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, 18 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Utilized for communication in a wide variety of contexts, scatterplots are simple representations of (usually) bivariate data. They were estimated in 1983 to account for between 70 and 80 percent of data visualizations

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.

in scientific publications [**?**], and while there is no doubt that the range of data visualizations employed in and beyond science is now far broader, scatterplots remain an important tool for the data visualization designer. The appeal of researching scatterplots lies in evidence that people generally interpret them in similar ways [16], that they are interpreted rapidly [25], and that they are ubiquitous in both academic [**?**] and non-academic contexts.

While most commonly used to communicate the linear correlation, or level of relatedness, between a pair of variables, scatterplots can also be designed to facilitate the detection of outliers, to convey differences between clusters, or to display non-linear correlations. The suitability of scatterplots to such a range of tasks, and the opportunity for designers to design with a multitude of tasks in mind, plays a large part in their popularity. Building on previous work, we elect to focus on the use of scatterplots for the communication of linear, bivariate, positive correlation. There is evidence that, while correlation perception in scatterplots experiences low levels of interindividual variance (especially when compared to other visualizations that communicate the same idea [14, 16]), our accuracy in interpretation is poor. Studies asking participants to numerically estimate correlation [4, 7, 9, 17, 18, 21, 27] or estimate it via a bisection task [26] find consistent levels of underestimation, particularly when $0.2 < r < 0.6$. If scatterplots were used solely for communication between those trained in statistics and data visualization, this would not be especially problematic, however this is not the case; lay people being expected to be able to use and interpret data visualizations on an almost daily basis. It is thus the duty of those who design such visualizations to design with the naive, inexperienced viewer in mind. Doing so requires us to understand *how* visualizations work, and to gain an appreciation for the hidden processes that allow pictorial representations to convey much more than words and numbers ever could.

Recent work has sought to address the correlation underestimation bias in positively correlated scatterplots through the use of novel point encodings. In 2023 and 2024, Strain et al. [28–30] began exploiting the idea that viewers use the width of a probability distribution conveyed by the arrangement of scatterplot points as a proxy for their judgements of correlation to successfully (albeit partially) correct for the underestimation bias. As of the time of writing, this work has only provided evidence about perceptual effects using a simple direct estimation paradigm, and while successful, has not investigated whether these techniques can influence cognition in the context of real-world data visualizations and the relatedness between variables. As doing this is crucial for providing designers with the tools to design visualizations that are adapted for the facilitation of more accurate correlation perception, we therefore present a two experiment study investigating the propensity for recently established scatterplot visualization techniques to bias participant's beliefs about levels of relatedness between variables.

2 RELATED WORK

In this section we briefly discuss related work on correlation perception and estimation, the history and current state of the use of point size and opacity adjustments in scatterplots, including how these visual features have been used with regards to correction for the underestimation bias, and perception and cognition in data visualization. We then review the literature around belief change as it pertains to data visualization, before ending with our hypotheses for the present study.

2.1 Correlation Perception

Correlation describes the level of relatedness between a number of variables. There are a number of different types of correlation, however in this work, we use the term to refer to Pearson's r, which takes a positive or negative value between 0 and 1 depending on the direction of the relationship being described. Mechanistically, evidence is inconclusive regarding what drives correlation perception in scatterplots, however some experimental results point towards the shape of the underlying probability distribution as a likely candidate. Scatterplots with smaller point clouds produce increased judgements of correlation [7], suggesting that it is the area of the point cloud that may influence perception. Work exploring the relationship between subjective and objective r values in scatterplots found that this relationship could be described by a power function that included the mean of the geometrical distances between scatterplot points and a regression line. Other work includes some representation of the shape of a scatterplot's point cloud in equations describing magnitude estimation and correlation discrimination [22, 26], and work on visual features as proxies for correlation found that a similar quantity again is predictive of performance. While we cannot say that this *is* the process of correlation perception in scatterplots, we can conclude that the shape of the point cloud is a good proxy for what is really occurring during judgements of correlation.

2.2 Scatterplots: Size, Opacity, and Recent Developments

Changing the opacities and sizes of points in scatterplots are standard practices. Regarding opacity, this is often uniformly lowered to address overplotting [19] issues that arise when visualizing very large datasets. Similarly, scatterplots of large datasets tend to have smaller points to maintain individual point discriminability. Point size has also been used to encode an additional third variable in what is known as *bubble charts*. Despite these techniques being established, there is relatively little experimental work on the effects of changing point opacities and sizes on correlation estimation. Some studies find that correlation estimation is invariant to changes in point opacities and sizes [25, 26], while more recent work reports strong effects of the systematic adjustment of each visual feature [28–30].

The idea that it is the shape of the point cloud, and the probability distribution it represents that is being used to inform judgements of correlation has received support from recent work exploiting visual features with the intention of making correlation estimation more accurate. Strain et al. [28–30], changed the sizes and opacities of points in scatterplots as a function of their distance from the regression line, and achieved success in biasing correlation estimates in positive and negative directions. When point opacities [29] or point sizes [28] were reduced with residual distance, participants were significantly more accurate on a correlation estimation task; employing both of these manipulations simultaneously [30] resulted in an overshoot of correction, biasing participants further. Figure 1 contains a summary of previously tested scatterplot manipulations and their effects on performance on a correlation estimation task.

Ultimately, it is not the aim of the present work to comment on what truly drives correlation perception in scatterplots, but rather to attempt to extend previously tested techniques to influence cognitions about relatedness.

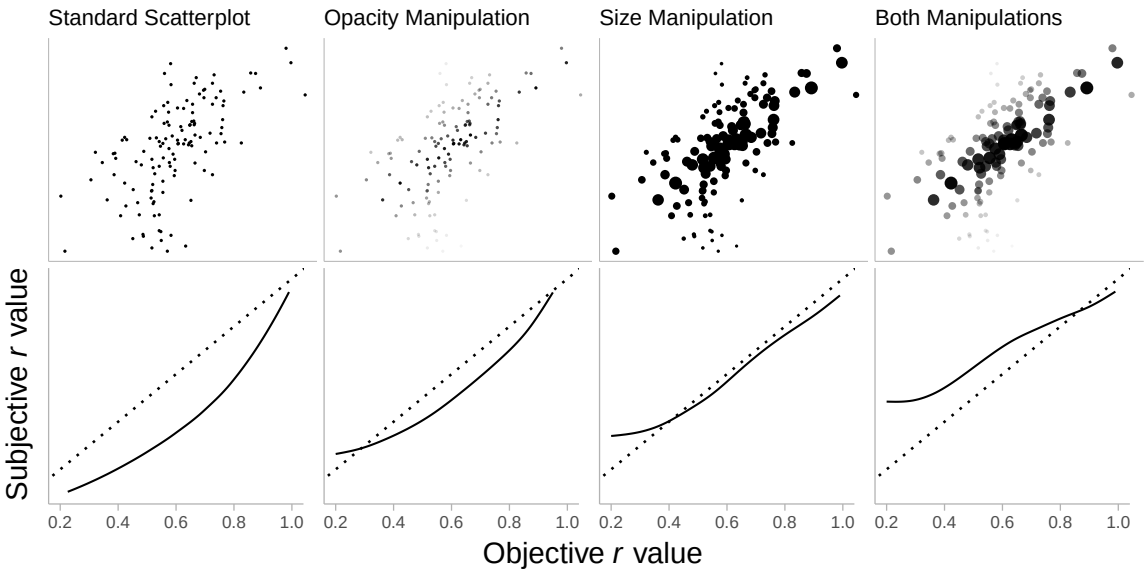

Fig. 1. Top row: Examples of scatterplot manipulations from previous work using an r value of 0.6. Bottom row: the corresponding correlation estimation behaviour across values of r between 0.2 and 0.99. The dashed diagonal line represents perfect estimation.

2.3 Perception & Cognition in Data Visualization

Interacting with data visualization is a complex process involving bottom-up and top-down mechanisms [**? ? ?**]. Previous work investigating alternative scatterplot designs has confined itself to exploring perception only; here we introduce the potential for top-down effects to bias participants. Despite this, we take measures to ensure that these potential top-down effects are minimised, namely by crowdsourcing the creation of our stimuli and by including by-participant effects in our modelling. Nevertheless, we acknowledge that we introduce the potential for murkiness regarding our conclusions; we argue this is necessary to gain a deeper understanding of how we interact with data visualizations and how designers can improve their craft.

2.4 Persuasion & Belief Change in Data Visualization

Data visualization is a powerful tool. After all, if numerical data were sufficient for understanding, there would be no need to visualize data beyond aesthetic preference. Patter recognition, attention, and familiarity are all aspects of human perception and cognition that can be exploited by visualization designers to facilitate more efficient, enjoyable, and effective communication [**?**]. This power is a double-edged sword, however; poor design, be it malevolent or misguided, can cause distrust, confusion, and misunderstanding amongst viewers. It is for both of these reasons that we choose to study belief change in scatterplots as a consequence of alternative designs. Scatterplots, like many other data visualizations, have been admitted as evidence in court cases [4], and play key roles in organizational decision making, including in healthcare [**?**]. It is a reasonable assumption to make that data visualizations, including scatterplots, are used to make decisions that result in positive or negative outcomes with regards to health and policy more generally. Studying the potential for new designs to alter beliefs about relatedness facilitates not only the exploitation

of these new designs for better visualizations, but also allows us to understand how these designs might be used by malevolent actors. To this end, we present a two experiment study. First, we first use crowdsourcing to select part of our experimental stimuli, then we test the propensity for previously established alternative scatterplot designs to alter beliefs about relatedness, taking into account a range of additional participant qualities.

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 [20] 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 [5, 6, 23], so we follow published guidelines [23] 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.

[1]The GitHub repository associated with this paper contains an R script that performs this analysis.
[2]removed for anon
[3]removed for anon

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.

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 those whose topics were both strong and neutral in emotional valence, we began by using ChatGPT4 [31] 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 [10]. 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 [12]), 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 [8, 11].

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

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.

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.

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 [24] 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] https://gitlab.pavlovia.org/Strain/beliefs_scatterplots_pretest

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 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

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 [28–30]. 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 [30]. 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, their mean defensive confidence score was 43.0 ($SD = 6.8$) out of 60, and their mean rating of topic emotionality was 2.9 ($SD = 1.3$) on a 7 point Likert scale. 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

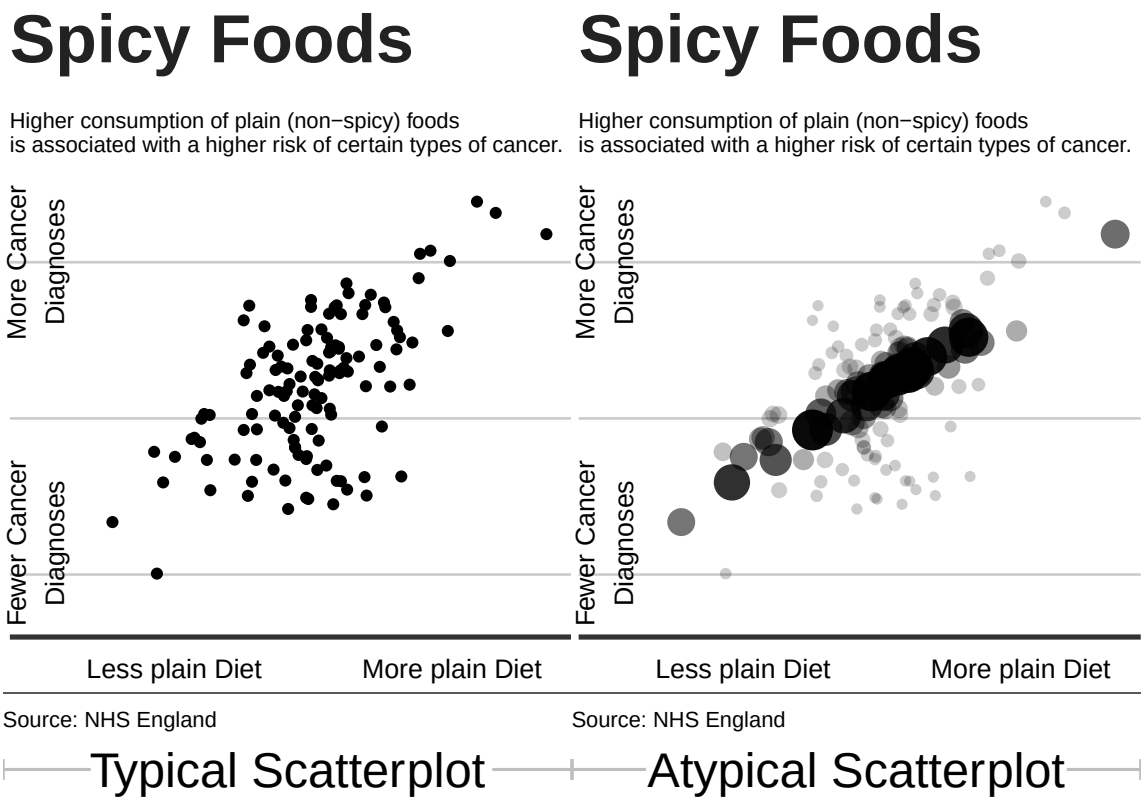

Fig. 2. Examples of the experimental stimuli used with an r value of 0.6.

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 [5] [6]

5.3.3 Procedure. We use PsychoPy [24] 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 [13][7]. 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

[5]The GitHub repository associated with this paper contains an R script that performs this analysis.
[6]removed for anon
[7]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

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 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 [32]) 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 [15]). 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.

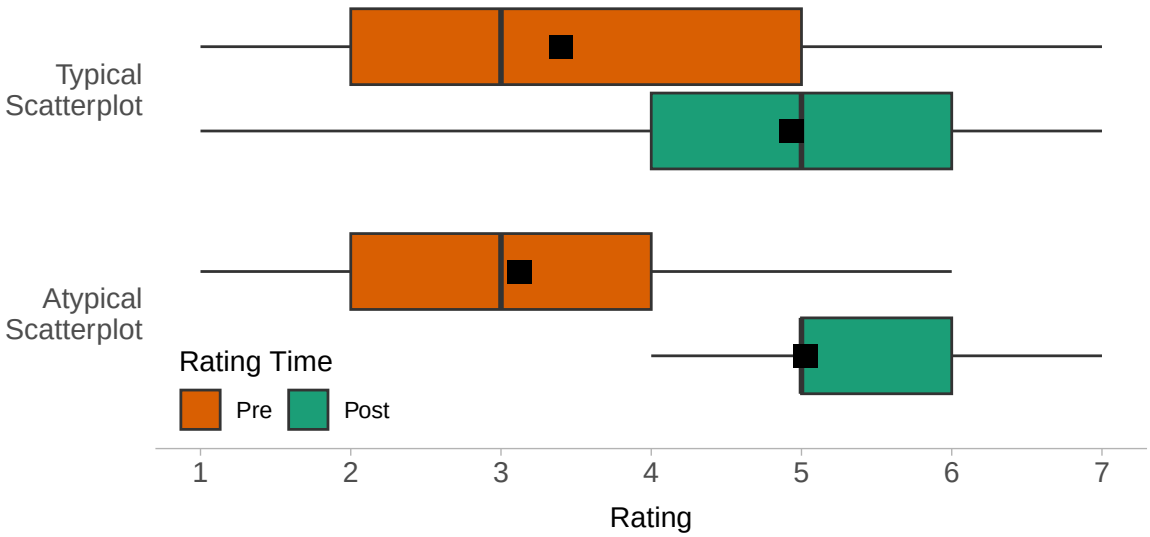

Fig. 3. Boxplots showing ranges, interquartile ranges, medians (vertical lines) and means for participants' ratings of strength of relatedness before and after viewing either typical or atypical scatterplots.

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

Table 5. Statistics for main effects and interactions when scores on the Subjective Graph Literacy test are included in the model.

	Estimate	Standard Error	df	t-value	p	R^2
(Intercept)	2.00	0.114	4	17.53	<0.001	
Condition	-1.36	0.228	4	-5.97	<0.001	0.006
Literacy	-0.01	0.005	4	-2.81	0.005	0.001
Condition * Literacy	0.08	0.010	5996	7.91	<0.001	0.01

Table 6. Statistics for main effects and interactions when scores on the Defensive Confidence test are included in the model.

	Estimate	Standard Error	df	t-value	p	R^2
(Intercept)	2.15	0.146	4	14.75	<0.001	
Condition	-1.51	0.292	4	-5.17	<0.001	0.004
Defensive Confidence	-0.01	0.003	4	-3.13	0.002	0.002
Condition * Defensive Confidence	0.04	0.007	5996	6.57	<0.001	0.007

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 3 shows 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. Test statistics, along with semi-partial R^2 can be seen in Table 4.

5.4.1 Additional Analyses. We find effects of participants' scores on a Defensive Confidence test ($F(2) = 23.23$, $p < .001$), participants' scores on a graph literacy test [13] ($F(2) = 39.23$, $p < .001$), and of how emotional participants' rated the chosen correlative statement before beginning the block of trials ($F(2) = 5.45$, $p = .004$). There are significant fixed and interactive effects present for both graph literacy and defensive confidence. Table 5 and Table 6 contain statistics, including partial R^2 for both factors respectively. We found no significant fixed effect of emotionality ratings, but an interaction was observed, which can be see in Table 7. Given the low estimates and effects sizes associated with the fixed effects of defensive

Table 7. Statistics for main effects and interactions when likert ratings of topic emotionality are included in the model.

	Estimate	Standard Error	df	t-value	p	R^2
(Intercept)	1.79	0.054	4	33.48	<0.001	
Condition	0.68	0.107	4	6.33	<0.001	0.007
Topic Emotionality Rating	-0.03	0.017	4	-1.58	0.115	0
Condition * Topic Emotionality Rating	-0.11	0.034	5996	-3.15	0.002	0.002

confidence and graph literacy, and the the absence of a fixed effect of emotionality, we instead choose to focus on the interactions themselves; see Section 5.5.1 for further discussion of these effects.

5.5 Discussion

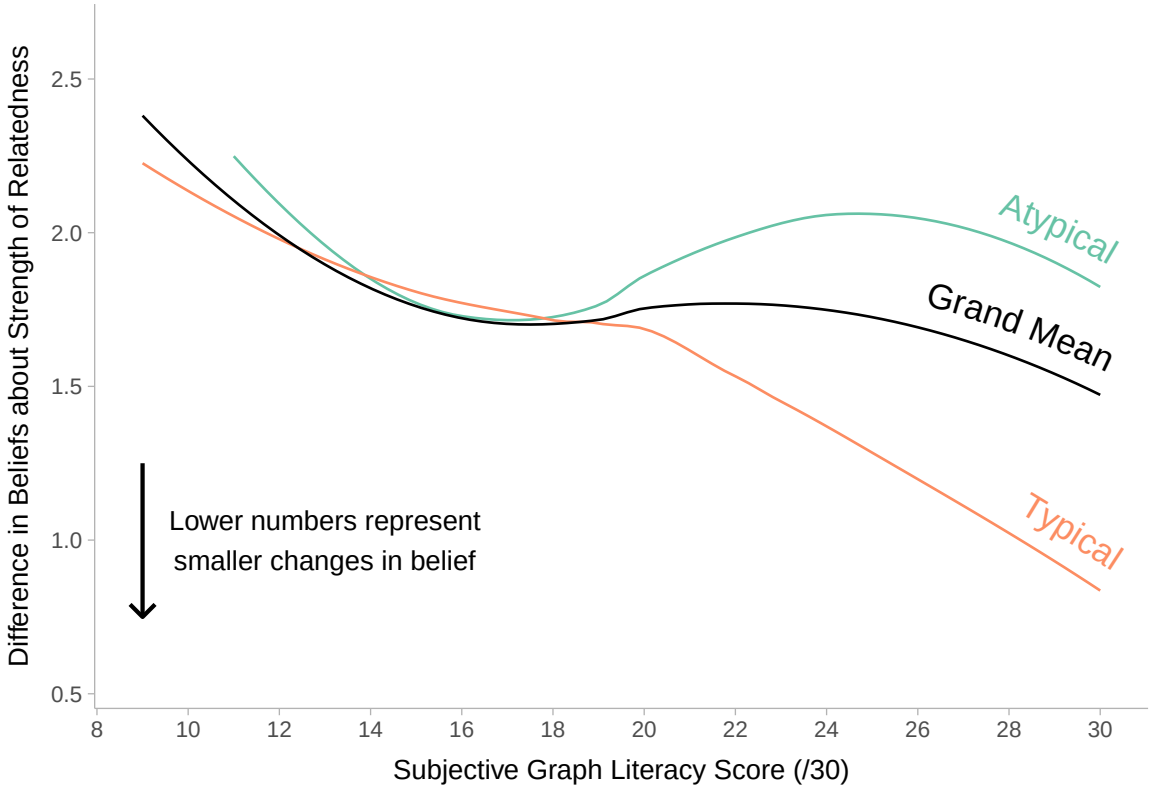

Fig. 4. Illustrating how differences in beliefs about strength of relatedness change as a function of partcipants' scores on the subjective graph literacy test; smoothed curves are shown for the grand mean, as well as separately for atypical and typical scatterplot presentation conditions.

5.5.1 Graph Literacy, Defensive Confidence, and Topic Emotionality. Mean differences in pre and post plot-viewing ratings of strength of relatedness can be seen in Figure 4. Generally, participants who scored higher on a graph literacy test experienced smaller changes in their ratings of strength of relatedness. While the

effect size associated with this interaction is small (~0.01, see Table 5), and does not alter the overall pattern of results, it is in line with previous work suggesting that those with higher levels of graph or visualization literacy show better performance in inference tasks related to visualizations [**?**], are more capable of describing effects that visualisations aim to communicate [**?**], and are able to preferentially attend to relevant features of visualizations to a greater degree [**?**], than those with lower levels of graph literacy. In the present study, we provide evidence that those with greater levels of graph literacy are *less susceptible* to having their beliefs changed by visualizations, although this difference is somewhat negated if participants viewed atypical scatterplots, in which case levels of belief change were relatively consistent. We hypothesize that this is due to the non-standard nature of the atypical scatterplots used in that condition.

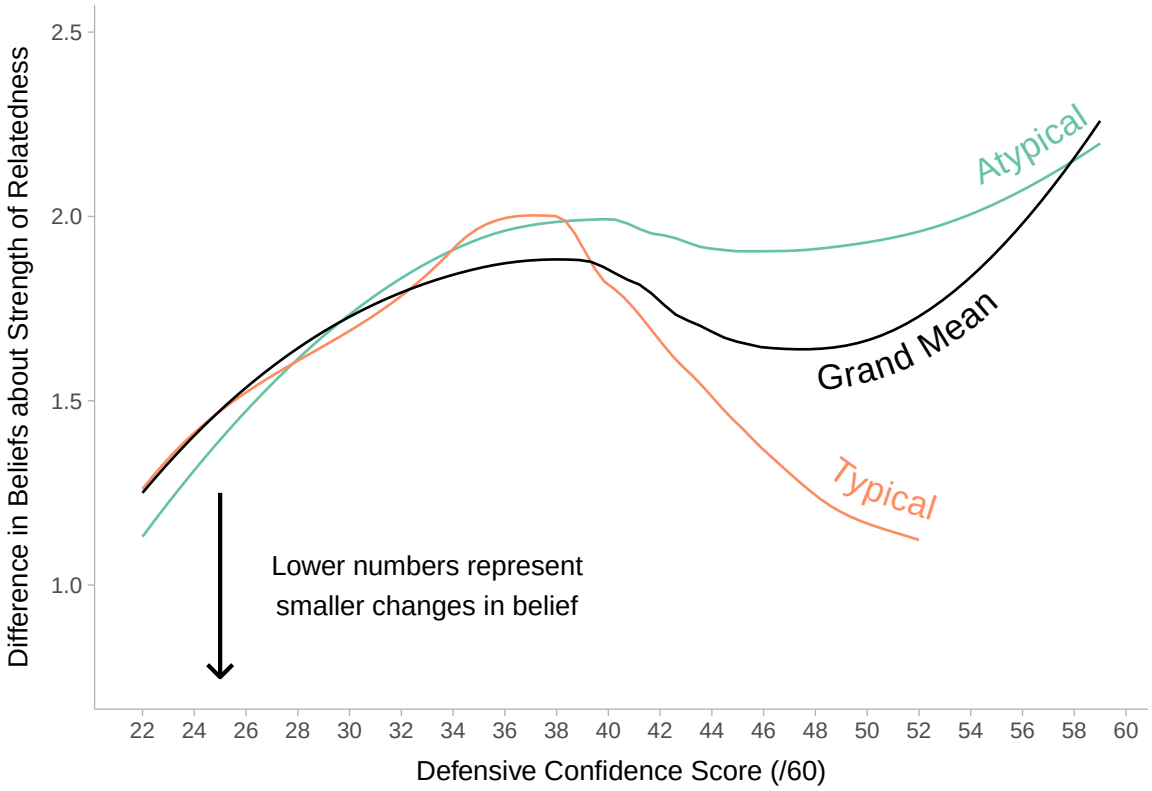

Fig. 5. Illustrating how differences in beliefs about strength of relatedness change as a function of partcipants' scores on the defensive confidence test; smoothed curves are shown for the grand mean, as well as separately for atypical and typical scatterplot presentation conditions.

We observe an opposing pattern of results when examining the effects of defensive confidence on participants' propensity for belief change. Generally, participants who scored more highly on the defensive confidence test experienced greater levels of belief change. This is in line with evidence that those who are more confidence in their ability to defend their own beliefs are more liable to having those beliefs changed in light of evidence [**?**]. An extended analysis of the defensive confidence data collected by Markant et al., [**?**] revealed a similar pattern of results; participants who scored more highly on a defensive confidence test

experienced greater levels of belief change after viewing scatterplot visualizations[8], although the effect in that case is much less pronounced due to differences in study design. This effect is explained as being due to those with a greater degree of confidence in their own ability to defend their ideas engage with information with lower levels of attention to the fact it opposes their beliefs. The present study provides evidence in favour of this phenomenon.

While the general pattern of results is expected based on previous work, the interaction present between defensive confidence and scatterplot condition is novel (see Figure 5). It would appear that despite following the normal pattern of results for low to moderate levels of defensive confidence, those participants who viewed the typical scatterplots experienced a drop in belief change as defensive confidence increased past ~ 36/60. There are several potential explanations in the data; firstly, that the 75 participants who formed the typical scatterplot presentation group had a more restricted range of defensive confidence scores, topping out at 52; and secondly, that the unfamiliar nature of the atypical scatterplots was protective against against an unexpected, standard behaviour whereby very high levels of defensive confidence decrease susceptibility to belief change.

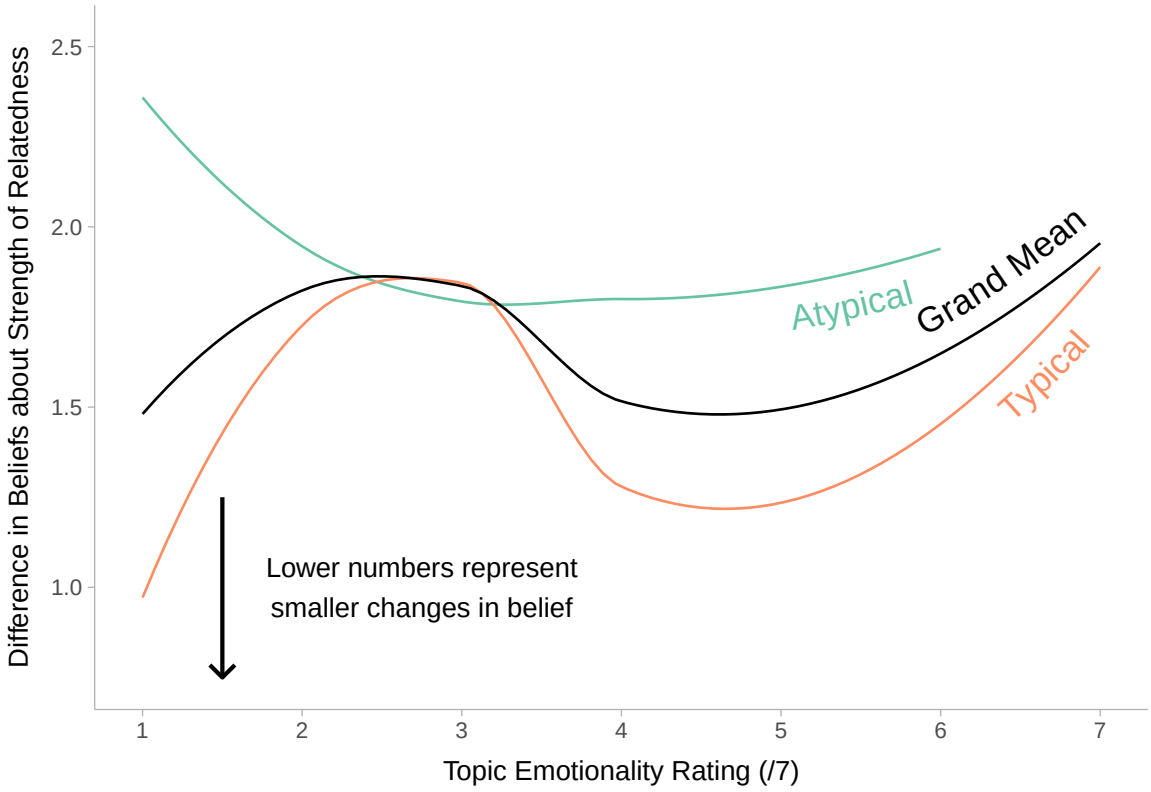

Fig. 6. Illustrating how differences in beliefs about strength of relatedness change as a function of partcipants' scores on the subjective graph literacy test; smoothed curves are shown for the grand mean, as well as separately for atypical and typical scatterplot presentation conditions.

[8]The GitHub repository associated with this paper contains an R script that performs this analysis.

Examining Figure 6, it would appear that the interaction we see between belief change and topic emotionality is driven by those participants who rated the statement low on emotionality having different levels of belief change between atypical and typical scatterplot presentation conditions; in this case levels of belief change were much higher for the group that viewed the atypical scatterplot designs. There is a broad research space regarding emotionality and data visualization [**?**], and it is clear from previous work that emotion affects perception, cognition, and behaviour [**? ?** , probably look for more] with regard to data visualization. Harrison et al. [**?**] _2013] found that participants who were positively primed performed better on a low-level visual judgement task. Comparison of this work to the current is difficult, as *success* on our task is hard to define. From our data, it is unclear why participants who rated the correlative statement as "strongly negative" experienced significantly higher levels of belief change when they viewed atypical scatterplots. We decided to capture the emotionality of the statement simply because we had done so in the pre-study; we made no predictions based on it, and chose a broad measure of emotionality, meaning we did not capture intensity separately from valence. It may be that the strong negative emotion amplified the effect of scatterplot condition in a way that strong positive emotion did not, although it is unclear why this has occurred. Further experimental work is required to provide concrete explanations for the interactive effects of graph literacy, defensive confidence, and statement emotionality in the current experimental paradigm; as the investigation of these is not the prime goal of the present study, however, we do not consider this a shortcoming.

6 GENERAL DISCUSSION

7 LIMITATIONS

8 FUTURE WORK

- maybe something about attitudes
- understanding how DC and literacy interact
- extending beliefs and attitudes to behavioural change (golden ticket really)

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] Philip Bobko and Ronald Karren. 1979. The Perception of Pearson Product Moment Correlations from Bivariate Scatterplots. *Personnel Psychology* 32, 2 (1979), 313–325. https://doi.org/10.1111/j.1744-6570.1979.tb02137.x

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

[6] 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

[7] W. S. Cleveland, P. Diaconis, and R. Mcgill. 1982. Variables on Scatterplots Look More Highly Correlated When the Scales Are Increased. *Science* 216, 4550 (June 1982), 1138–1141. https://doi.org/10.1126/science.216.4550.1138

[8] 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

[9] Charles E. Collyer, Kerrie A. Stanley, and Caroline Bowater. 1990. Psychology of the Scientist: LXIII. Perceiving Scattergrams: Is Visual Line Fitting Related to Estimation of the Correlation Coefficient? *Perceptual and Motor Skills* 71, 2 (Oct. 1990), 371–378E. https://doi.org/10.2466/pms.1990.71.2.371

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

[11] 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

[12] 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.

[13] 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

[14] L. Harrison, F. Yang, S. Franconeri, and R. Chang. 2014. Ranking Visualizations of Correlation Using Weber's Law. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 1943–1952. https://doi.org/10.1109/TVCG.2014.2346979

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

[16] Matthew Kay and Jeffrey Heer. 2015. Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation. *IEEE transactions on visualization and computer graphics* 22, 1 (Aug. 2015), 469–478. https://doi.org/10.1109/TVCG.2015.2467671

[17] David Lane, Craig Anderson, and Kathryn Kellam. 1985. Judging the Relatedness of Variables. The Psychophysics of Covariation Detection. *Journal of Experimental Psychology: Human Perception and Performance* 11 (Oct. 1985), 640–649. https://doi.org/10.1037/0096-1523.11.5.640

[18] Thomas W. Lauer and Gerald V. Post. 1989. Density in Scatterplots and the Estimation of Correlation. *Behaviour & Information Technology* 8, 3 (June 1989), 235–244. https://doi.org/10.1080/01449298908914554

[19] Justin Matejka, Fraser Anderson, and George Fitzmaurice. 2015. Dynamic Opacity Optimization for Scatter Plots. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 2707–2710. https://doi.org/10.1145/2702123.2702585

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

[21] Joachim Meyer and David Shinar. 1992. Estimating Correlations from Scatterplots. *Human Factors* 34, 3 (June 1992), 335–349. https://doi.org/10.1177/001872089203400307

[22] Joachim Meyer, Meirav Taieb, and Ittai Flascher. 1997. Correlation Estimates as Perceptual Judgments. *Journal of Experimental Psychology: Applied* 3, 1 (1997), 3–20. https://doi.org/10.1037/1076-898X.3.1.3

[23] 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

[24] 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

[25] Ronald A. Rensink. 2014. On the Prospects for a Science of Visualization. In *Handbook of Human Centric Visualization*, Weidong Huang (Ed.). Springer New York, New York, NY, 147–175. https://doi.org/10.1007/978-1-4614-7485-2_6

[26] Ronald A. Rensink. 2017. The Nature of Correlation Perception in Scatterplots. *Psychonomic Bulletin & Review* 24, 3 (2017), 776–797. https://doi.org/10.3758/s13423-016-1174-7

[27] Robert F. Strahan and Chris J. Hansen. 1978. Underestimating Correlation from Scatterplots. *Applied Psychological Measurement* 2, 4 (Oct. 1978), 543–550. https://doi.org/10.1177/014662167800200409

[28] 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

[29] 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

[30] 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
[31] version 4. 2024. ChatGPT. OpenAI.
[32] 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.