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

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

Data visualisation is the practice of pictorially presenting patterns otherwise described by numbers, and has been employed in one form or another for thousands of years [7]. Lists or matrices of numbers may be able to communicate simple trends, but for more complex patterns, harnessing the human visual system [**midway_2020**] through visualisation is crucial for the communication of science and data. Effective data visualisation is able to reduce cognitive and perceptual loads on viewers, regardless of their backgrounds levels of statistical knowledge or experience. This outsourcing, while efficient, leaves the viewer vulnerable to the design choices that were made when creating the visualisation. For this reason, understanding *how* design choices affect interpretation is crucial to designing better data visualisations and simultaneously allows for the inoculation of viewers against poor or malevolent design practices.

One cannot understate the importance and ubiquity of data visualisations. They are used to communicate with experts, with those with no background in science, and with everyone in between. They are relied upon to communicate vital public health information [37], to present evidence in court cases [10], and to facilitate collaboration and encourage engagement on climate change issues [46, 64]. Studying data visualisation therefore has the potential to change the nature of scientific study itself.

Interaction with a data visualisation involves steps from perception (viewing), to cognition (interpreting), to behaviour (deciding and acting). These stages must be examined both separately and together, as there are bottom-up and top-down interactions present. In this thesis, I therefore present a series of experiments whose aim was to investigate and address a long-standing perceptual bias in scatterplots, a common form of data visualisation [24]. Following attempts to address this bias, I further investigated the impacts these attempts may have had on cognition and behaviour.

1.1 Research Motivation

Data visualisations were once the preserve of the academic and professional classes. Now, however, visualisation is everywhere. This became especially apparent to me during the COVID-19 pandemic, when people such as my parents, professionals with no background in mathematics or statistics, were bombarded with data visualisations on a daily basis. Despite not being able to fully articulate the whole data story, they were still able to derive meaning from the visualisations they were seeing. The ability of data visualisations to transcend language mathematical prowess motivated my study of them; concepts such as exponential growth or the relatedness between variables can be displayed in simple

ways that do not rely on an underlying understanding of exponents or correlation coefficients.

Using data visualisations in such a way effectively provides a cognitive proxy for viewers. This is an efficient way to communicate, however still necessitates the presence of accurate and reliable perceptions. On further investigation into the reliance that those who design visualisations make on those who view them, I uncovered a historic perceptual bias that was still being described in the literature, seemingly with no attempt at correction. This bias was the underestimation of correlation in positively correlated scatterplots. In the literature, this robust effect had been observed in a number of experimental paradigms, including direct estimation [10, 19, 21, 33, 34, 43, 66] and estimation via bisection tasks [59], and as of 2021, no attempts had been made to correct for it. The goal of this thesis was therefore to collect empirical evidence on the perception of correlation in positively correlated scatterplots, to use that information to create novel visualisations with a view to correcting for the underestimation, and to further investigate the potential for these visualisations to effect what people think and believe.

1.2 Contributions

Through this thesis, I present a series of empirical experiments that provide knowledge about how changing visual features in scatterplots can affect how participants interpret the strength of the correlation displayed. I demonstrate that systematically reducing the opacities and sizes of points on scatterplots as a function of their distances from a regression line can significantly increase estimates of correlation and partially correct for a historic underestimation bias. Following this, I show that consequences of employing these techniques are not limited to perception, but in fact can be extended into a cognitive space to change what people think and believe. I utilise large-sample, controlled experiments with lay populations to provide generalisable conclusions and recommendations for visualisation designers and researchers. Through this thesis, I provide a framework for the large-scale testing of data visualisations with lay audiences. Additionally, I hope to provide an example of a project conducted entirely in an open and reproducible manner.

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 is 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* [68] is reproduced in Chapter 4. Sections 4.5, 4.6.2, 4.7.2, 4.6.3, 4.7.3, 4.6.4, 4.7.4, and 4.8 contain minimally altered parts of the published article.

- *Adjusting Point Size to Facilitate More Accurate Correlation Perception in Scatterplots* [67] 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* [69] 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

In Chapter 2, I conduct a thorough review of the literature and provide the necessary context for the following experimental chapters. I begin by exploring the history and significance of data visualisation, before discussing concepts of correlation and relatedness, and their visualisation using scatterplots. I discuss a long-standing perceptual bias in the interpretation of correlation in scatterplots, and thereby motivate novel experimental work that seeks to address this bias. I provide real-world motivation by exploring the potential for perceptual biases to have effects on people's lives and livelihoods. Finally, I provide a discussion on issues of graph and statistical literacy.

Chapter 3 explores the general methodological approaches taken by this thesis, with a particular focus on transparency and reproducibility. This thesis embodies these principles, and justification for them along practical and pedagogical lines is presented. This Chapter also contains detailed instructions for the full reproduction of this thesis, and provides context for the approaches to experimental design, recruitment, and analysis featured in the following chapters.

Chapter 4 presents a pair of experiments that establish the effects of point opacity on correlation perception in positively correlated scatterplots. The first experiment demonstrates that lowering point opacity in a uniform manner can increase participants' levels of error on a correlation estimation task. The second experiment describes how employing a function relating point opacity to residual error can correct for the correlation underestimation bias by shifting estimates upwards. Specifically, lowering point contrast as a function of residual error provides significant levels of correction for the underestimation bias, and demonstrates that, in contrast to previous work, changing visual features can alter perceptions of correlation. This finding was foundational for the remainder of the work described in this thesis. This Chapter also includes a discussion of a very early pilot study; I describe why this experiment was not included in published work, and offer reflections on what I learnt from conducting it in Section 4.2.

In Chapter 5, I present an experiment that expands the technique described in the previous chapter to point size. I show that reducing the sizes of scatterplot points as a function of residual error is able to alter participants' estimates of correlation to a greater degree than functions relating point opacity and residual error.

Chapter 6 describes an experiment in which point opacity and size adjustments are combined. This experiment includes conditions in which point opacity and size are reduced and increased together with residual error (congruent conditions), and conditions in which one increases while another decreases

(incongruent conditions). I demonstrate that opacity and size adjustments interact in a non-linear fashion, and explore the potential for further tuning of these manipulations to create perceptually optimised scatterplots.

In Chapter 7, I extend the perceptual effects described in previous chapters to a cognitive space. In a single experiment, I explore the potential for scatterplots using point opacity and size manipulations to change participants' beliefs about levels of relatedness between variables. I show that using such manipulations is able to change beliefs to a significantly greater degree compared to standard, unaltered scatterplots. This Chapter also describes a pre-study that explores thoughts and feelings about relatedness to ascertain a variable pair that is particularly vulnerable to belief-change. The main experiment demonstrates that visual features can not only affect perceptions of correlation, but can also affect beliefs.

Chapter 8 presents a synthesis of the empirical work described in this thesis. I present discussions of the theoretical and practical implications of my findings, along with an exploration of future directions that my work could inform.

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 I describe my research methodologies. The experiments described in chapters 4, 5, and 6 share most aspects of experimental method, and are therefore described in full in this chapter. Chapter 7 features a different methodology, and is described therein. This chapter discusses experimental designs, the tools used to build and run the experiments, the approach to statistical analyses, and the computational methods and practices 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 research questions we can ask and the conclusions that we may draw. The decisions that lead us to designing experiments in certain ways must be based not only on theory, but also on the external constraints imposed on the research team. Concerns such as time, practicality, and cost must be addressed, and a compromise between research that is *valuable* and research that is *doable* must be reached.

3.2.1 Experimental Design

All but the final experiment utilised within-participants designs. In such a design, each participant is exposed to each level of the intervention. This is in contrast with between-participants designs, where separate groups are exposed to only a single level of the intervention each. Where possible, within-participants designs are preferred. These designs do not rely on random allocation, and as each participant is able to provide as many data points as there are experiment items in levels [17], offer a significant boost in statistical power over between-participant designs where each participant may only provide data points for a portion of the total experimental items. In experiments 1 to 3, each participant saw all experimental stimuli and provided a judgement of correlation using a sliding scale between 0 and 1 (see Figure 3.1). Experiment 1 featured a single factor of global scatterplot point opacity with 4 levels (see Figure 3.2). Experiment 2 featured a single factor of scatterplot point design regarding opacity with 4 levels (see Figure 3.3). Experiment 3 featured a single factor of scatterplot point design regarding size with 4 levels (see Figure 3.4). Experiment 4 featured a factorial 2×2 design; IV_1 was

Current rating: 0

Press the spacebar to continue when you have made your response.
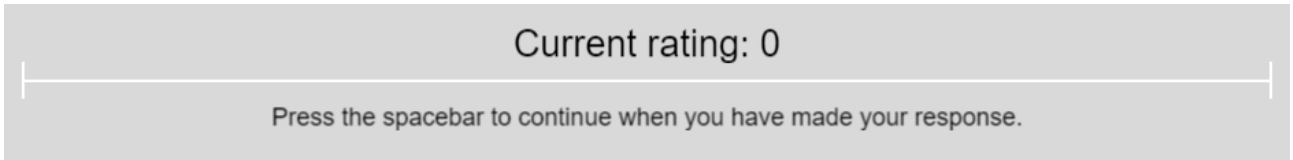
Figure 3.1. An example of the slider participants used to estimate correlation in experiments 1-4.

Opacity = 0.25   Opacity = 0.5   Opacity = 0.75   Opacity = 1
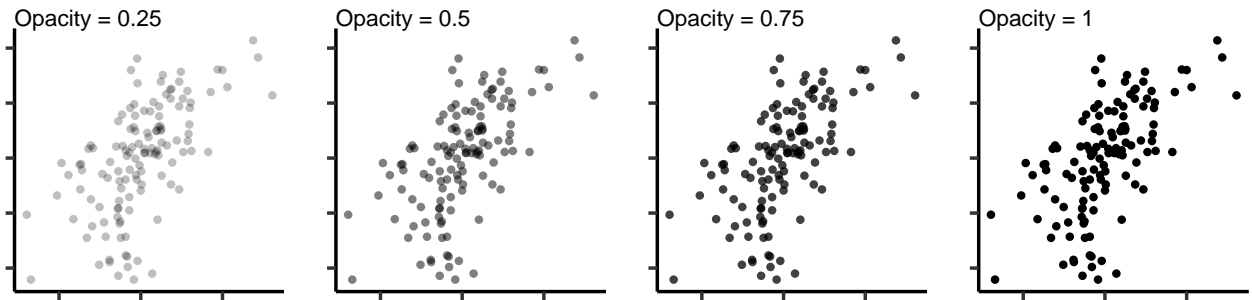
Figure 3.2. Examples of the stimuli used in experiment 1, demonstrated with an *r* value of 0.6. Here, "opacity" refers to the alpha value used by ggplot.

the scatterplot point opacity design used with 2 levels, and IV_2 was the scatterplot point size design used with 2 levels (see Figure 3.5). Experiment 5 is a departure from the shared experimental paradigm of the previous experiments, and features a 1 factor, 2 level between-participants design; group A saw scatterplots designed to ellict greater levels of belief change compared to typical scatterplots, which were shown to group b (see Figure 3.6).

3.2.2 Tools for Testing

However we design experiments, software plays a crucial role in allowing us to carry them out. Fortunately, there is a wealth of tools available to facilitate the testing of visualisations both in traditional lab-based tests and in online experiments. Following the principles of open and reproducible research [6], closed-source software, such as Gorilla [4] or E-prime [23] was discounted, as these rely on paid licences and do not allow for the sharing of code with future researchers. I settled on using PsychoPy [51] due to its open-source status, flexibility regarding graphical and code-based experimental design, and high level of timings accuracy [13]. Using such a open-source tool not only facilitated my own learning with regard to experiment building, but also enables the contribution of further examples of visualisation studies by hosting the resulting experiments online for use and modification by future

Full Opacity   Linear Decay   Non–linear Decay   Inverted Non–linear Decay
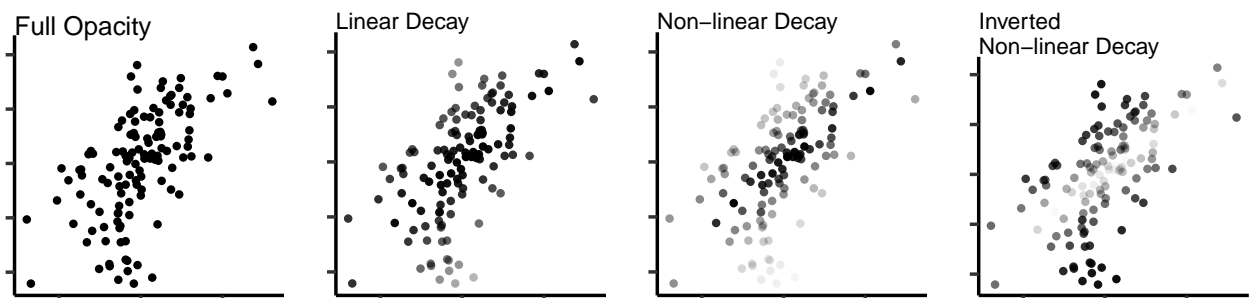
Figure 3.3. Examples of the stimuli used in experiment 2, demonstrated with an *r* value of 0.6.
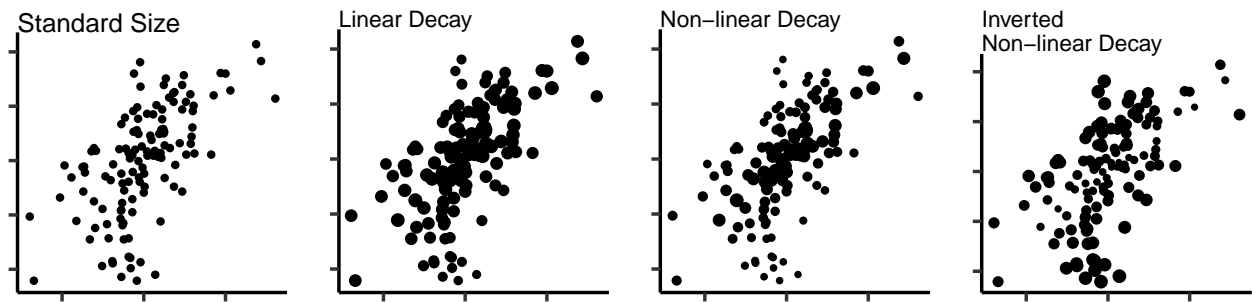
Figure 3.4. Examples of the stimuli used in experiment 3, demonstrated with an *r* value of 0.6.
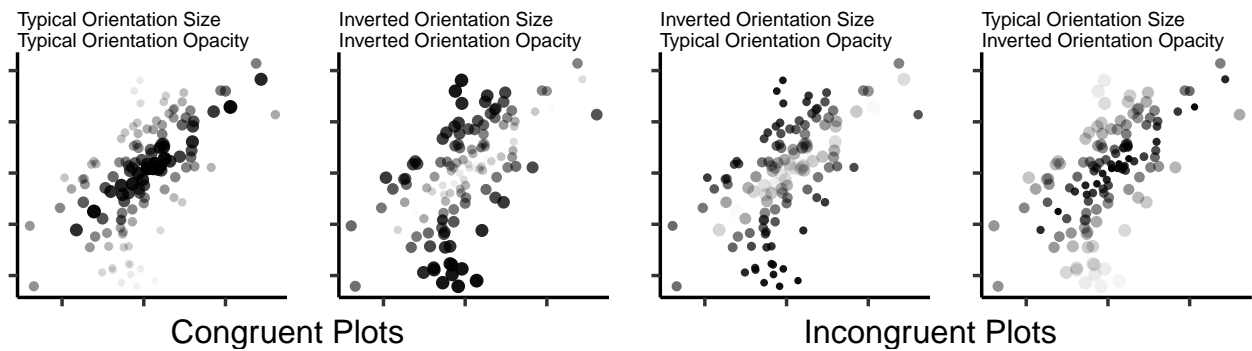
Figure 3.5. Examples of the stimuli used in experiment 4, demonstrated with an *r* value of 0.6.
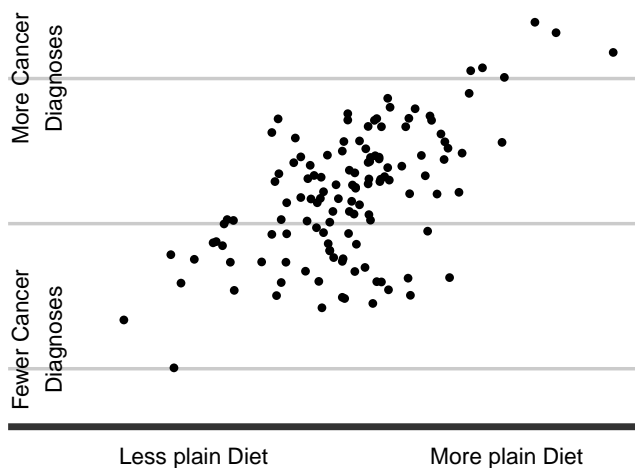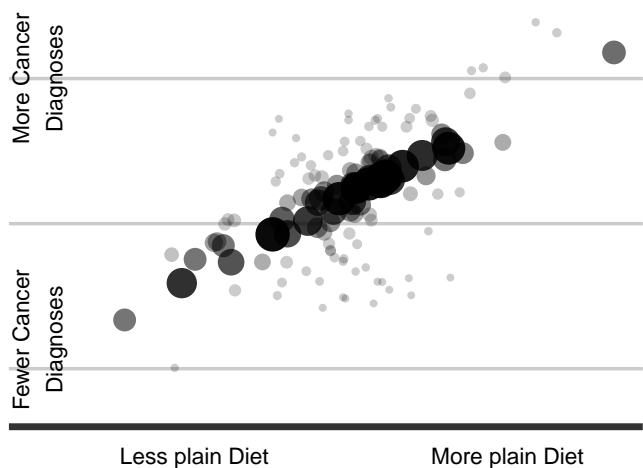
Figure 3.6. Examples of the experimental stimuli for experiment 5 using an *r* value of 0.6. Group A saw the alternative scatterplot presented on the right, while group B saw the typical design on the left.

researchers.

I elected to pursue online testing throughout this thesis. Doing so is much quicker than carrying out in-person lab-based testing, facilitating the collection of 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 (see Section 3.2.3). Online testing also affords 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, 26, 55], especially with large sample sizes. Due to its integration with PsychoPy, Pavlovia was used 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 [56] and CloudResearch provide the highest quality data for the lowest cost [22]; I elected to use the former due to my familiarity with the system. Despite these findings, there has also been evidence of low data quality and skewed demographics affecting both general crowdsourcing platforms, such as Amazon's MTurk, and those tailored specifically towards academic research. On the 24th of July, 2021, the Prolific.co platform went viral on social media [16], 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. This was addressed in the 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 the screened 1:1 ratio was maintained for the remainder of the experiments.

The first experiment 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, I was relatively naive to the intricacies of recruiting research participants online, and thus 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. The advert for each experiment stated 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 very low levels of participant engagement. For the following studies, published guidelines [50] were followed to address these issues; specifically, it was 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 the next experiment fell

---

[1] this is a more strict rate than the 95% recommended by Peer et al. [50].

to ~5%. Rejection rates were similar for the remainder of experiments. Exact numbers of accepted and rejected participants can be found in the **Participants** sections of each experiment.

Each experiment recruited until 150 participants had completed successfully. Due to the novelty of this work, it was difficult to get a sense of the size of the effect that would be seen. I assumed a small effect size (Cohen's d ~ 0.2), and aimed to recruit enough participants to adequately power the experiments [15]. NB: I did not conduct an *a priori* power analysis. A post-hoc power analysis of the first experiment revealed a power of 0.54. Effect sizes were larger in the subsequent experiments, however to facilitate comparison, it was decided that n = 150 would remain the target recruitment rate.

3.2.4 Creating Stimuli

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

Experiments were designed with the intention of isolating and addressing a perceptual effect; the underestimation of correlation in positively correlated scatterplots. To achieve this, confounding extraneous design factors were removed, including axis labels, tick labels, grid lines, and titles. The axis ticks themselves were preserved. Figure 3.7 demonstrates the basic design of the scatterplots used in experiments 1 to 4.

A single random seed was used to generate scatterplot datasets throughout this thesis. 45 r values were uniformly generated between 0.2 and 0.99, as there is evidence that little correlation is perceived below $r = 0.2$ [10, 19, 66]. The data that forms the scatterplots was randomly generated from bivariate normal distributions with standard deviations of 1 in each direction. Scatterplots always had a 1:1 aspect ratio, and were configured such that they occupied the same proportion of the experimental window regardless of the size or resolution of a participant's monitor. From Chapter 5 onwards, a measure of dot pitch is included, which facilitates the approximation of the physical size of scatterplot points on-screen; where available, this is included in analyses.

3.3 Analytical Methods

All analyses in this thesis were conducted using R (version 4.4.2 [**r_core**]). To investigate whether the experimental manipulations have actual effects on the interpretations participants provide, appropriate statistical testing must be employed. This involves taking into account the variability in responses that can be attributed to an experimental manipulation against the backdrop of other variability inherent in the dataset. Traditional analysis of the data collected throughout this thesis would involve the use of repeated measures analyses of variance (ANOVAs). This technique assesses whether there are significant differences 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 individual participant responses, nor do they allow for simultaneous consideration of by-item and by-participant variance. It is for these reasons that linear mixed-effects models were used throughout. Linear mixed-effects
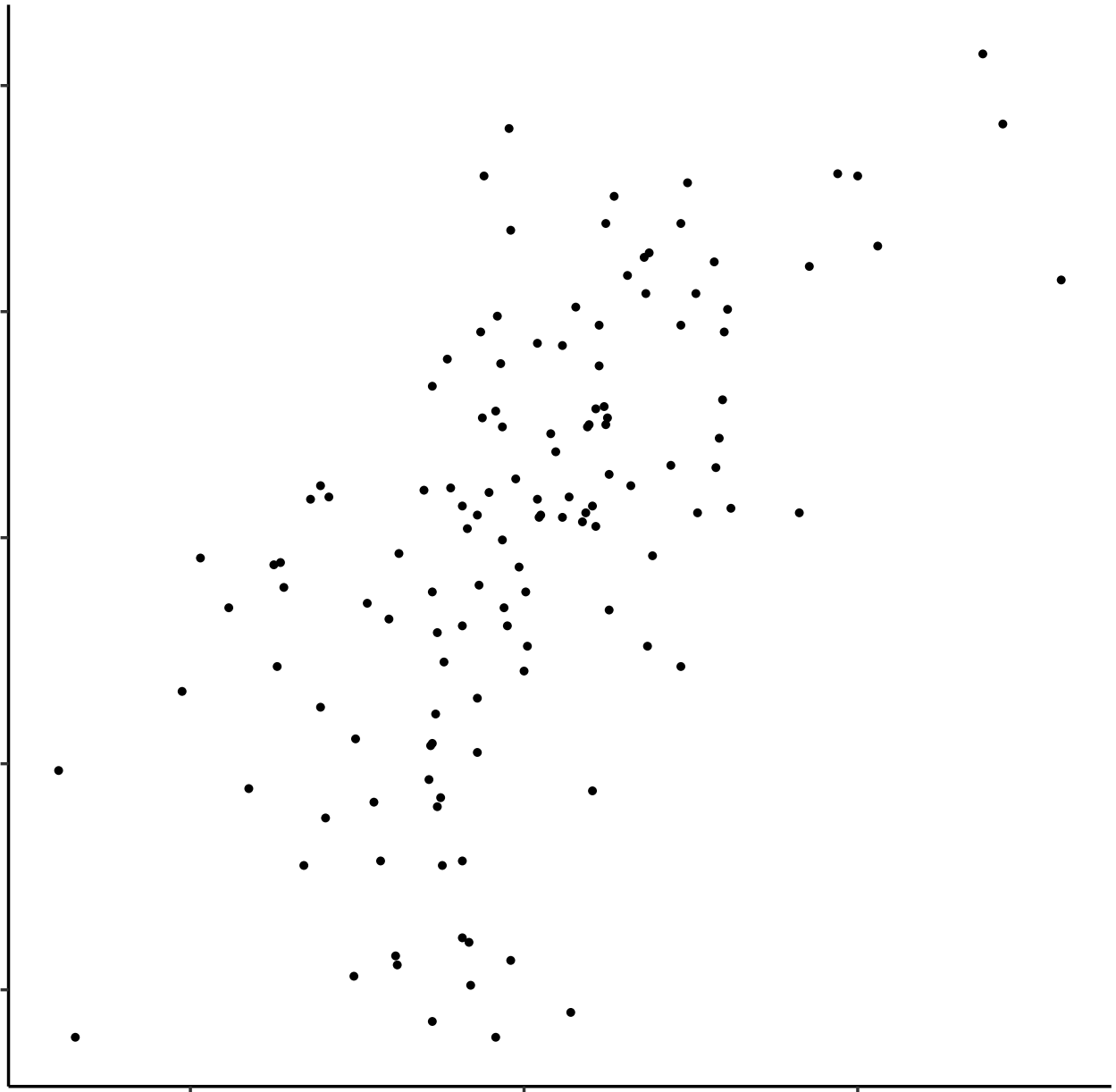
Figure 3.7. The basic design of scatterplots in experiments 1 to 4.

modelling is a reliable approach that is resistant to a variety of distributional assumption violations [63], and facilitates the appreciation of the data story in a broader and more detailed fashion.

3.3.1 Linear Mixed-Effects Models

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
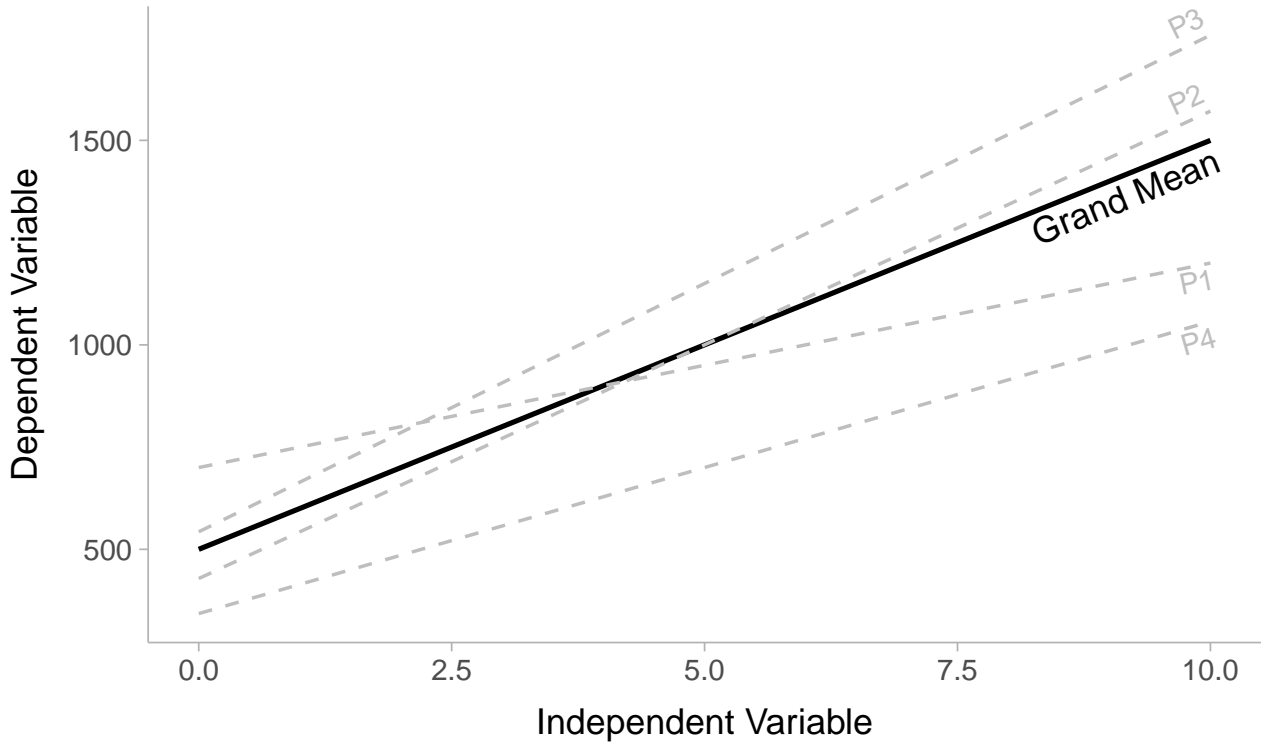
Figure 3.8. Visualising random intercepts and slopes for a theoretical experiment with 4 participants. The grand mean of the dependent variable is shown as a solid line, while each separate random intercept is drawn with dashed lines. Each line has a different gradient, representing different random slopes for each participant. This graphic was inspired by those featured in Brown, 2021 [14].

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 random effect [14]. Figure 3.8 visualises these concepts.

Throughout the course of this thesis, analyses attempt to model both random intercepts and slopes in order to capture the maximum amount of variability present in our datasets. In order to ascertain the goodness-of-fit of models, their ability to explain variance is compared to that of a nested null model [65]; such a model is identical bar the removal of the fixed effect of interest. The likelihood ratio test (LRT) is used here to assess goodness-of-fit. In cases where a model has in total more than two levels (here, all experiments bar experiment 5), the `emmeans` package [**lenth_2024**] is used to calculate estimated marginal means between levels of fixed effects.

3.3.2 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 back this assumption. Metric modelling is therefore considered inappropriate for modelling responses to Likert scales [38]. In light of these issues, the `ordinal` package [18] in R was used to build cumulative link mixed-effects models for the analysis of Likert scale data. This allows for the treatment of Likert responses as ordered factors as opposed to

continuous response scales.

3.3.3 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. [8] 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. [9] have argued that attempting to specify maximal models for insufficiently rich datasets may lead to overfitting and unreliable conclusions. In light of this I 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), the `buildmer` package [73] in R was used to automate the selection of 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 stepwise elimination of terms is evaluated using successive likelihood ratio tests. This results in a model that captures the maximal amount of feasible variability while minimising redundancy. Note that `buildmer` was not relied upon as a modelling *panacea*; models are still based on theoretical underpinnings and are evaluated critically.

3.3.4 Effects Sizes

My 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 featured a condition with no scatterplot manipulation present (henceforth referred to as a *baseline*); accordingly, the `EMAtools` package [30] was used to calculate equivalent Cohen's *d* effect sizes of manipulation-present conditions relative to the baseline. Experiment 4 did not feature a baseline condition, meaning Cohen's *d* was deemed inappropriate. The `r2glmm` package [28] was used instead to calculate semi-partial R^2. In lieu of a traditional measure of effect size, this demonstrated the unique variance in the dependent variable explained by each level of the independent [45]. Experiment 5 features a much simpler modelling situation, and returns to providing equivalent Cohen's *d* values for the pre- vs post- plot viewing conditions, this time calculated by converting odds ratios using the `effectsize` package [**effectsize**]. More details on specific calculations, measures, and conclusions can be found *in situ*.

3.3.5 Reporting Analyses

Throughout this thesis, a broad approach to the reporting of statistical analyses was taken; while I consider our analytical methods and conclusions valid, I also present a range of statistics to allow

the reader to draw their own conclusions should they wish. Statistical results are visualised where appropriate, and where visualisation aids understanding and interpretation. In addition, details about model structures and the issues I tackled when modelling are included for transparency [42].

3.4 Computational Methods

The approach to computational methods in this thesis sought to marry practicality, 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 literate programming [32] 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, these techniques, principles, and practices are shared to enable future researchers to learn, where I struggled. In this section, I detail my approach to computational methods, including how the idea of **executable papers** was utilised, and how containerised environments were used to capture a freeze-frame of the analyses.

3.4.1 Executable Reporting

Each paper published throughout this project, and this thesis, has been written to be executable. Packaging research in such a way means a lay person can follow simple instructions to recreate the work, while also facilitating and encouraging literate programming, or the close alignment of documentation and underlying code [54].

The use of a literate programming paradigm to generate reports (usually using LaTeX) has a rich history. This section focuses on this history as it pertains to the language used throughout this project, R. `Sweave` [35], written in 2002, allowed R code to be integrated into LaTeX documents. This was followed by Yihui Xie's `knitr` [75], which expanded `Sweave` functionality and improved integration with tools such as `pandoc` [40]. `knitr` uses `Rmarkdown` [76] to mix markdown-flavoured text with code chunks into a document that can be rendered into an appropriately-formatted conference or journal pdf; this workflow was used for the papers associated with experiments 1, 2, and 3. `Quarto` [2], released in 2022, further expands on `Rmarkdown` functionality, and removes reliance on R or Rstudio. `Quarto` was used 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 [27]. 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 [20, 61, 72]. 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.

These issues were addressed using containers, specifically, those created by Docker [11, 41]. 1979 saw the development of `chroot` (`change root`), which is able to isolate an application's file access to a 'chroot jail'. Since then, we have seen the 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. Recording software versions and dependencies avoids the potential for broken code in the future, and publicly hosting papers as GitHub repositories that build into Docker containers ensures that future researchers can interact with code and data in the same computational environment used when carrying out the research. While virtual machines make isolated sections of hardware available, containers abstract protected parts of the operating system [41]. This makes containers smaller and more lightweight than full virtual machines, while still conferring the advantages of virtualisation. For the Docker implementation here, portable R environments provided by the Rocker project [12] are used. These environments are agnostic regarding the host operating system, allowing the reader to reproduce the analyses featured here in a replica of the computational environment they were conducted in.

Building Docker containers is facilitated through a Dockerfile. This file instructs Docker to build a container with the appropriate version of R, the files required, and the correct package versions used during analysis. Below is the Dockerfile used to reproduce this thesis.

I first specify the Rocker image that will form the basis of the container. This includes the version of R required (version 4.4.2), the Rstudio Integrated Development Environment (IDE), Quarto, and the `tidyverse` package.

```
FROM rocker/version:4.4.2
```

Next, I add the files and folders required, including the Quarto document and related files, chapter folder, bibliography, additional scripts, LaTeX class file and template, the folders containing the cached models and raw data, and the R project file:

```
ADD thesis.qmd /home/rstudio/
```

```
ADD _quarto.yml /home/rstudio/
```

```
ADD chapters_quarto/ /home/rstudio/chapters_quarto/

ADD thesis.bib /home/studio/

ADD reformat_tex.R /home/studio/

ADD finalise_thesis.R /home/rstudio/

ADD helper_functions.R /home/rstudio/

ADD uom_thesis_casson.cls /home/rstudio/

ADD main.tex /home/rstudio/

ADD data/ /home/rstudio/data/

ADD cache/ /home/rstudio/cache/

ADD thesis.Rproj /home/rstudio/
```

Finally, I add the specific versions of the R packages used throughout the course of this thesis. For brevity, I only display the addition of the first three here:

```
RUN R -e "devtools::install_version('MASS', version = '7.3-60', dependencies = T)"

RUN R -e "devtools::install_version('buildmer', version = '2.10.1', dependencies
= T)"

RUN R -e "devtools::install_version('emmeans', version = '1.8.8', dependencies =
T)"
```

3.5 Reproducibility In This Thesis

Reproducibility is a broad spectrum [53] (see Figure 3.9. 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 [47] crystallised in the early 2010s [52], concerns had been voiced since at least the late 1960s [60]. 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 [62]). These issues led this project to strive for a gold standard [53] of reproducibility throughout. In this section, I detail how this was accomplished, and in doing so, expose my work to welcome critique.

3.5.1 Sharing Data and Code

The open and public sharing of data and code facilitates external assessment [3, 31] and secondary use of data [70], and guards against reproducibility issues [44]. Quite aside from external motivating factors, I found that developing and embedding the reproducibility practices described here have resulted

Not reproducible				Gold standard
Publication only	Code	Code and data	Linked and executable code and data	Full replication
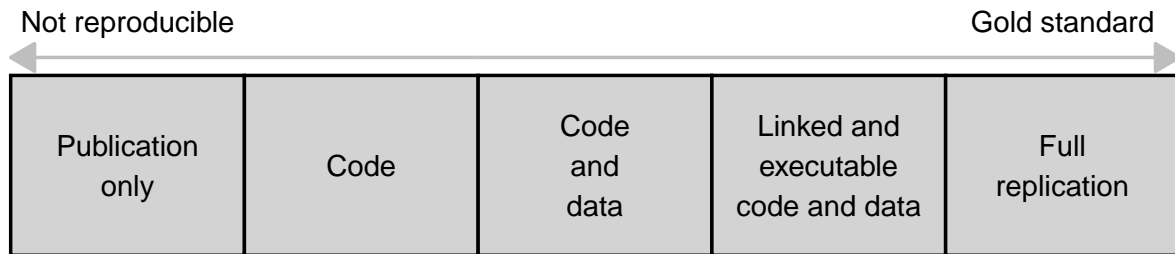
Figure 3.9. Peng's (2011) Reproducibility Spectrum. This figure has been reproduced from Peng (2011) [53].

in longer term savings in time and effort. Favouring a gold-standard reproducible approach is also a way of "paying it forward"; having come from a non-technical background, I found previous work that adhered to the same standard critical to learning and development. GitHub is used to host both this thesis and the papers associated with the project; links to these repositories can be found throughout. I favour permissive and lenient licencing, such as the MIT licence [71] for GitHub repositories and the CC-BY 4.0 license for pre-registrations. These enable future researchers to re-use data and code while providing clear guidance for appropriate use and facilitating long-term sustainability [29].

3.5.2 Executable Papers and Docker Containers

As detailed above, Quarto and Docker were used 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, a single implementation of Docker to is provided 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 [39], 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 [49]. All hypotheses and analysis plans were pre-registered with the Open Science Framework [48]. Pre-registrations are embargoed by the research team prior to data collection, and then made public in a frozen state following publication of the corresponding research. Deviations from registered analysis plans are detailed 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, I 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. I have provided justifications for the designs used, the methodological challenges faced, and how the use of a broad array of tools and techniques was able to overcome these challenges. Throughout, I have detailed how I have learnt from my mistakes. Open research and reproducibility is at the core of the work described here, and I hope this thesis can serve as an example for future work facing similar challenges and with similar commitments to open science. To this end, I have produced a template to facilitate future reproducible theses. We satisfy FAIR (**F**indable, **A**ccessible, **I**nteropable, and **R**eusable) data principles [74] through public sharing of data and code, literate programming, and containerisation.

Chapter 4

Adjusting the Opacities of Scatterplot Points Can Affect Correlation Estimates

4.1 Abstract

Scatterplots are common data visualisations utilised for communication with experts and lay people alike. Despite being widely studied, it is common for people to underestimate the level of correlation displayed in them. The weight of evidence points toward changes in the opacities of scatterplot points being unable to change perceptions of correlation, however this was not tested rigorously using systematic adjustments. Drawing on evidence that the shape of a scatterplot's point cloud may drive correlation perception, I conducted exploratory work addressing this underestimation bias. In two experiments (total $N = 300$), evidence is provided that changing the opacities of scatterplot points *can* have small effects on participants' performance on a correlation estimation task. The systematic adjustment of point opacity as a function of residual distance is able to alter estimates to a greater degree and correct for the underestimation bias. In this chapter, I also present an early pilot study that was ultimately not included in any published works.

4.2 Preface: Learning From an Early Pilot Study

The research proposal that kickstarted this project in 2021 set out a plan to investigate the perception of correlation in scatterplots as a function of screen size. This was prompted by recent research demonstrating consistent perceptual biases in scatterplots due to geometric scaling [**wei_2020**], the increasing prevalence of data visualisations in lay people's daily lives due to the COVID-19 pandemic, and the increasing adoption of wearable devices. The first experiment conducted therefore examined how perceptions of correlation changed according to the size of a scatterplot. Additionally, a very early version of the opacity decay factor from experiment 2 was included. The implementation of this factor was immature. In experiment 2 onwards, if a scatterplot point resided in a particular place, it would always have the same opacity or size. In the pilot study, the code that set the opacity of each point always scaled the opacity values such that the point with the highest residual had the lowest possible opacity, and vice versa, resulting in the plots seen in Figure 4.1.

This provided no consistency between different experimental stimuli, making it difficult to comment on the effects of changing levels of opacity in various parts of a plot on correlation estimation. This
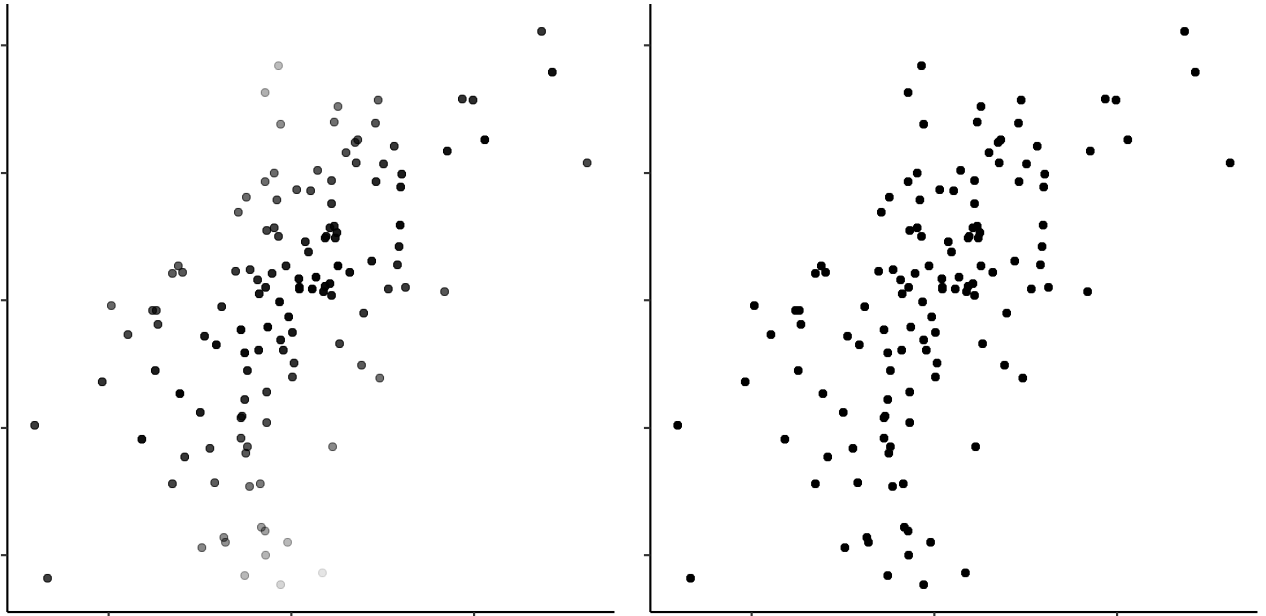
Figure 4.1. Examples of the experimental stimuli used in the pilot study (opacity decay factor). On the left, the opacity decay function is visible. Note the linear scaling used.

was later addressed by hardcoding residual size to a specific value of opacity or size. The pilot also suffered extensively from poor data quality. Of the 260 participants tested, data from only 118 was included in the final analyses due to failed attention checks. It is for this reason that the pre-screen requirements detailed in Section 3.2.3 were implemented.

Participants viewed 180 experimental plots in a 3x2 factorial design. The first independent variable, plot size, had three levels, 63%, 100%, and 252% scales. The second was the presence or absence of the opacity decay function (see Figure 4.1). I aimed to recruit 150 participants, but stopped after 118 due to data quality issues. Nevertheless, the results provided were crucial in informing the future direction of the research project. I present these results in brief below.

4.2.1 Pilot Study: Results

To investigate the effects of plot size and the presence or absence of an opacity decay manipulation on participants' estimates of correlation, a linear mixed effects model was built whereby participants' errors in correlation estimation was predicted by plot size and the presence of the opacity decay function. This model features random intercepts for participants and items, as well as random slopes for both participants and items relevant to the presence or absence of the opacity decay function. A likelihood ratio test between the experimental model and a null model with the fixed effects removed revealed that the experimental model explained significantly more variance than the null ($\chi^2(3) = 26.38$, $p <$ .001). There was no interaction between plot size and the presence or absence of the opacity decay function. The **emmeans** [36] package was used to explore estimated marginal means (see Table 4.1) and contrasts (see Table 4.2) separately for each factor.

Table 4.1. Estimated Marginal Means for plot size (left) and the presence of the opacity decay function (right).

Size	Mean	Standard Error
Large (252%)	0.12	0.014
Medium (100%)	0.12	0.014
Small (62%)	0.13	0.014

Decay	Mean	Standard Error
Absent	0.13	0.015
Present	0.11	0.013

Table 4.2. Contrasts between levels of the size factor (left) and opacity decay factor (right).

Contrasts		Statistics	
		Z ratio	p
Large (252%)	Medium (100%)	-0.94	0.618
Large (252%)	Small (52%)	-3.56	0.001
Medium (100%)	Small (52%)	-2.63	0.023

Contrasts		Statistics	
		Z ratio	p
Absent	Present	3.65	<0.001

4.2.2 Pilot Study: Discussion

For the factor of plot size, the effect observed was driven by significant differences in correlation estimation error between large and small plots. Participants estimated more accurately when the plot was large and when the decay function was present. Participants still underestimated correlation in all conditions. The finding that estimation error was lower for larger plots is in line with previous evidence that geometrically scaling a scatterplot up can increase perceptions of the strength of the correlation displayed [**wei_2020**]. Despite the statistical significance of this finding, we elected at this point to abandon the plot size factor due to the extremely small effect (see Table 4.1) and lack of novelty.

The impact of even an immature point opacity decay function on correlation estimation was a novel finding that we felt deserved further, and more rigorous, study. Its implementation was based on findings that changing the opacities of scatterplot points could bias estimates of means [**hong_2021**], and on limited evidence for the perception of correlation being based on the perceived width of a probability distribution represented by the arrangement of scatterplot points. I did not foresee the decay function, being novel, having a greater effect on correlation estimation than size. Once evidence had been found that changing the opacity of points in scatterplots could have effects on correlation estimation, in opposition to previous research [57, 58], the door was opened for a rigorous investigation into how this worked and how it could be used systematically to correct for the historic underestimation bias.

4.3 Introduction

4.3.1 Overview

In two experiments, the opacities of point in scatterplots was manipulated while participants were asked to make judgements of correlation. In the first, point opacity is changed in a uniform manner, while in the second, point opacity is systematically altered as a function of the size of a particular point's residual. By comparing participants' performance on a correlation estimation task for data-identical scatterplots that vary only in the opacities of their points, it is demonstrated that; lower global

point opacity results in greater errors in the estimation of positive correlation (experiment 1); and lowering point opacity as a function of the size of a point's residual is able to bias estimates of positive correlation upwards to

4.4 Related Work

4.4.1 Transparency, Contrast, Opacity, and Formal Definitions

- include the "formalising contrast" part of the original papers general methods section here
- also include justification for referring to "opacity" instead of contrast

4.4.2 Effects of Point Opacity on Correlation Estimation

4.5 General Methods

The experiments described in this chapter share multiple aspects of their procedures. Both experiments were built using PsychoPy [51] and are hosted on pavlovia.org. Both use 1-factor, 4-level designs. Ethical approval for both experiments was granted by the University of Manchester's Computer Science Departmental Panel (Ref: 2022-14660-24397). In each experiment, participants were shown the respective Participant Information Sheet (henceforth PIS) and provided consent through key presses in response to consent statements. Participants were asked to provide their age and gender identity, after which they completed the 5-item Subjective Graph Literacy test described by Garcia-Retamero et al. [25] and discussed in Section 2.9 of the literature review. Early piloting with a graduate student in humanities suggested the potential for participants to be unfamiliar with the visual nature of different values of Pearson's r. Participants were therefore shown examples of $r = 0.2$, 0.5, 0.8, and 0.95 (see Figure 4.2); a discussion of the effects of this training is provided in Section 4.8.1. Participants were given two practice trials to familiarise themselves with the response slider.

Each trial was preceded by text that either told the participant:

- Please look at the following plot and use the slider to estimate the correlation (n = 180).
- Please IGNORE the correlation displayed and set the slider to 1 (n = 3) or 0 (n = 3).

The latter instructions were attention checks, and were formatted with red text to increase their visibility. Each experimental trial was preceded by a visual mask (see Figure 4.3) that was displayed for 2.5 seconds. Participants were instructed to make their judgements as quickly and accurately as possible, but there was no time limit per trial. Both experiments described here use a fully repeated-measures, within-participants design. Participants saw all 180 experimental items, corresponding to ~27,000 individual judgements per experiment, in a fully randomised order.

Both experiments were conducted according to principles of open and reproducible research. All data
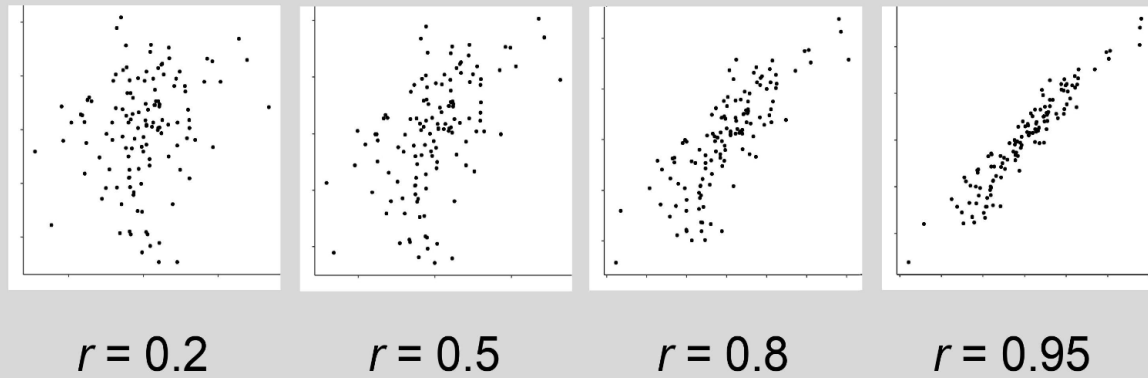
Figure 4.2. Participants viewed these plots for at least eight seconds before being allowed to continue to the practice trials.

and analysis code for the origin paper is available on GitHub [1]. Experiment 1 [2] and 2 [3] are hosted on Pavlovia.org, while the Open Science Framework hosts pre-registrations [4]. It is important to note at this point that experiment 2 was conducted prior to experiment 1; when the original paper was written, the order of presentation of the experiments was swapped to make the narrative more cohesive. I preserve this order in the present chapter.

4.6 Experiment 1: Uniform Opacity Adjustments

4.6.1 Introduction

Owing to the robust effects of altering stimulus opacity on perception described above [**wehrhahn_1990**, **champion_2017**], it was hypothesised that there would be a greater spread of estimates of correlation for plots with lower global opacity compared to higher opacity plots.

4.6.2 Method

Participants

150 participants were recruited using the Prolific platform [56]. Normal to corrected-to-normal vision and English fluency were required. Participants who had completed the pre-study were prevented from participating. Data were collected from 158 participants. 8 failed more than 2 out of 6 attention check questions, and, as per the pre-registration, had their submissions rejected from the study. The data

---

[1] https://github.com/gjpstrain/contrast_and_scatterplots
[2] https://gitlab.pavlovia.org/Strain/exp_uniform_adjustments
[3] https://gitlab.pavlovia.org/Strain/exp_spatially_dependent
[4] Experiment 1 - https://osf.io/tuexh. Experiment 2 - https://osf.io/6f5ev

Figure 4.3. An example of a visual mask displayed for 2.5 seconds before each experimental trial.

from the remaining 150 participants were included in the full analysis (50.67% male, 47.33 % female, and 1.33% non-binary). Participants' mean age was 28.29 (SD = 8.59). Mean graph literacy score was 21.79 (SD = 4.47). The mean time taken to complete the experiment was 33 minutes (SD = 10 minutes).

Design

For each of the 45 r values, there were four versions of each plot corresponding to the four levels of point opacity. Examples of each of these can be seen in Figure 4.4, demonstrated with an r value of 0.6.

Figure 4.4. Examples of the stimuli used in experiment 1, demonstrated with an *r* value of 0.6. Here, "opacity" refers to the alpha value used by ggplot.

Figure 4.5. Estimated marginal means for the four conditions tested in experiment 1. 95% confidence intervals are shown. The vertical dashed line represents no estimation error. The overestimation zone is included to facilitate comparison to later work.

4.6.3 Analysis

To investigate the effects of opacity condition on participants' estimates of correlation, a linear mixed effects model was built whereby opacity condition is a predictor for the difference between objective *r* values for each plot and participants' estimates of *r*. This model has random intercepts for items and participants. A likelihood ratio test revealed that the mode including global opacity as a fixed effect explained significantly more variance than a null model ($\chi^2(3) = 223.13$, $p < .001$). Figure 4.5 shows the mean errors in correlation estimation for each opacity condition, along with 95% confidence intervals.

Table 4.3. Contrasts between different levels of the opacity factor in experiment 1.

Contrasts		Statistics	
		Z ratio	p
Full Opacity (alpha = 1.0)	High Opacity (alpha = 0.75)	-1.363	0.523
Full Opacity (alpha = 1.0)	Medium Opacity (alpha = 0.5)	-6.809	<0.001
Full Opacity (alpha = 1.0)	Low Opacity (alpha = 0.25)	-13.439	<0.001
High Opacity (alpha = 0.75)	Medium Opacity (alpha = 0.5)	-5.443	<0.001
High Opacity (alpha = 0.75)	Low Opacity (alpha = 0.25)	-12.071	<0.001
Medium Opacity (alpha = 0.5)	Low Opacity (alpha = 0.25)	-6.631	<0.001

Table 4.4. Cohen's *d* effect sizes for levels of the opacity factor in experiment 1. Each effect size is compared to the reference level, full contrast (alpha = 1).

Effect	Cohen's *d*
Full Opacity (alpha = 1.0)	
High Opacity (alpha = 0.75)	0.02
Medium Opacity (alpha = 0.5)	0.08
Low Opacity (alpha = 0.25)	0.16

This effect was driven by significant difference between means of correlation estimation error between all conditions bar high and full opacity. Statistical tests for contrasts were performed using the **emmeans** package [36], and are shown in Table 4.3. To test whether the observed results could be explained by difference in participants' levels of graph literacy, an additional model was built. This model is identical to the experimental model, but also includes graph literacy as a fixed effect. Including graph literacy as a fixed effect explained no additional variance ($\chi^2(1) = .002$, $p = .962$), indicating that the differences observed in participants' correlation estimation performance were not as a result of differences in levels of graph literacy.

A function from the now archived **EMAtools** package [30] was used to calculate an approximation of Cohen's *d* between the reference level (full contrast, alpha = 1.0) and each other level of the opacity factor. These statistics can be seen in Table 4.4. The largest effect size observed ($d \sim 0.16$) is between the low and full opacity conditions, and is small. This was unsurprising given the lack of previously reported effects on correlation perception of global point opacity [57].

4.6.4 Discussion

4.7 Experiment 2: Spatially-Dependent Opacity Adjustments

4.7.1 Introduction

4.7.2 Methods

Participants

150 participants were recruited using the Prolific platform [56]. Normal to corrected-to-normal vision and English fluency were required. Participants who had completed the pre-study were prevented from participating. Data were collected from 158 participants. 7 failed more than 2 out of 6 attention check questions, and, as per the pre-registration, had their submissions rejected from the study. The data from the remaining 150 participants were included in the full analysis (51.33% male, 46.00 % female, and 2.67% non-binary). Participants' mean age was 27.05 ($SD = 7.37$). Mean graph literacy score was 21.71 ($SD = 4.06$). The average time taken to complete the experiment was 33 minutes (SD = 10 minutes).
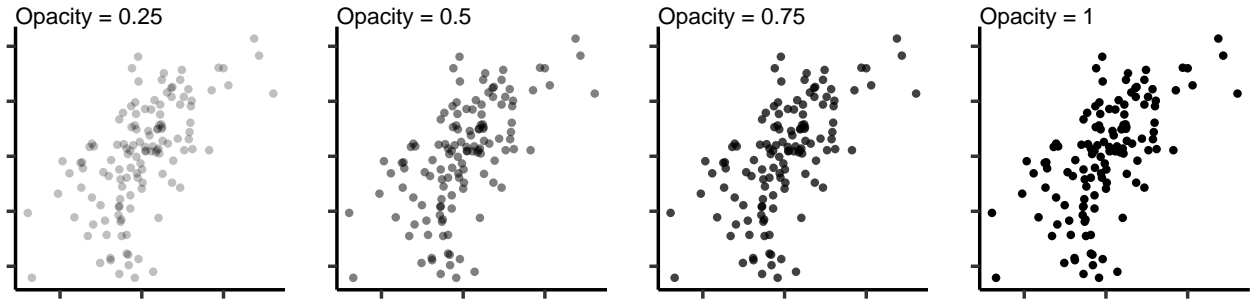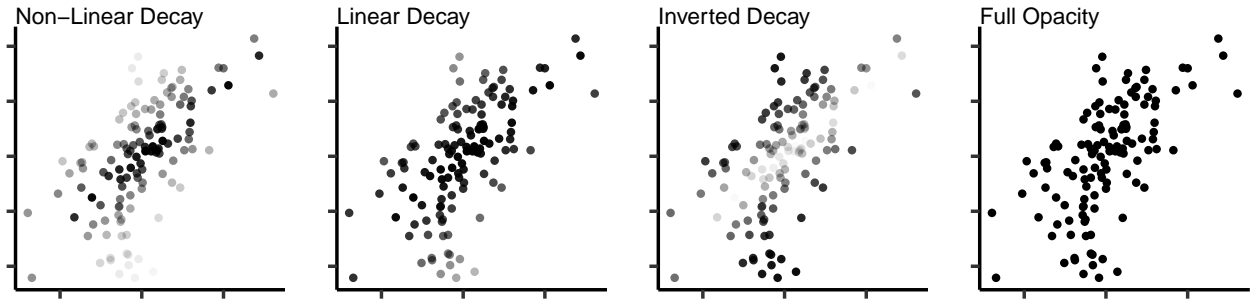
Figure 4.6. Examples of the stimuli used in experiment 2, demonstrated with an *r* value of 0.6. Here, "opacity" refers to the alpha value used by ggplot.
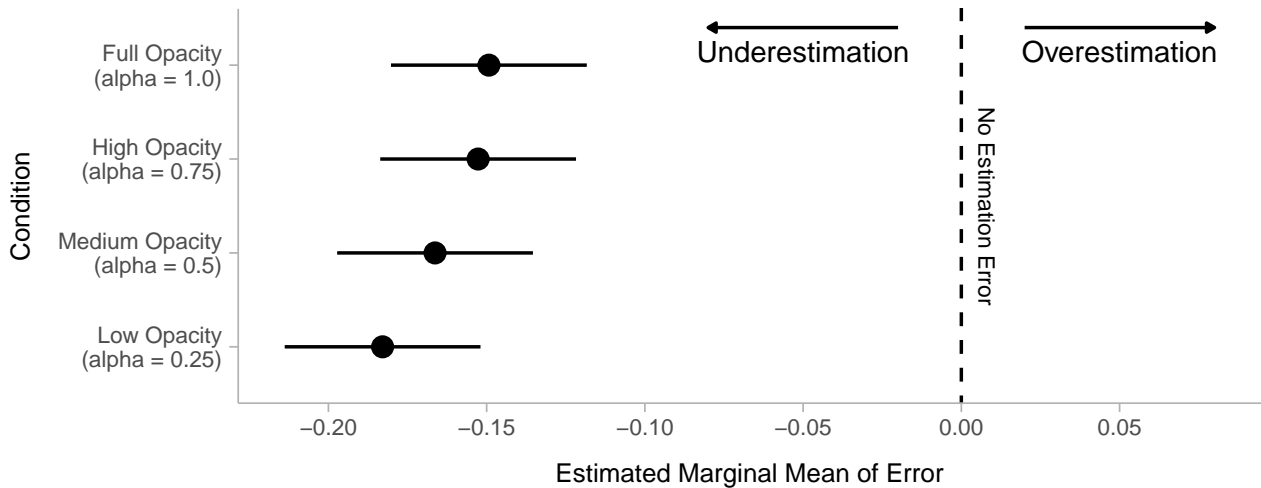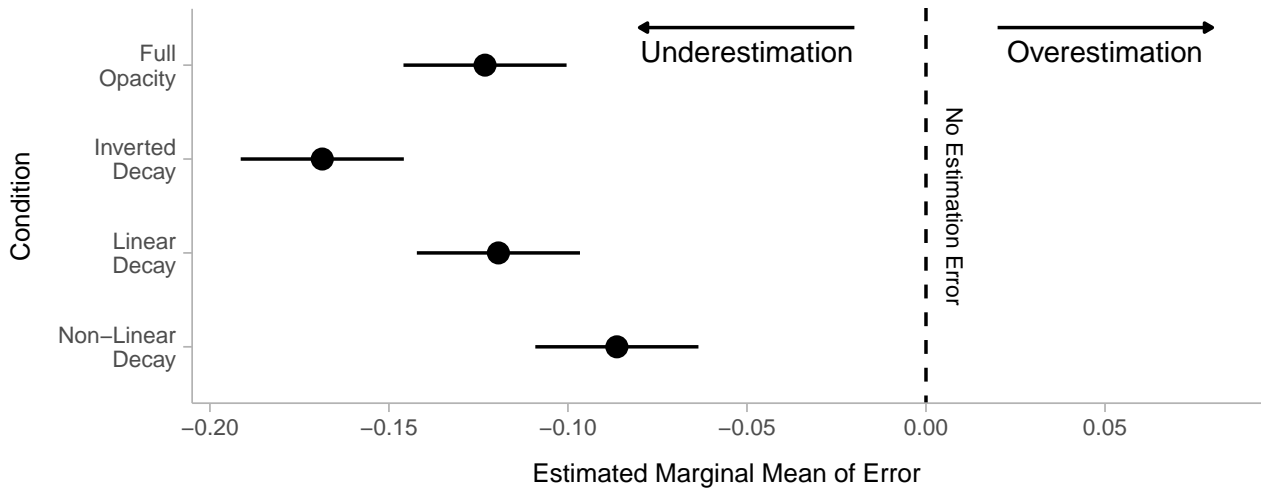

Figure 4.7. Estimated marginal means for the four conditions tested in experiment 2. 95% confidence intervals are shown. The vertical dashed line represents no estimation error.

Design

For each of the 45 *r* values in experiment 2, there were four versions of each plot corresponding to the three levels of point opacity decay function and the baseline global full opacity condition. Examples of each of these can be seen in Figure 4.6, demonstrated with an *r* value of 0.6.

4.7.3 Analysis

To investigate the effects of the opacity decay functions on participants' estimates of correlation, a linear mixed effects model was built whereby decay function condition is a predictor for the difference between objective *r* values for each plot and participants' estimates of *r*. This model has random intercepts for items and participants. A likelihood ratio test revealed that the mode including global opacity as a fixed effect explained significantly more variance than a null model ($\chi^2(3) = 1{,}157.62$, $p < .001$). Figure 4.5 shows the mean errors in correlation estimation for each opacity decay function condition, along with 95% confidence intervals.

The effect seen in experiment 2 was driven by significant differences in means of correlation estimation error between all levels of opacity decay function condition bar full opacity and linear decay. Statistical testing for contrasts was performed using the **emmeans** package [36], and are shown in Table 4.5. To

Table 4.5. Contrasts between different levels of opacity decay function in experiment 2.

Contrasts		Statistics	
		Z ratio	p
Full Opacity	Inverted Decay	-18.9	<0.001
Full Opacity	Linear Decay	1.6	0.405
Full Opacity	Non-Linear Decay	15.3	<0.001
Inverted Decay	Linear Decay	20.4	<0.001
Inverted Decay	Non-Linear Decay	34.2	<0.001
Linear Decay	Non-Linear Decay	13.7	<0.001

Table 4.6. Cohen's *d* effect sizes for levels of the opacity decay function factor in experiment 2. Each effect size is compared to the reference level, full contrast (alpha = 1).

Effect	Cohen's *d*
Full Opacity	
Inverted Decay	0.23
Linear Decay	-0.02
Non-Linear Decay	-0.19

test whether the observed results could be explained by differences in graph literacy, a model including participants' graph literacy scores as a fixed effect was built. Including graph literacy as a fixed effect again explained no additional variance ($\chi^2(1) = .242$, $p = .623$).

Approximated Cohen's *d* effect sizes between the baseline (global full opacity) and each other condition can be seen in Table 4.6. The largest effect size observed (~0.23 between full opacity and inverted decay conditions) is small to moderate, and the effect size between the baseline and the non-linear decay condition (~0.19) is small.

4.7.4 Discussion

4.8 General Discussion

4.8.1 Training

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. 28).

[2] JJ Allaire and Christophe Dervieux. *Quarto: R Interface to 'quarto' Markdown Publishing System*. Manual. 2024 (cited on p. 26).

[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. 28).

[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. 19).

[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. 21).

[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. 19).

[7] Tarek Azzam, Stephanie Evergreen, Amy A. Germuth, and Susan J. Kistler. "Data Visualization and Evaluation". In: *New Directions for Evaluation* 2013.139 (2013), pp. 7–32. ISSN: 1534-875X. DOI: 10.1002/ev.20065. (Visited on 09/08/2022) (cited on p. 13).

[8] 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. 25).

[9] 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. 25).

[10] Philip Bobko and Ronald Karren. "The Perception of Pearson Product Moment Correlations from Bivariate Scatterplots". In: *Personnel Psychology* 32.2 (1979), pp. 313–325. ISSN: 1744-6570. DOI: 10.1111/j.1744-6570.1979.tb02137.x. (Visited on 06/14/2022) (cited on pp. 13, 14, 22).

[11] 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. 27).

[12] Carl Boettiger and Dirk Eddelbuettel. "An Introduction to Rocker: Docker Containers for R". In: *The R Journal* 9.2 (2017), p. 527. ISSN: 2073-4859. DOI: 10.32614/RJ-2017-065. (Visited on 10/16/2024) (cited on p. 27).

[13] 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. 19).

[14] 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 p. 24).

[15] Marc Brysbaert. "How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables". In: *Journal of cognition* 2.1 (2019) (cited on p. 22).

[16] 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. 21).

[17] Gary Charness, Uri Gneezy, and Michael A. Kuhn. "Experimental Methods: Between-subject and within-Subject Design". In: *Journal of Economic Behavior & Organization* 81.1 (Jan. 2012), pp. 1–8. ISSN: 0167-2681. DOI: 10.1016/j.jebo.2011.08.009. (Visited on 10/14/2024) (cited on p. 18).

[18] Rune H. B. Christensen. *Ordinal—Regression Models for Ordinal Data*. Manual. 2023 (cited on p. 24).

[19] W. S. Cleveland, P. Diaconis, and R. Mcgill. "Variables on Scatterplots Look More Highly Correlated When the Scales Are Increased". In: *Science* 216.4550 (June 1982), pp. 1138–1141. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.216.4550.1138. (Visited on 02/08/2021) (cited on pp. 14, 22).

[20] 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. 27).

[21] Charles E. Collyer, Kerrie A. Stanley, and Caroline Bowater. "Psychology of the Scientist: LXIII. Perceiving Scattergrams: Is Visual Line Fitting Related to Estimation of the Correlation Coefficient?" In: *Perceptual and Motor Skills* 71.2 (Oct. 1990), 371–378E. ISSN: 0031-5125. DOI: 10.2466/pms.1990.71.2.371. (Visited on 09/13/2022) (cited on p. 14).

[22] 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. 21).

[23] *E-Prime*. Psychology Software Tools. 2020 (cited on p. 19).

[24] M. Friendly and Daniel J. Denis. "The Early Origins and Development of the Scatterplot." In: *Journal of the history of the behavioral sciences* (2005). DOI: 10.1002/JHBS.20078 (cited on p. 13).

[25] Rocio Garcia-Retamero, Edward T. Cokely, Saima Ghazal, and Alexander Joeris. "Measuring Graph Literacy without a Test: A Brief Subjective Assessment". In: *Medical Decision Making* 36.7 (2016), pp. 854–867. ISSN: 0272-989X. DOI: 10.1177/0272989X16655334. (Visited on 04/30/2021) (cited on p. 34).

[26] 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. 21).

[27] 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. 26).

[28] Byron Jaeger. *R2glmm: Computes R Squared for Mixed (Multilevel) Models*. Manual. 2017 (cited on p. 25).

[29] Rafael C. Jiménez, Mateusz Kuzak, Monther Alhamdoosh, Michelle Barker, Bérénice Batut, Mikael Borg, Salvador Capella-Gutierrez, Neil Chue Hong, Martin Cook, Manuel Corpas, Madison Flannery, Leyla Garcia, Josep Ll Gelpí, Simon Gladman, Carole Goble, Montserrat González Ferreiro, Alejandra Gonzalez-Beltran, Philippa C. Griffin, Björn Grüning, Jonas Hagberg, Petr Holub, Rob Hooft, Jon Ison, Daniel S. Katz, Brane Leskošek, Federico López Gómez, Luis J. Oliveira, David Mellor, Rowland Mosbergen, Nicola Mulder, Yasset Perez-Riverol, Robert

Pergl, Horst Pichler, Bernard Pope, Ferran Sanz, Maria V. Schneider, Victoria Stodden, Radosław Suchecki, Radka Svobodová Vařeková, Harry-Anton Talvik, Ilian Todorov, Andrew Treloar, Sonika Tyagi, Maarten van Gompel, Daniel Vaughan, Allegra Via, Xiaochuan Wang, Nathan S. Watson-Haigh, and Steve Crouch. *Four Simple Recommendations to Encourage Best Practices in Research Software*. June 2017. DOI: `10.12688/f1000research.11407.1`. F1000Research: `6:876`. (Visited on 10/15/2024) (cited on p. 29).

[30]   Evan Kleiman. *EMAtools: Data Management Tools for Real-Time Monitoring/Ecological Momentary Assessment Data*. Manual. 2021 (cited on pp. 25, 38).

[31]   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. 28).

[32]   D. E. Knuth. "Literate Programming". In: *The Computer Journal* 27.2 (Jan. 1984), pp. 97–111. ISSN: 0010-4620. DOI: `10.1093/comjnl/27.2.97`. (Visited on 10/11/2024) (cited on p. 26).

[33]   David Lane, Craig Anderson, and Kathryn Kellam. "Judging the Relatedness of Variables. The Psychophysics of Covariation Detection". In: *Journal of Experimental Psychology: Human Perception and Performance* 11 (Oct. 1985), pp. 640–649. DOI: `10.1037/0096-1523.11.5.640` (cited on p. 14).

[34]   Thomas W. Lauer and Gerald V. Post. "Density in Scatterplots and the Estimation of Correlation". In: *Behaviour & Information Technology* 8.3 (June 1989), pp. 235–244. ISSN: 0144-929X, 1362-3001. DOI: `10.1080/01449298908914554`. (Visited on 09/05/2023) (cited on p. 14).

[35]   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. 26).

[36]   Russell V. Lenth. *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*. Manual. 2024 (cited on pp. 32, 38, 39).

[37]   Rui Li. "Visualizing COVID-19 Information for Public: Designs, Effectiveness, and Preference of Thematic Maps". In: *Human Behavior and Emerging Technologies* 3.1 (2021), pp. 97–106. ISSN: 2578-1863. DOI: `10.1002/hbe2.248`. (Visited on 10/22/2024) (cited on p. 13).

[38]   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. 24).

[39] 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. 29).

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

[41] 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. 27).

[42] 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. 26).

[43] Joachim Meyer, Meirav Taieb, and Ittai Flascher. "Correlation Estimates as Perceptual Judgments". In: *Journal of Experimental Psychology: Applied* 3.1 (1997), pp. 3–20. ISSN: 1939-2192. DOI: 10.1037/1076-898X.3.1.3 (cited on p. 14).

[44] 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. 28).

[45] 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. 25).

[46] Thomas Nocke, Till Sterzel, Michael Böttinger, Markus Wrobel, et al. "Visualization of Climate and Climate Change Data: An Overview". In: *Digital earth summit on geoinformatics* (2008), pp. 226–232 (cited on p. 13).

[47] 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. 28).

[48] *OSF*. https://osf.io/. (Visited on 10/10/2024) (cited on p. 29).

[49] 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. 29).

[50] 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. 21).

[51] 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 pp. 19, 34).

[52] 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. 28).

[53] 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 pp. 28, 29).

[54] 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. 26).

[55] 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. 21).

[56] *Prolific.Co*. Prolific. 2024 (cited on pp. 21, 35, 38).

[57] Ronald Rensink. "Invariance of Correlation Perception". In: *Journal of Vision*. Vol. 12. Vision Sciences Society, May 2012, pp. 433–433. DOI: 10.1167/12.9.433 (cited on pp. 33, 38).

[58] Ronald A. Rensink. "On the Prospects for a Science of Visualization". In: *Handbook of Human Centric Visualization*. Ed. by Weidong Huang. New York, NY: Springer New York, 2014, pp. 147–175. ISBN: 978-1-4614-7484-5 978-1-4614-7485-2. DOI: 10.1007/978-1-4614-7485-2_6. (Visited on 06/15/2022) (cited on p. 33).

[59] Ronald A. Rensink. "The Nature of Correlation Perception in Scatterplots". In: *Psychonomic Bulletin & Review* 24.3 (2017), pp. 776–797. ISSN: 1069-9384. DOI: 10.3758/s13423-016-1174-7. (Visited on 10/20/2021) (cited on p. 14).

[60] 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. 28).

[61] 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. 27).

[62] 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. 28).

[63] 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. 23).

[64] Regina Schuster, Kathleen Gregory, Torsten Möller, and Laura Koesten. ""Being Simple on Complex Issues" – Accounts on Visual Data Communication About Climate Change". In: *IEEE Transactions on Visualization and Computer Graphics* 30.9 (Sept. 2024), pp. 6598–6611. ISSN: 1941-0506. DOI: 10.1109/TVCG.2024.3352282. (Visited on 10/22/2024) (cited on p. 13).

[65] Henrik Singmann and David Kellen. "An Introduction to Mixed Models for Experimental Psychology". In: *New Methods in Cognitive Psychology*. Routledge, 2019, pp. 4–31 (cited on p. 24).

[66] Robert F. Strahan and Chris J. Hansen. "Underestimating Correlation from Scatterplots". In: *Applied Psychological Measurement* 2.4 (Oct. 1978), pp. 543–550. ISSN: 0146-6216. DOI: 10.1177/014662167800200409. (Visited on 06/29/2022) (cited on pp. 14, 22).

[67] 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: 979-8-3503-2984-1. DOI: 10.1109/VisXVision60716.2023.00006. (Visited on 02/15/2024) (cited on p. 14).

[68] 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).

[69] 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: 979-8-4007-0330-0. DOI: 10.1145/3613904.3642127. (Visited on 05/29/2024) (cited on p. 15).

[70] 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. 28).

[71] *The MIT License*. https://opensource.org/license/mit. (Visited on 10/15/2024) (cited on p. 29).

[72] 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. 27).

[73] Cesko C. Voeten. *Buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression*. Manual. 2023 (cited on p. 25).

[74] 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. 30).

[75] Yihui Xie. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman and Hall/CRC, 2015 (cited on p. 26).

[76] 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. 26).

Appendices

Appendix A

First appendix

A.1 Section in Appendix