The Effects of Visual and Design Features on the Perception of Correlation in Scatterplots

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Science and Engineering

2024

Gabriel Strain
Department of Computer Science

Contents

Word count: 1000

List of figures

List of tables

List of publications

Publications go here.

Abstract

put abstract here

Lay abstract

This is lay abstract text.

Declaration of originality

I hereby confirm that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

i The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made *only* in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see `http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420`), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see `http://www.library.manchester.ac.uk/about/regulations/`) and in The University's policy on Presentation of Theses.

Acknowledgements

Acknowledgements go here.

Chapter 1

Introduction

1.1 Research Motivation

1.2 Contributions

1.3 Included Publications

The research described in chapters 4, 5, 6, and 7 in this thesis is adapted from earlier publications, the last of which under review as of writing. To avoid repetition, information and discussion that would be repeated has been consolidated into the literature review and general methodology chapters. *Gabriel Strain* is the primary author of all included papers.

- *The Effects of Contrast on Correlation Perception in Scatterplots* [46] is reproduced in Chapter 4. Sections 4.5.2, 4.6.2, 4.5.3, 4.6.3, 4.5.4, 4.6.4, and4.7 contain minimally altered parts of the published article.
- *Adjusting Point Size to Facilitate More Accurate Correlation Perception in Scatterplots* [45] is reproduced in Chapter 5. Sections 5.4.2, 5.4.3, 5.4.4, and 5.5 contain minimally altered parts of the published article.
- *Effects of Point Size and Opacity Adjustments in Scatterplots* [47] is reproduced in Chapter 6. Sections 6.4.2, 6.4.3, 6.4.4, and 6.5 contain minimally altered parts of the published article.
- *Effects of Alternative Scatterplot Designs on Belief* (*under review*) is reproduced in Chapter 7. Sections 7.4, 7.5.2, 7.5.3, 7.5.4, and 7.6 contain minimally altered parts of the published article.

1.4 Overview of Thesis

Chapter 2

Literature Review

2.1 Data Visualisation: A Brief History

2.2 Measuring Relatedness

2.3 Conceptions of Correlation

2.4 Visualising Correlation

2.4.1 History

2.4.2 Present Landscape

2.4.3 Scatterplots

2.5 Correlation Perception

2.6 Correlation Cognition

2.7 Underestimation: What's Really Going On?

2.8 Underestimation: Potential Consequences

2.9 Data Visualisation Literacy

2.10 Objectives and Contributions

Chapter 3

General Methodology

3.1 Introduction

In this chapter we describe our research methodologies. Chapters 4, 5, and 6 share most aspects of experimental method, while the experiment described in chapter 7 differs substantially. Throughout this chapter, the reader should assume that we are referring to the entire body of experimental work this thesis describes. Methods that differ regarding the final experiment in chapter 7 are detailed along the way. In this chapter, we discuss our experimental designs, the tools we use to build and run our experiments, our approach to statistical analyses, and the computational methods and practices we employed, particularly with regards to reproducibility and open science.

3.2 Experimental Methods

It is important to acknowledge that the way in which we conduct experiments influences what we find and the conclusions that we may draw from those findings. The decisions that lead us to designing experiments in certain ways must be based not only on theory, but also on the practical constraints imposed by external factors on the research team. Concerns such as time, convenience, and cost must be addressed, and a compromise between research that is *valuable* and research that is *doable* must be reached. We focused on pragmatism and impact throughout the course of this research project; happily, the research journey we embarked on resulted in methodologies that satisfied both principles. It is for this reason that we consider the framework we present to be a key contribution of this thesis.

3.2.1 Experimental Design

All but our final experiment utilised within-participants designs. Each participant saw all experimental stimuli and provided a judgement of correlation using a sliding scale between 0 and 1 (see Figure 3.1). Experiments 1 to 3 featured featured a single experimental factor of design, each with 4 levels corresponding to scatterplots with different design features. Experiment 4 employed a factorial 2×2 design. Experiment 5 is a departure from the shared experimental paradigm of the previous experiments, and features a 1 factor, 2 level between-participants design.

Figure 3.1. An example of the slider participants used to estimate correlation in experiments 1-4.

3.2.2 Tools for Testing

Whatever the design of our experiments, software plays a crucial role in allowing us to carry them out. Fortunately, at the time of writing, there is a wealth of tools available to facilitate the testing of visualisations both in traditional lab-based tests and in online experiments. As we adhere to the principles of open and reproducible research [6], we discount closed-source software, such as Gorilla [4] or E-prime [17], as these rely on paid licenses and do not allow us to share code with future researchers. We settled on using PsychoPy [35] due to its open-source status, flexibility regarding graphical and code-based experimental design, and high level of timings accuracy [11]. Using such a open-source tool not only facilitated our own learning with regard to experiment building, but also enables to contribute further examples of visualisation studies by hosting the resulting experiments online for use and modification by future researchers.

We elected to pursue online testing throughout this thesis. Doing so is much quicker than carrying out in-person lab-based testing, meaning we can collect data from a much larger number of participants. This reduces the chances of detecting false positives during analysis and ensures adequate levels of power despite the potential for small effects sizes. Online testing also affords us access to diverse groups of participants across our populations of interest, especially when compared to the relatively homogeneous student populations usually accessed by doctoral researchers. Research has identified online experimentation as producing reliable results that closely match those found in traditional lab-based experiments [5, 18, 39], especially with large sample sizes. Due to its integration with PsychoPy, we chose to use Pavlovia to host all the experiments described in this thesis. Section 3.5.4 contains links to all experiments publicly hosted on Pavlovia's GitLab instance.

3.2.3 Recruitment & Participants

Recruitment of participants online is possible through a range of service providers, each with advantages and disadvantages. Research evaluating a number of these providers recently found that Prolific [40] and CloudResearch provide the highest quality data for the lowest cost [16]; we elected to use the former due to familiarity with the system. Despite these findings, there has also been evidence of low data quality and skewed demographics affecting even high quality platforms tailored towards academic research. On the 24th of July, 2021, the Prolific.co platform went viral on social media [13], leading to a participant pool heavily skewed towards young people identifying as female. At the time, Prolific did not manually balance the participants recruited for a study. We addressed this in our pilot study (see Section 4.2) by preventing participants who joined after this date from participating, in addition to manually requesting a 1:1 ratio of male to female participants. The demographic issues settled quickly, however we maintained our screened 1:1 ratio for the remainder of the experiments.

The first experiment we conducted was a pilot study (see Section 4.2 for full details) investigating a very early iteration of the point opacity manipulation in combination with exploratory work around plot size and correlation estimation. At the time, the author was relatively naive to the intricacies of recruiting research participants online, and thus we experienced issues regarding participant engagement. Each experiment included attention check questions in which participants were instructed to ignore the stimulus and provide a specific answer. We stated in the advert for each experiment that failure of more than 2 attention check items would result in a submission being rejected. This pilot study suffered from a rejection rate of 57.5%, indicating that we were experiencing very low levels of participant engagement. For our following studies, we therefore followed published guidelines [34] to address these issues; specifically, we required that participants:

- Had previously completed at least 100 studies on Prolific.
- Had an acceptance rate of at least 99% for those studies. [1]

Following implementation of these pre-screen criteria, the rejection rate for our next experiment fell to ~5%. Rejection rates were similar for the remainder of our experiments. Exact numbers of accepted and rejected participants can be found in the **Participants** sections of each experiment.

3.2.4 Creating Stimuli

All our stimuli were created using `ggplot2` in R. Specific versions are cited separately with regard to the specific visualisations produced for each experiment. We followed identical principles regarding data visualisation design for each experiment bar the last, which is discussed *in situ*.

We designed with the intention of isolating and addressing a perceptual effect; the underestimation of correlation in positively correlated scatterplots. For this reason, we sought to remove the potential for other design factors to have effects on correlation estimation. To this end, we removed most of the conventionally present visual features of scatterplots, including axis labels, tick labels, grid lines, and titles. We elected to preserve the axis ticks themselves. Figure 3.2 demonstrates the basic design of the scatterplots used in experiments 1 to 4.

3.3 Analytical Methods

3.3.1 Linear Mixed-Effects Models

To investigate whether the experimental manipulations we test have actual effects on the interpretations participants provide, we must employ appropriate statistical testing. This involves taking into account the variability in responses that can be attributed to an experimental experimental against the backdrop of other variability inherent in the dataset. To accomplish this, we utilise linear mixed-effects modelling, a broadly applicable and reliable approach that is also resistant to a variety of distributional assumption violations [44].

[1] this is a more strict rate than the 95% recommended by Peer et al. [34].

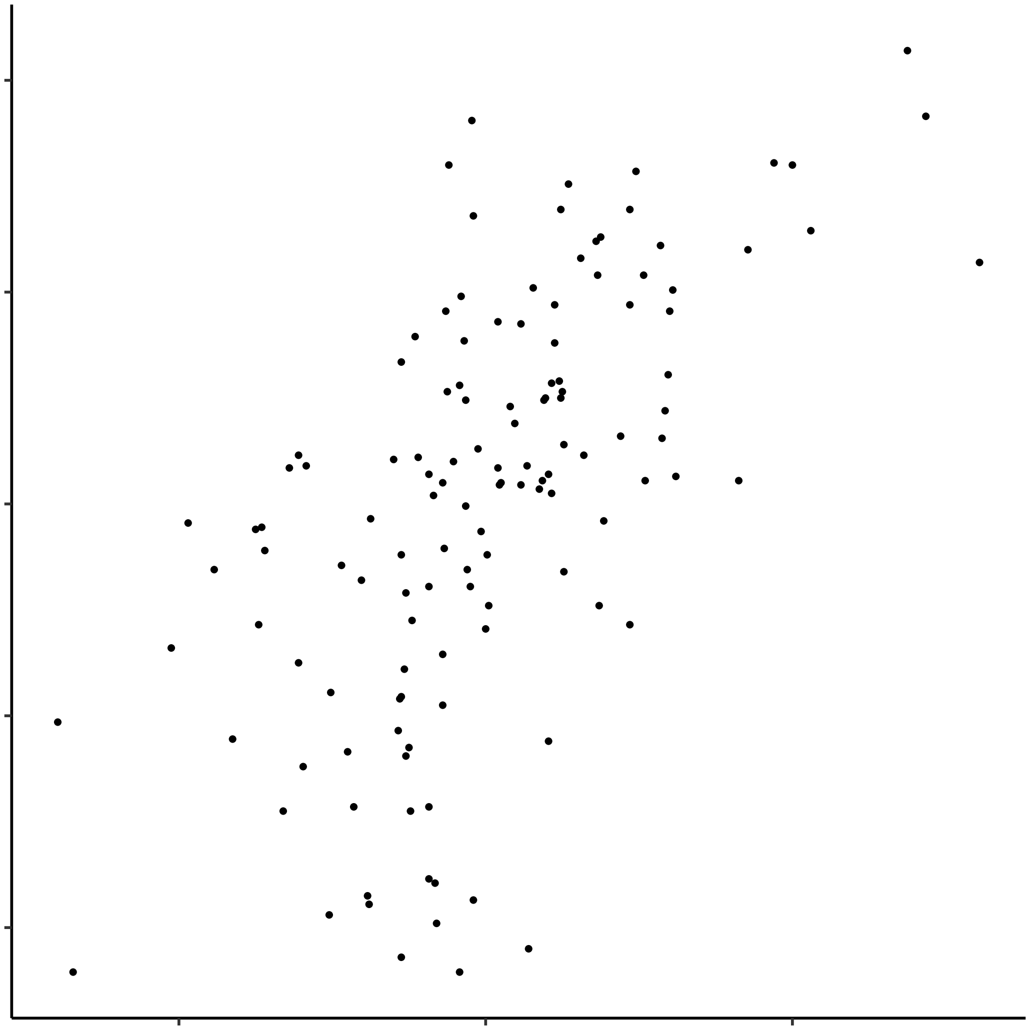

Figure 3.2. The basic design of scatterplots in experiments 1 to 4.

In a mixed-effects modelling paradigm, a distinction is made between variability that is thought to arise as a result of an experimental manipulation (fixed effects), and that which arises due to differences between, for example, participants or particular experimental items (random effects). When a variable is manipulated by a researcher in an experiment, each level of that variable is present, meaning it is appropriate to be modelled as a fixed effect. When only a *subset* of levels of a variable is present, such as a sample of all possible participants or experimental items, then this variable is appropriate for modelling as a random effect. Typically, mixed-effects models require the specification of *intercepts*; these are different baselines for each participant or item that reflect random deviations from the mean of the dependent variable. Mixed-effects models may also specify random *slopes*; these are differences in the magnitude of the difference between levels of the independent variable for each participant or experimental item [12]. Figure 3.3 visualises these concepts.

Throughout the course of this thesis we attempt to model both random intercepts and slopes in order to capture the maximum amount of variability present in our datasets.

3.3.2 Advantages Over Aggregate-Level Statistical Tests

Traditional analysis of the data that we collect throughout this thesis would involve the use of repeated measures analyses of variance (ANOVAs). This technique assesses whether there are significant differ-

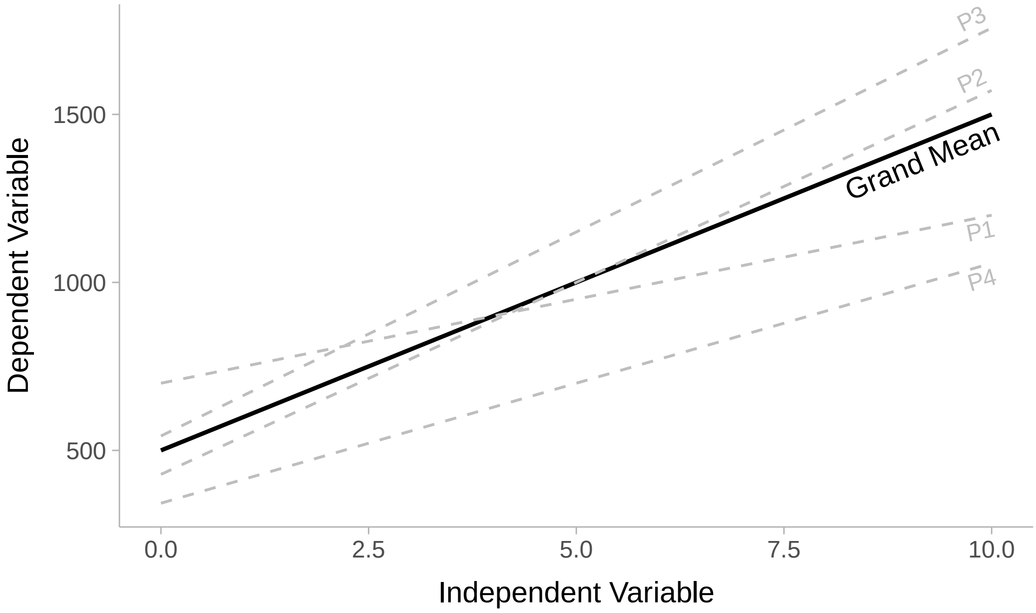

Figure 3.3. Visualising random intercepts and slopes for a theoretical experiment with 4 participants. The grand mean of the dependent variable is shown as a solid line, while each separate random intercept is drawn with dashed lines. Each line has a different gradient, representing different random slopes for each participant. This graphic was inspired by those featured in Brown, 2021 [12].

ences in means of dependent variables between conditions. While these techniques are commonplace, they do not allow for comparisons of differences across the full range of participant responses, nor do they allow for simultaneous consideration of by-item and by-participant variance. It is for these reasons that we employ linear mixed-effects models throughout this thesis; doing so simply allows us to appreciate the data story in a broader and more detailed fashion.

3.3.3 Ordinal Modelling

In experiment 5, participants used Likert scales to provide responses. These scales capture whether one rating is higher or lower than another, however they do not quantify the magnitude of the difference between levels of rating. Metric modelling, such as linear regression, treats the response options to a Likert scale as if they were numeric. Doing so assumes equal levels of difference between ratings, when in reality there is no theoretical reason to assume so. Metric modelling is therefore considered inappropriate for modelling responses to Likert scales [24]. In light of these issues, we use the `ordinal` package [14] in R to build cumulative link mixed effects models for the analysis of Likert scale data. This allows us to treat our Likert responses as an ordered factor as opposed to a continuous response scale.

3.3.4 Model Construction

Choices are inherent in every type of statistical analysis, and can play a large role in the conclusions that are drawn from them. In linear mixed-effects modelling, deciding *what* is a fixed or random effect is straightforward; deciding *how to specify* random effects is a more complicated matter. Barr et al. [7] argue that for fully repeated measures designs, we should prefer a maximal model; one with random intercepts and slopes for each participant and experimental item. More recently, Bates et al. [8] have argued that attempting to specify maximal models for insufficiently rich datasets may lead to overfitting and unreliable conclusions. In light of this we sought a more systematic approach to selecting the random effects structure of a given model.

In an attempt to balance simplicity, explanatory power, and model convergence (whether or not a solution can be found), we chose to use the `buildmer` package [50] in R to automate the selection of our model specifications. Having been provided with a maximal model, `buildmer` uses stepwise regression to select the most complex model structure that successfully converges. Following this, random effects terms that fail to explain a significant amount of variance in the dataset are dropped. This results in a model that captures the maximal amount of feasible variability while minimising redundancy. Note that we do not rely on `buildmer` as a modelling *panacea*; models are still based on theoretical underpinnings and are evaluated critically.

3.3.5 Effects Sizes

Our approach to effects sizes evolved throughout the course of the research project due to reviewer feedback and a growing appreciation of the complexities of effect sizes when discussing linear mixed-effects models. Experiments 1, 2, and 3 used the `EMAtools` package [21] to calculate equivalent Cohen's d effect sizes. Experiment 4 did not feature a theoretically sound baseline condition, meaning Cohen's d was inappropriate. We therefore used the `r2glmm` package [20] to calculate semi-partial R^2. We use this in lieu of a traditional measure of effect size to demonstrate the unique variance in our dependent variable explained by each level of our independent [30]. Experiment 5 features a much simpler modelling situation, and returns to providing equivalent Cohen's d values, this time calculated by converting odds ratios using the `effectsize` package [9]. More details on specific calculations, measures, and conclusions can be found *in situ*.

3.3.6 Reporting Analyses

Throughout this thesis, we take a broad approach to the reporting of statistical analyses; while we consider our analytical methods and conclusions to be sound, we also present a range of statistics to allow the reader to draw their own conclusions should they wish. Statistics are visualised where appropriate, and where visualisation aids understanding and interpretation. In addition, we include details about model structures and the issues we tackled when modelling, for transparency [28].

3.4 Computational Methods

We took an approach to computational methods that sought to marry convenience, simplicity, and reproducibility. Often, this meant that what would otherwise be a makeshift script followed by copy-pasting of results into Overleaf ended up being an involved exercise in functional programming and code wrangling. This involved effort and time, particularly in the early stages of the project, however has yielded a number of benefits. Many of the techniques developed early in the project proved to be instrumental later on, resulting in time-savings overall. Additionally, we share these techniques, principles, and practices to enable future researchers to learn, where we had to struggle. In this section, we detail our approaches to computational methods, including how we utilised the idea of **executable papers**, and how we used containerised environments to capture a freeze-frame of our analyses.

3.4.1 Executable Reporting

Each paper published throughout this project, and this thesis, has been written to be executable. Doing so allows us to package our research such that a lay person can follow simple instructions to recreate our work, while also facilitating and encouraging literate programming, or the close alignment of documentation and underlying code [38].

The use of a literate programming paradigm to generate reports (usually using LaTeX) has a rich history. Here, we focus on that history specifically with regards to the language used throughout the course of this project, R. `Sweave` [23], written in 2002, allowed R code to be integrated into LaTeX documents. This was followed by Yihui Xie's `knitr` [52], which expanded `Sweave` functionality and improved integration with tools such as `pandoc` [26]. `knitr` uses `Rmarkdown` [53] to mix markdown-flavoured text with code chunks into a document that can easily be rendered into an appropriately-formatted conference or journal pdf; this workflow was used for the paper associated with experiments 1 and 2. `Quarto` [2], released in 2022, further expands on `Rmarkdown` functionality, and removes reliance on R or Rstudio. We used `Quarto` for the remainder of the papers associated with this project, and for the present thesis.

Writing executable or dynamic documents allows results to be re-generated whenever the document is rendered. This includes any associated data visualisations and statistical modelling. Structuring documents like this effectively "open up" research by allowing others to view the code that performed the analysis and generated the data visualisations, in addition to guarding against accusations of questionable research practices through high levels of transparency [19]. This paradigm also allows for the caching of computationally expensive statistical models.

3.4.2 Containerised Environments

Providing the code associated with a project, even when that code is integrated into a literately programmed executable paper, is necessary, but not sufficient, for enabling adequate reproducibility. Previous work has found many instances where publicly-accessible code could not reproduce the results included in the corresponding document or failed to run entirely [15, 42, 49]. Poor programming practices accounted for a significant portion of these problems, highlighting the issue of researchers without technical backgrounds being expected to produce high quality technical documentation. Elsewhere, differences in computational environment, package versions, and operating systems have been identified as responsible for the non-replication of results. Large research projects, such as this, can include hundreds of functions from scores of packages, meaning that small changes can critically break code.

To address these issues, we elected to use containers, specifically, those created by Docker [10, 27].

1979 saw the development of `chroot`, which is able to isolate an application's file access. The next 50 years saw rapid development and uptake of containerisation software, mostly within the software development and security communities. Docker, released in 2014, is a popular, lightweight containerisation tool that enables a precise recreation of computational environments. By recording software versions and dependencies, we avoid the potential for broken code in the future, and by publicly hosting papers as GitHub repositories that build into Docker containers, we ensure that future researchers can interact with our code and data in the same computational environment we did when carrying out the research.

3.5 Reproducibility In This Thesis

Reproducibility is a broad spectrum [37]. As discussed above, even when code and data are provided, results are often not replicable, and this is before issues around poor research practice, inappropriate analysis, and dishonest science even rear their heads. While for most, the reproducibility crisis [31] crystallised in the early 2010s [36], concerns had been voiced since at least the late 1960s [41]. Since coming into the wider academic conscious, numerous studies have identified reasons for the crisis, ranging from poor practice (e.g Potti et al., 2006 [1]) to outright deception and fabrication (e.g the Woo-Suk Hwang scandal [43]). These issues led us to strive for a gold standard [37] of reproducibility throughout this project. In this section, we detail how we accomplished this, and in doing so, expose our work to welcome critique.

3.5.1 Sharing Data and Code

The open and public sharing of data and code facilitates external assessment [3, 22], secondary use of data [48], and guards against reproducibility issues [29]. Quite aside from external motivating factors, we have found that developing and embedding the reproducibility practices described here have resulted in longer term savings in time and effort. We also favour a gold-standard reproducible approach as a way of "paying it forward"; having come from a non-technical background, this author found previous work that adhered to the same standard critical to learning and development. We use GitHub to host both this thesis and the papers associated with the project; links to these repositories can be found throughout.

3.5.2 Executable Papers and Docker Containers

As detailed above, we used Quarto and Docker to produce executable journal/conference papers for each of the published works this thesis describes. For simplicity, all analyses from these papers have been repeated using up to date packages here. Accordingly, we provide a single implementation of Docker to reproduce this thesis. All statistics have been checked against those provided in the original analyses, and repositories for the corresponding papers are provided complete with separate Docker implementations.

3.5.3 Pre-Registration of Hypotheses and Analysis Plans

Often touted as a low-cost entry point into reproducible research practices [25], pre-registration is the practice of clarifying hypotheses and analysis plans prior to data collection. While this may not be able to prevent research fraud and QRPs entirely, it does lend credibility to the researcher [33]. We pre-registered all hypotheses and analysis plans with the Open Science Framework [32]. Pre-registrations are embargoed, and we detail where we deviate from our analysis plans in the methods sections of the corresponding experiments.

3.5.4 Experimental Resources

Everything needed to run each experiment is included in the corresponding GitLab repository. Links to these repositories are also provided in the sections concerning each experiment.

Chapter 4

Experiment 1: https://gitlab.pavlovia.org/Strain/exp_uniform_adjustments

Experiment 2: https://gitlab.pavlovia.org/Strain/exp_spatially_dependent

Chapter 5

Experiment 3: https://gitlab.pavlovia.org/Strain/exp_size_only

Chapter 6

Experiment 4: https://gitlab.pavlovia.org/Strain/size_and_opacity_additive_exp

Chapter 7

Experiment 5 Pre-Study: https://gitlab.pavlovia.org/Strain/beliefs_scatterplots_pretest

Experiment 5 Main Study (Group A): https://gitlab.pavlovia.org/Strain/atypical_scatterplots_main_a

Experiment 5 Main Study (Group B): https://gitlab.pavlovia.org/Strain/atypical_scatterplots_main_t

3.6 Conclusion

In this chapter, we have established the broad methodological approach taken by this thesis. This project sought to investigate novel ways of visualising data and their effects on perception and cognition. We have detailed justifications for the designs we used, the methodological challenges we faced, and how we have used a broad array of tools and techniques to overcome these challenges, including how we learnt from our mistakes. We have endeavoured to keep open research and reproducibility at the core of our work throughout this project, and we hope this thesis can serve as an example for future work facing similar challenges and with similar commitments to open science. To this end, we have produced a template to facilitate future reproducible theses. Through public sharing of data and code, literate programming, and containerisation, we satisfy FAIR (**F**indable, **A**ccessible, **I**nteropable, and **R**eusable) data principles [51].

Chapter 4

Adjusting the Opacities of Scatterplot Points Can Affect Correlation Estimates

4.1 Abstract

4.2 Preface: Learning From an Early Pilot Study

4.3 Introduction

4.3.1 Overview

4.4 Related Work

4.4.1 Transparency, Contrast, Opacity, and Formal Definitions

4.4.2 Effects of Point Opacity on Correlation Estimation

4.5 Experiment 1: Uniform Opacity Adjustments

4.5.1 Introduction

4.5.2 Methods

4.5.3 Analysis

4.5.4 Discussion

4.6 Experiment 2: Spatially-Dependent Opacity Adjustments

4.6.1 Introduction

4.6.2 Methods

4.6.3 Analysis

Chapter 5

Adjusting the Sizes of Scatterplot Points Can Correct for a Historic Correlation Underestimation Bias

5.1 Abstract

5.2 Overview

5.3 Related Work

5.3.1 Size and Perception

5.3.2 Scatterplot Point Size and Correlation Perception

5.4 Experiment: Adjusting Point Size to Facilitate More Accurate Correlation Perception in Scatterplots

5.4.1 Introduction

5.4.2 Methods

5.4.3 Analysis

5.4.4 Discussion

5.5 General Discussion

Chapter 6

Interactions of Opacity and Size Adjustments

6.1 Abstract

6.2 Overview

6.3 Related Work

6.3.1 Size and Opacity

6.4 Experiment: Adjusting Point Size and Opacity Together

6.4.1 Introduction

6.4.2 Methods

6.4.3 Analysis

6.4.4 Discussion

6.5 General Discussion

Chapter 7

Visual Features Affecting Perceptual Estimates Also Affect Beliefs About Correlations

7.1 Abstract

7.2 Overview

7.3 Related Work

7.3.1 From Perception to Cognition

7.3.2 From Cognition to Belief

7.4 Pre-Study: Investigating Beliefs About Relatedness Statements

7.4.1 Introduction

7.4.2 Methods

7.4.3 Analysis

7.4.4 Discussion

7.5 Experiment: Potential for Belief Change Using Atypical Scatterplots

7.5.1 Introduction

7.5.2 Methods

7.5.3 Analysis

7.5.4 Discussion

Chapter 8

Conclusion

8.1 Main Findings

8.2 Relationship to Prior Work

8.3 Reproducibility

8.4 Contributions

8.5 Implications

8.5.1 For Design

8.5.2 For Society

8.6 Limitations

8.7 Future Directions

8.8 Closing Remarks

References

[1] Potti A, Dressman Hk, Bild A, Riedel Rf, Chan G, Sayer R, Cragun J, Cottrill H, Kelley Mj, Petersen R, Harpole D, Marks J, Berchuck A, Ginsburg Gs, Febbo P, Lancaster J, and Nevins Jr. “Genomic Signatures to Guide the Use of Chemotherapeutics”. In: *Nature medicine* 12.11 (Nov. 2006). ISSN: 1078-8956. DOI: 10.1038/nm1491. (Visited on 10/10/2024) (cited on p. 23).

[2] JJ Allaire and Christophe Dervieux. *Quarto: R Interface to 'quarto' Markdown Publishing System*. Manual. 2024 (cited on p. 22).

[3] George Alter and Richard Gonzalez. “Responsible Practices for Data Sharing”. In: *The American psychologist* 73.2 (2018), pp. 146–156. ISSN: 0003-066X. DOI: 10.1037/amp0000258. (Visited on 10/10/2024) (cited on p. 23).

[4] Alexander L. Anwyl-Irvine, Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo K. Evershed. “Gorilla in Our Midst: An Online Behavioral Experiment Builder”. In: *Behavior Research Methods* 52.1 (Feb. 2020), pp. 388–407. ISSN: 1554-3528. DOI: 10.3758/s13428-019-01237-x. (Visited on 10/03/2024) (cited on p. 17).

[5] Antonio A. Arechar, Simon Gächter, and Lucas Molleman. “Conducting Interactive Experiments Online”. In: *Experimental Economics* 21.1 (Mar. 2018), pp. 99–131. ISSN: 1573-6938. DOI: 10.1007/s10683-017-9527-2. (Visited on 10/04/2024) (cited on p. 17).

[6] 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. “LIBER Open Science Roadmap”. In: (July 2018). DOI: 10.20350/digitalCSIC/15061. (Visited on 09/13/2023) (cited on p. 17).

[7] Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. “Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal”. In: *Journal of memory and language* 68.3 (Apr. 2013), 10.1016/j.jml.2012.11.001. ISSN: 0749-596X. DOI: 10.1016/j.jml.2012.11.001. (Visited on 08/23/2022) (cited on p. 20).

[8] Douglas Bates, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen. *Parsimonious Mixed Models*. https://arxiv.org/abs/1506.04967v2. 2018. (Visited on 10/09/2024) (cited on p. 21).

[9] Mattan S. Ben-Shachar, Daniel Lüdecke, and Dominique Makowski. "effectsize: Estimation of Effect Size Indices and Standardized Parameters". In: *Journal of Open Source Software* 5.56 (2020), p. 2815. DOI: 10.21105/joss.02815 (cited on p. 21).

[10] Carl Boettiger. "An Introduction to Docker for Reproducible Research". In: *SIGOPS Oper. Syst. Rev.* 49.1 (Jan. 2015), pp. 71–79. ISSN: 0163-5980. DOI: 10.1145/2723872.2723882. (Visited on 10/09/2024) (cited on p. 22).

[11] David Bridges, Alain Pitiot, Michael R. MacAskill, and Jonathan W. Peirce. "The Timing Mega-Study: Comparing a Range of Experiment Generators, Both Lab-Based and Online". In: *PeerJ* 8 (July 2020), e9414. ISSN: 2167-8359. DOI: 10.7717/peerj.9414. (Visited on 10/03/2024) (cited on p. 17).

[12] Violet A. Brown. "An Introduction to Linear Mixed-Effects Modeling in R". In: *Advances in Methods and Practices in Psychological Science* 4.1 (Jan. 2021), p. 2515245920960351. ISSN: 2515-2459. DOI: 10.1177/2515245920960351. (Visited on 10/08/2024) (cited on pp. 19, 20).

[13] Nick Charalambides. *We Recently Went Viral on TikTok - Here's What We Learned*. https://www.prolific.com/resources/we-recently-went-viral-on-tiktok-heres-what-we-learned. Aug. 2021. (Visited on 10/04/2024) (cited on p. 17).

[14] Rune H. B. Christensen. *Ordinal—Regression Models for Ordinal Data*. Manual. 2023 (cited on p. 20).

[15] Christian Collberg and Todd A. Proebsting. "Repeatability in Computer Systems Research". In: *Communications of the ACM* 59.3 (Feb. 2016), pp. 62–69. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/2812803. (Visited on 10/09/2024) (cited on p. 22).

[16] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. "Data Quality in Online Human-Subjects Research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA". In: *PLOS ONE* 18.3 (Mar. 2023), e0279720. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0279720. (Visited on 10/04/2024) (cited on p. 17).

[17] *E-Prime*. Psychology Software Tools. 2020 (cited on p. 17).

[18] Naoyasu Hirao, Koyo Koizumi, Hanako Ikeda, and Hideki Ohira. "Reliability of Online Surveys in Investigating Perceptions and Impressions of Faces". In: *Frontiers in Psychology* 12 (Sept. 2021). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.733405. (Visited on 10/04/2024) (cited on p. 17).

[19] Daniel T. Holmes, Mahdi Mobini, and Christopher R. McCudden. "Reproducible Manuscript Preparation with RMarkdown Application to JMSACL and Other Elsevier Journals". In: *Journal of Mass Spectrometry and Advances in the Clinical Lab* 22 (Nov. 2021), pp. 8–16. ISSN: 2667145X. DOI: 10.1016/j.jmsacl.2021.09.002. (Visited on 10/09/2024) (cited on p. 22).

[20] Byron Jaeger. *R2glmm: Computes R Squared for Mixed (Multilevel) Models*. Manual. 2017 (cited on p. 21).

[21] Evan Kleiman. *EMAtools: Data Management Tools for Real-Time Monitoring/Ecological Momentary Assessment Data*. Manual. 2021 (cited on p. 21).

[22] Olivier Klein, Tom E. Hardwicke, Frederik Aust, Johannes Breuer, Henrik Danielsson, Alicia Hofelich Mohr, Hans IJzerman, Gustav Nilsonne, Wolf Vanpaemel, and Michael C. Frank. "A Practical Guide for Transparency in Psychological Science". In: *Collabra: Psychology* 4.1 (June 2018). Ed. by Michéle Nuijten and Simine Vazire, p. 20. ISSN: 2474-7394. DOI: 10.1525/collabra.158. (Visited on 10/10/2024) (cited on p. 23).

[23] Friedrich Leisch. "Sweave, Part I: Mixing R and LaTeX". In: *R News* 2.3 (Dec. 2002), pp. 28–31. ISSN: 1609-3631. (Visited on 10/09/2024) (cited on p. 22).

[24] Torrin M. Liddell and John K. Kruschke. "Analyzing Ordinal Data with Metric Models: What Could Possibly Go Wrong?" In: *Journal of Experimental Social Psychology* 79 (Nov. 2018), pp. 328–348. ISSN: 0022-1031. DOI: 10.1016/j.jesp.2018.08.009. (Visited on 09/03/2024) (cited on p. 20).

[25] Jennifer M. Logg and Charles A. Dorison. "Pre-Registration: Weighing Costs and Benefits for Researchers". In: *Organizational Behavior and Human Decision Processes* 167 (Nov. 2021), pp. 18–27. ISSN: 0749-5978. DOI: 10.1016/j.obhdp.2021.05.006. (Visited on 10/10/2024) (cited on p. 24).

[26] John MacFarlane. *Pandoc*. https://pandoc.org/index.html. (Visited on 10/09/2024) (cited on p. 22).

[27] Dirk Merkel. "Docker: Lightweight Linux Containers for Consistent Development and Deployment". In: *Linux J.* 2014.239 (Mar. 2014), 2:2. ISSN: 1075-3583 (cited on p. 22).

[28] Lotte Meteyard and Robert A. I. Davies. "Best Practice Guidance for Linear Mixed-Effects Models in Psychological Science". In: *Journal of Memory and Language* 112 (June 2020), p. 104092. ISSN: 0749-596X. DOI: 10.1016/j.jml.2020.104092. (Visited on 10/09/2024) (cited on p. 21).

[29] Tsuyoshi Miyakawa. "No Raw Data, No Science: Another Possible Source of the Reproducibility Crisis". In: *Molecular Brain* 13.1 (Feb. 2020), p. 24. ISSN: 1756-6606. DOI: 10.1186/s13041-020-0552-2. (Visited on 10/10/2024) (cited on p. 23).

[30] Shinichi Nakagawa and Holger Schielzeth. "A General and Simple Method for Obtaining R2 from Generalized Linear Mixed-Effects Models". In: *Methods in Ecology and Evolution* 4.2 (2013), pp. 133–142. ISSN: 2041-210X. DOI: 10.1111/j.2041-210x.2012.00261.x. (Visited on 09/07/2023) (cited on p. 21).

[31] Open Science Collaboration. “Estimating the Reproducibility of Psychological Science”. In: *Science* 349.6251 (Aug. 2015), aac4716. DOI: 10.1126/science.aac4716. (Visited on 10/10/2024) (cited on p. 23).

[32] *OSF*. https://osf.io/. (Visited on 10/10/2024) (cited on p. 24).

[33] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. “Pre-Registration: Why and How”. In: *Journal of Consumer Psychology* 31.1 (2021), pp. 151–162. ISSN: 1532-7663. DOI: 10.1002/jcpy.1208. (Visited on 10/10/2024) (cited on p. 24).

[34] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. “Data Quality of Platforms and Panels for Online Behavioral Research”. In: *Behavior Research Methods* 54.4 (Sept. 2021), pp. 1643–1662. ISSN: 1554-3528. DOI: 10.3758/s13428-021-01694-3. (Visited on 07/05/2022) (cited on p. 18).

[35] Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. “PsychoPy2: Experiments in Behavior Made Easy”. In: *Behavior Research Methods* 51.1 (Feb. 2019), pp. 195–203. ISSN: 1554-3528. DOI: 10.3758/s13428-018-01193-y (cited on p. 17).

[36] Roger Peng. “The Reproducibility Crisis in Science: A Statistical Counterattack”. In: *Significance* 12.3 (2015), pp. 30–32. ISSN: 1740-9713. DOI: 10.1111/j.1740-9713.2015.00827.x. (Visited on 10/10/2024) (cited on p. 23).

[37] Roger D. Peng. “Reproducible Research in Computational Science”. In: *Science* 334.6060 (Dec. 2011), pp. 1226–1227. DOI: 10.1126/science.1213847. (Visited on 10/09/2024) (cited on p. 23).

[38] Stephen R Piccolo and Michael B Frampton. “Tools and Techniques for Computational Reproducibility”. In: *GigaScience* 5.1 (Dec. 2016), s13742-016-0135–4. ISSN: 2047-217X. DOI: 10.1186/s13742-016-0135-4. (Visited on 10/09/2024) (cited on p. 22).

[39] Benjamin Prissé and Diego Jorrat. “Lab vs Online Experiments: No Differences”. In: *Journal of Behavioral and Experimental Economics* 100 (Oct. 2022), p. 101910. ISSN: 2214-8043. DOI: 10.1016/j.socec.2022.101910. (Visited on 10/04/2024) (cited on p. 17).

[40] *Prolific.Co*. Prolific. 2024 (cited on p. 17).

[41] Felipe Romero. “Philosophy of Science and the Replicability Crisis”. In: *Philosophy Compass* 14.11 (2019), e12633. ISSN: 1747-9991. DOI: 10.1111/phc3.12633. (Visited on 10/10/2024) (cited on p. 23).

[42] Sheeba Samuel and Daniel Mietchen. “Computational Reproducibility of Jupyter Notebooks from Biomedical Publications”. In: *GigaScience* 13 (Jan. 2024), giad113. ISSN: 2047-217X. DOI: 10.1093/gigascience/giad113. (Visited on 10/09/2024) (cited on p. 22).

[43] R. Saunders and J. Savulescu. “Research Ethics and Lessons from Hwanggate: What Can We Learn from the Korean Cloning Fraud?” In: *Journal of Medical Ethics* 34.3 (Mar. 2008), pp. 214–221. ISSN: 1473-4257. DOI: 10.1136/jme.2007.023721 (cited on p. 23).

[44] Holger Schielzeth, Niels J. Dingemanse, Shinichi Nakagawa, David F. Westneat, Hassen Allegue, Céline Teplitsky, Denis Réale, Ned A. Dochtermann, László Zsolt Garamszegi, and Yimen G. Araya-Ajoy. “Robustness of Linear Mixed-Effects Models to Violations of Distributional Assumptions”. In: *Methods in Ecology and Evolution* 11.9 (2020), pp. 1141–1152. ISSN: 2041-210X. DOI: 10.1111/2041-210X.13434. (Visited on 11/29/2023) (cited on p. 18).

[45] Gabriel Strain, Andrew J. Stewart, Paul Warren, and Caroline Jay. “Adjusting Point Size to Facilitate More Accurate Correlation Perception in Scatterplots”. In: *2023 IEEE Vis X Vision*. Melbourne, Australia: IEEE, Oct. 2023, pp. 1–5. ISBN: 9798350329841. DOI: 10.1109/VisXVision60716.2023.00006. (Visited on 02/15/2024) (cited on p. 14).

[46] Gabriel Strain, Andrew J. Stewart, Paul Warren, and Caroline Jay. “The Effects of Contrast on Correlation Perception in Scatterplots”. In: *International Journal of Human-Computer Studies* 176 (Aug. 2023), p. 103040. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2023.103040. (Visited on 04/11/2023) (cited on p. 14).

[47] Gabriel Strain, Andrew J. Stewart, Paul A. Warren, and Caroline Jay. “Effects of Point Size and Opacity Adjustments in Scatterplots”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI ’24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–13. ISBN: 9798400703300. DOI: 10.1145/3613904.3642127. (Visited on 05/29/2024) (cited on p. 14).

[48] Tsaone Tamuhla, Eddie T Lulamba, Themba Mutemaringa, and Nicki Tiffin. “Multiple Modes of Data Sharing Can Facilitate Secondary Use of Sensitive Health Data for Research”. In: *BMJ Global Health* 8.10 (Oct. 2023), e013092. ISSN: 2059-7908. DOI: 10.1136/bmjgh-2023-013092. (Visited on 10/10/2024) (cited on p. 23).

[49] Ana Trisovic, Matthew K. Lau, Thomas Pasquier, and Mercè Crosas. “A Large-Scale Study on Research Code Quality and Execution”. In: *Scientific Data* 9.1 (Feb. 2022), p. 60. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01143-6. (Visited on 10/09/2024) (cited on p. 22).

[50] Cesko C. Voeten. *Buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression*. Manual. 2023 (cited on p. 21).

[51] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. ’t Hoen, Rob

Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. "The FAIR Guiding Principles for Scientific Data Management and Stewardship". In: *Scientific Data* 3 (2016). DOI: 10.1038/sdata.2016.18. (Visited on 10/10/2024) (cited on p. 25).

[52] Yihui Xie. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman and Hall/CRC, 2015 (cited on p. 22).

[53] Yihui Xie, Christophe Dervieux, and Emily Riederer. *R Markdown Cookbook*. Boca Raton, Florida: Chapman and Hall/CRC, 2020. ISBN: 978-0-367-56383-7 (cited on p. 22).

Appendices

Appendix A

First appendix

A.1 Section in Appendix