Eye Fitting Straight Lines in the Modern Era

Emily A. Robinson 1
Department of Statistics, University of Nebraska - Lincoln
and
Susan VanderPlas 2
Department of Statistics, University of Nebraska - Lincoln
and
Reka Howard 3
Department of Statistics, University of Nebraska - Lincoln

February 2, 2022

Abstract

How do statistical regression results compare to intuitive, visually fitted results? Fitting lines by eye through a set of points has been explored since the 20th century. Common methods of fitting trends by eye involve maneuvering a string, black thread, or ruler until the fit is suitable, then drawing the line through the set of points. In 2015, the New York Times introduced an interactive feature, called 'You Draw It', where readers are asked to input their own assumptions about various metrics and compare how these assumptions relate to reality. In this paper, we validate 'You Draw It' as a method for graphical testing, comparing results to the less technological method utilized in Mosteller et al. (1981) and extending that study with formal statistical analysis methods. Results were consistent with those found in the previous study; when shown points following a linear trend, participants tended to fit the slope of the first principal component over the slope of the least-squares regression line. This trend was most prominent when shown data simulated with larger variances. This study reinforces the differences between intuitive visual model fitting and statistical model fitting, providing information about human perception as it relates to the use of statistical graphics.

Keywords: Cognitive Bias, Graph Perception, Graphical Testing, Linear Regression, Statistical Graphics

1 Introduction

We all use statistical graphics, but how do we know that the graphics we use are communicating properly? When creating a graphic, we must consider the design choices most effective for conveying the intended result. For instance, we may decide to highlight the relationship between two variables in a scatter-plot by including a trend line, or adding color to highlight clustering (VanderPlas & Hofmann 2017). These design choices require that we understand the perceptual and visual biases that come into play when creating graphics, and as graphics are evaluated visually, we must use human testing to ground our understanding in empiricism.

Much of the research on the perception of visual features in charts has been conducted in psychophysics and tests for accuracy and quantitative comparisons when understanding a plot. Cleveland and McGill conducted a series of cognitive tasks designed to establish a hierarchy of visual components for making comparisons (Cleveland & McGill 1984). For example, it is more effective to display information on an x or y axis rather than using color in order to reduce the visual effort necessary to make numerical comparisons. Cleveland & McGill (1985) found that assessing the position of points along an axis is easier than determining the slope of a line. XXX we probably could include a few more papers and discussions here - the Cleveland 1984 and 1985 papers are based on the same study and are basically duplicates.

- Viewers often mentally exaggerate the magnitude of correlations in scatterplots when the data appear dense [@cleveland1982variable; @lauer1989density].
- Similarly, the aspect ratio of a graph can have an influence on the patterns that viewers identify. @cleveland1993visualizing, for example, argues that viewers can most easily detect cyclical patterns when an aspect ratio that makes the curve closest to 45◦ is used.
- Gestalt principles of perceptual organization play a crucial role in data extraction [@kosslyn2006graph]: for example, data points that are closer in space are more likely to be perceived as grouped [@ciccione2020grouping], and vertical and horizontal lines are easier to discriminate than oblique ones [@appelle1972perception].

The results of these cognitive tasks provide some consistent guidance for chart design; however, other methods of visual testing can further evaluate design choices and help us understand cognitive biases related to the evaluation of statistical charts.

1.1 Testing Statistical Graphics

We need human testing of graphics in order to draw broad conclusions, develop guidelines for graphical design, and improve graphical communication. Studies might ask participants to identify differences in graphs, read information off of a chart accurately, use data to make correct real-world decisions, or predict the next few observations. All of these types of tests require different levels of use and manipulation of the information being presented in the chart. Early researchers studied graphs from a psychological perspective (Spence 1990, Lewandowsky & Spence 1989), testing participants' abilities to detect a stimulus or a difference between two stimuli. Psychophysical methods have been used to test graphical perception, as in VanderPlas & Hofmann (2015*a*), which used the method of adjustment to estimate the magnitude of the impact of the sine illusion. However, there are more modern testing methods that have been developed since the heyday of psychophysics.

One major development in statistical graphics which led to more advanced testing methods is Wilkinson's Grammar of Graphics (Wilkinson 2013). The grammar of graphics serves as the fundamental framework for data visualization with the notion that graphics are built from the ground up by specifying exactly how to create a particular graph from a given data set. Visual representations are constructed through the use of "tidy data" which is characterized as a data set in which each variable is in its own column, each observation is in its own row, and each value is in its own cell (Wickham & Grolemund 2016). Graphics are viewed as a mapping from variables in a data set (or statistics computed from the data) to visual attributes such as the axes, colors, shapes, or facets on the canvas in which the chart is displayed. Software, such as Hadley Wickham's ggplot2 (Wickham 2011), aims to implement the framework of creating charts and graphics as the grammar of graphics recommends.

Combining the grammar of graphics with another tool for statistical graphics testing, the statistical lineup, yields interesting results. Buja et al. (2009) introduced the lineup protocol

to provide a framework for inferential testing. Through experimentation, methods such as the lineup protocol allow researchers to conduct studies geared at understanding human ability to conduct tasks related to the perception of statistical charts such as differentiation, prediction, estimation, and extrapolation (VanderPlas & Hofmann 2017, 2015*b*, Hofmann et al. 2012). The advancement of graphing software provides the tools necessary to develop new methods of testing statistical graphics.

While these testing methods are excellent, there is one particular subset of statistical graphics testing methods which we intend to develop further in this paper: assessing graphics by fitting statistical models "by eye".

1.2 Fitting Trends by Eye

Initial studies in the 20th century explored the use of fitting lines by eye through a set of points (Finney 1951, Mosteller et al. 1981). Common methods of fitting trends by eye involved maneuvering a string, black thread, or ruler until the fit is suitable, then drawing the line through the set of points. Recently, Ciccione & Dehaene (2021) conducted a comprehensive set of studies investigating human ability to detect trends in graphical representations using these types of manipulations. XXX this study was conducted online, but asked users to use the arrows on their keyboard to adjust the slope or select whether the shown slope was positive or negative.

Finney (1951) used graphical testing for computational purposes: to determine the effect of stopping iterative maximum likelihood calculations after one iteration. Many techniques in statistical analysis are performed with the aid of iterative calculations such as Newton's method or Fisher's scoring. The author was interested in whether one iteration of calculations was sufficient in the estimation of parameters connected with dose-response relationships. One measure of interest is the relative potency between a test preparation of doses and standard preparation of does; relative potency is calculated as the ratio of two equally effective doses between the two preparation methods. In this study, twenty-one scientists were recruited via postal mail and asked to "rule two lines" in order to judge by eye the positions for a pair of parallel probit regression lines in a biological assay. The author then computed one iterative calculation of the relative potency based on

starting values as indicated by the pair of lines provided by each participant and compared these relative potency estimates to that which was estimated by the full probit technique (reaching convergence through multiple iterations). Results indicated that one cycle of iterations for calculating the relative potency was sufficient based on the starting values provided by eye from the participants.

Mosteller et al. (1981) sought to understand the properties of least squares and other computed lines by establishing one systematic method of fitting lines by eye. Participants were asked to fit lines by eye to four scatter-plots using an 8.5 x 11 inch transparency with a straight line etched completely across the middle. A latin square design with packets of the set of points stapled together in four different sequences was used to determine if there is an effect of order of presentation. It was found that order of presentation had no effect and that participants tended to fit the slope of the principal axis (error minimized orthogonally, both horizontal and vertical, to the regression line) over the slope of the least squares regression line (error minimized vertically to the regression line). These results support previous research on "ensemble perception" indicating the visual system can compute averages of various features in parallel across the items in a set (Chong & Treisman 2003, 2005, Van Opstal et al. 2011).

In Ciccione & Dehaene (2021), participants were asked to judge trends, estimate slopes, and conduct extrapolation. To estimate slopes, participants were asked to report the slope of the best-fitting regression line using a track-pad to adjust the tilt of a line on screen. Results indicated the slopes participants reported were always in excess of the ideal slopes, both in the positive and in the negative direction, and those biases increase with noise and with number of points. This supports the results found in Mosteller et al. (1981) and suggest that participants might use Deming regression when fitting a line to a noisy scatter-plot (Deming 1943, Linnet 1998, Martin 2000) xxx added citation.

While not explicitly intended for perceptual testing, in 2015, the New York Times introduced an interactive feature, called You Draw It (Aisch et al. 2015, Buchanan et al. 2017, Katz 2017). Readers are asked to input their own assumptions about various metrics and compare how these assumptions relate to reality. The New York Times team utilizes Data Driven Documents (D3) (Bostock et al. 2011) xxx added citation that allows readers

to predict these metrics through the use of drawing a line on their computer screen with their computer mouse. After the reader has completed drawing the line, the actual observed values are revealed and the reader may check their estimated knowledge against the actual reported data. While this interactive feature is designed to get readers to confront their own intuitions about data in the news, we feel that the interactivity of this method may be useful for another purpose.

In this paper, we establish 'You Draw It', adapted from the New York Times feature, as a tool for graphical testing, validating the 'You Draw It' method by replicating the study conducted by Mosteller et al. (1981). Based on previous research surrounding "ensemble perception," we hypothesize that visual regression tends to mimic principle component regression rather than an ordinary least squares regression. In order to assess this hypothesis, we introduce a method for statistically modeling the participant drawn lines using generalized additive mixed models.

2 Methods

2.1 Participants

Participants were recruited through through Twitter, Reddit, and direct email in May 2021. A total of 35 individuals completed 119 unique 'You Draw It' task plots; all completed you draw it task plots were included in the analysis. All participants had normal or corrected to normal vision and signed an informed consent form. The experimental tasks took approximately 15 minutes to complete. While this study does utilize a convenience sample, as this is primarily a perceptual task, previous results have found few differences between expert and non-expert participants in this context (VanderPlas & Hofmann 2015*b*). These data were collected to validate this method of graphical testing, with the hopes of providing a new tool to assess graphical perception interactively. Participants completed the experiment on their own computers in an environment of their choosing. The experiment was conducted and distributed through a Shiny application (Chang et al. 2021) xxx added citation found here. We should probably make a rstudioconnect version that saves data to e.g. a google spreadsheet? Just in case my server goes down or something? XXX linked to

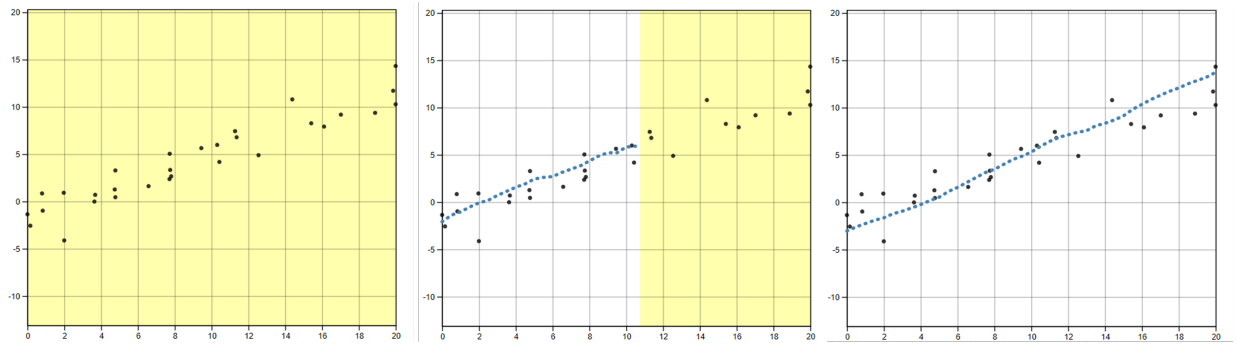

Figure 1: 'You Draw It' task plot as shown to particpants during the study. The first frame (left) illustrates what particpants first see with the prompt "Use your mouse to fill in the trend in the yellow box region." The second frame (middle), illustrates what the particpant sees while completing the task; the yellow region provides a visual cue for participants indicating where the participant still needs to complete a trend-line. The last frame (right) illustrates the participants finished trend-line before submission.

updated shinyapps.io through new repository - does not currently write data anywhere.

2.2 'You Draw It' Task

In the study, participants are shown an interactive scatter-plot (Fig. 1) along with the prompt, "Use your mouse to fill in the trend in the yellow box region." The yellow box region moves along as the user draws their trend-line, providing a visual cue which indicates where the user still needs to complete a trend line. After the entire domain has been visually estimated or predicted, the yellow region disappears, indicating the participant has completed the task. Data Driven Documents (D3), a JavaScript-based graphing framework that facilitates user interaction, is used to create the 'You Draw It' visual. In order to allow for user interaction and data collection, we integrate the D3 visual into Shiny using the `r2d3` package. While the interface is highly customized to this particular task, we hope to generalize the code and provide a Shiny widget in an R package soon.

Table 1: Designated model equation parameters for simulated data.

Parameter Choice	$y_{\bar{x}}$	β_1	σ
S	3.88	0.66	1.30
F	3.90	0.66	1.98
V	3.89	1.98	1.50
N	4.11	-0.70	2.50

2.3 Data Generation

All data processing was conducted in R statistical software. A total of $N = 30$ points $(x_i, y_i), i = 1, ...N$ were generated for $x_i \in [x_{min}, x_{max}]$ where x and y have a linear relationship. Data were simulated based on a linear model with additive errors:

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{1}$$
$$\text{with } e_i \sim N(0, \sigma^2).$$

Model equation parameters, β_0 and β_1, were selected to reflect the four data sets (F, N, S, and V) used in Mosteller et al. (1981) (Table 1). We obtain β_0 from the point-slope equation of a line using the mean of the generated x values and the predefined y value at $\bar{x}$, denoted $y_{\bar{x}}$. Parameter choices F, N, and S simulated data across a domain of 0 to 20. Parameter choice F produces a trend with a positive slope and a large variance while N has a negative slope and a large variance. In comparison, S shows a trend with a positive slope with a small variance and V yields a steep positive slope with a small variance over the domain of 4 to 16. Fig. 2 illustrates an example of simulated data for all four parameter choices intended to reflect the trends in Mosteller et al. (1981). Aesthetic design choices were made consistent across each of the interactive 'You Draw It' task plots. The y-axis range extended 10% beyond (above and below) the range of the simulated data points to allow for users to draw outside the simulated data set range and avoid anchoring their lines to the corners of the plot.

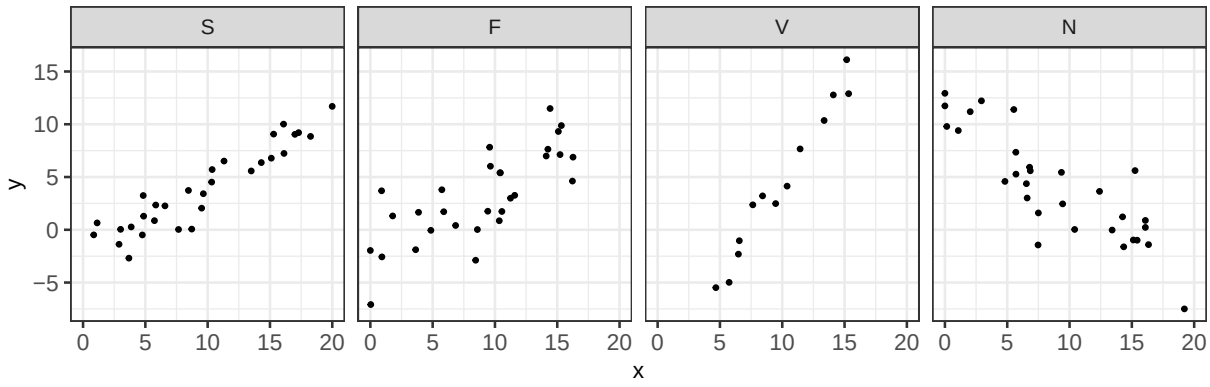

Figure 2: Example of simulated data points displayed in a scatter-plot illustrating the trends associated with the four selected parameter choices.

2.4 Study Design

This experiment was conducted as part of a larger study of the perception of log and linear scales; for simplicity, we focus on the study design and methods related to the current study. Each scatter-plot was the graphical representation of a data set that was generated randomly and independently for each participant at the start of the experiment. Participants in the study are shown two 'You Draw It' practice plots in order to train participants in the skills associated with executing the task - in particular, the responsiveness of the applet requires that participants draw a line at a certain speed, ensuring that all of the evenly spaced points along the hand-drawn line are filled in. During the practice session, participants are provided with instruction prompts accompanied by a .gif and a practice plot. Instructions guide participants to start at the edge of the yellow box, to make sure the yellow region is moving along with their mouse as they draw, and that they can draw over their already drawn line. Practice plots are then followed by four 'You Draw It' task plots associated with the current study. The order of the task plots was randomly assigned for each individual in a completely randomized design.

3 Results

3.1 Fitted Regression Lines

We compare the participant drawn line to two regression lines determined by ordinary least squares (OLS) regression and regression based on the principal axis (PCA). Fig. 3 illustrates the difference between an OLS regression line which minimizes the vertical distance of points from the line and a regression line based on the principal axis which minimizes the Euclidean distance of points (orthogonal) from the line.

Due to the randomness in the data generation process, the actual slope of the linear regression line fit through the simulated points could differ from the predetermined slope. Therefore, we fit an OLS regression to each scatter-plot to obtain estimated parameters $\hat{\beta}_{0,OLS}$ and $\hat{\beta}_{1,OLS}$. Fitted values, $\hat{y}_{k,OLS}$, are then obtained every 0.25 increments across the domain from the OLS regression equation, $\hat{y}_{k,OLS} = \hat{\beta}_{0,OLS} + \hat{\beta}_{1,OLS} x_{k.}$, for $k = 1, ..., 4x_{max} + 1$. The regression equation based on the principal axis was determined by using the `princomp` function in the stats package in base R to obtain the rotation of the coordinate axes from the first principal component (direction which captures the most variance). The estimated slope, $\hat{\beta}_{1,PCA}$, is determined by the ratio of the axis rotation in y and axis rotation in x of the first principal component with the y-intercept, $\hat{\beta}_{0,PCA}$ calculated by the point-slope equation of a line using the mean of of the simulated points, $(\bar{x}_i, \bar{y}_i)$. Fitted values, $\hat{y}_{k,PCA}$, are then obtained every 0.25 increment across the domain from the PCA regression equation, $\hat{y}_{k,PCA} = \hat{\beta}_{0,PCA} + \hat{\beta}_{1,PCA} x_k$.

3.2 Residual Trends

For each participant, the final data set used for analysis contains $x_{ijk}, y_{ijk,drawn}, \hat{y}_{ijk,OLS}$, and $\hat{y}_{ijk,PCA}$ for parameter choice $i = 1, 2, 3, 4$, j $= 1, ...N_{participant}$, and x_{ijk} value $k = 1, ..., 4x_{max} + 1$. Using both a linear mixed model and a generalized additive mixed model, comparisons of vertical residuals in relation to the OLS fitted values ($e_{ijk,OLS} = y_{ijk,drawn} - \hat{y}_{ijk,OLS}$) and PCA fitted values ($e_{ijk,PCA} = y_{ijk,drawn} - \hat{y}_{ijk,PCA}$) were made across the domain. Fig. 4 displays an example of all three fitted trend lines for parameter choice F.

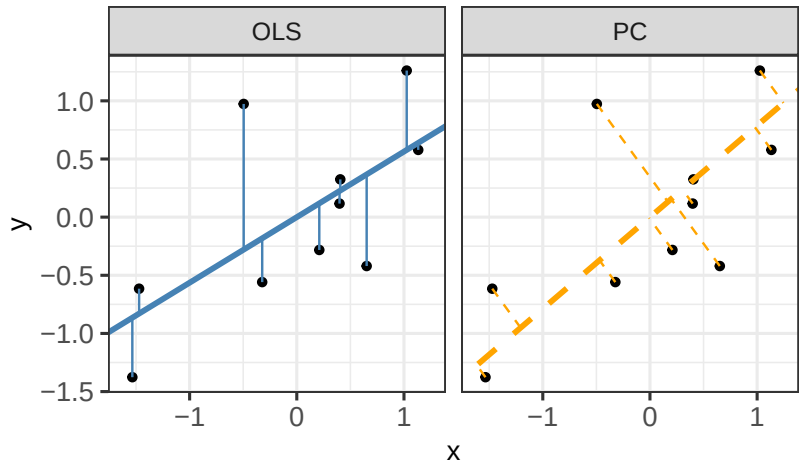

Figure 3: Comparison between an OLS regression line which minimizes the vertical distance of points from the line and a regression line based on the principal axis which minimizes the Euclidean distance of points (orthogonal) from the line.

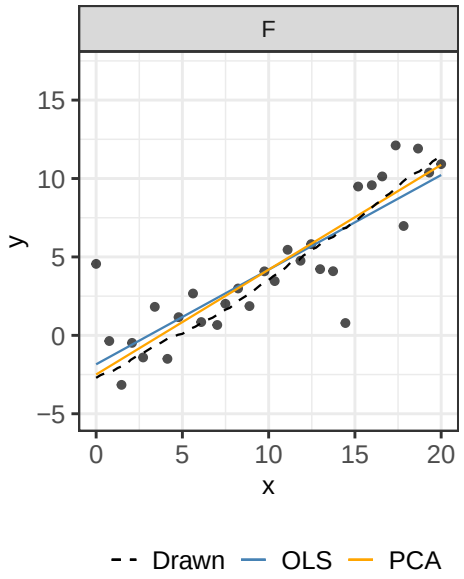

Figure 4: Illustrates the data associated with and collected for one 'You Draw It' task plot. Trend-lines include the participant drawn line (dashed black), the OLS regression line (solid steelblue) and the PCA regression line based on the principal axis (solid orange).

3.2.1 Linear Trend Constraint

Using the `lmer` function in the lme4 package (Bates et al. 2015), a linear mixed model (LMM) is fit separately to the OLS residuals and PCA residuals, constraining the fit to a linear trend. Parameter choice, x, and the interaction between x and parameter choice were treated as fixed effects with a random participant effect accounting for variation due to participant. The LMM equation for each fit (OLS and PCA) is given by:

$$y_{ijk,drawn} - \hat{y}_{ijk,fit} = e_{ijk,fit} = [\gamma_0 + \alpha_i] + [\gamma_1 x_{ijk} + \gamma_{2i} x_{ijk}] + p_j + \epsilon_{ijk} \tag{2}$$

where

- $y_{ijk,drawn}$ is the drawn y value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x value
- $\hat{y}_{ijk,fit}$ is the fitted y value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x value corresponding to either the OLS or PCA fit
- $e_{ijk,fit}$ is the residual between the drawn and fitted y values for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x value corresponding to either the OLS or PCA fit
- γ_0 is the overall intercept
- α_i is the effect of the i^{th} parameter choice (F, S, V, N) on the intercept
- γ_1 is the overall slope for x
- γ_{2i} is the effect of the parameter choice on the slope
- x_{ijk} is the x value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment
- $p_j \sim N(0, \sigma^2_{participant})$ is the random error due to the j^{th} participant's characteristics
- $\epsilon_{ijk} \sim N(0, \sigma^2)$ is the residual error.

Constraining the residual trend to a linear fit, Fig. 5 shows the estimated trend line of the residuals between the participant drawn points and fitted values for both the OLS regression line and PCA regression line. Estimated residual trend lines are overlaid on the observed individual participant residuals. Results indicate the estimated trends of PCA residuals (orange) appear to align closer to the $y = 0$ horizontal (dashed) line than the OLS residuals (blue). In particular, this trend is more prominent in parameter choices

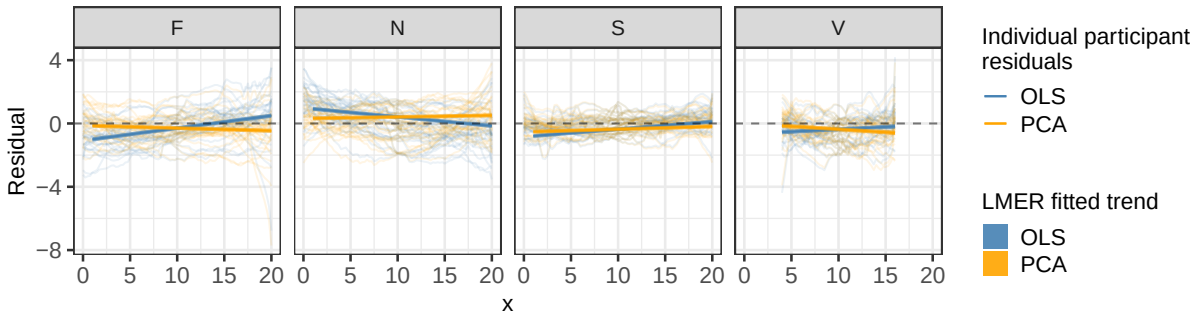

Figure 5: Estimated trend line of the residuals between the participant drawn points and fitted values for both the OLS (blue) regression line and PCA (orange) regression line constrained to a linear fit modeled by a linear mixed model. Estimated residual trends with 95% confidence bands are overlaid on the observed individual participant residuals.

with large variances (F and N). These results are consistent to those found in Mosteller et al. (1981) indicating participants fit a trend-line closer to the estimated regression line with a slope based on the first principal axis than the estimated OLS regression line thus, providing support for "ensemble perception".

3.2.2 Smoothing Spline Trend

Eliminating the linear trend constraint, the `bam` function in the mgcv package (Wood 2011, Wood et al. 2016, Wood 2004, 2017, 2003) is used to fit a generalized additive mixed model (GAMM) separately to the OLS residuals and PCA residuals to allow for estimation of smoothing splines. Parameter choice was treated as a fixed effect with no estimated intercept and a separate smoothing spline for x was estimated for each parameter choice. A random participant effect accounting for variation due to participant and a random spline for each participant accounted for variation in spline for each participant. The GAMM equation for each fit (OLS and PCA) residuals is given by:

$$y_{ijk,drawn} - \hat{y}_{ijk,fit} = e_{ijk,fit} = \alpha_i + s_i(x_{ijk}) + p_j + s_j(x_{ijk}) \tag{3}$$

where

- $y_{ijk,drawn}$ is the drawn y value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x value
- $\hat{y}_{ijk,fit}$ is the fitted y value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x value corresponding to either the OLS or PCA fit
- $e_{ijk,fit}$ is the residual between the drawn and fitted y values for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x value corresponding to either the OLS or PCA fit
- α_i is the intercept for the parameter choice i
- s_i is the smoothing spline for the i^{th} parameter choice
- x_{ijk} is the x value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment
- $p_j \sim N(0, \sigma^2_{participant})$ is the error due to participant variation
- s_j is the random smoothing spline for each participant.

Allowing for flexibility in the residual trend, Fig. 6 shows the estimated trend line of the residuals between the participant drawn points and fitted values for both the OLS regression line and PCA regression line. Estimated residual trends are overlaid on the observed individual participant residuals. The results of the GAMM align with those shown in Fig. 5 providing support that for scatter-plots with more noise (F and N), estimated trends of PCA residuals (orange) appear to align closer to the $y = 0$ horizontal (dashed) line than the OLS residuals (blue). By fitting smoothing splines, we can determine whether participants naturally fit a straight trend-line to the set of points or whether they deviate throughout the domain. In particular, in scatter-plots with smaller variance (S and V), we can see that participants began at approximately the correct starting point then deviated away from the fitted regression lines and corrected for their fit toward the end of their trend-line. In scatter-plots with larger variance (F and N), participants estimated their starting value in the extreme direction of the OLS regression line based on the increasing or decreasing trend but more accurately represented the starting value of the PCA regression line. As participants continued their trend-line, they crossed through the OLS regression line indicating they estimated the slope in the extreme direction. These results provide further insight into the curvature humans perceive in a set of points.

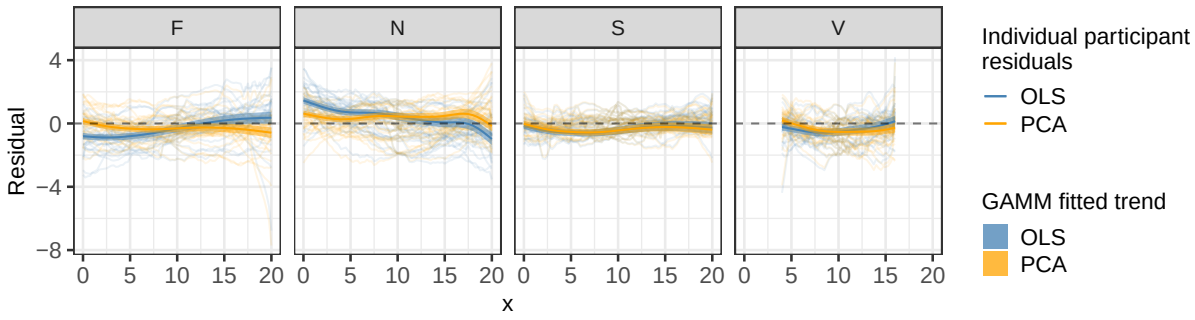

Figure 6: Estimated trend line of the residuals between the participant drawn points and fitted values for both the OLS (blue) regression line and PCA (orange) regression line determined by smoothing splines fit by a generalized additive mixed model. Estimated residual trends with 95% confidence bands are overlaid on the observed individual participant residuals.

4 Discussion and Conclusion

The intent of this research was to adapt 'You Draw It' from the New York Times feature as a tool and method for testing graphics and introduce a method for statistically modeling the participant drawn lines. We provided support for the validity of the 'You Draw It' method by replicating the study found in Mosteller et al. (1981). Using generalized additive mixed models, we assessed the deviation of the participant drawn lines from the statistically fitted regression lines. Our results found that when shown points following a linear trend, participants visually fit a regression line that mimics the first principle component regression as opposed to ordinary least squares regression. Data simulated with a larger variance provided strong support for a participants tendency to visually fit the first principle component regression. Our results indicate that participants minimized the distance from the their regression line over both the x and y axis simultaneously. XXX some sentance that says that our results are much stronger than the original results because our analysis methods allow for more sophistication (and we have a slight technological advantage on the original paper's authors as well, after 40 years). These results provide support that humans perform "ensemble perception" in a statistical graphic setting. We allowed

participants to draw trend lines that deviated from a straight line and gained an insight into the curvature the human eye perceives in a set of points.

5 Future Work

This study provided a basis for the use of 'You Draw It' as a tool for testing statistical graphics and introduced a method for statistically modeling participant drawn lines using generalized additive mixed models. Further investigation is necessary to implement this method in non-linear settings and with real data in order to facilitate scientific communication. This tool could also be used to evaluate human ability to extrapolate data from trends. We intend to create an R package designed for easy implementation of 'You Draw It' task plots in order to make this tool accessible to other researchers.

SUPPLEMENTARY MATERIAL

- **Participant Data:** De-identified participant data collected in the study and used for analyses are available to be downloaded from GitHub here.
- **Data Analysis Code:** The code used to replicate the analysis in this paper can be found here.
- **Study Applet:** The shiny app used to conduct the study can be accessed here.
- **RShiny Applet Code:** The code used to create the RShiny Applet for data collection can be found here.

References

Aisch, G., Cox, A. & Quealy, K. (2015), 'You draw it: How family income predicts children's college chances'.
URL: *https://www.nytimes.com/interactive/2015/05/28/upshot/you-draw-it-how-family-income-affects-childrens-college-chances.html*

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015), 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software* **67**(1), 1–48.

Bostock, M., Ogievetsky, V. & Heer, J. (2011), 'D^3 data-driven documents', *IEEE transactions on visualization and computer graphics* **17**(12), 2301–2309.

Buchanan, L., Park, H. & Pearce, A. (2017), 'You draw it: What got better or worse during obama's presidency'.
URL: *https://www.nytimes.com/interactive/2017/01/15/us/politics/you-draw-obama-legacy.html*

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F. & Wickham, H. (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.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A. & Borges, B. (2021), *shiny: Web Application Framework for R*. R package version 1.7.1.
URL: *https://CRAN.R-project.org/package=shiny*

Chong, S. C. & Treisman, A. (2003), 'Representation of statistical properties', *Vision research* **43**(4), 393–404.

Chong, S. C. & Treisman, A. (2005), 'Statistical processing: Computing the average size in perceptual groups', *Vision research* **45**(7), 891–900.

Ciccione, L. & Dehaene, S. (2021), 'Can humans perform mental regression on a graph? accuracy and bias in the perception of scatterplots', *Cognitive Psychology* **128**, 101406.

Cleveland, W. S. & McGill, R. (1984), 'Graphical perception: Theory, experimentation, and application to the development of graphical methods', *Journal of the American statistical association* **79**(387), 531–554.

Cleveland, W. S. & McGill, R. (1985), 'Graphical perception and graphical methods for analyzing scientific data', *Science* **229**(4716), 828–833.

Deming, W. E. (1943), 'Statistical adjustment of data.'.

Finney, D. (1951), 'Subjective judgment in statistical analysis: An experimental study', *Journal of the Royal Statistical Society: Series B (Methodological)* **13**(2), 284–297.

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.

Katz, J. (2017), 'You draw it: Just how bad is the drug overdose epidemic?'.
URL: *https://www.nytimes.com/interactive/2017/04/14/upshot/drug-overdose-epidemic-you-draw-it.html*

Lewandowsky, S. & Spence, I. (1989), 'The perception of statistical graphs', *Sociological Methods & Research* **18**(2-3), 200–242.

Linnet, K. (1998), 'Performance of deming regression analysis in case of misspecified analytical error ratio in method comparison studies', *Clinical chemistry* **44**(5), 1024–1031.

Martin, R. F. (2000), 'General deming regression for estimating systematic bias and its confidence interval in method-comparison studies', *Clinical chemistry* **46**(1), 100–104.

Mosteller, F., Siegel, A. F., Trapido, E. & Youtz, C. (1981), 'Eye fitting straight lines', *The American Statistician* **35**(3), 150–152.

Spence, I. (1990), 'Visual psychophysics of simple graphical elements.', *Journal of Experimental Psychology: Human Perception and Performance* **16**(4), 683.

Van Opstal, F., de Lange, F. P. & Dehaene, S. (2011), 'Rapid parallel semantic processing of numbers without awareness', *Cognition* **120**(1), 136–147.

VanderPlas, S. & Hofmann, H. (2015*a*), 'Signs of the sine illusion—why we need to care', *Journal of Computational and Graphical Statistics* **24**(4), 1170–1190.

VanderPlas, S. & Hofmann, H. (2015*b*), 'Spatial reasoning and data displays', *IEEE Transactions on Visualization and Computer Graphics* **22**(1), 459–468.

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

Wickham, H. (2011), 'ggplot2', *Wiley Interdisciplinary Reviews: Computational Statistics* **3**(2), 180–185.

Wickham, H. & Grolemund, G. (2016), *R for data science: import, tidy, transform, visualize, and model data*, " O'Reilly Media, Inc.".

Wilkinson, L. (2013), *The grammar of graphics*, Springer Science & Business Media.

Wood, S. (2017), *Generalized Additive Models: An Introduction with R*, 2 edn, Chapman and Hall/CRC.

Wood, S. N. (2003), 'Thin-plate regression splines', *Journal of the Royal Statistical Society (B)* **65**(1), 95–114.

Wood, S. N. (2004), 'Stable and efficient multiple smoothing parameter estimation for generalized additive models', *Journal of the American Statistical Association* **99**(467), 673–686.

Wood, S. N. (2011), 'Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models', *Journal of the Royal Statistical Society (B)* **73**(1), 3–36.

Wood, S., N., Pya & Saefken, B. (2016), 'Smoothing parameter and model selection for general smooth models (with discussion)', *Journal of the American Statistical Association* **111**, 1548–1575.