Changing Beliefs About Correlations in Atypical Scatterplots

GABRIEL STRAIN, Department of Computer Science, Faculty of Science and Engineering, University of Manchester, United Kingdom

ANDREW J. STEWART, Department of Computer Science, Faculty of Science and Engineering, University of Manchester, United Kingdom

PAUL WARREN, Division of Psychology, Communication and Human Neuroscience, School of Health Sciences, Faculty of Biology, Medicine, and Health, University of Manchester, United Kingdom

CHARLOTTE RUTHERFORD, Division of Psychology Communication and Human Neuroscience, School of Health Sciences, Faculty of Biology, Medicine, and Health, University of Manchester, United Kingdom

CAROLINE JAY, Department of Computer Science, Faculty of Science and Engineering, University of Manchester, United Kingdom

abstract goes here

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: belief change, correlation perception, scatterplot, crowdsourced

ACM Reference Format:
Gabriel Strain, Andrew J. Stewart, Paul Warren, Charlotte Rutherford, and Caroline Jay. 2018. Changing Beliefs About Correlations in Atypical Scatterplots. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

Authors' addresses: Gabriel Strain, Department of Computer Science, Faculty of Science and Engineering, University of Manchester, Oxford Road, Manchester, United Kingdom, M13 9PL; Andrew J. Stewart, Department of Computer Science, Faculty of Science and Engineering, University of Manchester, Oxford Road, Manchester, United Kingdom, M13 9PL; Paul Warren, Division of Psychology, Communication and Human Neuroscience, School of Health Sciences, Faculty of Biology, Medicine, and Health, University of Manchester, Oxford Road, Manchester, United Kingdom, M13 9PL; Charlotte Rutherford, Division of Psychology Communication and Human Neuroscience, School of Health Sciences, Faculty of Biology, Medicine, and Health, University of Manchester, Oxford Road, Manchester, United Kingdom, M13 9PL; Caroline Jay, Department of Computer Science, Faculty of Science and Engineering, University of Manchester, Oxford Road, Manchester, United Kingdom, M13 9PL.

Manuscript submitted to ACM

1 INTRODUCTION

2 RELATED WORK

3 GENERAL METHODS

In this section we discuss our general research methods, including our implementations of open research practices, our approach to and justification for crowdsourcing, and our use of the ChatGPT4 LLM in preparing parts of our stimuli.

3.1 Open Research

Both our pre and main studies were conducted according to the principles of open and reproducible research [2]. We pre-registered hypotheses and analysis plans with the Open Science Framework (OSF) for the pre-study[1] and the main experiment[2], and there were no deviations from them. All data and analysis code are included in a GitHub repository[3]. This repository contains instructions for building a Docker container [12] that reproduces the computational environment the paper was written in. This allows for full replication of stimuli, figures, analysis, and the paper itself. Ethical approval was granted by the (removed for anon).

3.2 Crowdsourcing

While much prior work into correlation perception in scatterplots has taken place in person, there is precedence for work that explores cognition to take place online using crowdsourced participants [**?**]. Crowdsourcing not only affords us recruitment of samples from across our lay population of interest, it is considerably quicker and less expensive than in-person testing. We therefore choose to crowdsource all participants. Previous work has reported issues of data quality and skewed demographics [4, 5, 13], so we follow published guidelines [13] to give us the best chance of collecting high quality data. We use the Prolific.co platform [1] with strict pre-screening criteria; participants were required to have completed at least 100 studies using Prolific, and were required to have a Prolific score of 100, representing a 99% approval rate.

3.3 Use of Large Language Models

- issues regarding stimulus generation normally
- advantages conferred by using ChatGPT
- reproducibility issues?

4 PRE-STUDY: INVESTIGATING BELIEFS ABOUT RELATEDNESS STATEMENTS

4.1 Introduction

4.1.1 Testing Beliefs.

4.1.2 Preparation of Stimuli. Due to previous evidence suggesting effects of prior belief strength and topic emotionality on the propensity for belief change, we first aim to build a picture of people's thoughts and feelings along these dimensions in our population of interest. With the intention of testing the potential for changes in beliefs about correlations displayed in scatterplots depicting weak and strong correlations, and

[1] https://osf.io/xuf4d

[2] tbc

[3] https://github.com/gjpstrain/beliefs_attitudes_atypical

Table 1. Pre-test statements that were rated as being strongly correlated.

Item Number	Statement - Strong Correlation Depicted
1	Increased exposure to sunlight is correlated with higher vitamin D levels.
2	As caffeine consumption increases, so does the average heart rate.
3	Greater frequency of exercise is linked to a lower risk of depression.
4	Greater use of helmets is associated with a lower incidence of head injuries in cyclists.
5	As the quality of healthcare improves, life expectancy tends to increase.
6	As access to clean water improves, the incidence of waterborne diseases decreases.
7	Higher levels of empathy are linked to stronger interpersonal relationships.
8	As soil quality degrades, agricultural productivity tends to decrease.
9	Higher levels of civic engagement are linked to a stronger sense of community.
10	Higher sugar consumption is associated with an increased risk of dental cavities.
11	Higher attendance at preventive health screenings is linked to earlier detection of diseases.
12	Increased use of energy-efficient appliances is associated with lower electricity bills.
13	As pedestrian-friendly infrastructure improves, urban walkability tends to increase.
14	Greater regularity in sleep patterns is associated with improved mental health.

those whose topics were both strong and neutral in emotional valence, we began by using ChatGPT4 [18] to generate 100 correlation statements using the following prompt:
“Generate 100 statements that describe the correlation between two variables, such as :
”X is associated with a higher level of Y” or
”As X increases, Y increases”.
Try to match all the statements on emotionality.“

The full list of these statements can be found in the supplementary materials. Note that we cite our use of ChatGPT according to the AI Code of Conduct developed by Iliada Eleftheriou and Ajmal Mubarik and the University of Manchester [7]. Two authors rated each statement on topic emotionality and strength of correlation using Likert scales from 1 to 7. Topic emotionality had a midpoint at 4, whereas strength of correlation varied between 1 (Not Related At All) and 7 (Strongly Related). We calculated a quadratic weighted Cohen’s Kappa between the two raters using the **irr** package (version 0.84.1 [9]), in order to penalise larger magnitude disagreements more harshly. We found agreement above chance for both topic emotionality ($\kappa = 0.49$, $p < .001$) and strength of correlation ($\kappa = 0.51$, $p < .001$), indicating moderate levels of agreement in both cases [6, 8].

Following this, we selected strongly and weakly correlated statements with the highest level of absolute agreement, resulting in the 14 strongly correlated statements that can be seen in Table 1 and the 11 weakly correlated statements that can be seen in Table 2. We then tested these 25 statements with a representative UK sample in order to ascertain consensus on both topic emotionality and strength of correlation. Doing so allows us to effectively exclude these factors when we analyse the effects of our atypical scatterplot designs on the propensity for belief change in our main experiment.

Table 2. Pre-test statements that were rated as being weakly correlated.

Item Number	Statement - Weak Correlation Depicted
15	Greater water consumption is linked to improved kidney function.
16	As the amount of sleep decreases, the risk of obesity increases.
17	Greater intake of omega-3 fatty acids is associated with lower inflammation levels.
18	Greater exposure to music education is linked to enhanced cognitive development in children.
19	Higher exposure to air conditioning is associated with increased respiratory issues.
20	Higher frequency of family meals is linked to better eating habits in children.
21	As participation in community arts programs increases, local cultural engagement tends to rise.
22	Higher consumption of spicy foods is associated with a lower risk of certain types of cancer.
23	Greater adherence to a Mediterranean diet is linked to a lower risk of neurodegenerative diseases.
24	Higher consumption of nuts and seeds is associated with reduced risk of cardiovascular diseases.
25	As cultural preservation efforts increase, community identity and cohesion tend to strengthen.

4.2 Method

4.2.1 Participants. 100 participants were recruited using the Prolific.co platform [1]. English fluency and residency was required for participation, as our main experiment relied on familiarity with data visualizations from a popular British news source. In addition to 25 experimental items, we included six attention check items, which asked participants to provide specific answers. No participants failed more than 2 out of 6 attention check items, and therefore data from all 100 were included in the full analysis (52.0% male and 48.0% female. Participants' mean age was 41.1 (SD = 12.3). The average time taken to complete the survey was 7.6 minutes (SD = 2.9 minutes).

4.2.2 Design. Each participant saw all survey items (Table 1 and Table 2), along with the six attention check items, in a fully randomised order. All experimental code, materials, and instructions are hosted on GitLab[4].

4.2.3 Procedure. The experiment was built using Psychopy [14] and hosted on Pavlovia.org. Participants were permitted to complete the experiment using a phone, tablet, desktop, or laptop computer. Participants were first shown the participant information sheet and were asked to provide consent through key presses in response to consent statements. They were asked to provide their age in a free text box, followed by their gender identity. Participants were told that they would be asked to read statements about the relatedness between a pair of variables, after which they would have to indicate their beliefs about topic emotionality and the strength of correlation suggested using a pair of sliders. To familiarize themselves with the sliders, they were asked to complete a practice round in response to the statement "As participation in online experiments increases, society becomes happier."

4.3 Results

All analyses were conducted using R (version 4.4.1). We use the **irr** package to calculate Fleiss' Kappa to measure interrater agreement on topic emotionality and strength of correlation for the 25 experimental

[4] https://gitlab.pavlovia.org/Strain/beliefs__scatterplots__pretest

items. This analysis revealed that participants agreed above chance for both topic emotionality ($\kappa = 0.07$, $p < .001$) and strength of correlation ($\kappa = 0.06$, $p < .001$).

4.4 Selecting Statements for the Main Experiment

To control for any potential effects of topic emotionality in the main experiment, we first select statements that represent neutral emotional valence. Statements with average topic emotionality ratings between 3 and 5 are statements 2, 10, 22, 16, and 23. To ascertain which statements represent the greatest consensus, we add standard deviations in ratings for topic emotionality and strength of correlation. Due to concerns about experimental power, and in line with evidence that propensity for belief change is highest when prior beliefs are not strongly held [**?**], we elected at this point to test only the statement corresponding to weak beliefs about the strength of correlation between the variables in question. We therefore test statement number 22, "Higher consumption of spicy foods is associated with a lower risk of certain types of cancer.", however we modify the wording so that both variables (food consumption and cancer risk) are positively correlated, as previous work indicates that the manipulations we use in the atypical scatterplot condition are able to change estimates of correlation in positively correlated scatterplots; no work regarding the effects of these manipulations in negatively correlated scatterplots has been completed.

4.5 Discussion

Fleiss' Kappa values for interrater agreement on both topic emotionality and strength of correlation scales are low ($\kappa = 0.07$ and $\kappa = 0.06$ respectively), however do exceed that which would be expected by chance. We suggest this may be due to Fleiss' Kappa not being designed with ordinal (Likert scales in this case) data in mind. In light of this we do not make decisions regarding which statement to use based on the values of Fleiss' Kappa observed, but rather on the standard deviations of ratings across all raters. Regardless, we do not consider this to be a particular weakness, as we also test topic emotionality and strength of correlation with participants in the main study and include these ratings as part of our analysis.

5 MAIN STUDY: POTENTIAL FOR BELIEF CHANGE USING ATYPICAL SCATTERPLOTS

We test the statement that exhibited the lowest average level of belief about correlation, and the 2nd highest level of consensus. Modified for directionality, this statement is therefore: "Higher consumption of plain (non-spicy) foods is associated with a lower risk of certain types of cancer."

5.1 Introduction

- hypotheses
-

5.1.1 Defensive Confidence. In line with evidence that those who are more confident in their ability to defend their own positions are more susceptible to having those positions changed [**?**], we test participants' defensive confidence using a 12-item scale. This scale is replicated from previous work in the supplemental material, and has additionally been utilized more recently [**?**] to explore the potential for attitude change specifically with regards to correlations in scatterplots. Participants provide answers to the 12 scale items using a 5 point Likert scale ranging from 1 (*not at all characteristic of me*) to 5 (*extremely characteristic of me*). Analysis including participants' defensive confidence scores is included in **?@sec-additional-analyses**.

5.2 Stimuli

Recent work has shown that estimates of correlation can be altered when point opacities and sizes are systematically varied in scatterplots [15–17]. These manipulations have been used in an attempt to correct for a long-standing underestimation bias observed in correlation perception as it pertains to scatterplots. As we now aim to test the propensity of these manipulations to affect participants' beliefs about levels of relatedness, we choose the set of manipulations that has previously produced the most dramatic effect on correlation estimates; namely, the combination of typical orientation size and opacity manipulations provided by Strain et al [17]. Here, the size and opacity of a certain scatterplot point is lowered as a function of that point's residual error using equation 1:

$$point_{size/opacity} = 1 - b^{residual} \tag{1}$$

In order to facilitate comparison to this work, we use the same protocol to produce scatterplots for our atypical condition. This includes creating scatterplots with 128 points, using $b = 0.25$, employing a scaling factor and constant for point size, and using an opacity floor to ensure point visibility, as this has been an issue in previous work. As there is evidence that people are less likely to update strongly held beliefs following viewing scatterplot visualizations [**?** , look for more], we selected a correlative statement that was judged as representing weakly correlated variables. Thus, in order to induce belief change, participants only viewed scatterplots representing **strongly** correlated variables ($0.6 > r > 0.99$). We used **ggplot** in R to create plots that echoed the style of a British news broadcaster and fictitiously claimed the data to be provided by the British National Health service. Examples of typical and atypical data-identical scatterplots can be seen in **?@fig-main-examples**.

5.3 Method

5.3.1 Participants. Participants were recruited using Prolific.co [1]. English fluency and UK residency was required for participation, as well as normal or corrected-to-normal vision, and having not participated in any of our previous studies regarding correlation perception in scatterplots [**?**], as these represented earlier testing of the alternative designs we employ in the atypical scatterplot condition. Data were collected from 77 participants for each condition. 2 participants failed more than 2 out of 4 attention check questions for each condition, meaning their data were excluded per pre-registration stipulations. Data from the remaining 150 participants were included in the full analysis (48.7% male, 48.7% female, and 2.7% non-binary). Participants' mean age was 39.3 (SD = 11.5). Participants' mean graph literacy score was 21.3 (SD = 4.3) out of 30. On average, participants took 14.2 minutes to complete the experiment (SD = 6.41).

5.3.2 Design. We employed a between-participants design. Each participants was randomly assigned to either group A, in which case they viewed typical scatterplots, or group B, in which they viewed atypical scatterplots designed deliberately to elicit higher levels of belief change. Participant saw all experimental items for their group, along with 4 attention check items, in a fully randomised order. All experimental code, materials, and instructions are hosted on GitLab as two separate experiments[insert]

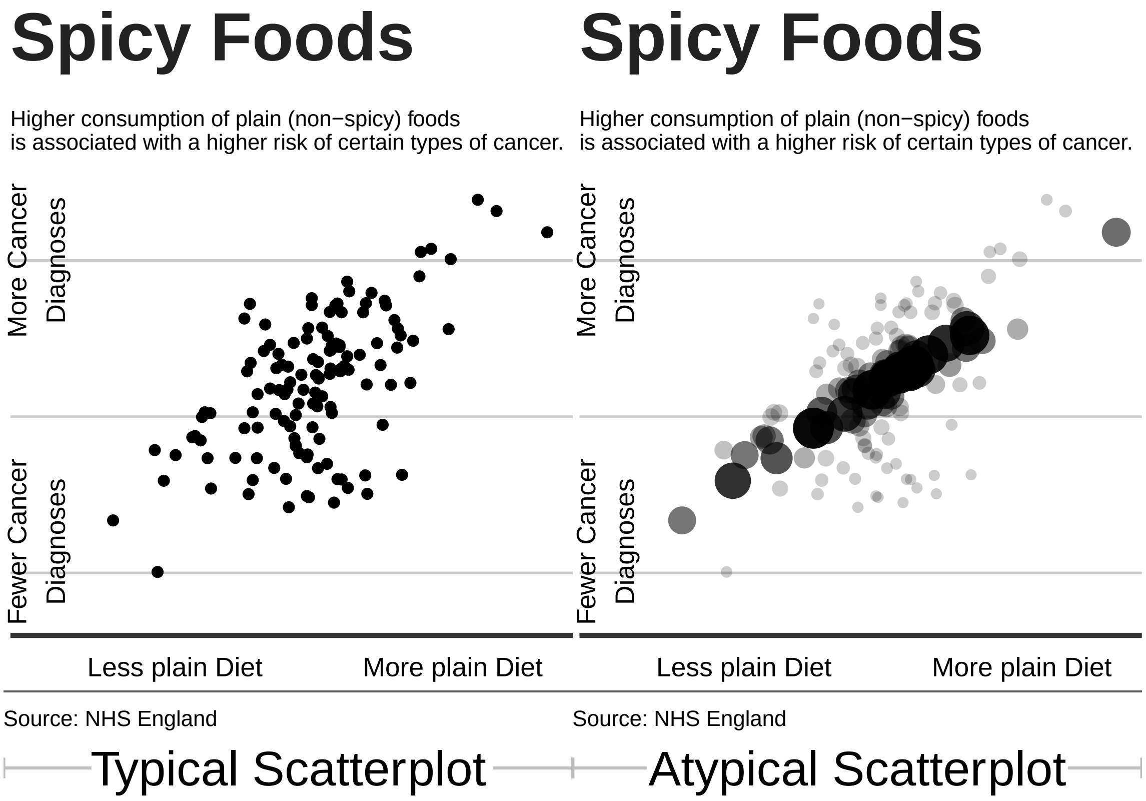

Fig. 1. Examples of the experimental stimuli used with an r value of 0.6.

5.3.3 Procedure. We use PsychoPy [14] to build our experiment and Pavlovia.org to host it. Participants were permitted to complete the experiment on a desktop or laptop computer. We elected to prevent participants from using a phone or tablet to complete the experiment in line with evidence that differences in on-screen sizes of data visualizations can alter perceptions [cleveland_1982]. Participants were first shown the participant information sheet and asked to provide consent through key presses in response to consent statements. They were, again, asked to provide their age and gender identity. Participants then completed the 12-item Defensive Confidence scale described by Albarracín and Mitchell [**?**] and the 5-item Subjective Graph Literacy scale [10][5]. Following instructions, which included descriptions of scatterplots and Pearson's r. In order to give legitimacy to our data visualizations with the hope of maximizing any potential belief change, participants were told that the graphs were taken from a well-known British news source, but that the identity of this source had been obscured. In order to promote participant engagement with the visualizations, participants were instructed to use a slider to estimate the correlation displayed in each scatterplot; no hypotheses were made based on these data, however they are discussed in light of recent literature in **?@sec-correlation-ratings**. Participants then had a chance to practice using the slider, before being asked their belief about topic emotionality and the relatedness between variables described in

[5] The inclusion of this scale was not specified in the pre-registration.

Table 3. Statistics for the significant main effect of rating time. Semi-partial R^2 is also incuded.

	Estimate	Standard Error	df	t-value	p	R^2
(Intercept)	4.98	0.083	151.69	59.76	<0.001	
Rating Time	-1.72	0.016	11849.00	-109.29	<0.001	0.295

our chosen statement. Following the completion of the experimental trials, participants were tested again on their beliefs about relatedness, and then debriefed that the data they saw were fictional. Interspersed among the experimental items were 4 attention check trials which explicitly asked participants to set the slider to 0 or 1.

5.4 Results

All analyses were conducted using R (version 4.4.1). The **buildmer** (version 2.11 [19]) and **lme4** (version 1.1-35.3 [3]) packages were used to build linear mixed effects models. Semi-partial R^2, which is presented in lieu of traditional measures of effect size, was calculated using the **r2glmm** package (version 0.1.2 [11]). To test the first hypothesis, that ratings of strength of relatedness would be different before and after participants viewed experimental items, we build a model whereby the rating is predicted by the point in time the participant made it. Our first hypothesis was supported; there was a significant difference in ratings of strength of relatedness made before and after participants viewed the experimental plots.

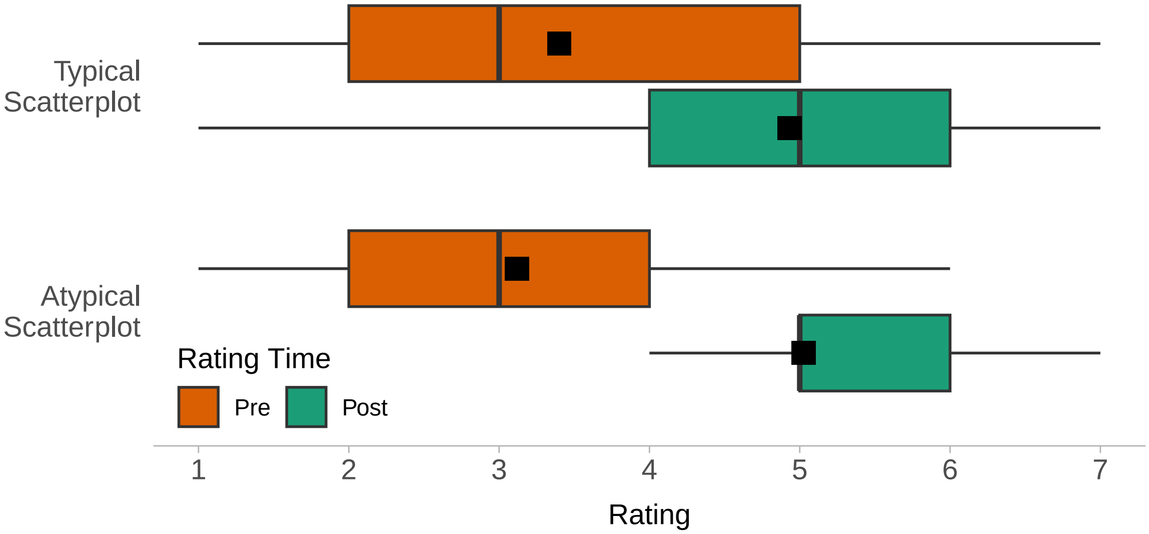

Fig. 2. Boxplots showing ranges, interquartile ranges, medians (vertical lines) and means for participants' ratings of strength of relatedness before and after viewing either typical or atypical scatterplots.

A likelihood ratio test revealed that the model including time of rating as a predictor explained significantly more variance than the null ($\chi^2(1) = 8{,}261.07$, $p < .001$). This model had random intercepts for participants. Statistical testing providing support for this hypothesis is shown in Table 3. Figure 2 shows

Table 4. Statistics for the significant main effect of condition on the difference between pre and post scatterplot viewing ratings for typical and atypical plots. Semi-partial R^2 is also incuded.

	Estimate	Standard Error	df	t-value	p	R^2
(Intercept)	1.72	0.022	2	78.22	<0.001	
Condition	0.37	0.044	5998	8.49	<0.001	0.012

means and boxplots for ratings of strength of relatedness before and after viewing scatterplots in either the typical or atypical condition.

Our second hypothesis, that the difference between ratings of strength of relatedness before and after viewing experimental plots would be greater when participants were assigned to the atypical scatterplot condition, also received support. We built a linear mixed effects model whereby the difference was predicted by the condition the participant was assigned to. A likelihood ratio test revealed that the model including condition as a predictor explained significantly more variance than the null ($F(1) = 72.07$, $p < .001$). This model contains no random effects terms. Statistics, along with semi-partial R^2 can be seen in Table 4.

5.4.1 Additional Analyses. We find no effect of participant's scores on a Defensive Confidence test ($F(1) = 3.22$, $p = .073$). We find a significant effect of participants' scores on a graph literacy test [10] ($F(1) = 15.78$, $p < .001$)

5.5 Discussion

6 GENERAL DISCUSSION

7 LIMITATIONS

8 FUTURE WORK

9 CONCLUSION

REFERENCES

[1] 2023. Prolific. https://www.prolific.co

[2] Paul Ayris, Isabel Bernal, Valentino Cavalli, Bertil Dorch, Jeannette Frey, Martin Hallik, Kistiina Hormia-Poutanen, Ignasi Labastida i Juan, John MacColl, Agnès Ponsati Obiols, Simone Sacchi, Frank Scholze, Birgit Schmidt, Anja Smit, Adam Sofronijevic, Jadranka Stojanovski, Martin Svoboda, Giannis Tsakonas, Matthijs van Otegem, Astrid Verheusen, Andris Vilks, Wilhelm Widmark, and Wolfram Horstmann. 2018. LIBER Open Science Roadmap. (July 2018). https://doi.org/10.20350/digitalCSIC/15061

[3] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[4] Nick Charalambides. 2021. We Recently Went Viral on TikTok - Here's What We Learned.

[5] Michael Chmielewski and Sarah C. Kucker. 2020. An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. *Social Psychological and Personality Science* 11, 4 (May 2020), 464–473. https://doi.org/10.1177/1948550619875149

[6] Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin* 70, 4 (1968), 213–220. https://doi.org/10.1037/h0026256

[7] Iliada Eleftheriou and Ajmal Mubarik. 2023. AI Code of Conduct. https://www.iliada-eleftheriou.com/AICodeOfConduct/#how-to-cite-and-reference-chatgpt.

[8] Joseph L. Fleiss, Jacob Cohen, and B. S. Everitt. 1969. Large Sample Standard Errors of Kappa and Weighted Kappa. *Psychological Bulletin* 72, 5 (Nov. 1969), 323–327. https://doi.org/10.1037/h0028106

[9] Matthias Gamer, Jim Lemon, and Ian Fellows Puspendra Singh <puspendra.pusp22@gmail.com>. 2019. *irr: Various Coefficients of Interrater Reliability and Agreement.* https://CRAN.R-project.org/package=irr R package version

0.84.1.

[10] Rocio Garcia-Retamero, Edward T. Cokely, Saima Ghazal, and Alexander Joeris. 2016. Measuring Graph Literacy without a Test: A Brief Subjective Assessment. *Medical Decision Making* 36, 7 (2016), 854–867. https://doi.org/10.1177/0272989X16655334

[11] Byron Jaeger. 2017. *r2glmm: Computes R Squared for Mixed (Multilevel) Models.* https://github.com/bcjaeger/r2glmm R package version 0.1.2.

[12] Dirk Merkel. 2014. Docker. *Linux Journal* (March 2014). https://doi.org/10.5555/2600239.2600241

[13] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2021. Data Quality of Platforms and Panels for Online Behavioral Research. *Behavior Research Methods* 54, 4 (Sept. 2021), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

[14] Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in Behavior Made Easy. *Behavior Research Methods* 51, 1 (Feb. 2019), 195–203. https://doi.org/10.3758/s13428-018-01193-y

[15] Gabriel Strain, Andrew J. Stewart, Paul Warren, and Caroline Jay. 2023. Adjusting Point Size to Facilitate More Accurate Correlation Perception in Scatterplots. In *2023 IEEE Vis X Vision*. IEEE, Melbourne, Australia, 1–5. https://doi.org/10.1109/VisXVision60716.2023.00006

[16] Gabriel Strain, Andrew J. Stewart, Paul Warren, and Caroline Jay. 2023. The Effects of Contrast on Correlation Perception in Scatterplots. *International Journal of Human-Computer Studies* 176 (Aug. 2023), 103040. https://doi.org/10.1016/j.ijhcs.2023.103040

[17] Gabriel Strain, Andrew J. Stewart, Paul A. Warren, and Caroline Jay. 2024. Effects of Point Size and Opacity Adjustments in Scatterplots. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3613904.3642127

[18] version 4. 2024. ChatGPT. OpenAI.

[19] Cesko C. Voeten. 2023. *buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression.* https://CRAN.R-project.org/package=buildmer R package version 2.11.