

Appendix to “A Plot is Worth a Thousand Tests: Assessing Residual Diagnostics with the Lineup Protocol” published in the Journal of Computational and Graphical Statistics

Weihao Li^a, Dianne Cook^a, Emi Tanaka^{a,b,c}, Susan VanderPlas^d

^aDepartment of Econometrics and Business Statistics, Monash University, Clayton, VIC, Australia; ^bBiological Data Science Institute, Australian National University, Acton, ACT, Australia; ^cResearch School of Finance, Actuarial Studies and Statistics, Australian National University, Acton, ACT, Australia; ^dDepartment of Statistics, University of Nebraska, Lincoln, Nebraska, USA

ARTICLE HISTORY

Compiled March 25, 2024

Appendix A. Additional details of testing procedures

A.1. *Statistical significance*

Within the context of visual inference, with K independent observers, the visual p -value can be seen as the probability of having as many or more participants detect the data plot than the observed result.

The approach used in Majumder, Hofmann, and Cook (2013) is as follows. Define $X_j = \{0, 1\}$ to be a Bernoulli random variable measuring whether participant j detected the data plot, and $X = \sum_{j=1}^K X_j$ be the total number of observers who detected the data plot. Then, by imposing a relatively strong assumption that all K evaluations are fully independent, under H_0 $X \sim \text{Binom}_{K, 1/m}$. Therefore, the p -value of a lineup of size m evaluated by K observer is estimated with $P(X \geq x) = 1 - F(x) + f(x)$, where

CONTACT Weihao Li. Email: weihao.li@monash.edu, Dianne Cook. Email: dicoock@monash.edu, Emi Tanaka. Email: emi.tanaka@anu.edu.au, Susan VanderPlas. Email: susan.vanderplas@unl.edu

$F(\cdot)$ is the binomial cumulative distribution function, $f(\cdot)$ is the binomial probability mass function and x is the realization of number of observers choosing the data plot.

As pointed out by VanderPlas et al. (2021), this basic binomial model is deficient. It does not take into account the possible dependencies in the visual test due to repeated evaluations of the same lineup, or account for when participants are offered the option to select one or more “most different” plots, or none, from a lineup. They suggest three common lineup scenarios: (1) K different lineups are shown to K participants, (2) K lineups with different null plots but the same data plot are shown to K participants, and (3) the same lineup is shown to K participants. Scenario 3 is the most feasible to apply, but has the most dependencies to accommodate for the p -value calculation. For Scenario 3, VanderPlas et al propose modelling the probability of plot i being selected from a lineup as θ_i , where $\theta_i \sim \text{Dirichlet}(\alpha)$ for $i = 1, \dots, m$ and $\alpha > 0$. The number of times plot i being selected in K evaluations is denoted as c_i . In case participant j makes multiple selections, $1/s_j$ will be added to c_i instead of one, where s_j is the number of plots participant j selected for $j = 1, \dots, K$. This ensures $\sum_i c_i = K$. Since we are only interested in the selections of the data plot i , the marginal model can be simplified to a beta-binomial model and thus the visual p -value is given as

$$P(C \geq c_i) = \sum_{x=c_i}^K \binom{K}{x} \frac{B(x + \alpha, K - x + (m - 1)\alpha)}{B(\alpha, (m - 1)\alpha)}, \quad \text{for } c_i \in \mathbb{Z}_0^+ \quad (\text{A1})$$

where $B(\cdot)$ is the beta function defined as

$$B(a, b) = \int_0^1 t^{a-1} (1 - t)^{b-1} dt, \quad \text{where } a, b > 0. \quad (\text{A2})$$

We extend the equation to non-negative real number c_i by applying a linear approximation

$$P(C \geq c_i) = P(C \geq \lceil c_i \rceil) + (\lceil c_i \rceil - c_i)P(C = \lceil c_i \rceil), \quad \text{for } c_i \in \mathbb{R}_0^+, \quad (\text{A3})$$

where $P(C \geq \lceil c_i \rceil)$ is calculated using Equation A1 and $P(C = \lfloor c_i \rfloor)$ is calculated by

$$P(C = c_i) = \binom{K}{c_i} \frac{B(c_i + \alpha, K - c_i + (m - 1)\alpha)}{B(\alpha, (m - 1)\alpha)}, \quad \text{for } c_i \in \mathbb{Z}_0^+. \quad (\text{A4})$$

The parameter α used in Equation A1 and A4 is usually unknown and will need to be estimated from the survey data. An interpretation of α is that when it is low only a few plots are attractive to the observers and tend to be selected, and when high, most plots are equally likely to be chosen. VanderPlas et al define c -interesting plot to be if c or more participants select the plot as the most different. The expected number of plots selected at least c times, $E[Z_c]$, is then calculated as

$$E[Z_c(\alpha)] = \frac{m}{B(\alpha, (m - 1)\alpha)} \sum_{\lceil c \rceil}^K \binom{K}{x} B(x + \alpha, K - x + (m - 1)\alpha). \quad (\text{A5})$$

With Equation A5, α can be estimated using maximum likelihood estimation. Precise estimation of α , is aided by evaluation of Rorschach lineups, where all plots are null plots. In a Rorschach, in theory all plots should be equally likely, but in practice some (irrelevant) visual elements may be more eye-catching than others. This is what α captures, the capacity for extraneous features to distract the observer for a particular type of plot display.

A.2. *Effect size derivation*

Effect size can be defined as the difference of a parameter for a particular model or distribution, or a statistic derived from a sample. Importantly, it needs to reflect the treatment we try to measure. Centred on a conventional statistical test, we usually can deduce the effect size from the test statistic by substituting the null parameter value. When considering the diagnostics of residual departures, there exist many possibilities of test statistics for a variety of model assumptions. Meanwhile, diagnostic plots such as the residual plot have no general agreement on measuring how strong a model violation

pattern is. To build a bridge between various residual-based tests, and the visual test, we focus on the shared information embedded in the testing procedures, which is the distribution of residuals. When comes to comparison of distribution, Kullback-Leibler divergence (Kullback and Leibler 1951) is a classical way to represent the information loss or entropy increase caused by the approximation to the true distribution, which in our case, the inefficiency due to the use of false model assumptions.

Following the terminology introduced by Kullback and Leibler (1951), P represents the measured probability distribution, and Q represents the assumed probability distribution. The Kullback-Leibler divergence is defined as $\int_{-\infty}^{\infty} \log(p(x)/q(x))p(x)dx$, where $p(\cdot)$ and $q(\cdot)$ denote probability densities of P and Q .

Let \mathbf{X} denotes the $p + 1$ predictors with n observations, $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ denotes the OLS solution, $\mathbf{R} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ denotes the residual operator, and let $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ denotes the error. The residual vector

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \mathbf{X}\mathbf{b} \\ &= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} \\ &= \mathbf{R}\mathbf{y} \\ &= \mathbf{R}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \mathbf{R}\boldsymbol{\varepsilon}.\end{aligned}$$

Because $\text{rank}(\mathbf{R}) = n - p - 1 < n$, \mathbf{e} follows a degenerate multivariate normal distribution and does not have a density. Since the Kullback-Leibler divergence requires a proper density function, we need to simplify the covariance matrix of \mathbf{e} by setting all the off-diagonal elements to 0. Then, the residuals will be assumed to follow $N(\mathbf{0}, \text{diag}(\mathbf{R}\sigma^2))$ under the null hypothesis that the model is correctly specified. If the model is however misspecified due to omitted variables \mathbf{Z} , or a non-constant variance \mathbf{V} , the distribution of residuals can be derived as $N(\mathbf{R}\mathbf{Z}\boldsymbol{\beta}_z, \text{diag}(\mathbf{R}\sigma^2))$ and $N(\mathbf{0}, \text{diag}(\mathbf{R}\mathbf{V}\mathbf{R}'))$ respectively.

By assuming both P and Q are multivariate normal density functions, the Kullback-Leibler divergence can be rewritten as

$$KL = \frac{1}{2} \left(\log \frac{|\Sigma_p|}{|\Sigma_q|} - n + \text{tr}(\Sigma_p^{-1}\Sigma_q) + (\mu_p - \mu_q)' \Sigma_p^{-1} (\mu_p - \mu_q) \right).$$

Then, we can combine the two residual departures into one formula

$$KL = \frac{1}{2} \left(\log \frac{|\text{diag}(\mathbf{R}\mathbf{V}\mathbf{R}')|}{|\text{diag}(\mathbf{R}\sigma^2)|} - n + \text{tr}(\text{diag}(\mathbf{R}\mathbf{V}\mathbf{R}')^{-1}\text{diag}(\mathbf{R}\sigma^2)) + \boldsymbol{\mu}'_z(\mathbf{R}\mathbf{V}\mathbf{R}')^{-1}\boldsymbol{\mu}_z \right).$$

When there are omitted variables but constant error variance, the formula can be reduced to

$$KL = \frac{1}{2} (\boldsymbol{\mu}'_z(\text{diag}(\mathbf{R}\sigma^2))^{-1}\boldsymbol{\mu}_z).$$

And when the model equation is correctly specified but the error variance is non-constant, the formula can be reduced to

$$KL = \frac{1}{2} \left(\log \frac{|\text{diag}(\mathbf{R}\mathbf{V}\mathbf{R}')|}{|\text{diag}(\mathbf{R}\sigma^2)|} - n + \text{tr}(\text{diag}(\mathbf{R}\mathbf{V}\mathbf{R}')^{-1}\text{diag}(\mathbf{R}\sigma^2)) \right).$$

To compute the effect size for each lineup we simulate a sufficiently large number of samples from the same model with the number of observations n fixed for each sample. We then compute the effect size for each sample and take the average as the final value. This ensures lineups constructed with the same experimental factors will share the same effect size.

A.3. *Sensitivity analysis for α*

The parameter α used for the p -value calculation needs to be estimated from responses to null lineups. With a greater value of $\hat{\alpha}$, the p -value will be smaller, resulting in more lineups being rejected. However, The way we generate Rorschach lineup is not

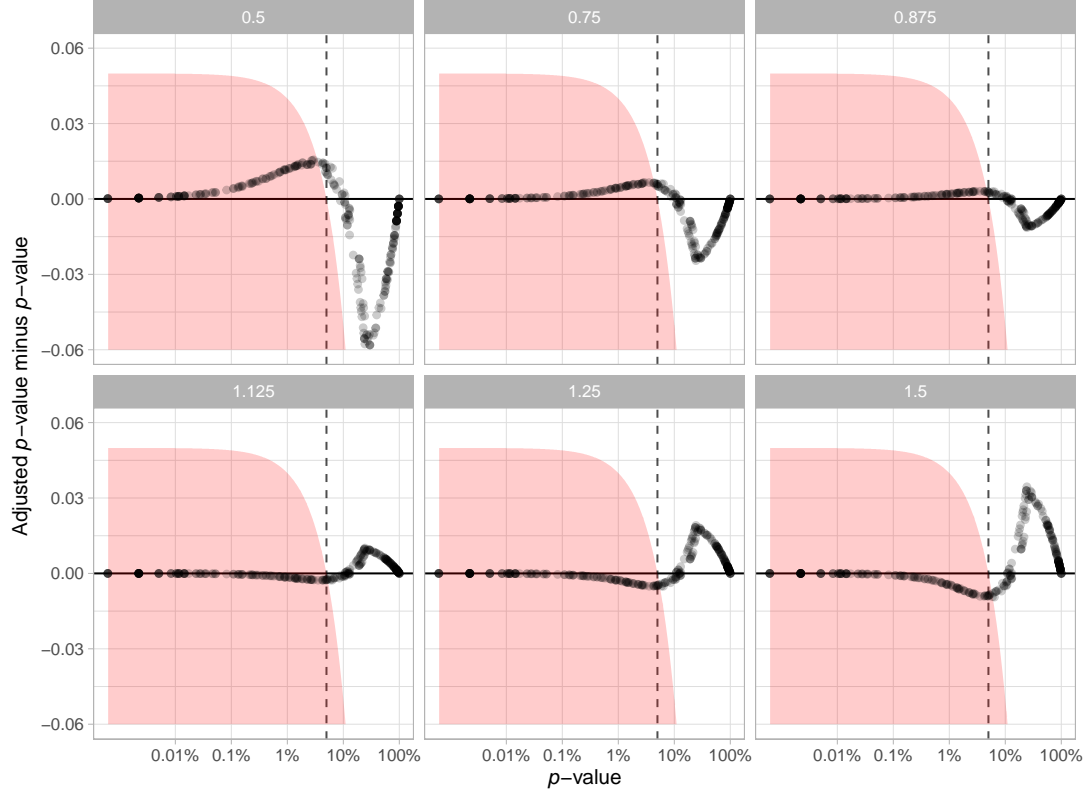


Figure A1. Change of p -values with $\hat{\alpha}$ multiplied by 0.5, 0.75, 0.875, 1.125, 1.25 and 1.5. Only lineups with uniform fitted value distribution is used. The vertical dashed line indicates $p\text{-value} = 0.05$. The area coloured in red indicates adjusted $p\text{-value} < 0.05$. The x-axis is drawn on logarithmic scale. For multipliers smaller than 1, the adjusted p -value will initially increase and decline when the p -value increases. The trend is opposite with multipliers greater than 1, but the difference eventually reaches 0.

strictly the same as what suggested in VanderPlas et al. (2021) and Buja et al. (2009). Therefore, we conduct a sensitivity analysis in this section to examine the impact of the variation of the estimator α on our primary findings.

The analysis is conducted by setting up several scenarios, where the α is under or overestimated by 12.5%, 25% and 50%. Using the adjusted $\hat{\alpha}$, we recalculate the p -value for every lineup and show the results in Figure A1. It can be observed that there are some changes to p -values, especially when the $\hat{\alpha}$ is multiplied by 50%. However, Table A1 shows that adjusting $\hat{\alpha}$ will not result in a huge difference in rejection decisions. There are only a small percentage of cases where the rejection decision change. It is very unlikely the downstream findings will be affected because of the estimate of α .

Table A1. Examining how decisions might change if $\hat{\alpha}$ was different. Percentage of lineups that where the p -value would switch to above or below 0.05, when $\hat{\alpha}$ is multiplied by a multiplier.

| Multiplier | Reject to not reject | % | Not reject to reject | % |
|------------|----------------------|------|----------------------|------|
| 0.500 | 7 | 2.51 | 0 | 0.00 |
| 0.750 | 4 | 1.43 | 0 | 0.00 |
| 0.875 | 3 | 1.08 | 0 | 0.00 |
| 1.125 | 0 | 0.00 | 3 | 1.08 |
| 1.250 | 0 | 0.00 | 4 | 1.43 |
| 1.500 | 0 | 0.00 | 5 | 1.79 |

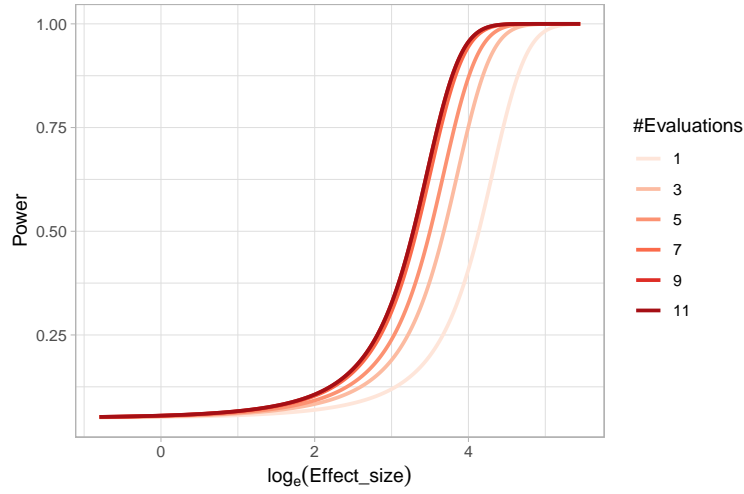


Figure A2. Change of power of visual tests for different number of evaluations on lineups with uniform fitted value distribution. The power will increase as the number of evaluations increases, but the margin will decrease.

A.4. *Effect of number of evaluations on the power of a visual test*

When comparing power of visual tests across different fitted value distributions, we have discussed the number of evaluations on a lineup will affect the power of the visual test. Using the lineups with uniform fitted value distribution, we show in Figure A2 the change of power of visual tests due to different number of evaluations. It can be learned that as the number of evaluations increases, the power will increase but the margin will decrease. Considering we have eleven evaluations on lineups with uniform fitted value distribution, and five evaluations on other lineups, it is necessary to use the same number of evaluations for each lineup in comparison.

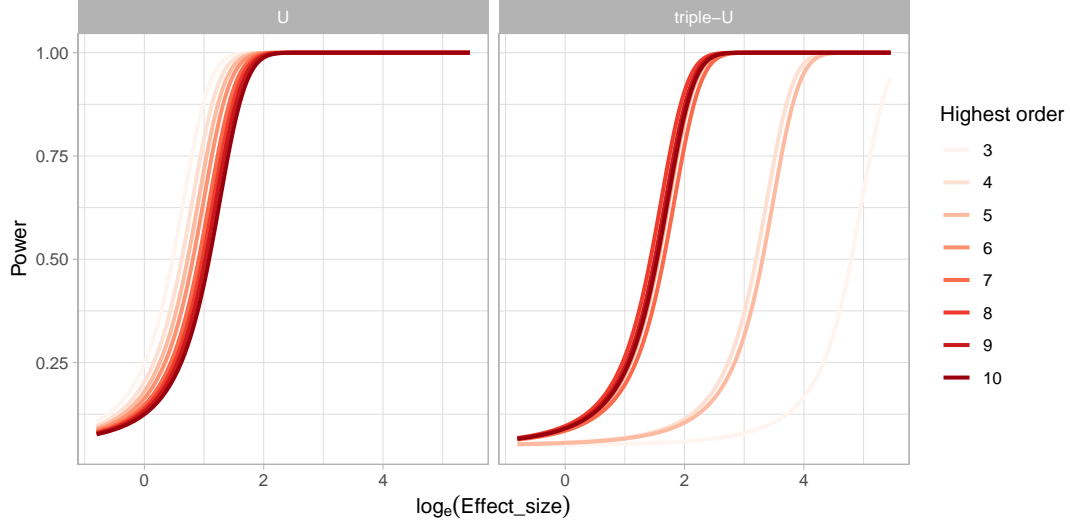


Figure A3. Change of power of RESET tests for different orders of fitted values included in the auxiliary formula. The left panel is the power of testing the "U" shape and the right panel is the power of testing the "triple-U" shape. The power will not be greatly affected by the highest order in the case of testing the "U" shape. In the case of testing the "triple-U" shape, the highest order needs to be set to at least six to avoid the loss of power.

A.5. *Power of a RESET test under different auxiliary regression formulas*

It is found in the result that the power of a RESET test will be affected by the highest order of fitted values included in the auxiliary formula. And we suspect that the current recommendation of the highest order - four, is insufficient to test complex non-linear structures such as the "triple-U" shape designed in this paper. Figure A3 illustrates the change of power of RESET test while testing the "U" shape and the "triple-U" shape with different highest orders. Clearly, when testing a simple shape like the "U" shape, the highest order has very little impact on the power. But for testing the "triple-U" shape, there will be a loss of power if the recommended order is used. To avoid the loss of power, the highest order needs to be set to at least six.

A.6. *Conventional test rejection rate for varying significance levels*

In the main paper, Sections 5.1 and 5.2 compared the power, and the decisions made by the conventional tests and the visual test. The power curves for the visual test is effectively a right-shift from the conventional test. The effect is that the visual test rejects less often than the conventional test, at the same significance level. We also saw

that the visual test rejected a subset of those that the conventional tests rejected. This means that they agreed quite well - only residual plots rejected by the conventional tests were rejected by the visual test. There was little disagreement, where residual plots not rejected by the conventional test were rejected by the visual test. The question arises whether the decisions made conventional test could be made similar to that of the visual test by reducing the significance level. Reducing the significance level from 0.05, to 0.01, 0.001, ... will have the effect of rejecting fewer of the residual plots.

It would be interesting if a different conventional test significance level results in both the visual tests and conventional tests reject only the same residual plots, and fails to reject the same residual plots. This would be a state where both systems agree perfectly. Figure A4 examines this. Plot A shows the percentage of residual plots rejected by the visual test, given the conventional test rejected (solid lines) or failed to reject (dashed lines). The vertical grey line marks significance level 0.05. When the significance level gets smaller, it is possible to see that the visual tests reject (nearly) 100% of the time that the conventional test rejects. However, there is no agreement, because the visual tests also increasingly reject residual plots where the conventional test failed to reject. Plot B is comparable to an ROC curve, where the percentage visual test rejection conditional on conventional test decision is plotted: Reject conditional on reject is plotted against reject conditional on fail to reject, for different significance levels. The non-linearity pattern results are close to being ideal, that the percentage of reject relative to fail to reject increases very slowly as the reject relative to reject converges to 100. The heteroskedasticity pattern is more problematic, and shows that the cost of rejecting less with the conventional test is disagreement with the visual test.

Appendix B. Additional details of experiment setup

B.1. *Non-linearity Model*

The non-linearity model used in the experiment includes a non-linear term \mathbf{z} constructed using Hermite polynomials on random vector \mathbf{x} formulated as

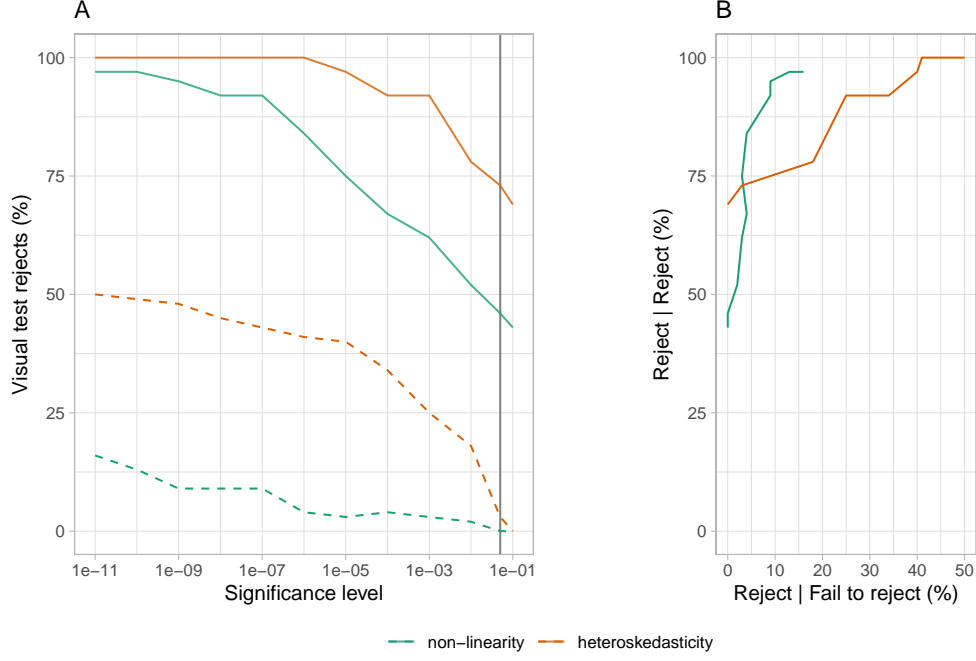


Figure A4. Changing the significance level of the conventional test will change the rejection rate. A: Percentage of conventional tests also rejected by the visual test: rejected (solid), not rejected (dashed). As the significance level is reduced, the percentage rejected by the visual test that has been rejected by the conventional test approaches 100. The percentage of tests not rejected by the conventional test also increases, as expected, but the percentage of these that are rejected by the visual test increases too. B: ROC curve shows that forcing the conventional test to not reject creates a discrepancy with the visual test. Many of the residual plots not rejected by the conventional test are rejected by the visual test. It is not possible to vary the significance level of the conventional test to match the decisions made by the visual test.

$$\mathbf{y} = \mathbf{1}_n + \mathbf{x} + \mathbf{z} + \boldsymbol{\varepsilon},$$

$$\mathbf{x} = g(\mathbf{x}_{raw}, 1),$$

$$\mathbf{z} = g(\mathbf{z}_{raw}, 1),$$

$$\mathbf{z}_{raw} = He_j(g(\mathbf{x}, 2)),$$

where \mathbf{y} , \mathbf{x} , $\boldsymbol{\varepsilon}$, \mathbf{x}_{raw} , \mathbf{z}_{raw} are vectors of size n , $\mathbf{1}_n$ is a vector of ones of size n , $He_j(\cdot)$ is the j th-order probabilist's Hermite polynomials (Hermite 1864; originally by Laplace 1820), $\boldsymbol{\varepsilon} \sim N(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$, and $g(\mathbf{x}, k)$ is a scaling function to enforce the support of the random vector to be $[-k, k]^n$ defined as

$$g(\mathbf{x}, k) = 2k \cdot \frac{\mathbf{x} - x_{\min} \mathbf{1}_n}{x_{\max} - x_{\min}} - k, \quad \text{for } k > 0. \quad (\text{B1})$$

where $x_{\min} = \min_{i \in \{1, \dots, n\}} x_i$, $x_{\max} = \max_{i \in \{1, \dots, n\}} x_i$ and x_i is the i -th entry of \mathbf{x} . The function `hermite` from the R package `mpoly` (Kahle 2013) is used to simulate \mathbf{z}_{raw} to generate Hermite polynomials.

The null regression model used to fit the realizations generated by the above model is formulated as

$$\mathbf{y} = \beta_0 \mathbf{1}_n + \beta_1 \mathbf{x} + \mathbf{u}, \quad (\text{B2})$$

where $\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. Here \mathbf{z} is a higher order term of \mathbf{x} and leaving it out in the null regression results in model misspecification.

B.2. *Heteroskedasticity Model*

The heteroskedasticity model used in the experiment is formulated as

$$\begin{aligned} \mathbf{y} &= \mathbf{1}_n + \mathbf{x} + \boldsymbol{\varepsilon}, \\ \mathbf{x} &= g(\mathbf{x}_{raw}, 1), \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}_n, \mathbf{1}_n + (2 - |a|)(\mathbf{x} - a\mathbf{1}_n)'(\mathbf{x} - a\mathbf{1}_n)b\mathbf{I}_n), \end{aligned}$$

where \mathbf{y} , \mathbf{x} , $\boldsymbol{\varepsilon}$ are vectors of size n and $g(\cdot)$ is the scaling function defined in Equation B1. The null regression model used to fit the realizations generated by the above model is formulated exactly the same as Equation B2.

For $b \neq 0$, the variance-covariance matrix of the error term $\boldsymbol{\varepsilon}$ is correlated with the predictor \mathbf{x} , which will lead to the presence of heteroskedasticity.

Since $\text{supp}(X) = [-1, 1]$, choosing a to be -1 , 0 and 1 can generate “left-triangle”,

“butterfly” and “right-triangle” shapes. The term $(2 - |a|)$ maintains the magnitude of residuals across different values of a .

B.3. *Controlling the signal strength*

The three parameters n , σ and b are important for controlling the strength of the signal to generate lineups with a variety difficulty levels. This will ensure that estimated power curves will be smooth and continuous, and that participants are allocated a set of lineups with a range of difficulty.

Parameter $\sigma \in \{0.5, 1, 2, 4\}$ and $b \in \{0.25, 1, 4, 16, 64\}$ are used in data collection periods I and II respectively. A large value of σ will increase the variation of the error of the non-linearity model and decrease the visibility of the visual pattern. The parameter b controls the variation in the standard deviation of the error across the support of the predictor. A larger value of b generates a larger ratio between it is smallest and highest values, making the visual pattern more obvious. The sample size n sharpens (or blurs) a pattern when it is larger (or smaller).

B.4. *Lineup allocation to participants*

There are a total of $4 \times 4 \times 3 \times 4 = 192$ and $3 \times 5 \times 3 \times 4 = 180$ combinations of parameter values for the non-linearity model and heteroskedasticity models respectively. Three replications for each combination results in $192 \times 3 = 576$ and $180 \times 3 = 540$ lineups, respectively.

Each lineup needs to be evaluated by at least five participants. From previous work, and additional pilot studies for this experiment, we decided to record evaluations from 20 lineups for each participant. Two of the 20 lineups with clear visual patterns were used as attention checks, to ensure quality data. Thus, $576 \times 5 / (20 - 2) = 160$ and $540 \times 5 / (20 - 2) = 150$ participants were needed to cover the experimental design for the data collection periods I and II, respectively. The factor levels and range of difficulty was assigned relatively equally among participants.

Data collection period III was primarily to obtain Rorschach lineup evaluations to estimate α , and also to obtain additional evaluations of lineups made with uniform

fitted value distribution to ensure consistency in the results. To construct a Rorschach lineup, the data is generated from a model with zero effect size, while the null data are generated using the residual rotation technique. This procedure differs from that of the canonical Rorschach lineup, where all 20 plots are generated directly from the null model, so the method suggested in VanderPlas et al. (2021) for typical lineups containing a data plot is used to estimate α . The $3 \times 4 = 12$ treatment levels of the common factors, replicated three times results in 36 lineups. And 6 more evaluations on the 279 lineups with uniform fitted value distribution, results in at least $(36 \times 20 + 279 \times 3 \times 6)/(20 - 2) = 133$ participants needed.

B.5. Mapping of participants to experimental factors

Mapping of participants to experimental factors is an important part of experiment design. Essentially, we want to maximum the difference in factors exposed to a participant. For this purpose, we design an algorithm to conduct participant allocation. Let L be a set of available lineups and S be a set of available participants. According to the experimental design, the availability of a lineup is associated with the number of participants it can assign to. For lineups with uniform fitted value distribution, this value is 11. And other lineups can be allocated to at most five different participants. The availability of a participant is associated with the number of lineups that being allocated to this participant. A participant can view at most 18 different lineups.

The algorithm starts from picking a random participant $s \in S$ with the minimum number of allocated lineups. It then tries to find a lineup $l \in L$ that can maximise the distance metric D and allocate it to participant s . Set L and S will be updated and the picking process will be repeated until there is no available lineups or participants.

Let F_1, \dots, F_q be q experimental factors, and f_1, \dots, f_q be the corresponding factor values. We say f_i exists in L_s if any lineup in L_s has this factor value. Similarly, $f_i f_j$ exists in L_s if any lineup in L_s has this pair of factor values. And $f_i f_j f_k$ exists in L_s if any lineup in L_s has this trio of factor values. The distance metric D is defined between a lineup l and a set of lineups L_s allocated to a participant s if L_s is non-empty:

$$D = C - \sum_{1 \leq i \leq q} I(f_i \text{ exists in } L_s) - \sum_{\substack{1 \leq i \leq q-1 \\ i < j \leq q}} I(f_i f_j \text{ exists in } L_s) - \sum_{\substack{1 \leq i \leq q-2 \\ i < j \leq q-1 \\ j < k \leq q}} I(f_i f_j f_k \text{ exists in } L_s)$$

where C is a sufficiently large constant such that $D > 0$. If L_s is empty, we define $D = 0$.

The distance measures how different a lineup is from the set of lineups allocated to the participant in terms of factor values. Thus, the algorithm will try to allocate the most different lineup to a participant at each step.

In Figure B1 and Figure B2, we present examples of lineup allocations generated by our algorithm for data collection periods I and II, respectively. These figures illustrate the factor values presented to the first 20 participants recruited during each period. It can be observed that the algorithm effectively distributed almost all possible one-way and two-way factor combinations among the participants, thereby ensuring a diverse set of lineups for each participant.

B.6. Data collection process

The survey data is collected via a self-hosted website designed by us. The complete architecture is provided in Figure B3. The website is built with the **Flask** (Grinberg 2018) web framework and hosted on **PythonAnywhere** (PythonAnywhere LLP 2023). It is configured to handle HTTP requests such that participants can correctly receive webpages and submit responses. Embedded in the resources sent to participants, the **jsPsych** front-end framework (De Leeuw 2015) instructs participants' browsers to render an environment for running behavioural experiments. During the experiment, this framework will automatically collect common behavioural data such as response time and clicks on buttons. participants' responses are first validated by a scheduled **Python** script run on the server, then push to a Github repository. Lineup images shown to users are saved in multiple Github repositories and hosted in corresponding Github pages. The URLs to these images are resolved by **Flask** and bundled in HTML files.

Once the participant is recruited from Prolific (Palan and Schitter 2018), it will

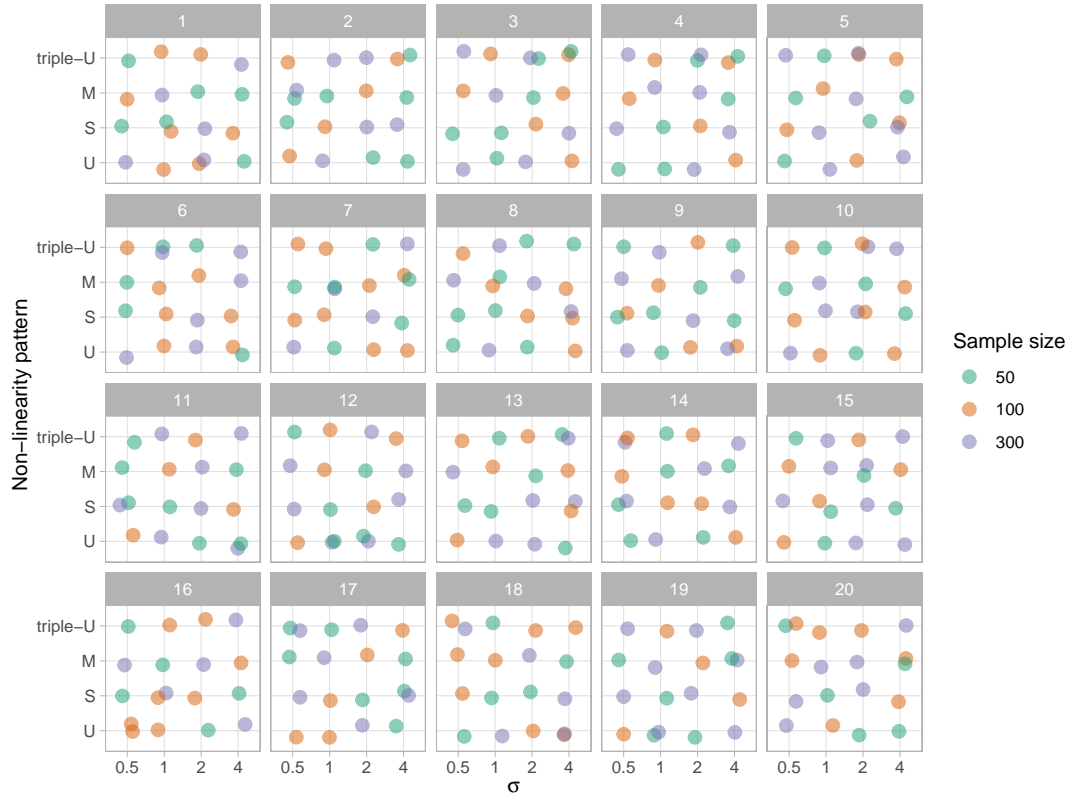


Figure B1. Factor values assigned to the first 20 participants recruited during data collection period I, with data points slightly jittered to prevent overlap. Each participant have been exposed to all one-way and two-way factor combinations.

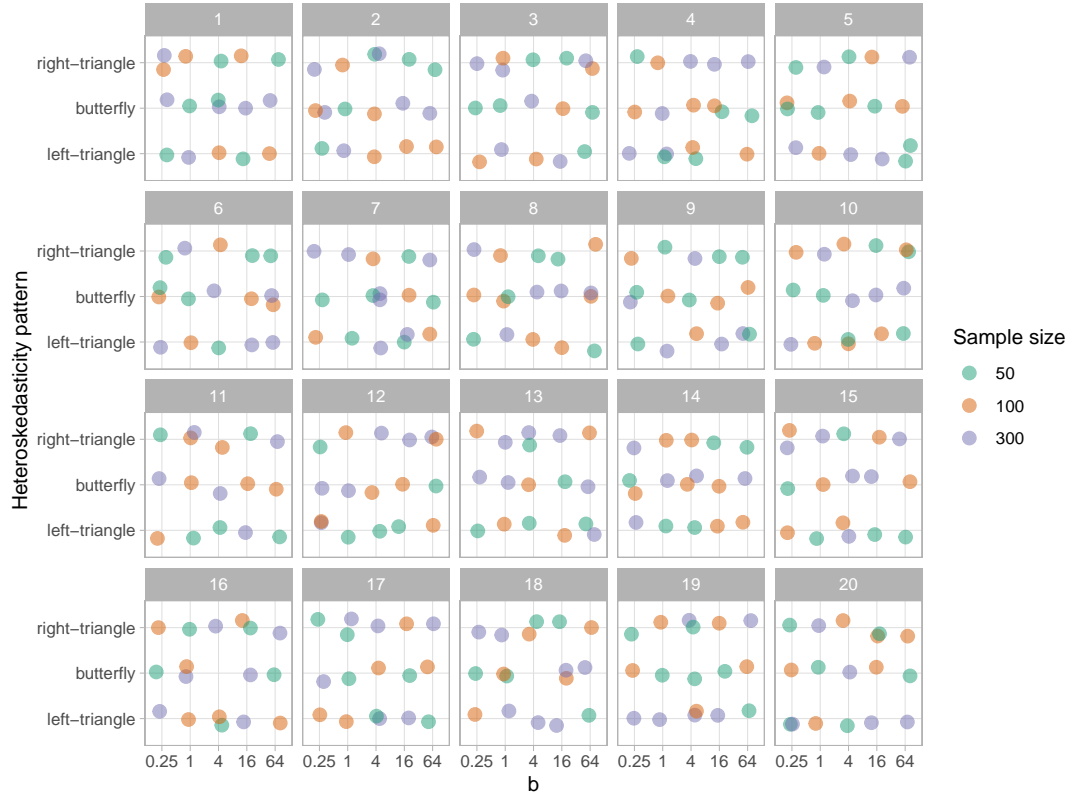


Figure B2. Factor values assigned to the first 20 participants recruited during data collection period II, with data points slightly jittered to prevent overlap. Each participant appears to have been exposed to almost all one-way and two-way factor combinations. However, two participants, namely participant 16 and participant 18, were not presented with lineups containing a butterfly shape with $b = 4$. Additionally, participant 7 did not encounter a lineup featuring a butterfly shape with $b = 1$.

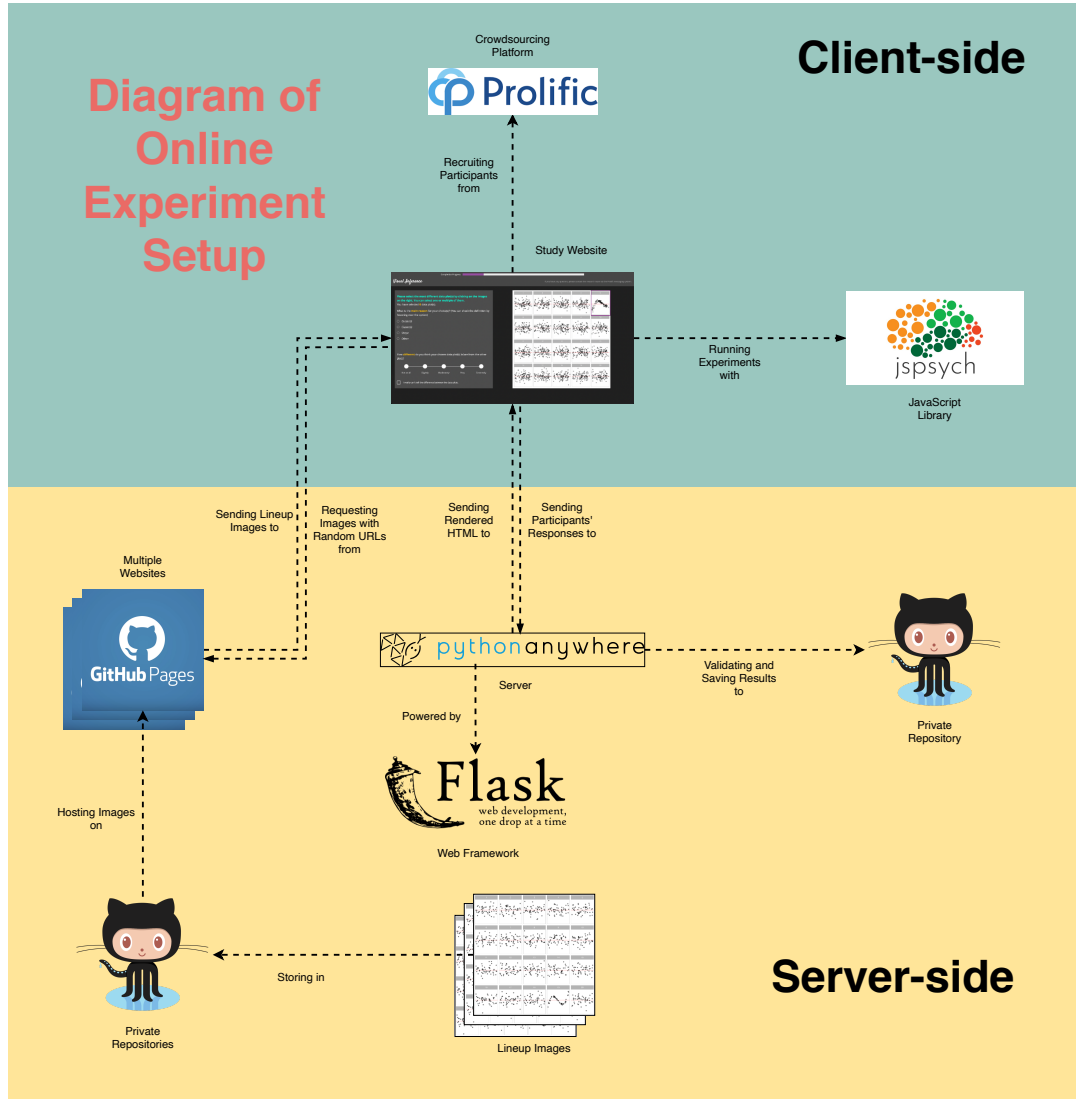


Figure B3. Diagram of online experiment setup. The server-side of the study website uses Flask as backend hosted on PythonAnywhere. And the client-side uses jsPsych to run experiment.

be redirected to the entry page of our study website. An image of the entry page is provided in Figure B4. Then, the participant needs to submit the online consent form and fill in the demographic information as shown in B5 and B6 respectively. Before evaluating lineups, participant also need to read the training page as provide in Figure B7 to understand the process. An example of the lineup page is given in Figure B8. A half of the page is taken by the lineup image to attract participant's attention. The button to skip the selections for the current lineup is intentionally put in the corner of the bounding box with smaller font size, such that participants will not misuse this functionality.

Visual Inference

Completion Progress

Please do not refresh the page. If you have any question, please contact the research team via the Prolific messaging system.

Please read the **Explanatory Statement**. To continue, click the checkbox below and hit "Continue".

This study will take you **10-15 minutes**. You must complete the study in **1 hour**.

Please do not refresh the page during the study.

EXPLANATORY STATEMENT
(Prolific users)

Project ID: 29199
Project title: Advances in Artificial Intelligence for Data Visualization: Developing Computer Vision Models to Automate Reading of Data Plots, with Application to Predictive Model Diagnostics

| | |
|---|---|
| Chief Investigator's name: Emi Tanaka Department of Econometrics and Business Statistics, Monash University email: emi.tanaka@monash.edu | Student's name: Weihao Li Department of Econometrics and Business Statistics, Monash University email: Patrick.li@monash.edu |
| Other Investigator's name: Dianne Cook Department of Econometrics and Business Statistics, Monash University email: dicook@monash.edu | |

You are invited to take part in this study. Please read this Explanatory Statement in full before deciding whether or not to participate in this research. If you would like further information regarding any aspect of this project, you are encouraged to contact the researchers via the email addresses listed above.

☐ I have read and understood the Explanatory Statement.

Figure B4. The entry page of the study website.

Visual Inference

Completion Progress

Please do not refresh the page. If you have any question, please contact the research team via the Prolific messaging system.

To continue, click the checkbox below and hit "Continue".

Consent Form

Project ID: 29199
Project title: Advances in Artificial Intelligence for Data Visualization: Developing Computer Vision Models to Automate Reading of Data Plots, with Application to Predictive Model Diagnostics

Chief Investigator: Emi Tanaka
Other Investigator's name: Dianne Cook
Student's name: Weihao Li

I consent to the following:

☒ The data that I provide during this research may be used by the research team or other researchers in future research projects.

☒ I have been asked to take part in the Monash University research project specified above. I have read and understood the **Explanatory Statement** and I hereby consent to participate in this project.

Continue

Figure B5. The consent form provided in the study website.

Visual Inference Completion Progress Please do not refresh the page. If you have any question, please contact the research team via the Prolific messaging system.

Survey Questions

Please enter your Prolific ID:

Prolific ID

Please select your age category:

18-24 25-39 40-54 55-64 65 or above

Please select your highest level of education:

High school or below Diploma and Bachelor Degree Honours Degree Masters Degree Doctoral Degree

Please select your preferred pronoun:

☐ He
☐ She
☐ They
☐ Other

Have you participated in any research that requires reading data graphs?

☐ Yes
☐ No

Figure B6. The form to provide demographic information.

Visual Inference Completion Progress Please do not refresh the page. If you have any question, please contact the research team via the Prolific messaging system.

Please read the **Training Page**. To continue, click the checkbox below and hit "Continue".

Training Page (3 min read)

This document will provide you with the essential knowledge to finish the study.

1 Webpage layout






When you start the study, you will see a webpage like this:

Visual Inference Completion Progress If you have any question, please contact the research team via the Prolific messaging system.

Please select the most different data plot(s) by clicking on the images on the right. You can select one or multiple of them. You have selected 0 data plot(s).

What is the **main reason** for your choice(s)? (You can check the definition by hovering over the option)

☐ Outlier(s)

☐ I have read and understood the Training Page.

Figure B7. The training page of the study website.

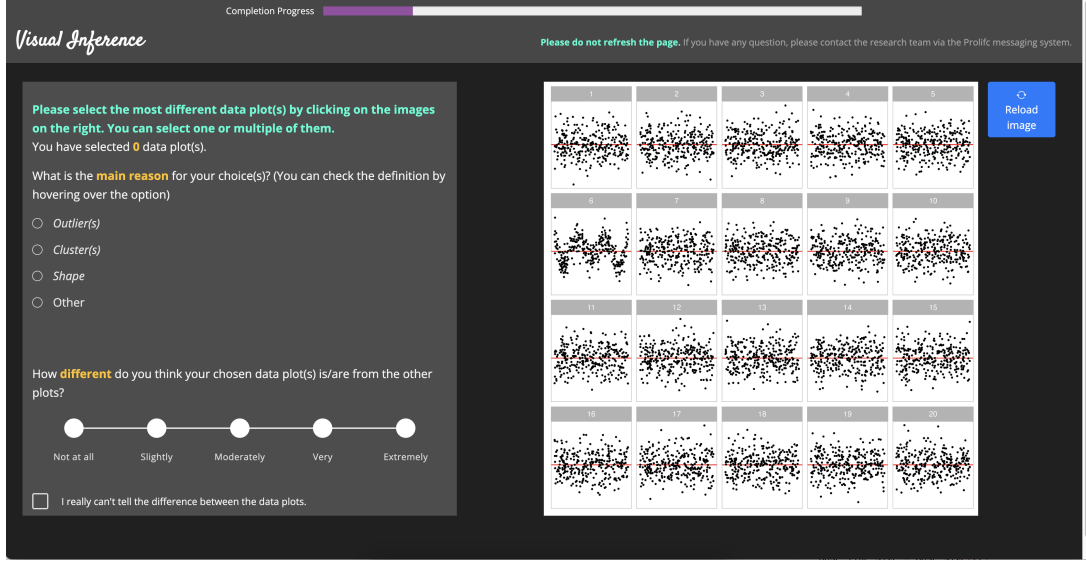


Figure B8. The lineup page of the study website.

Table C1. Count of lineups, evaluations and participants over departure types and data collection periods.

| Number | Non-linearity | | | Heteroskedasticity | | | Total |
|--------------|---------------|----|-----|--------------------|------|-----|-------|
| | I | II | III | I | II | II | |
| Lineups | 576 | 0 | 144 | 0 | 540 | 135 | 1116 |
| Evaluations | 2880 | 0 | 864 | 0 | 2700 | 810 | 7254 |
| Participants | 160 | 0 | 123 | 0 | 160 | 123 | 443 |

Appendix C. Analysis of results relative to data collection process

C.1. Demographics

Throughout the study, we have collected 7254 evaluations on 1116 non-null lineups. Table C1 gives further details about the number of evaluations, lineups and participants over pattern types and data collection periods.

Along with the responses to lineups, we have collected a series of demographic information including age, pronoun, education background and previous experience in studies involved data visualization. Table C2, C3, C4 and C5 provide summary of the demographic data.

It can be observed from the tables that most participants have Diploma or Bachelor degrees, followed by High school or below and the survey data is gender balanced. Majority of participants are between 18 to 39 years old and there are slightly more

Table C2. Summary of pronoun distribution of participants recruited in this study.

| Pronoun | Period I | % | Period II | % | Period III | % | Total | % |
|---------|----------|------|-----------|------|------------|------|-------|-------|
| He | 77 | 17.4 | 79 | 17.8 | 61 | 13.8 | 217 | 49.0 |
| She | 78 | 17.6 | 77 | 17.4 | 61 | 13.8 | 216 | 48.8 |
| Other | 5 | 1.1 | 4 | 0.9 | 1 | 0.2 | 10 | 2.3 |
| | 160 | 36.1 | 160 | 36.1 | 123 | 27.8 | 443 | 100.0 |

Table C3. Summary of age distribution of participants recruited in this study.

| Age group | Period I | % | Period II | % | Period III | % | Total | % |
|-----------|----------|------|-----------|------|------------|------|-------|-------|
| 18-24 | 83 | 18.7 | 86 | 19.4 | 51 | 11.5 | 220 | 49.7 |
| 25-39 | 69 | 15.6 | 63 | 14.2 | 63 | 14.2 | 195 | 44.0 |
| 40-54 | 6 | 1.4 | 8 | 1.8 | 6 | 1.4 | 20 | 4.5 |
| 55-64 | 2 | 0.5 | 3 | 0.7 | 3 | 0.7 | 8 | 1.8 |
| | 160 | 36.1 | 160 | 36.1 | 123 | 27.8 | 443 | 100.0 |

participants who do not have previous experience than those who have.

C.2. Data collection periods

We have the same type of model collected over different data collection periods, that may lead to unexpected batch effect. Figure C1 and C2 provide two lineups to examine whether there is an actual difference across data collection periods for non-linearity model and heteroskedasticity model respectively. To emphasize the tail behaviour and display fewer outliers, we use the “letter-value” boxplot (Hofmann, Wickham, and Kafadar 2017) which is an extension of the number of “letter value” statistics to check the weighed proportion of detect over different data collection period. The weighted

Table C4. Summary of education distribution of participants recruited in this study.

| Education | Period I | % | Period II | % | Period III | % | Total | % |
|-----------------------------|----------|------|-----------|------|------------|------|-------|-------|
| High School or below | 41 | 9.3 | 53 | 12.0 | 33 | 7.4 | 127 | 28.7 |
| Diploma and Bachelor Degree | 92 | 20.8 | 79 | 17.8 | 66 | 14.9 | 237 | 53.5 |
| Honours Degree | 6 | 1.4 | 15 | 3.4 | 6 | 1.4 | 27 | 6.1 |
| Masters Degree | 21 | 4.7 | 13 | 2.9 | 16 | 3.6 | 50 | 11.3 |
| Doctoral Degree | 0 | 0.0 | 0 | 0.0 | 2 | 0.5 | 2 | 0.5 |
| | 160 | 36.1 | 160 | 36.1 | 123 | 27.8 | 443 | 100.0 |

Table C5. Summary of previous experience distribution of participants recruited in this study.

| Previous experience | Period I | % | Period II | % | Period III | % | Total | % |
|---------------------|----------|------|-----------|------|------------|------|-------|-------|
| No | 96 | 21.7 | 88 | 19.9 | 67 | 15.1 | 251 | 56.7 |
| Yes | 64 | 14.4 | 72 | 16.3 | 56 | 12.6 | 192 | 43.3 |
| | 160 | 36.1 | 160 | 36.1 | 123 | 27.8 | 443 | 100.0 |

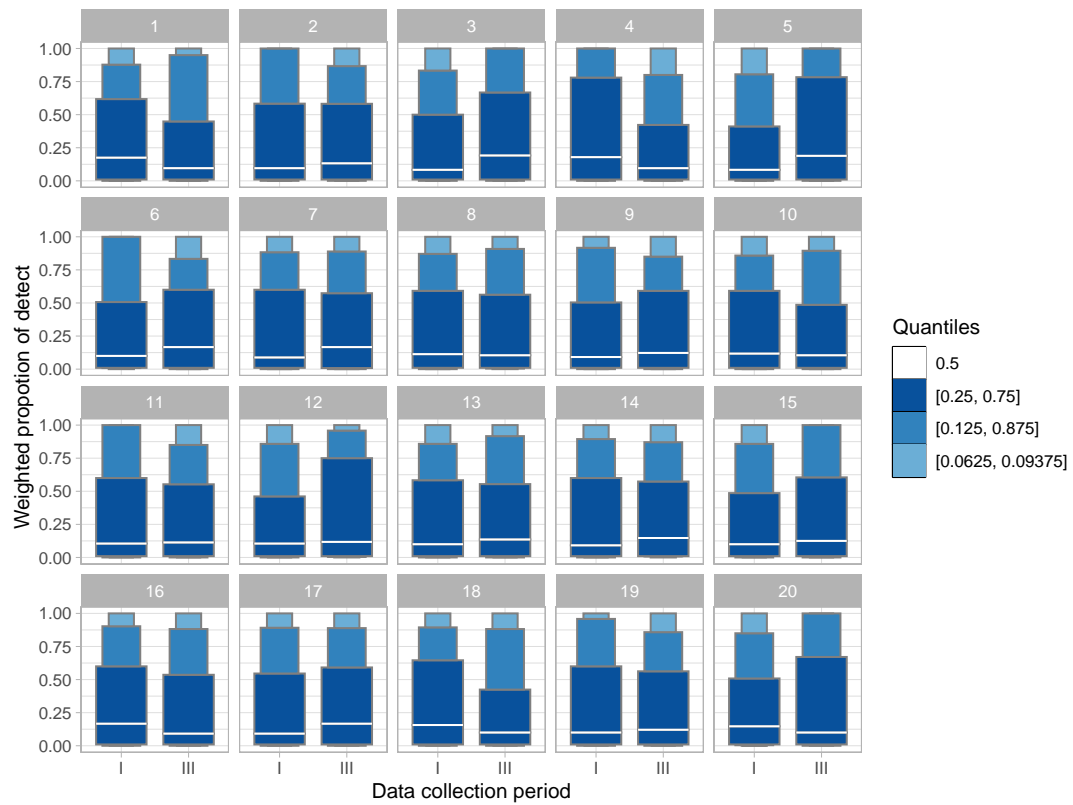


Figure C1. A lineup of "letter-value" boxplots of weighted proportion of detect for lineups over different data collection periods for non-linearity model. Can you find the most different boxplot? The data plot is positioned in panel $2^3 - 1$.

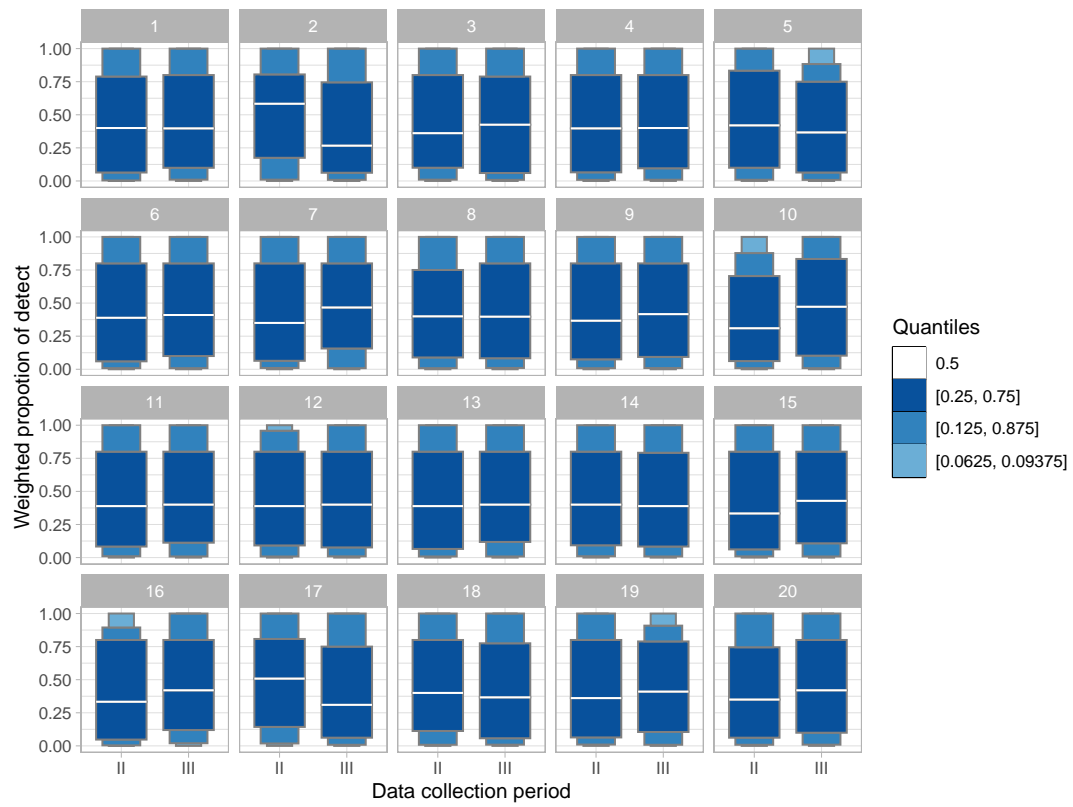


Figure C2. A lineup of "letter-value" boxplots of weighted propotion of detect for lineups over different data collection periods for heteroskedasticity model. Can you find the most different boxplot? The data plot is positioned in panel $2^4 - 2$.

proportion of detect is calculated by taking the average of c_i of a lineup over a data collection period. Within our research team, we can not identify the data plot from the null plots for these two lineups, result in p -values much greater than 5%. Thus, there is no clear evidence of batch effect.

References

- Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F Swayne, and Hadley Wickham. 2009. “Statistical inference for exploratory data analysis and model diagnostics.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367 (1906): 4361–4383.
- De Leeuw, Joshua R. 2015. “jsPsych: A JavaScript library for creating behavioral experiments in a web browser.” *Behavior Research Methods* 47: 1–12.
- Grinberg, Miguel. 2018. *Flask web development: Developing web applications with Python*. ” O’Reilly Media, Inc.”.
- Hermite, M. 1864. *Sur un nouveau développement en série des fonctions*. Imprimerie de Gauthier-Villars.
- Hofmann, Heike, Hadley Wickham, and Karen Kafadar. 2017. “value plots: Boxplots for large data.” *Journal of Computational and Graphical Statistics* 26 (3): 469–477.
- Kahle, David. 2013. “mpoly: Multivariate Polynomials in R.” *The R Journal* 5 (1): 162–170.
- Kullback, Solomon, and Richard A Leibler. 1951. “On information and sufficiency.” *The Annals of Mathematical Statistics* 22 (1): 79–86.
- Laplace, Pierre-Simon. 1820. *Théorie analytique des probabilités*. Vol. 7. Courcier.
- Majumder, Mahbubul, Heike Hofmann, and Dianne Cook. 2013. “Validation of visual statistical inference, applied to linear models.” *Journal of the American Statistical Association* 108 (503): 942–956.
- Palan, Stefan, and Christian Schitter. 2018. “Prolific. ac—A subject pool for online experiments.” *Journal of Behavioral and Experimental Finance* 17: 22–27.
- PythonAnywhere LLP. 2023. “PythonAnywhere.” <https://www.pythonanywhere.com>.
- VanderPlas, Susan, Christian Röttger, Dianne Cook, and Heike Hofmann. 2021. “Statistical significance calculations for scenarios in visual inference.” *Stat* 10 (1): e337.