any problems with the null plot generation method before the full study is conducted.

With methods for calculating p-values for visual inference studies conducted under Scenario 1 or Scenario 3, we can now approach the variety of study designs that might fall under Scenario 2 as described in this paper. Scenario 2 is an intermediate option between the two extremes of no dependence between successive lineup evaluations and complete dependence between successive lineup evaluations. As a result, by describing the calculation of p-values for Scenario 3, this paper lays a foundation for a comprehensive assessment of the visual p-values of different lineups under intermediate scenarios, too.

Supporting Information

The code and data necessary to reproduce this article are available at `https://github.com/srvanderplas/visual-inference-alpha`.

The data that support the findings of this study are also openly available in figshare at `https://doi.org/10.6084/m9.figshare.12894932`, reference number 12894932.

References

Amazon. (2005-2015). *Mechanical Turk*. `https://www.mturk.com/mturk/welcome`, Accessed: 2020-08-30.

Anscombe, A. J. (1972). Graphs in Statistical Analysis. *The American Statistician*, *27:1*, 17-21.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E., Swayne, D., & Wickham, H. (2009). Statistical Inference for Exploratory Data Analysis and Model Diagnostics. *Royal Society Philosophical Transactions A*, *367*(1906), 4361-4383.

Connor, R. J., & Mosimann, J. E. (1969, March). Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution. *Journal of the American Statistical Association*, *64*(325), 194. doi: 10.2307/2283728

Correll, M., & Heer, J. (2017). Regression by Eye: Estimating Trends in Bivariate Visualizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17* (pp. 1387–1396). Denver, Colorado, USA: ACM Press. Retrieved from `http://dl.acm.org/citation.cfm?doid=3025453.3025922` doi: 10.1145/3025453.3025922

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, *2*(2), 183–201.

Exner, J., & Erdberg, P. (2003). *The rorschach, basic foundations and principles of interpretation*. Wiley. Retrieved from https://books.google.com/books?id=4F5qAAAAMAAJ tex.lccn: 20229613.

Hamilton, N. E., & Ferry, M. (2018). ggtern: Ternary diagrams using ggplot2. *Journal of Statistical Software, Code Snippets*, *87*(3), 1–17. doi: 10.18637/jss.v087.c03

Hofmann, H., Follett, L., Majumder, M., & Cook, D. (2012). Graphical tests for power comparison of competing designs. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2441–2448.

Lawrence, M., & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes*, *43*(2), 172 – 187. Retrieved from http://www.sciencedirect.com/science/article/pii/0749597889900496 doi: https://doi.org/10.1016/0749-5978(89)90049-6

Loy, A., Follett, L., & Hofmann, H. (2015). Variations of Q-Q Plots – the Power of our Eyes! *The American Statistician*, *2015*(ja), 1-36. Retrieved from http://dx.doi.org/10.1080/00031305.2015.1077728 doi: 10.1080/00031305.2015.1077728

Loy, A., Follett, L., & Hofmann, H. (2016). Variations of q–q plots: The power of our eyes! *The American Statistician*, *70*(2), 202–214.

Loy, A., & Hofmann, H. (2013). Diagnostic tools for hierarchical linear models. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(1), 48–61. Retrieved 2019-05-10, from http://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1238 doi: 10.1002/wics.1238

Loy, A., & Hofmann, H. (2015, October). Are You Normal? The Problem of Confounded Residual Structures in Hierarchical Linear Models. *Journal of Computational and Graphical Statistics*, *24*(4), 1191–1209. Retrieved from http://www.tandfonline.com/doi/full/10.1080/10618600.2014.960084 doi: 10.1080/10618600.2014.960084

Loy, A., Hofmann, H., & Cook, D. (2017, July). Model Choice and Diagnostics for Linear Mixed-Effects Models Using Statistics on Street Corners. *Journal of Computational and Graphical Statistics*, *26*(3), 478–492. Retrieved 2019-05-10, from https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1330207 doi: 10.1080/10618600.2017.1330207

Majumder, M., Hofmann, H., & Cook, D. (2013). Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association*, *108*(503), 942–956.

Majumder, M., Hofmann, H., & Cook, D. (2014, August). Human Factors Influencing Visual Statistical Inference. *arXiv:1408.1974 [stat]*. Retrieved 2019-07-26, from http://arxiv.org/abs/1408.1974

Matejka, J., & Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 chi conference on human factors in computing systems* (p. 1290âĂŞ1294). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3025453.3025912 doi: 10.1145/3025453.3025912

Meyer, J., & Shinar, D. (1992, June). Estimating Correlations from Scatterplots. *Human Factors*, *34*(3), 335–349. Retrieved from https://doi.org/10.1177/001872089203400307 Publisher: SAGE Publications Inc. doi: 10.1177/001872089203400307

Minka, T. P. (2012). Estimating a dirichlet distribution [Computer software manual]. Technical Report. Retrieved from https://tminka.github.io/papers/dirichlet/minka-dirichlet.pdf

Mosteller, F., Siegel, A. F., Trapido, E., & Youtz, C. (1981, August). Eye Fitting Straight Lines. *The American Statistician*, *35*(3), 150–152. Retrieved from https://amstat.tandfonline.com/doi/abs/10.1080/00031305.1981.10479335 doi: 10.1080/00031305.1981.10479335

Robitzsch, A. (2020). sirt: Supplementary item response theory models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=sirt R package version 3.9-4.

Roy Chowdhury, N., Cook, D., Hofmann, H., & Majumder, M. (2012). *Where's Waldo: Looking Closely at a Lineup* (Tech. Rep. No. 2). Iowa State University, Department of Statistics.

Roy Chowdhury, N., Cook, D., Hofmann, H., Majumder, M., Lee, E.-K., & Toth, A. L. (2015). Using visual statistical inference to better understand random class separations in high dimension, low sample size data. *Computational Statistics*, *30*(2), 293-316. Retrieved from http://dx.doi.org/10.1007/s00180-014-0534-x doi: 10.1007/s00180-014-0534-x

Sakowicz, A., & Wesołowski, J. (2014, August). Dirichlet distribution through neutralities with respect to two partitions. *Journal of Multivariate Analysis*, *129*, 1–15. doi: 10.1016/j.jmva.2014.04.004

VanderPlas, S., & Hofmann, H. (2017). Clusters beat trend!? testing feature hierarchy in statistical graphics. *Journal of Computational and Graphical Statistics*, *26*(2), 231–242.

Vanderplas, S., Hofmann, H., & Cook, D. (2020, Aug). *Statistical significance calculations for scenarios in visual inference*. figshare. Retrieved from https://figshare.com/articles/dataset/_/12894932/0 doi: 10.6084/m9.figshare.12894932

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Wilkinson, L. (1999). *The Grammar of Graphics*. New York: NY: Springer.

Yin, T., Majumder, M., Roy Chowdhury, N., Cook, D., Shoemaker, R., & 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(139). Retrieved from https://www.omicsonline.org/visual-mining-methods-for-rnaseq-data-data-structure-dispersion-estimation-and-significance-testing-2153-0602.1000139.php?aid=17041 doi: http://dx.doi.org/10.4172/2153-0602.1000139

Zhao, Y., Cook, D., Hofmann, H., Majumder, M., & Chowdhury, N. R. (2013). Mind reading: Using an eye-tracker to see how people are looking at lineups. *International Journal of Intelligent Technologies and Applied Statistics*, *6*(4), 393–413.

APPENDIX

A VISUAL P-VALUE DISTRIBUTION

Assume we have a lineup of size m with K evaluations resulting in c_t target plot evaluations. We defined the BB model in Equation (5) leading to densities given as:

$$f(\theta \mid \alpha) = \frac{1}{B(\alpha, (m-1)\alpha)} \cdot \theta^{\alpha-1}(1-\theta)^{(m-1)\alpha-1}$$

$$P(C = c_t \mid K, \theta) = \binom{K}{c_t}\theta^{c_t}(1-\theta)^{K-c_t}$$

We are interested in the probability of observing at least c_t picks of the target plot assuming that the target plot is not inconsistent with the null plots generated from the null model, i.e. we are interested in the (unconditional) distribution of counts C. We get there by integrating over the rate parameter θ. From the theorem of total probability we know that

$$P(C = c) = \int_0^1 P(C = c \mid \theta) f(\theta) d\theta$$

Now we use that $C \mid \theta \sim \text{Binom}_{\theta,K}$ and $\theta \sim \text{Beta}_{\alpha,(m-1)\alpha}$:

$$\begin{aligned} P(C = c) &= \int_0^1 \binom{K}{c}\theta^c(1-\theta)^{K-c} \cdot \frac{1}{B(\alpha,(m-1)\alpha)}\theta^{\alpha-1}(1-\theta)^{(m-1)\alpha-1} d\theta \\ &= \binom{K}{c}\frac{1}{B(\alpha,(m-1)\alpha)} \underbrace{\int_0^1 \theta^{c+\alpha-1}(1-\theta)^{K-c+(m-1)\alpha-1} d\theta}_{\text{Beta function}} \\ &= \binom{K}{c}\frac{B(c+\alpha, K-c+(m-1)\alpha)}{B(\alpha,(m-1)\alpha)}. \end{aligned}$$

Thus, the visual p-value for a lineup with c_t target selections out of K evaluations is

$$P(C \geq c_t) = \frac{1}{B(\alpha,(m-1)\alpha)} \sum_{x=c_t}^{K} \binom{K}{x} B(x+\alpha, K-x+(m-1)\alpha). \tag{A1}$$

A similar derivation holds in the full Dirichlet-Multinomial model.

B SCENARIO 1 IS BINOMIAL

Under Scenario 1 a lineup is shown to only one participant. Let Z_i be the binary variable capturing whether lineup i resulted in a data pick. Then C, the number of data picks in K evaluations of K lineups is given as the sum of K independent evaluations Z_i, i.e. $C = Z_1 + ... + Z_k$. Assuming (independent) Dirichlet samples for each lineup, assume θ_i is the probability (under the null hypothesis) to pick the data plot. That is, each evaluation is a Bernoulli trial with success probability θ_i. Then $C \mid \theta_1, ..., \theta_K \sim$ Poisson-Binomial distribution:

$$P(C = c \mid \theta_1, ..., \theta_K) = \sum_{A \in A_c} \prod_{i \in A} \theta_i \prod_{j \notin A} (1-\theta_j)$$

where A_c is the set of all subsets of c integers that can be selected from $\{1, 2, 3, ..., K\}$.

For the distribution of C therefore holds:

$$\begin{aligned} P(C = c) &= \sum_{A \in A_c} \prod_{i \in A} \underbrace{\int_0^1 \theta_i f(\theta_i) d\theta_i}_{E[\theta_i]} \cdot \prod_{j \notin A} \underbrace{\int_0^1 (1-\theta_j) f(\theta_j) d\theta_j}_{1-E[\theta_j]} = \\ &= \sum_{A \in A_c} \prod_{i \in A} \frac{1}{m} \cdot \prod_{j \notin A} (1 - \frac{1}{m}) = \\ &= \binom{K}{c} \left(\frac{1}{m}\right)^c \left(1 - \frac{1}{m}\right)^{K-c} \end{aligned}$$

This shows that under Scenario 1 the null distribution of the number of data picks simplifies to a Binomial distribution $B_{\frac{1}{m},K}$.

C EXPECTED NUMBER OF PANELS PICKED

Define $C = (C_1, ..., C_m) \sim Mult_{\theta,K}$ to be a (simulated) lineup that is evaluated K times, and $\theta = (\theta_1, ..., \theta_m) \sim Dir_\alpha = (\alpha, ..., \alpha)$ with $\sum_i \theta_i = 1$. With indicator function I, defined as 1 for true statements and 0 for false statements, we define:

$$Z_c(\alpha) = \sum_{i=1}^{m} I(C_i \geq c),$$

where Z_c is the number of panels in a lineup that were picked at least c times. We express this random variable as a function in α - the dependency becomes clear, once we look at the expected value of Z_c:

$$E[Z_c(\alpha)] = \sum_{i=1}^{m} E\left[I(C_i \geq c)\right] = \sum_{i=1}^{m} P(C_i \geq c).$$

The probabilities $P(C_i \geq c)$ are derived in the previous section as marginal Beta-binomials:

$$E[Z_c(\alpha)] = \frac{m}{B(\alpha, (m-1)\alpha)} \cdot \sum_{x=\lceil c \rceil}^{K} \binom{K}{x} B(x + \alpha, K - x + (m-1)\alpha). \tag{C2}$$

D PARTITIONS OF DIRICHLET DISTRIBUTIONS

Let $\boldsymbol{\theta} = (\theta_1, ...\theta_m)$ with $\boldsymbol{\theta} \sim Dir(\alpha)$ for some real-valued $\alpha > 0$. The Dirichlet distribution is neutral with respect to all possible partitions of the corresponding index set (Connor & Mosimann 1969; Doksum 1974; Sakowicz & Wesołowski 2014). This means that θ_m is independent of the (normalized) random variable $\boldsymbol{\theta}^{-m} := (\frac{\theta_1}{1-\theta_m}, ..., \frac{\theta_{m-1}}{1-\theta_m})$. Further, this implies that the conditional distribution $\boldsymbol{\theta}^{-m} \mid \theta_m \sim Dir(\alpha)$. In particular, the density of the joint distribution of θ_m and θ^{-m} factors as a Gamma density for θ_m and an $m-1$-dimensional symmetric Dirichlet density with the same parameter α for θ^{-m}.

In the context of lineups this means that given the same null model generation we can assume the probabilities to select a panel from a Rorschach lineup and the probabilities to select a panel corresponding to a null plot from a lineup follow symmetric Dirichlet distributions with the same parameter $\alpha > 0$.

List of Figures

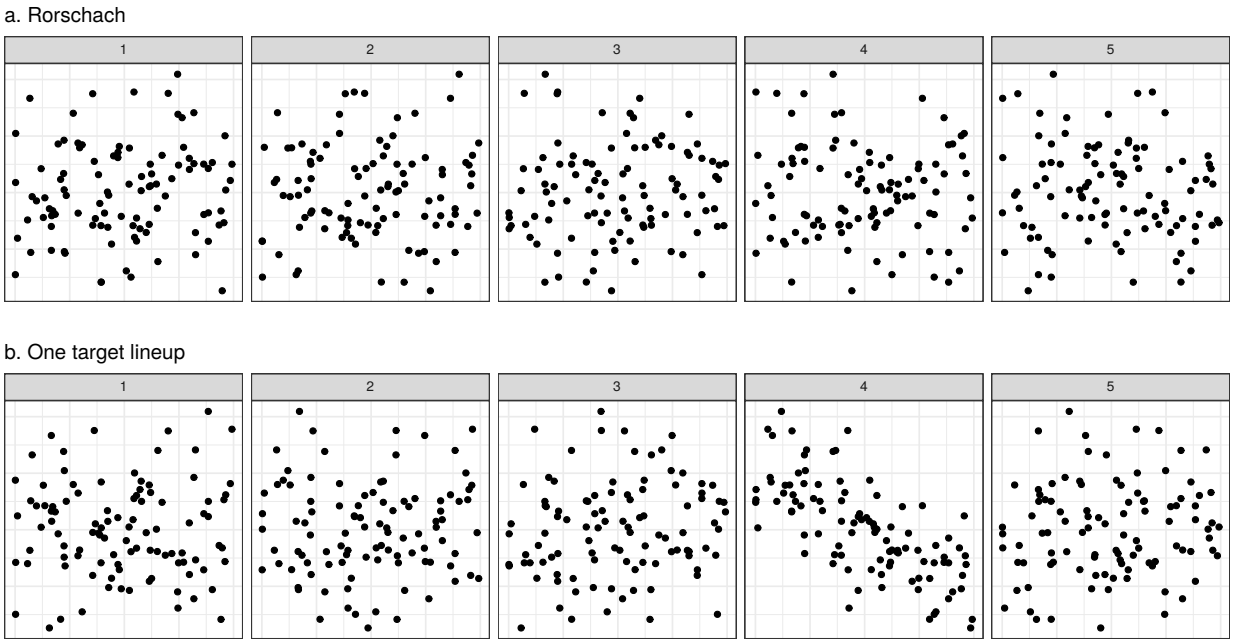

FIGURE D1 Types of lineups. In a Rorschach, all plots are null plots, with a purpose of understanding patterns that may occur by chance. A one target lineup has one data plot, and the remaining are null plots.

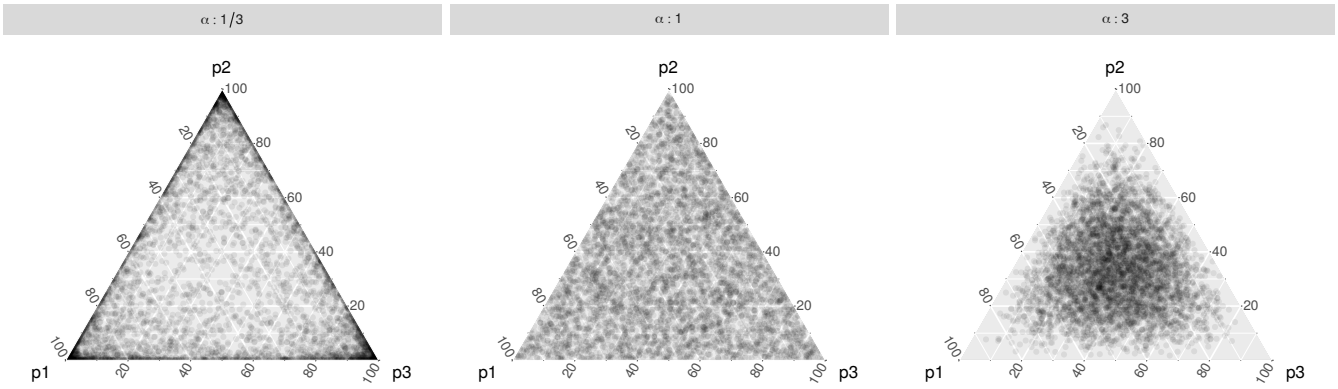

FIGURE D2 Dirichlet distributed samples on the 2-dimensional simplex. $\alpha = 1$ is a uniform distribution on the simplex, $\alpha < 1$ up-weighs the (hyper)faces of the simplex, $\alpha > 1$ puts more weight towards the center of the simplex.

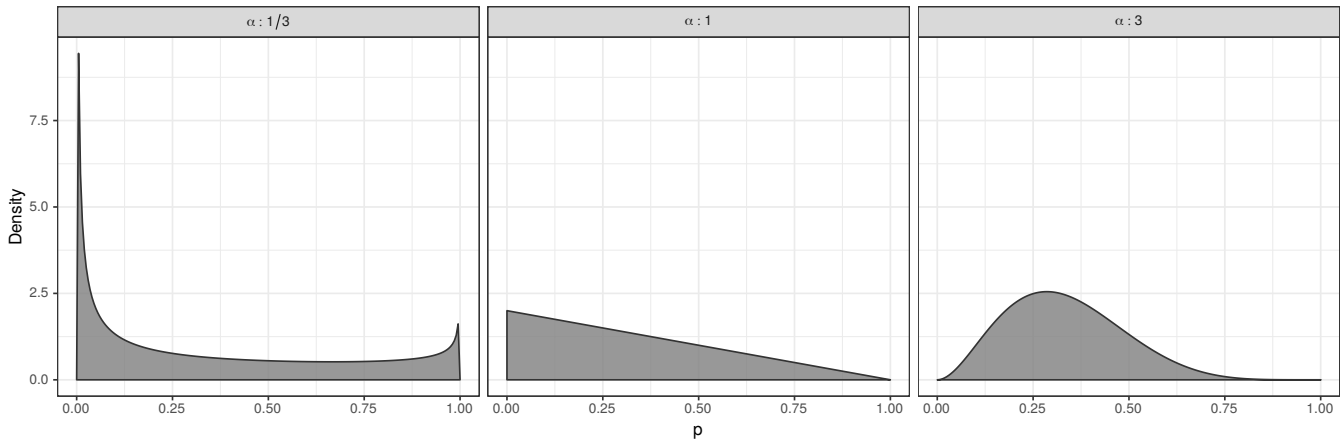

FIGURE D3 Marginal Beta($\alpha, 2\alpha$) densities corresponding to the above Dirichlet densities.

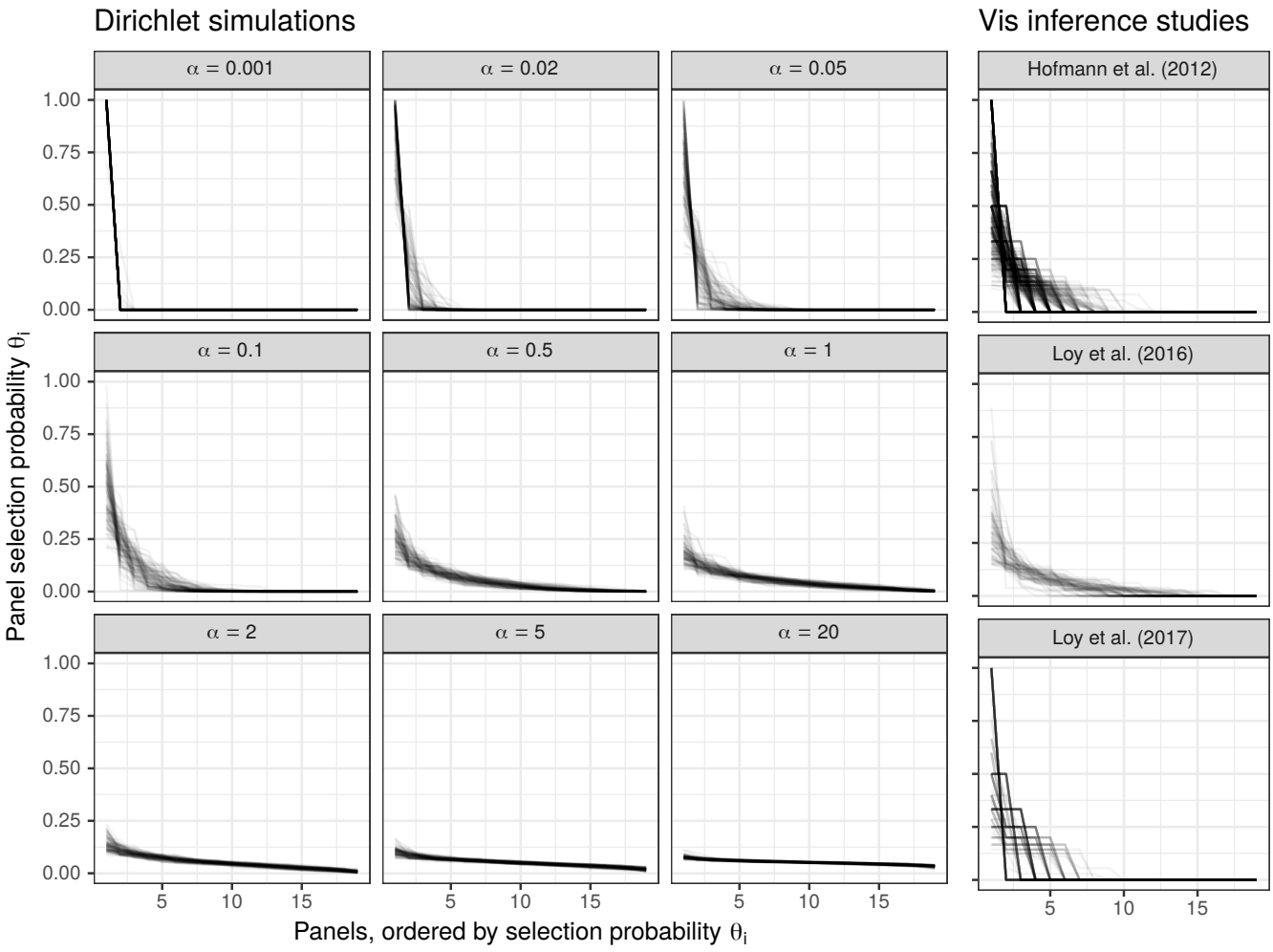

FIGURE D4 Simulation of the panel selection probabilities θ_i, sorted, for different values of α (left). In the right column, panel selection probabilities estimated from lineups used in past visual inference experiments; these probabilities are calculated from lineups with finite evaluations, so the calculated probabilities are not continuous. For low values of α, plot selections are concentrated on only a few plots, while higher values of α show a wider spread of selections among more plots. The included selection probabilities from previous studies(Vanderplas et al. 2020) show that not all panels in each lineup are selected, even when evaluated multiple times; it is also clear that the selection probabilities are more concentrated in the first study, where in some lineups only one panel was selected; in the third study, at least 4 panels were selected in every lineup.

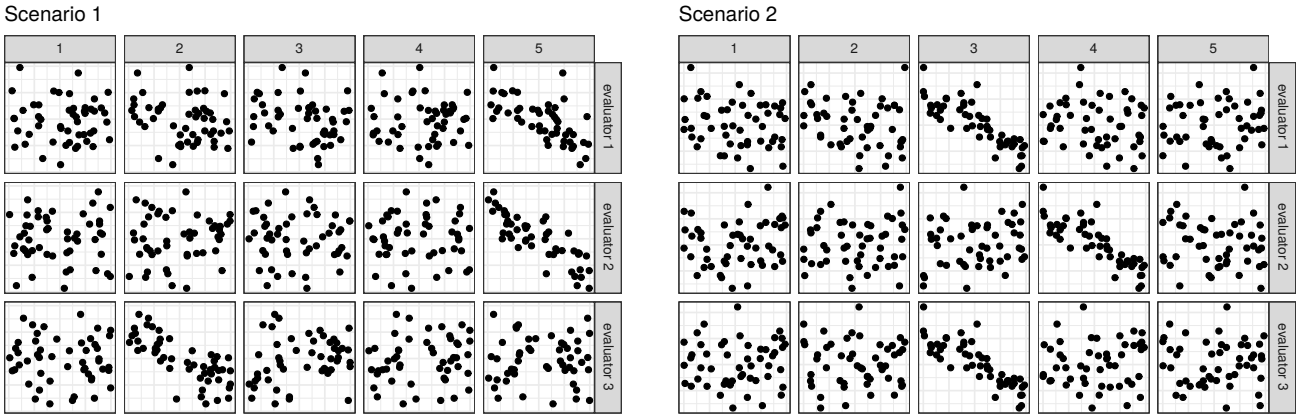

FIGURE D5 Illustrations of scenarios 1 and 2. Left column shows what three different evaluators would see under scenario 1, and the right column shows what three different evaluators would see under scenario 2. In scenario 1, all nulls and data plots are different, while in scenario 2 only the nulls differ. (Scenario 3 would have all three lineups the same.) The location of the data plot may vary from one evaluator to another.

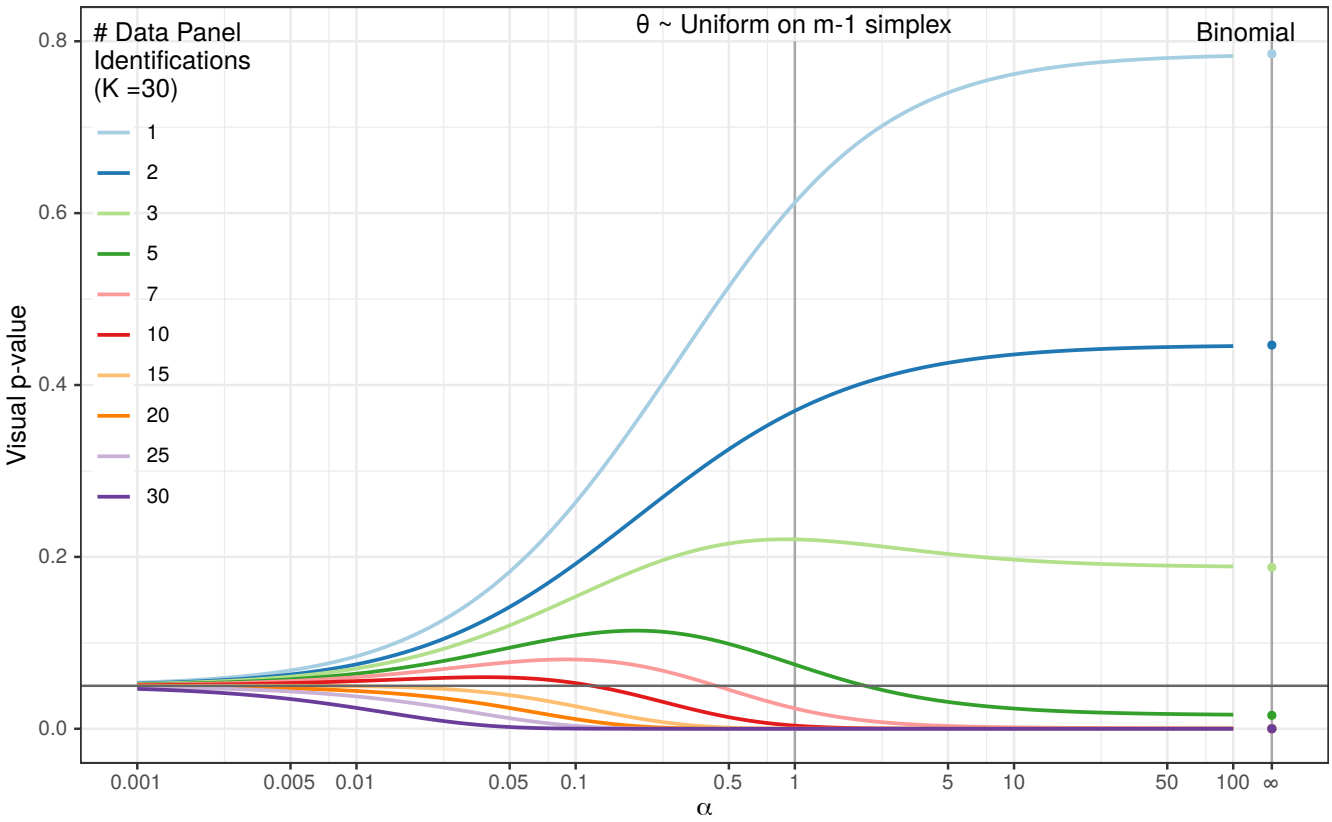

FIGURE D6 Sensitivity of visual p-value to selection of α under the marginal beta-binomial model. Corresponding values for the binomial model are shown on the right side of the plot; as $\alpha \to \infty$, the beta-binomial p-values converge to the binomial model p-value.

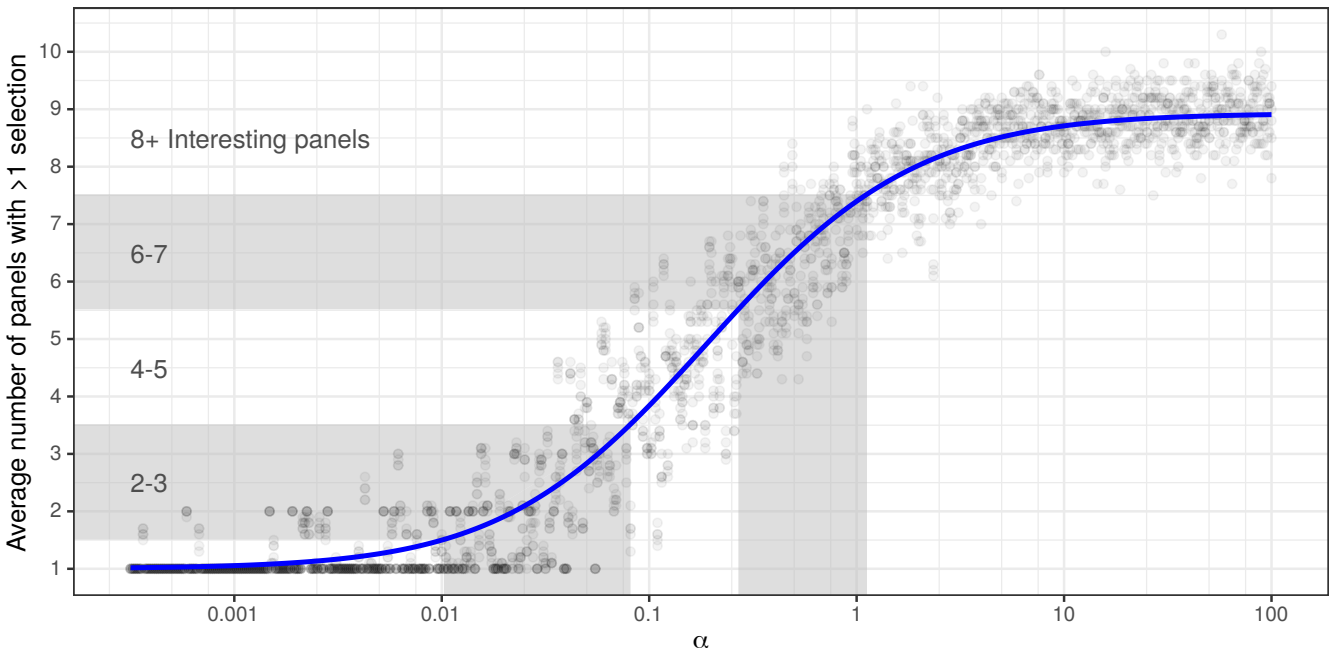

FIGURE D7 Average number of panels selected more than once for a range of α values. Each point represents 10 simulations of lineups with K $= 30$ evaluations. The line in blue shows the expected number of panels selected more than once as given in Equation (7). Bands are shown in alternating grey and white corresponding to a discretized heuristic for selection of α when K $= 30$.

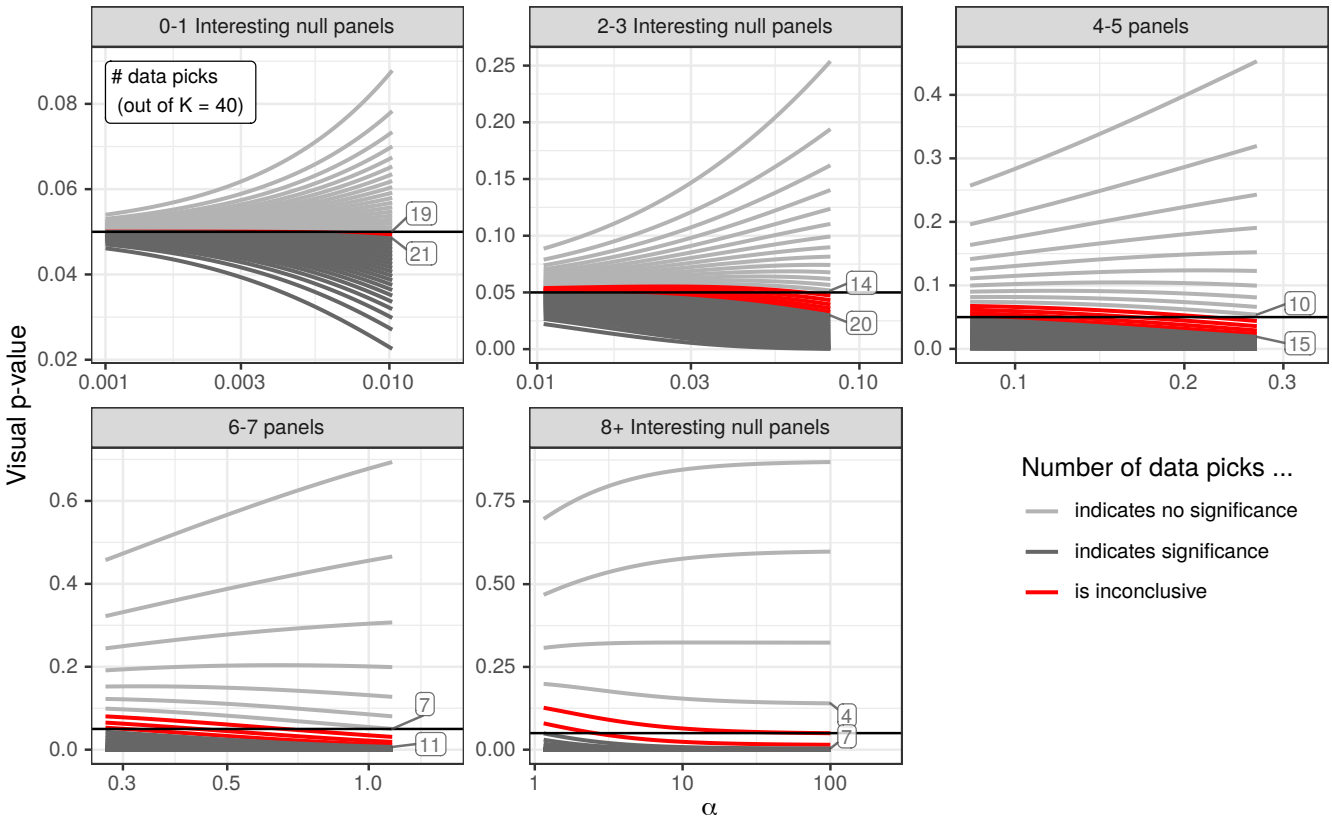

FIGURE D8 Sensitivity of visual p-value to selection of α under the beta-binomial model, for a m $= 20$ panel lineup with K $= 40$ evaluations. Data picks which result in ambiguous results under the discretized bands of α values are highlighted in red. At any particular α level, there are only a few values which result in an inconclusive results.

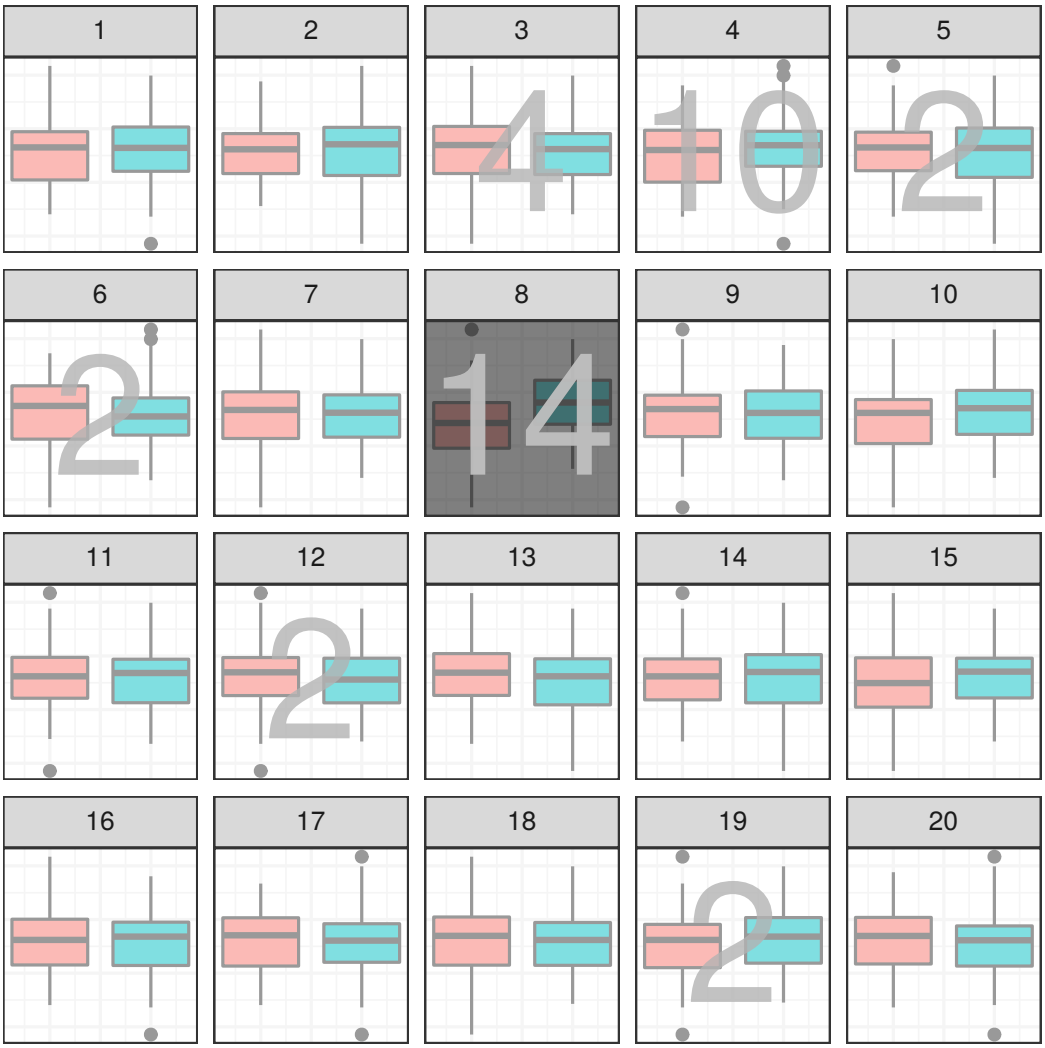

FIGURE D9 A lineup designed to test the utility of boxplots for detecting distributional differences. Selection counts are provided in light grey on top of each panel which was selected at least once. The data panel is highlighted in dark grey (panel 8).

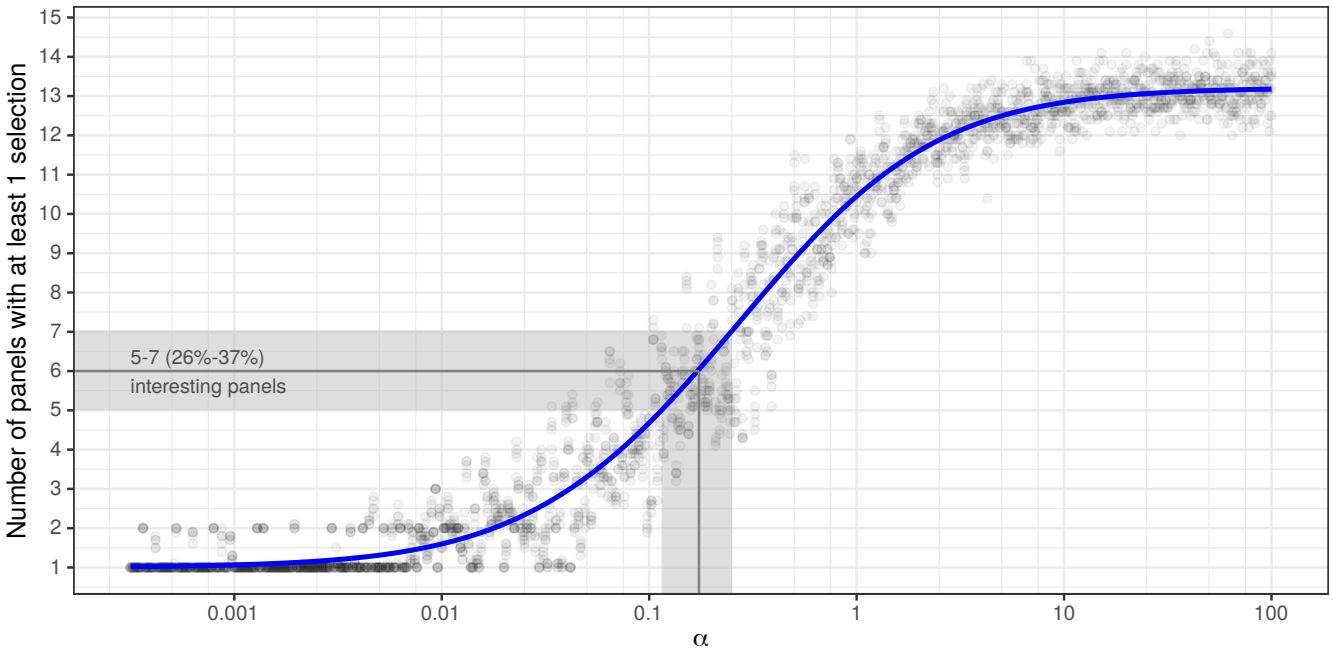

FIGURE D10 Number of expected panels with at least one selection in the 22 null plot selections for the lineup shown in Figure D9.