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

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 approach to stimulus generation.

Authors' Contact Information: Gabriel Strain, Department of Computer Science, Faculty of Science and Engineering, University of Manchester, Manchester, United Kingdom; Andrew J. Stewart, Department of Computer Science, Faculty of Science and Engineering, University of Manchester, 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, 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, Manchester, United Kingdom; Caroline Jay, Department of Computer Science, Faculty of Science and Engineering, University of Manchester, Manchester, United Kingdom.

Manuscript submitted to ACM

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 [9] 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 [3, 4, 10], so we follow published guidelines [10] 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.

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 those whose topics were both strong and neutral in emotional valence, we began by using ChatGPT4 [12] 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 [6]. 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 [8]), in order to penalise larger magnitude disagreements more harshly. We found agreement

[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.

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.

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 [5, 7].

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.

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 visualisations 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 [11] 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 4.4.0). 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] https://gitlab.pavlovia.org/Strain/beliefs_scatterplots_pretest

4.5 Discussion

Fleiss' Kappa values for interrater agreement on both topic emotionality and strength of correlation scales are low (κ = 0.07 and κ = 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.

- don't forget about reverse scoring items

5.2 Stimuli

- build experiment and pilot to see how many people can do
- then this section can be written

5.3 Method

5.3.1 Participants.

5.3.2 Design. We employ 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.

5.3.3 Procedure.

5.4 Results

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

[4] 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

[5] 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

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

[7] 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

[8] 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.

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

[10] 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

[11] 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

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