Supplmentary material - A plot is worth a thousand tests: assessing residual diagnostics with visual inference

Weihao Li, Dianne Cook, Emi Tanaka and Susan VanderPlas

A Appendix

A.1 Experiment setup

A.1.1 Mapping of subjects to experimental factors

Mapping of subjects to experimental factors is an important part of experiment design. Essentially, we want to maximum the difference in factors exposed to a subject. For this purpose, we design an algorithm to conduct subject allocation. Let L be a set of available lineups and S be a set of available subjects. According to the experimental design, the availability of a lineup is associated with the number of subjects 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 subjects. The availability of a subject is associated with the number of lineups that being allocated to this subject. A subject can view at most 18 different lineups.

The algorithm starts from picking a random subject $s \in S$ 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 subject s. Set L and S will be updated and the picking process will be repeated until there is no available lineups or subjects.

Let $F_1, ..., F_q$ be q experimental factors, and $f_1, ..., f_q$ 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 subject s if L_s is non-empty:

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

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 subject in terms of factor values. Thus, the algorithm will try to allocate the most different lineup to a subject at each step.

A.1.2 Data collection process

The survey data is collected via a self-hosted website designed by us. The complete architecture is provided in Figure 1. 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 subjects can correctly receive webpages and submit responses. Embedded in the resources sent to subjects, the `jsPsych` front-end framework (De Leeuw 2015) instructs subjects' browsers to render an environment for running behavioral experiments. During the experiment, this framework will automatically collect common behavioral data such as response time and clicks on buttons. Subjects' 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.

Table 1: Summary of pronoun distribution of subjects recuritted 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 2: Summary of age distribution of subjects recuritted in this study.

Age group	Period I	Period II	Period III	Total
18-24	83	86	51	220
25-39	69	63	63	195
40-54	6	8	6	20
55-64	2	3	3	8

Once the participant is recruited from Prolific (Palan and Schitter 2018), it will be redirected to the entry page of our study website. An image of the entry page is provided in Figure 2. Then, the participant needs to submit the online consent form and fill in the demographic information as shown in **??** and 4 respectively. Before evaluating lineups, participant also need to read the training page as provide in Figure 5 to understand the process. An example of the lineup page is given in Figure 6. 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.

A.2 Demographics

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 1, 2, 3 and 4 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 participants who do not have previous experience than those who have.

A.3 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

Table 3: Summary of education distribution of subjects recuritted in this study.

Education	Period I	Period II	Period III	Total
High School or below	41 (9.3%)	53 (12%)	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%)	2 (0.5%)	2 (0.5%)
	160 (36.1%)	160 (36.1%)	123 (27.8%)	443 (100%)

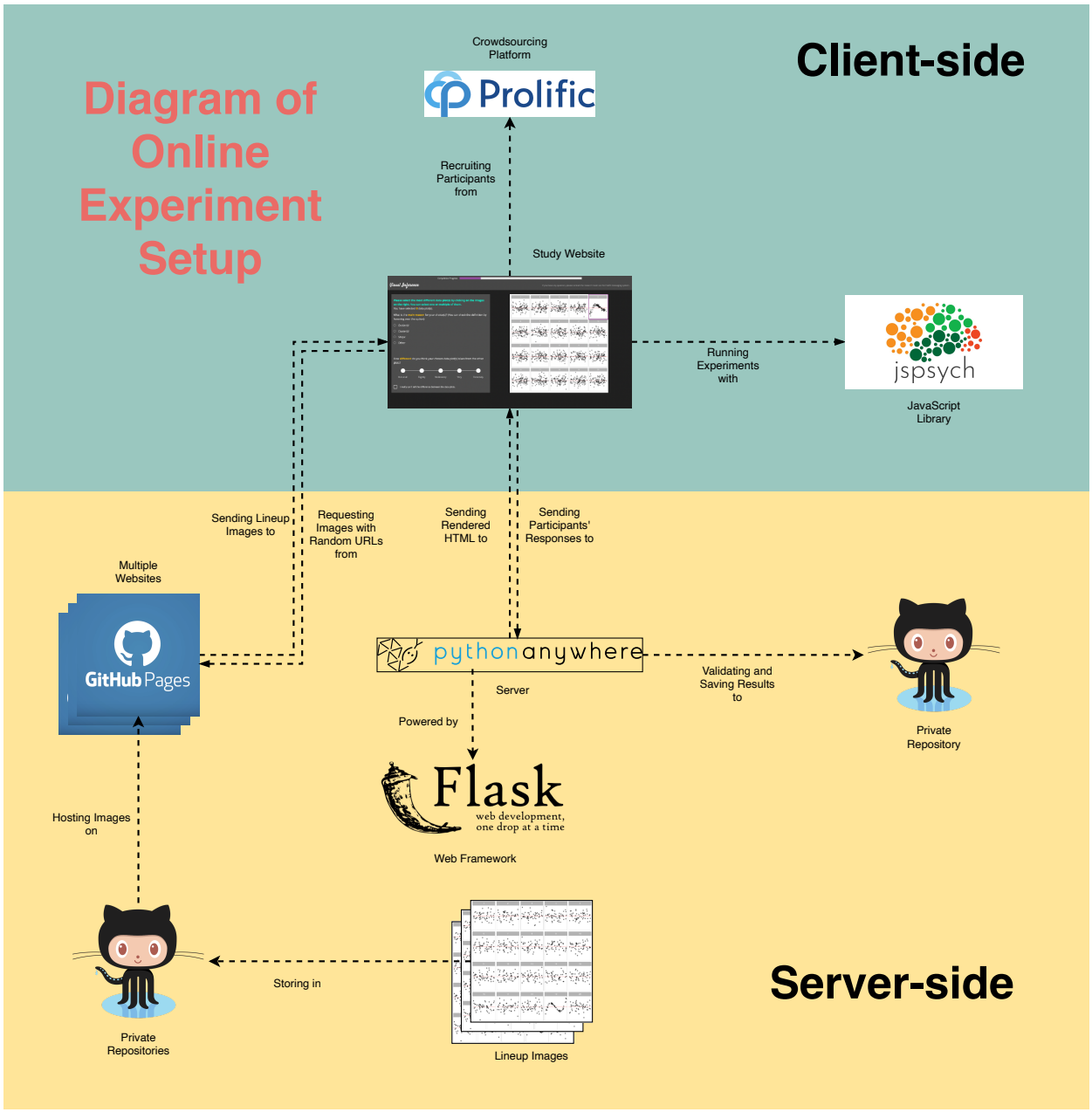

Figure 1: Diagram of online experiment setup. The server-side of the study website uses Flask as backend hosted on PythonAnywhere. And the client-side uses jsPsych to run experiment.

Table 4: Summary of previous experience distribution of subjects recuritted in this study.

Previous experience	Period I	Period II	Period III	Total
No	96	88	67	251
Yes	64	72	56	192

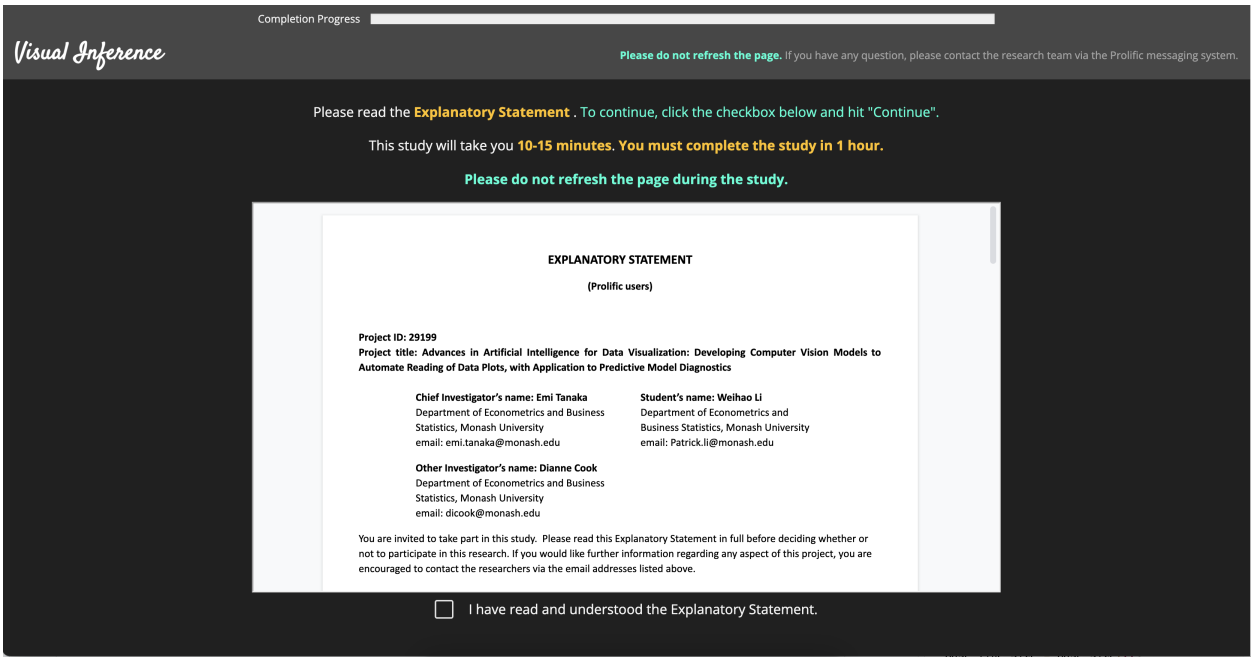

Figure 2: The entry page of the study website.

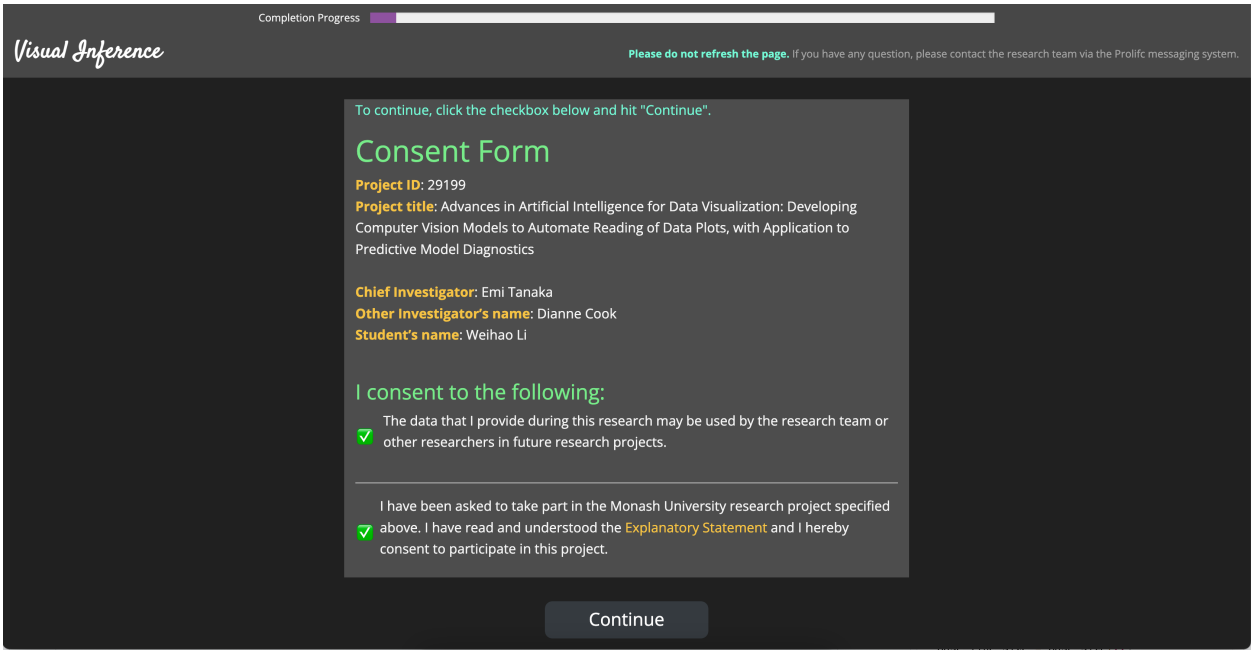

Figure 3: The consent form provided in the study website.

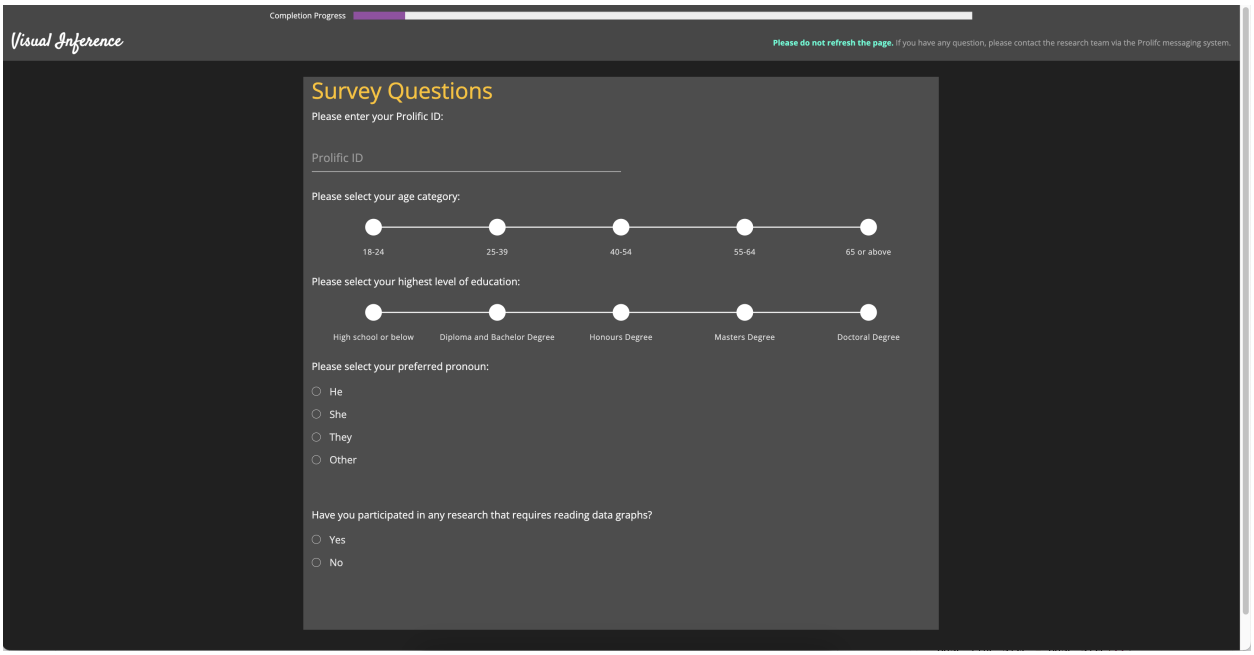

Figure 4: The form to provide demographic information.

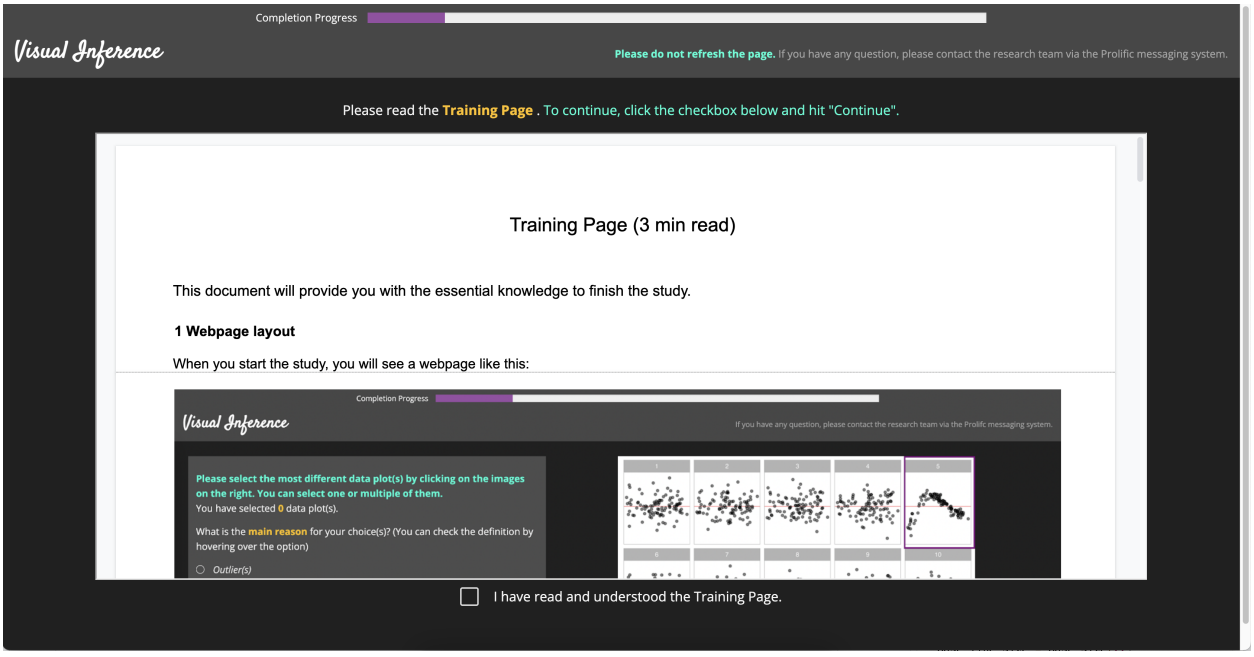

Figure 5: The training page of the study website.

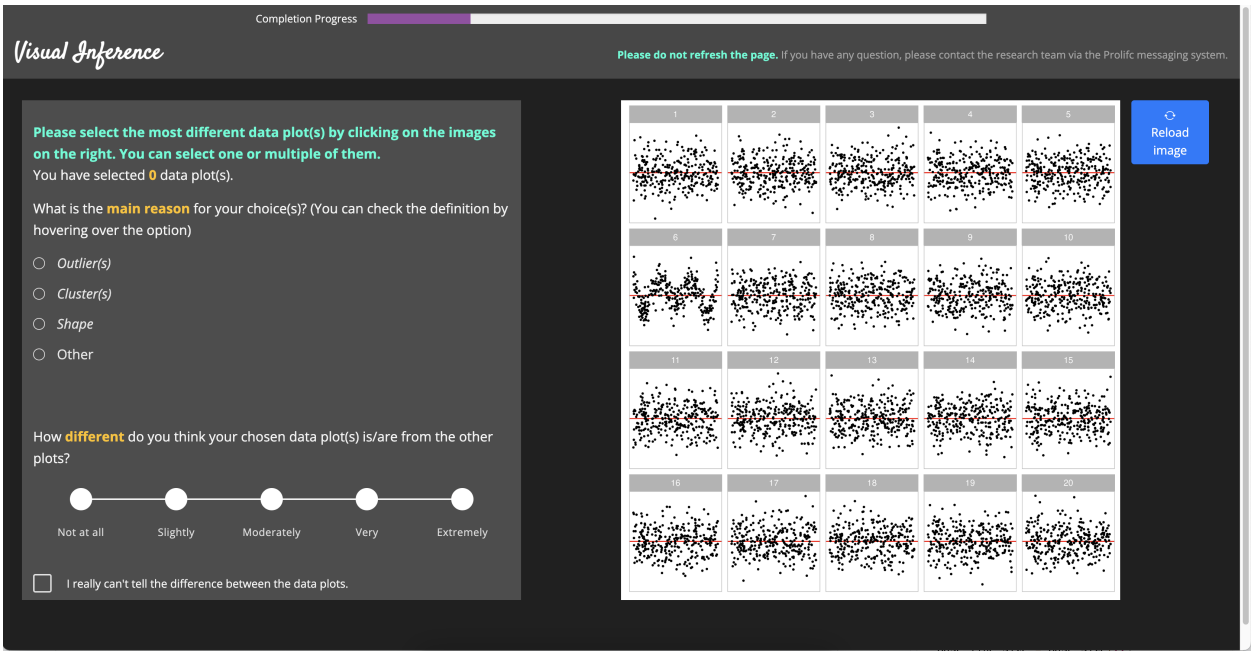

Figure 6: The lineup page of the study website.

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 $p(.)$ and $q(.)$ denote probability densities of P and Q.

Let $\boldsymbol{X}_a = (\mathbf{1}, \boldsymbol{X})$ denotes the p regressors with n observations, $\boldsymbol{R}_a = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ denotes the residual operator, and let $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$ denotes the error. Using the Frisch–Waugh–Lovell theorem, residuals $\boldsymbol{e} = \boldsymbol{R}_a\boldsymbol{\varepsilon}$. Because $rank(\boldsymbol{R}_a) = n - p < n$, 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 $\boldsymbol{e}$ by setting all the off-diagonal elements to 0. Then, the residuals will assumed to follow $N(\mathbf{0}, diag(\boldsymbol{R}_a\sigma^2))$ under the null hypothesis that the model is correctly specified. If the model is however misspecified due to omitted variables $\boldsymbol{Z}$, or a non-constant variance $\boldsymbol{V}$, the distribution of residuals can be derived as $N(\boldsymbol{R}_a\boldsymbol{Z}\boldsymbol{\beta}_z, diag(\boldsymbol{R}_a\sigma^2))$ and $N(\mathbf{0}, diag(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a'))$ respectively.

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

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

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

$$KL = \frac{1}{2}\left(log\frac{|diag(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a')|}{|diag(\boldsymbol{R}_a\sigma^2)|} - n + tr(diag(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a')^{-1}diag(\boldsymbol{R}_a\sigma^2)) + \boldsymbol{\mu}_z^T(\boldsymbol{R}_a\boldsymbol{V}\boldsymbol{R}_a')^{-1}\boldsymbol{\mu}_z\right). \quad (1)$$

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

$$KL = \frac{1}{2}\left(\boldsymbol{\mu}_z^T (diag(\boldsymbol{R}_a\sigma^2))^{-1}\boldsymbol{\mu}_z\right).$$

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

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

Since we assume $\sigma = 1$ for the heteroskedasticity model, the final form of the formula is

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

A.4 Effect of data collection period

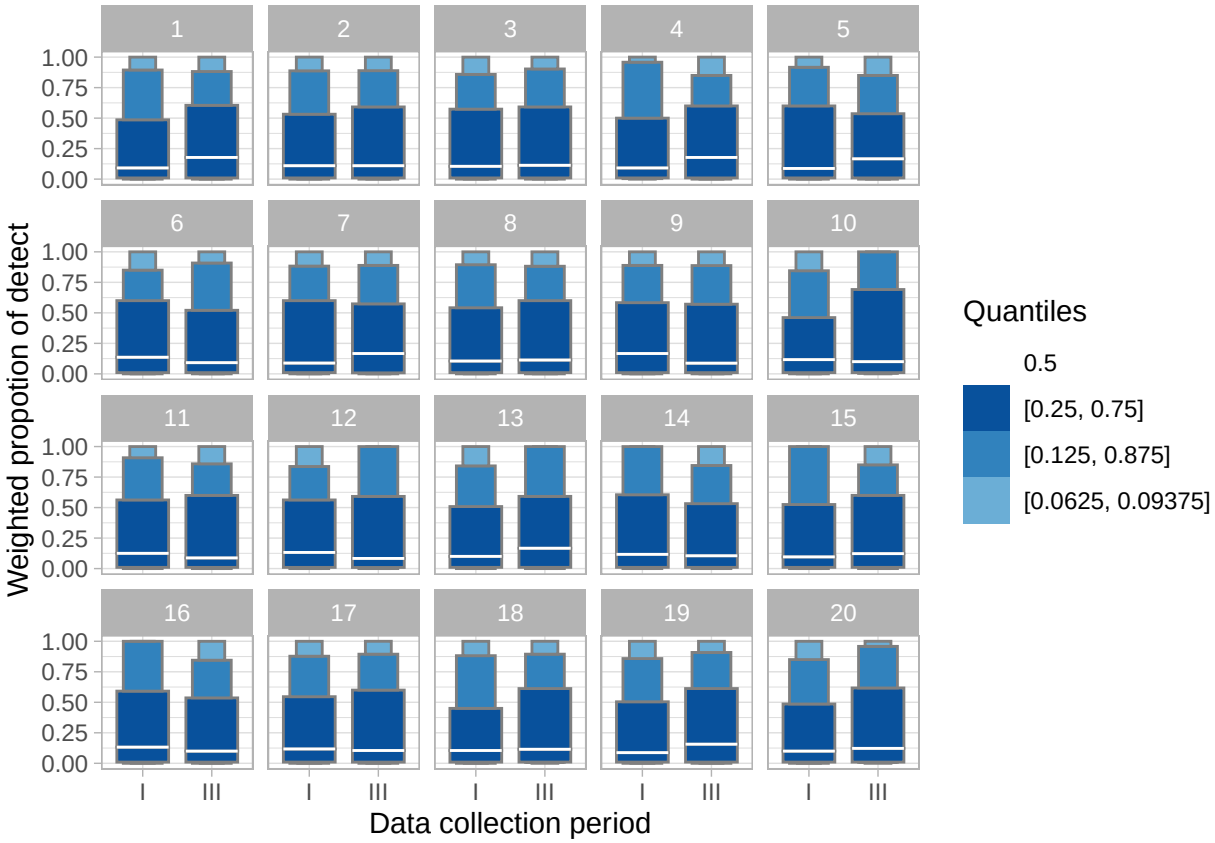

Figure 7: A lineup of "letter-value" boxplots of weighted propotion of detect for lineups over different data collection periods for non-linearity model. Can you find the most different boxplot? The data plot is positioned in panel $2^3 - 1$.

We have the same type of model collected over different data collection periods, that may lead to unexpected batch effect. Figure 7 and 8 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

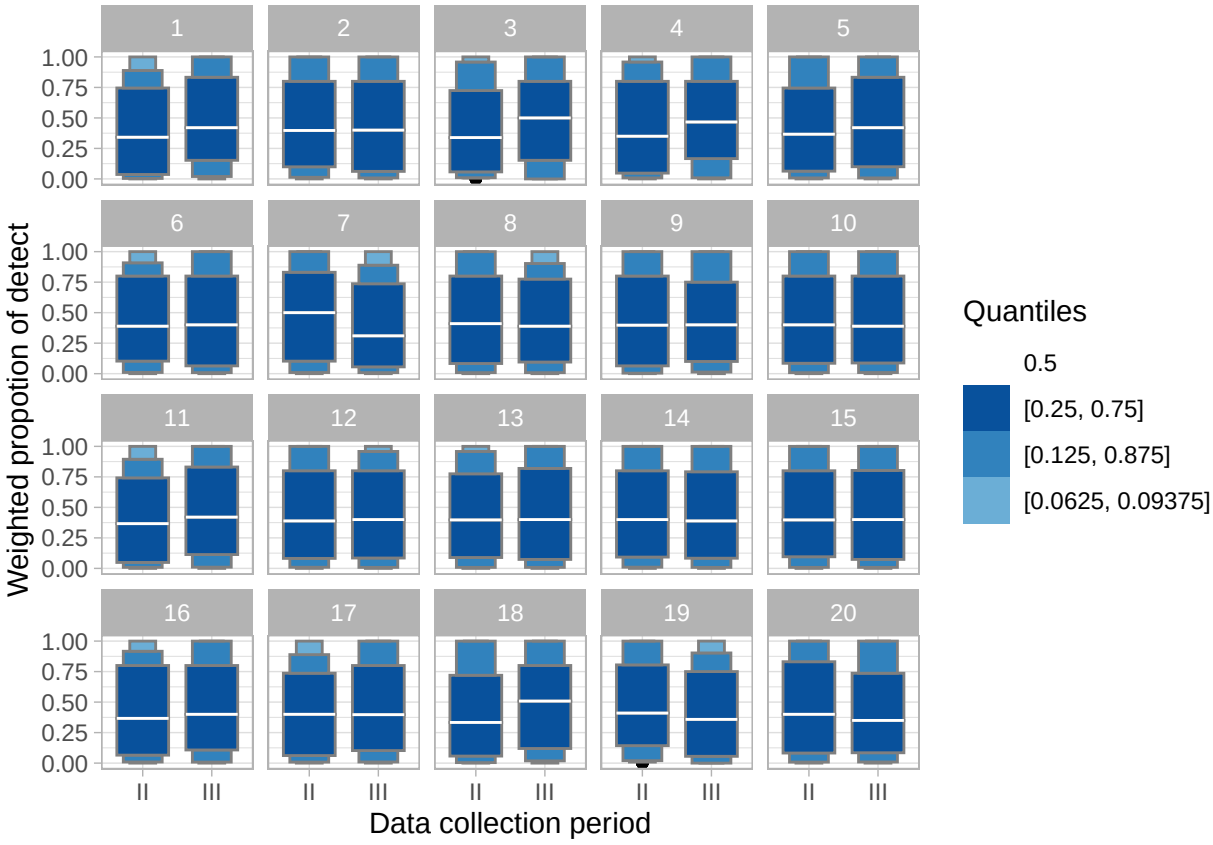

Figure 8: A lineup of "letter-value" boxplots of weighted propotion of detect for lineups over different data collection periods for heteroskedasticity model. Can you find the most different boxplot? The data plot is positioned in panel $2^4 - 2$.

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 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.

A.5 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 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 9. It can be observed that there are some changes to p-values, especially when the $\hat{\alpha}$ is multiplied by 50%. However, Table 5 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.

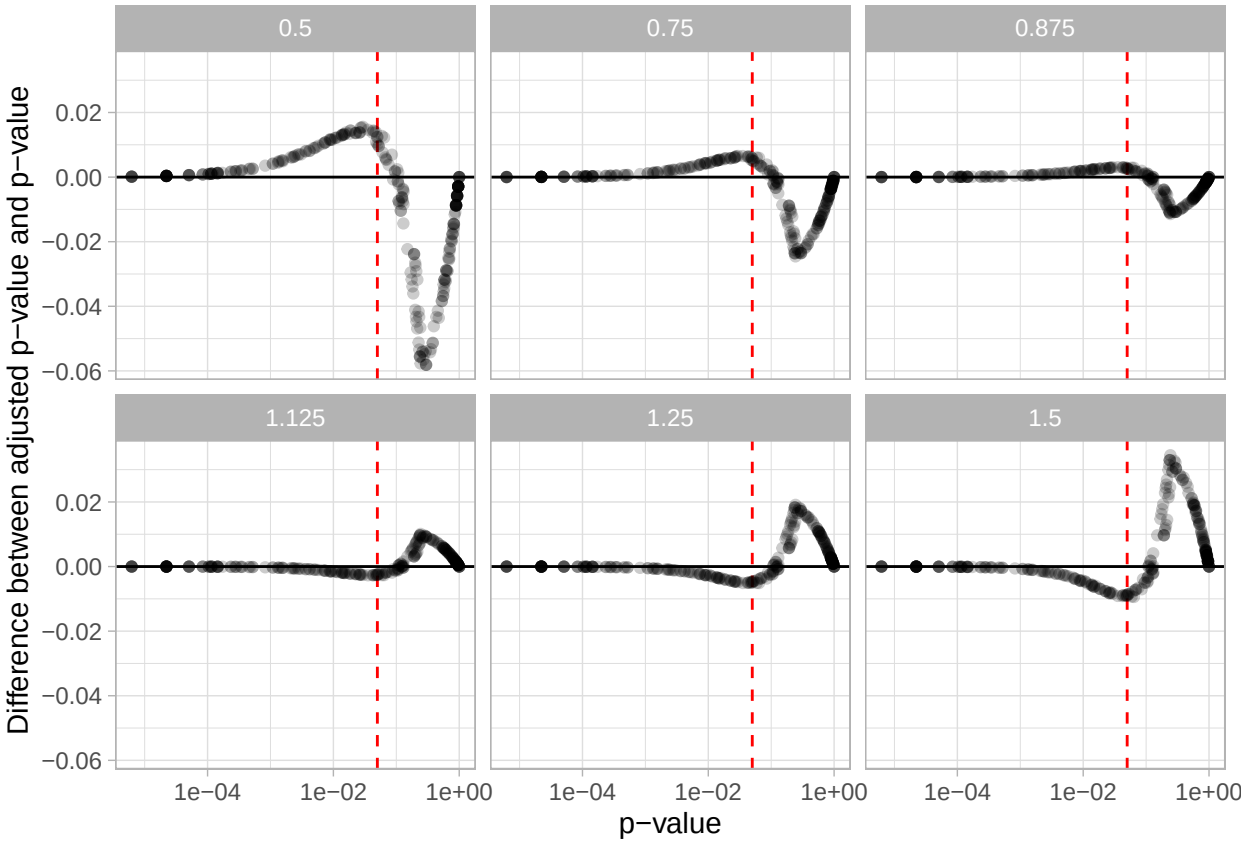

Figure 9: Change of p-values with $\hat{\alpha}$ multiplied by 0.5, 0.75, 0.875, 1.125, 1.25 and 1.5. The vertical dashed line is to indicate p-value = 0.05. The x-axis is drawn on logarithmic scale. For multiplier smaller than 1, the adjusted p-value will incrase then decrase as p-value increases. The trend is the opposite for multiplier greater than 1, but the difference will eventually reach 0.

Table 5: Change of rejection decision because of the modification of $\hat{\alpha}$.

multiplier	Proportion of lineups transforms to "not reject"	Proportion of lineups transforms to "reject"
0.500	2.51%	0%
0.750	1.43%	0%
0.875	1.08%	0%
1.125	0%	1.08%
1.250	0%	1.43%
1.500	0%	1.79%

A.6 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 **??** 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.

A.7 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 **??** 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.

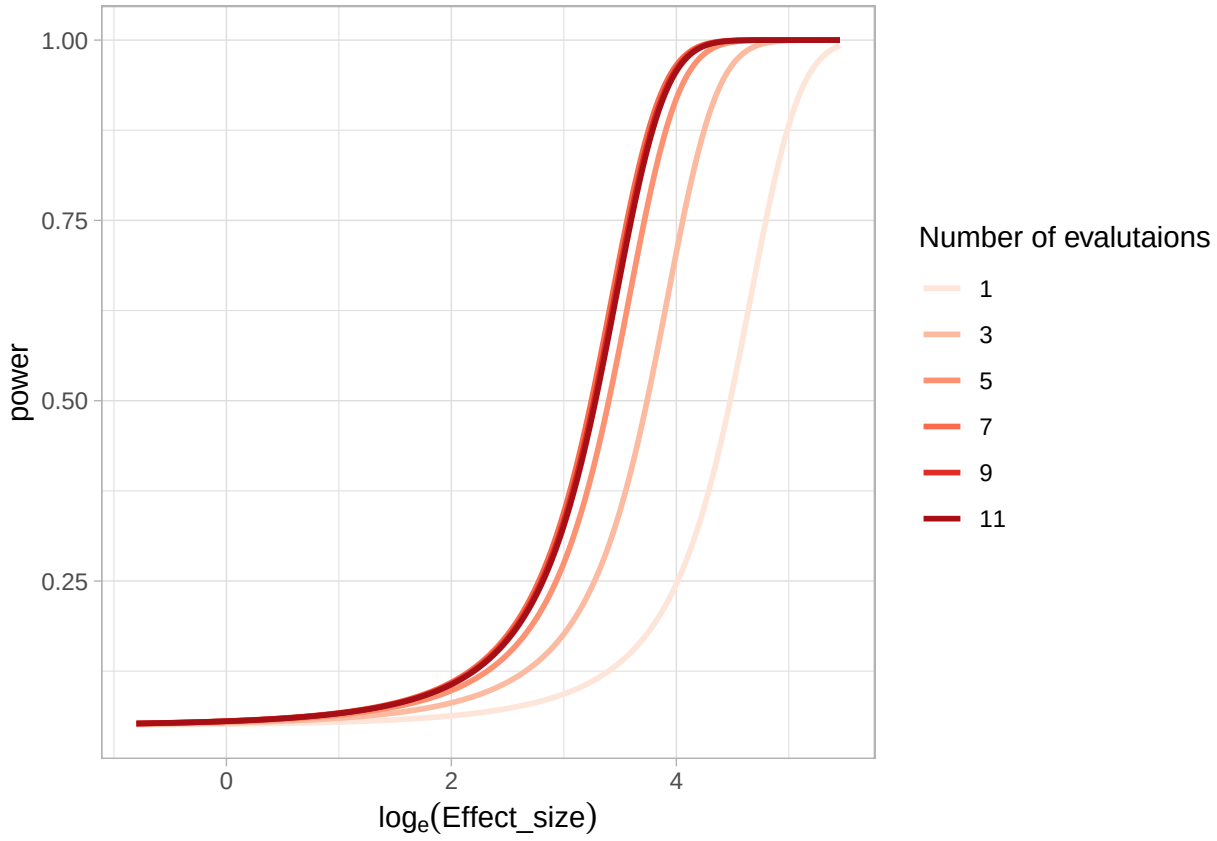

Figure 10: Change of power of visual tests for different number of evalutions on lineups with uniform fitted value distribution. The power will incrase as the number of evaluations increases, but the margin will decrase.

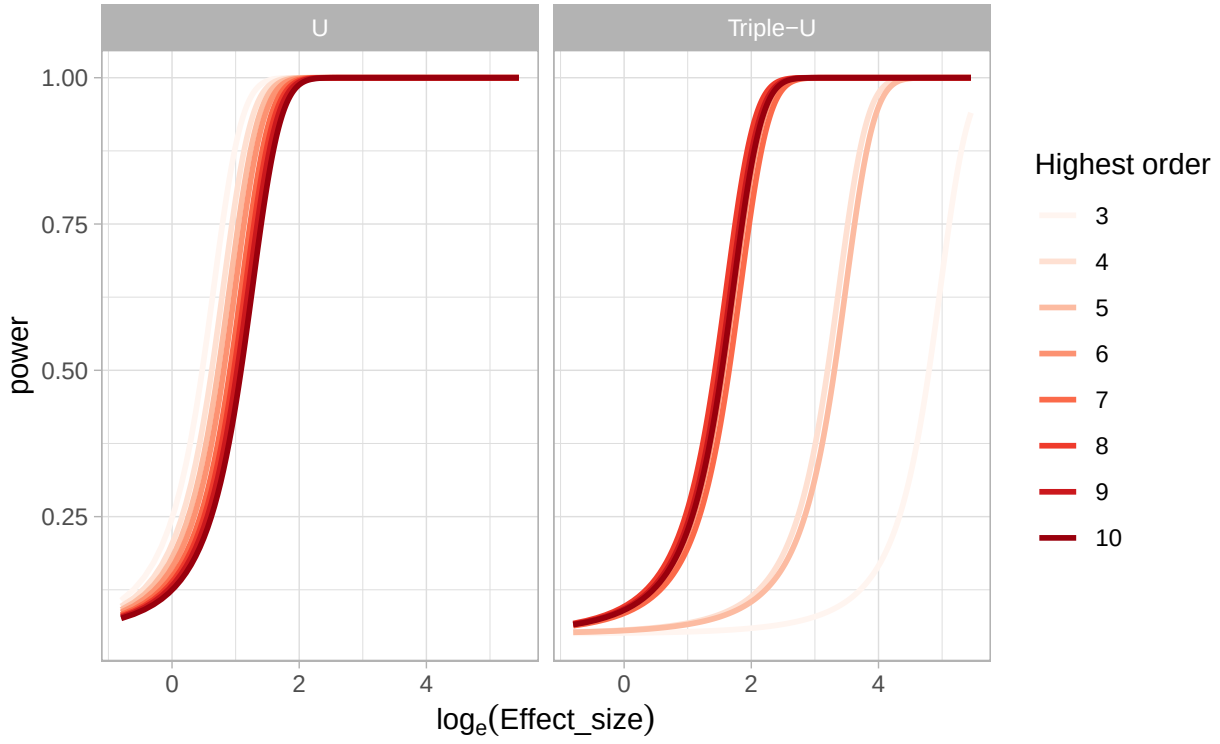

Figure 11: Change of power of RESET tests for different orders of fitted values included in the auxiliary formula. The left panel is the power of testing the "U" shape and the right panel is the power of testing the "Triple-U" shape. The power will not be greatly affected by the highest order in the 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.

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): 4361–83. https://doi.org/10.1098/rsta.2009.0120.

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.".

Hofmann, Heike, Hadley Wickham, and Karen Kafadar. 2017. "Value Plots: Boxplots for Large Data." *Journal of Computational and Graphical Statistics* 26 (3): 469–77.

Kullback, Solomon, and Richard A Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22 (1): 79–86.

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.